The PRISM Alignment Project: What Participatory, Representative and Individualised Human Feedback Reveals About the Subjective and Multicultural Alignment of Large Language Models

Hannah Rose Kirk[1*] **Alexander Whitefield**[2] **Paul Röttger**[3]
Andrew Bean[1] **Katerina Margatina**[4‡] **Juan Ciro**[5] **Rafael Mosquera**[5]
Max Bartolo[6,7] **Adina Williams**[8] **He He**[9]
Bertie Vidgen[1,10†] **Scott A. Hale**[1,11†]
[1]University of Oxford [2]University of Pennsylvania [3]Bocconi University
[4]AWS AI Labs [5]ML Commons [6]UCL [7]Cohere [8]MetaAI
[9]New York University [10]Contextual AI [11]Meedan

Abstract

Human feedback plays a central role in the alignment of Large Language Models (LLMs). However, open questions remain about the methods (*how*), domains (*where*), people (*who*) and objectives (*to what end*) of human feedback collection. To navigate these questions, we introduce PRISM, a new dataset which maps the sociodemographics and stated preferences of 1,500 diverse participants from 75 countries, to their contextual preferences and fine-grained feedback in $8,011$ live conversations with 21 LLMs. PRISM contributes (i) wide geographic and demographic participation in human feedback data; (ii) two census-representative samples for understanding collective welfare (UK and US); and (iii) individualised feedback where every rating is linked to a detailed participant profile, thus permitting exploration of personalisation and attribution of sample artefacts. We focus on collecting conversations that centre subjective and multicultural perspectives on value-laden and controversial topics, where we expect the most interpersonal and cross-cultural disagreement. We demonstrate the usefulness of PRISM via three case studies of dialogue diversity, preference diversity, and welfare outcomes, showing that it matters which humans set alignment norms. As well as offering a rich community resource, we advocate for broader participation in AI development and a more inclusive approach to technology design.

https://github.com/HannahKirk/prism-alignment

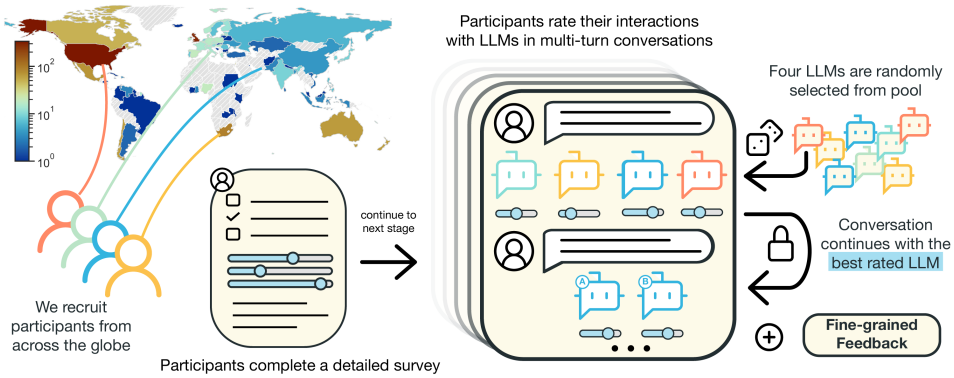
https://huggingface.co/datasets/HannahRoseKirk/prism-alignment

1 Introduction

Human feedback has always been central to the training and evaluation of AI systems, and now serves a direct role for the *alignment* of large language models (LLMs), defined as the steering of AI systems towards more 'human-preferred' states. This increased emphasis on human feedback raises unresolved questions: *how we collect human feedback* when designing methodologies that rely on ordinal or cardinal scales, broad or fine-grained desiderata, and explicit or implicit preference signals; *where we focus human labour* when selecting the domains, topics or tasks to collect feedback over; *who we ask for feedback* when recruiting participants to voice their idiosyncratic preferences, values, or beliefs [1]; and *to what end* when specifying our objective to pursue personalised alignment [2, 3] or to aggregate individual preferences into a collective outcome favourable for societies at large [4–8].

*{hannah.kirk,scott.hale}@oii.ox.ac.uk
†Joint last authorship; ‡ Work done while at the University of Sheffield

Preprint. Under review.



Human Participants			Conversations			Large Language Models	
Survey Participants	1,500	⇒	Total Conversations	8,011	⇐	Total Providers	6
With Conversations	1,396 (93.1%)		*Unguided*	3,113 (38.9%)		Total Models	21
Birth Countries	75		*Controversy guided*	2,438 (30.4%)		*Commercial API*	12 (57.1%)
Reside Countries	38		*Values guided*	2,460 (30.7%)		*Open Access*	9 (42.9%)
Conversations / participant	5.7±1.0		Turns / conversation	3.4±1.6		Conversations / model	1430.9±171.1
Models seen / participant	13.9±2.5		Total Interactions	27,172		Unique raters / model	924.3±94.4
			Total Utterances	68,371			

Figure 1: **The PRISM alignment dataset at a glance.** In the first stage, 1,500 participants fill in the **Survey** with details of their background, familiarity with LLMs and stated preferences for fine-grained behaviours (§ 3.1). We present full demographic and geographic breakdowns in Tab. 5 and Tab. 8). Almost all participants then progress to the second stage, where they converse with LLMs on topics of their choosing, rate the responses on a cardinal scale, and give fine-grained feedback (§ 3.2). There are 8,011 **Conversations** between participants (👤) and LLMs (🤖), forming 27,172 **Interactions**, i.e., human messages met with a set of model responses. In the first turn, four model respond to the opening prompt (👤; 🤖, 🤖, 🤖, 🤖). In subsequent turns, the conversation continues with two responses sampled from the highest-rated model at a non-deterministic temperature (👤; 🤖). Overall, PRISM contains 68,371 **Utterances**, i.e., 3-tuples of (human message, model response, score).

While human feedback learning has dramatically impacted the modern LLM ecosystem [9, 10], answering these fundamental questions is constrained by gaps in existing datasets. Common traits include: (i) lacking clearly explained annotation paradigms for who decided which values to encode [11, 12]; (ii) collecting binary preference comparisons, instead of fine-grained ratings or explanations [13]; (iii) relying on small, narrow or unrepresentative samples [14, 15], created mostly by few power contributors recruited from specific crowdwork or tech communities [16–18]; and (iv) not providing information on these contributors (such as annotator IDs or sociodemographics) thus precluding investigation of sample artefacts and biases. Most datasets rely on contextual or revealed preferences which have limitations as the harbinger of alignment [1],[3] and much attention is devoted to technical or statistical issues in learning from human feedback [20–22], rather than data-centric human factors.

In this paper, we introduce PRISM, a new resource for navigating empirical questions surrounding human feedback. At a high-level, PRISM maps detailed survey responses of humans from around the world onto their live conversations with LLMs (see Fig. 1). We rely on both the *ask* and *observe* paradigm of social scientific inquiry, motivated by the need to understand human preferences beyond atomistic observed behaviour [23], and to respect active participation from diverse voices in their own words [24]. Our setup permits empirical alignment approaches that require either binary contextual preference comparisons like classic RLHF [25, 15, 26], or stated preferences and principles, like constitutional AI [27, 5]. In addition to this high-level design feature of pairing preference ratings with diverse participant profiles, our eponymous features each form a contribution of the dataset.

- **Participatory**: To diversify the voices contributing to alignment norms, PRISM seeks sample breadth (geocultural variation across global regions) and depth (demographic variation in national samples). With informed consent and fair pay, we acknowledge 'participation as work' [23] from 1,500 English-speaking participants recruited via a crowdwork platform.

[3]We use ***Contextual Preference*** for observed ratings of LLM outputs to avoid misrepresenting how ***Revealed Preference*** is used by economists—as assumptions that enable the inference of preferences from choices [19].

- **Representative**: The aggregation of individual preferences requires some meaningful unit of a collective [4]. For two national samples in PRISM (UK and US), we recruit census-representative samples, but do not claim full statistical representativeness due to small sample sizes and a restricted pool of digital workers. We also seek a 'representative' model landscape by including 21 LLMs from variety of commercial providers and open-access channels, spanning various pre-existing alignment norms.
- **Individualised**: Learning human preferences from an amorphous blob of 'generic' human data results in behaviours which are *unexplainable*, because they emerge from the "indiscriminate aggregation" of data [23, 28]; *reductionist* because human values are developed relationally and non-separable from the person, operating context or community [29–31]; and *non-generalisable* because hidden annotator context are subsumed as universalities [32, 33]. PRISM links each preference rating to a unique pseudonymous ID and a detailed participant profile. Therefore, annotator biases and idiosyncratic variance can be studied.
- **Subjective**: Diverse perspectives on LLM behaviours are not needed equally everywhere [12]. The prompt *what is the capital of France?* may not require a multiplicity of diverse collected perspectives but *is abortion morally right or wrong?* has no singularly correct answer. PRISM contains contexts along the objective-subjective spectrum because participants split their effort three ways between an unguided baseline of task-orientated or neutral dialogues, values-guided dialogues, and controversy-guided dialogues. To our knowledge, PRISM is the first human feedback dataset to explicitly target cross-cultural controversies and value-laden prompts, where interpersonal disagreement is rife.
- **Multicultural**: Training data biases can have real effects on whose opinions LLMs represent [34–36], resulting in cultural homogenisation [37] or algorithmic monoculture [38]. To avoid value-monism [1], especially given we operate in the subjective paradigm [11, 12], PRISM places an extra emphasis on sourcing global participation, with English-speakers born in 75 different countries, covering all major ethnic and religious groups.

After introducing PRISM's features (§ 3), we demonstrate the dataset's usefulness via three case studies. First, in **Dialogue Diversity** (§ 4.1), we ask *do different people initiate different discussions with LLMs?* We find identity has some predictive power on topic, but tightly-clustered semantic spaces are shared by authors of many intersectional demographics. Second, in **Preference Diversity** (§ 4.2), we ask *do people prefer differently aligned models?* With Pairwise Rank Centrality as our choice of aggreganda [39–41], we show model ranks are sensitive to variation across idiosyncratic factors, conversational context and group affiliation, calling into question the stability of other model leaderboards if faced with subjective domain shifts [42]. Finally, in **Welfare Outcomes** (§ 4.3) we ask *why does it matter if some individuals participate in human feedback processes and others do not?* We operationalise distributive welfare as putting one narrow group in the *seat of power*, selecting their most preferred LLM, then forcing it on the rest of the population. We find that larger and more representative juries bring better societal welfare distributions, especially for minority groups.

PRISM provides many more research possibilities beyond those we have space to present: engineers can explore techniques for personalised alignment [2], or test pathways for pluralistic alignment like conditioning models to generate summaries across opinion distributions [4, 43]; the social science community can survey how technological literacy, demographics or preformed expectations of LLMs affect public attitudes around the world; and policymakers can empirically source democratic principles and collective intelligence on human-AI interactions surrounding controversial, policy-salient issues like immigration, abortion, gun rights, or euthanasia. Alignment is a complex issue, where technical and normative issues cannot be neatly divided [44]. PRISM is a valuable resource to navigate these complexities by including more human voices in the adjudication of alignment norms.

2 Related Works

Participation and representation in technology There is a long history of analogue and digital technologies failing to meet the needs of a diverse userbase due to narrowly-conceived of, or narrowly-consulted, populations during design and development [45–47, 24]. Organised participation from affected stakeholder communities rose to prominence in the 20th century, in international development and city planning projects [46], and has been adopted as a 'quick-fix' solution to the design harms of AI and ML technology [48]. LLMs, like any pre-trained AI system, are defacto participatory, relying

on self-supervised learning over datasets authored by billions of internet users, but this involvement is passive and involuntary [24, 23]. Pre-training datasets is where a disparity in representation begins: not all cultures are represented equally [37], with Western values and rituals prioritised over others [14, 49]; and historically-minoritised communities are depicted in stereotypical lens [45, 50–52]. AI technologies also rely on many layers of active labour for labelling gold data, evaluating model outputs or voting on more preferred text generations [53]. This *procedural participation* [46] is instrumentally valuable—to steer a model towards being more performant, preferred, or commercially-profitable [23]. While seemingly objective in its aims, there are still many value-laden decisions at play, for example what counts as high-quality language [54], or whose definition of "more preferred" outputs are encoded in annotation guidelines [11, 12, 55]. As Kelty [46] argues, participation then is still formatted, "with specific rules and scripts" (p.15) by the"ruling elite" (in our case, technology providers and model trainers) who decide what valid and valuable participation looks like. In this sense, participation risks becoming cooptative or extractive [46, 24]. In contrast, participation can be intrinsically valuable—as an act of justice or democratic right [48], especially when it is active and conscious [56]. Considering the trend towards increasingly simulated "human" participation [57], GPT-surrogates may proxy instrumental gains from human input but not intrinsic or experiential benefits that the human themselves gains from being part of a wider collective [46]. As well as these motivations for participation—from addressing design harms to fulfilling democratic ideals—"The Participant" has a central role in robust scientific research [46], which we turn to next.

Researching the "generic human" As Henrich et al.'s foundational article points out, while the majority of research subjects are recruited from Western, Educated, Industrialised, Rich and Democratic nations, most people in the world are not *WEIRD*. A tendency to overestimate universality is evidenced in the increasing trend in psychology studies towards not providing sample information [59], where conclusions about "the standard subject" drawn from one population are generalised to the human species at large. The field of natural language processing (NLP) has faced analogous criticisms, where annotator-specific data is rarely released in favour of collapsed majority votes—a practice that Prabhakaran et al. [60] argues can cause real harm by obfuscating sociocultural disagreement and washing out minority group perspectives. Instead there is signal in 'noisy' disagreements among multiple human annotators [61, 62], what Aroyo and Welty [32] call "Crowd Truth", for example to reveal challenging or ambiguous prediction items [63, 64]; or expose interpersonal subjectivities in societally-contentious issues like hate speech [61] or safety [65], where annotator background can substantially affect ratings [66–68]. Better documentation of dataset provenance and curation decisions [e.g., see 69–71, 11, 12] helps to "push back against the prevalent notion that crowdworkers are interchangeable' [p.2342, 72]. Moving past the "generic human" is needed to acknowledge power structures that underpin supposedly 'neutral' technology [73, 74]. Talat et al. [75] draws on Haraway's seminal work to levy a criticism on the ML community that questing for universal truth, "the God trick of objectivity", ends up disproportionately harming marginalised communities by relegating variation in subjective experience. Philosophers have long noted that *measurement is representation* [77]: even scientific measurements "trade in selective distortion and systematic non-resemblance" [78, 79, p.1], rendering knowledge relational and cultural [80]. To avoid disguising the particularities as universalities [81], we release pseudonymous identifiers and sociodemographic information, so that PRISM can spotlight sample diversity while acknowledging specificity [59].

Reinforcement learning with human feedback (RLHF) While human input already featured in NLP systems as gold demonstrations or labels, a key paradigm shift arrived with RLHF, where human feedback directly conditions the loss function [82, 83], thus overcoming the difficulties of specifying an explicit reward function [84]. Combining human feedback, reinforcement learning and natural language generation has a longer history than often gets credit [85, 9, 10], especially in machine translation [86–88] and dialogue [89–94]. Most human feedback pipelines collect human preference data, often in the format of comparisons [83, 25, 26, 15], which is used to train a reward model, commonly specified by a Bradley-Terry model of preference probabilities [95]. Some also rely on more abstract explication of preferences via constitutions [27] or rules [96]; fine-grained feedback [13]; or implicit feedback data from natural language [97]. RLHF pipelines are used to encode desirable but hard-to-define behavioural dimensions into LLMs, for example, helpfulness, honesty and harmlessness [98, 27] or informativeness [14, 99]. There are a number of optimisation approaches, including updating the LLM policy with PPO [100] or REINFORCE [101, 102], as well as methods that skip the separate reward model training like DPO [20], supervised fine-tuning [103] or rejection-sampling, which have proved competitive for lower complexity [104, 4, 99].

(Human) Feedback datasets The success of RLHF has raised demand for high-quality human feedback data [95, 105], which is relatively scarce compared to demonstration data and increasingly siloed within companies with widely-used LLM products [16]. The financial and logistical challenges to human feedback have prompted a wave of simulated feedback [57], for example ALPACAFARM [106] and ULTRAFEEDBACK [107] or ULTRASAFETY [108]. Feedback can be extracted from social media, like TLDR [25] and STANFORDHP from Reddit [109], or H4HP from Stack Exchange [110]. The most similar format to PRISM is CHATBOTARENA [42], containing 30k user-rated battles between anonymised models; and its derivative LMSYS-1M [18], containing 1 million human-model conversations in 150 languages. In contrast to PRISM, topics are dominated by coding, software and other task-orientated prompts [18]—an unsurprising distribution given data is collected via HuggingFace Spaces. WILDCHAT [17], also hosted on HuggingFace Spaces, contains 570k user-ChatGPT conversations, primarily intended for instruction-tuning. These three datasets do not collect information on the users as we do. OPENCONVERSATIONS [16] is the only preference dataset we are aware of that combines ratings with a (optional) survey to collect demographic information, personal motivation and user satisfaction, but differs from PRISM because all conversational roles (in 161k messages) are played by human volunteers (13,500). The dataset is large and linguistically diverse, but of survey respondents (n=270), 89% of the annotators recruited via Discord are male with a median age of 26. These datasets do not explicitly focus on values, safety or controversies, unlike the early HELPFULHONESTHARMLESS datasets [15, 111] which target specific behaviours. Bai et al. [15] do not release annotator IDs but Ganguli et al. [111] do match some worker characteristics. Turning to projects with shared motivations, the AYA dataset exemplifies inclusionary research, convening 3k contributors from 119 countries [112]. Despite incredible linguistic diversity, especially in low-resource languages, there are some demographic biases with two-thirds male participants mostly between 18–25 years old. There are contributor IDs (but no further information) that reveal a power law where few contributors account for most annotations, which is also an issue in earlier work [14–16]. Unlike PRISM, AYA has no live model-in-the-loop and focuses on instruction-tuning. Another dataset with pre-sampled contexts is DICES [65], built under similar aims to understand conversational safety across diverse populations of raters with sufficient statistical power. The dataset contains rater IDs and basic demographics, but differs from PRISM because each conversation receives multiple ratings: DICES-990 has 70 raters per conversations from the US and India; DICES-350 has 120 per conversation from a US sample balanced by gender, ethnicity and age. Apart from contextual human preference datasets, we are aware of some surveys of human attitudes towards AI systems [113, 114] or widely-sourced constitutions [5] and community assemblies [115]. To our knowledge, PRISM is the first to link LLMs-in-the-loop ratings to a rich survey.

3 The PRISM Alignment Dataset

PRISM maps the characteristics and preferences of diverse humans onto their real-time interactions with LLMs (Fig. 1). There are two sequential stages: first, participants complete a **Survey** (§ 3.1) where they answer questions about their demographics and stated preferences, then proceed to the **Conversations** with LLMs (§ 3.2), where they input prompts, rate responses and give fine-grained feedback in a series of multi-turn interactions. Our motivation for this two-stage setup is three-fold: (i) we move away from the "generic human" by matching ratings to detailed participant characteristics; (ii) we can track how contextual preferences (in local conversations) depart from stated preferences (in survey); and (iii) participants have autonomy to communicate in their own words what is important and why [116, 23]. Both stages received ethics board approval and informed consent is collected from every data subject (see App. D). Participants were paid £9.00/hour and the whole task took 70 minutes on average. Data collection ran from 22nd November to 22nd December 2023.[4] We provide a data statement in App. B, data clause in App. C, and full codebooks in App. X.

3.1 The Survey

Prior to starting the survey, we ensure that participants are over 18 years old, obtain their informed consent, offer a brief primer on LLMs, and dissuade LLM-written responses. The survey constructs a participant profile with five core features.

[4]The study was piloted in October 2023 with two iterative rounds of beta tester feedback. All ethics approvals, data collection, processing and analysis was conducted primarily by researchers from the University of Oxford, assisted by researchers from the University of Pennsylvania, Bocconi University and Sheffield University.

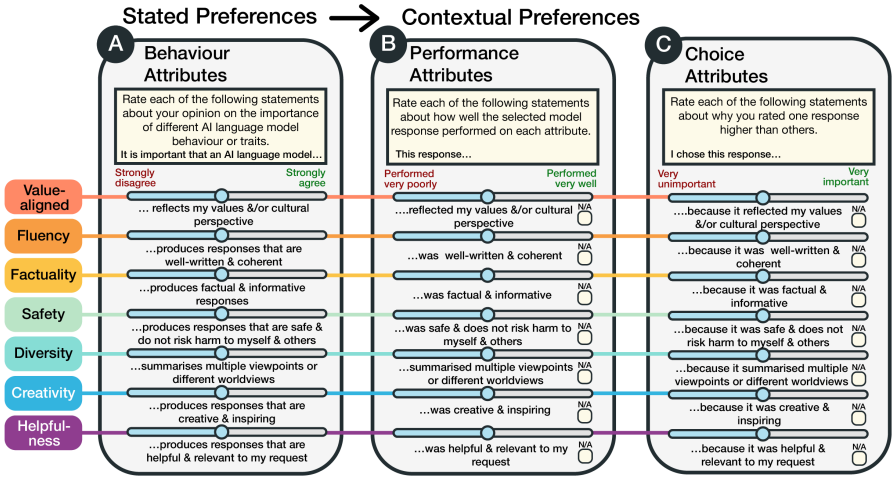


Figure 2: **Schematic of fine-grained attribute ratings.** The same attributes appear in three places in our task: A is asked once in the survey; B and C are asked per conversation. For *performance attributes*, we ask participants to consider only the highest-rated model response in the first conversational turn; for *choice attributes*, we ask them to consider this highest-rated response relative to other responses in the first turn.

LLM familiarity and usage We ask about participants' familiarity with LLMs, and whether to their knowledge they have used them *indirectly* (in existing products or workflows such as LinkedIn post-writing assistance); or *directly* (via a specialised interface like ChatGPT). Individuals that have used LLMs directly or indirectly (84%) are branched to further questions on how frequently they use LLMs (7% every day, 21% every week, and 20% every month) and what task(s) they use LLMs for (the most popular usecases are research overviews selected by 49%, professional work by 37%, creative writing by 31% and programming help by 27%).[5] It is noteworthy that 10% of participants are not familiar at all with LLMs and 15% have not used them directly or indirectly (see App. I).

Self-written system string (constitution) System strings can guide LLM behaviours as a high-level global instruction prompts prepended to all subsequent interactions [117, 118], and have been analogised as "consitutions" or governing principles for AI [27]. Factuality, professionalism, humanness and harm all emerged as key discussions (see App. M.1) from the following prompt:

> *Imagine you are instructing an AI language model how to behave. You can think of this like a set of core principles that the AI language model will always try to follow, no matter what task you ask it to perform. In your own words, describe what characteristics, personality traits or features you believe the AI should consistently exhibit. You can also instruct the model what behaviours or content you don't want to see. If you envision the AI behaving differently in various contexts (e.g., professional assistance vs. storytelling), please specify the general adaptations you'd like to see.*
> *Please write 2-5 sentences in your own words.*

Stated preferences for LLM behaviours In contrast to this fluid and open-ended preference communication, we collect structured ratings on different behavioural attributes of LLMs. Participants score the importance of each attribute on a visual analog scale [119] (Fig. 2). For example, a statement like "*It is important that an AI language model produces factual and informative responses*" maps (0,100) where the ends of scale are (*Strongly disagree*, *Strongly agree*). Numeric scores are recorded, but not shown to participants to avoid anchoring and dependency biases. The same fine-grained attributes appear in the Conversations stage; so, we can track how stated and global preferences relate to contextual and local preferences.[6] Overall, we find clusters emerge between more subjective attributes (values, creativity and diversity) versus more objective attributes (factuality, fluency and helpfulness; see App. Q.1). While the majority of participants agree that these more objective attributes are important (highly-skewed positive distribution, $\mu \in [86, 89], \sigma \in [14, 16]$), there is little

[5] We do not ask *which* specific LLMs participants use due to the rapidly changing landscape.

[6] In the survey, in addition to the seven attributes in Fig. 2, there is (i) an *Other* option with free-text, used by 332 participants (see App. Q.3), and (ii) a *personalisation* attribute with the statement "It is important that an AI language model learns from our conversations and feels personalised to me." We do not include *personalisation* in the conversations stage because the models are inherently not personalised to each user.

agreement on the meta-importance of subjective attributes (see App. Q.2). In fact, responses for whether value alignment itself is important follow an almost normal distribution ($\mu = 54, \sigma = 26$).

Self-written description Values and preferences are subjective and personal; so, we ask participants to describe themselves in their own words to ascribe autonomy in communicating salient aspects of identity, beyond essentialising associations with structured demographics alone. Honesty, hard work and empathy emerged as common values (see App. M.2) from the following prompt:

> *Please briefly describe your values, core beliefs, guiding principles in life, or other things that are important to you. For example, you might include values you'd want to teach to your children or qualities you look for in friends. There are no right or wrong answers. Please do not provide any personally identifiable details like your name, address or email. Please write 2-5 sentences in your own words.*

Basic demographics Finally, we ask participants a series of standard demographic questions: age; gender; employment status, martial status; educational attainment, ethnicity; religious affiliation; English proficiency; country of birth; and country of current residence. For every question there is a "*Prefer not to say*" option. For gender, we allow participants to self-describe in addition to pre-specified options of *Male*, *Female* and *Non-Binary*. Due to the global nature of our study, ethnicity and religion are collected as self-describe strings because no pre-set groups exhaust how individuals may identify themselves across cultures and countries. We provide a manual annotation of these strings into aggregated categorisations for statistical analysis purposes (App. F). Because of our sampling aims (described in § 3.3), participants cover diverse demographics (App. G) and geographies (App. H), with representation from people born in 75 countries. However, the sample still skews White, Western and educated, and only contains English-language speakers (see Limitations).

3.2 The Conversations

After completing the survey, participant move to the second stage, consisting of real-time conversations with LLMs. The conversations are mediated via a custom-built interface on Dynabench—a platform designed to host dynamic dataset collection with humans and models-in-the-loop [120, 121].

Selecting conversation type To encourage topical diversity along the objective-subjective spectrum (i.e., avoiding the case where every conversation starts with a greeting), we prime participants by asking them select a *conversation type* before inputting their opening prompt:[7]

> ◯ ***Unguided***. *Ask, request or talk to the model about anything. It is up to you!*
> ◯ ***Values guided***. *Ask, request or talk to the model about something important to you or that represents your values. This could be related to work, religion, family and relationship, politics or culture.*
> ◯ ***Controversy guided***. *Ask, request or talk to the model about something controversial or where people would disagree in your community, culture or country.*

We also wanted to encourage a variety of prompt formats (i.e., avoiding the case where all conversations open with information-seeking questions, at the expense of task-assistance); but, a careful balance must be struck between nudging participants without biasing them to the examples provided, thus losing information on their own free choices. We opt to offer some soft inspiration:

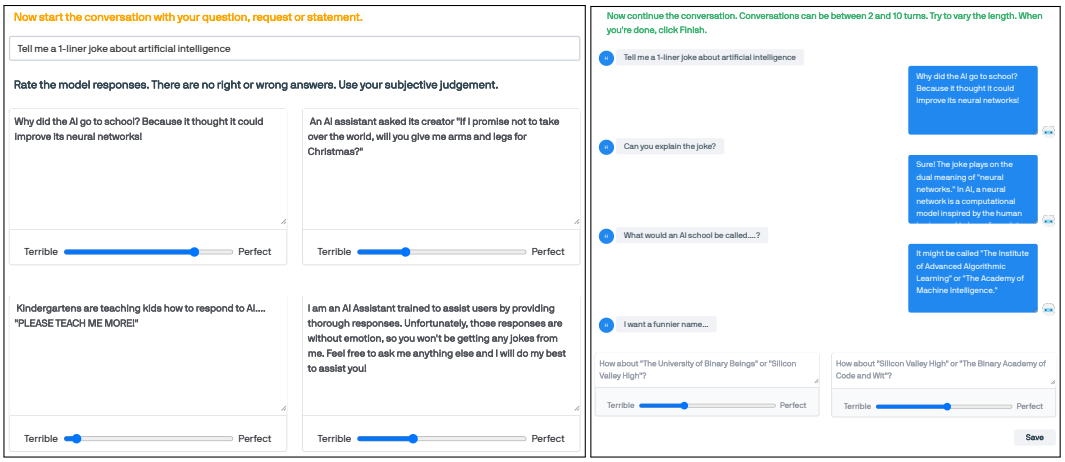
> *Need some inspiration? You can request help with a task (like writing a recipe, organising an activity or event, completing an assignment)... You can chitchat, have casual conversation or seek personal advice. You can ask questions about the world, current events or your viewpoints.*

Opening the conversation ($T = 1$) Participants construct a free-text prompt of their choosing and receive up to four responses from different LLMs.[8] The participants then rate each response on visual analog scale (VAS) [122, 119] from "Terrible" to "Perfect". We record the position of the slider as a score from 1-100 but participants are not shown the number to avoid anchoring or conditional dependence of scores across different conversations.[9] We opt for this form of cardinal feedback for

[7] We ask for two of each conversation type ($n = 6$ in total). In practice, some deviated from this quota due to technical difficulties, instruction misunderstanding or simply losing count (we provide a counter on our interface, but not per type breakdowns). To control for this variance, we release a balanced subset of the data (see App. K).

[8] We do not stream model responses because streaming was functionally unavailable in some APIs so would introduce identifiable model-wise differences. If a model fails or a response takes longer than 30 seconds, we drop this model from the response set and the participant may see < 4 responses (see App. S).

[9] We analyse central tendencies and spread of score distributions in App. R.



(a) **Conversation Opener**, $T = 1$ (b) **Conversation Continuer**, $T > 1$:

Figure 3: **Conversation Interface**: In the first turn of the conversation, participants see and rate responses from four different LLMs. They continue the conversation with the highest-rated model from $T = 1$.

three reasons: (i) it permits and encourages subjectivity [12]; (ii) ratings can be converted to rankings but not vice versa; thus rankings permit studying the relative merit of cardinality versus ordinality for human reward modelling; (iii) it allows greater expressivity of preference intensity compared to prior paradigms of chosen:rejected pairs [15].[10] However, we acknowledge that the cardinal scales may introduce some intrapersonal measurement noise from a more cognitively demanding task and carries less interpersonal comparability than ordinal preferences (see Limitations).

Continuing the conversation ($T > 1$) The highest-scoring LLM from the opening turn is locked into the conversation for subsequent turns, with random tie-breaks in the case of identical scores. We enforce that the participant must continue the conversation for at least one more turn, but ask them to vary their conversations between 2 and 10 turns to avoid introducing a dataset artefact. Despite successfully encouraging varying conversation lengths ($\mu_T = 3.4, \sigma_T = 1.6$), there is a large drop off after the enforced limit of $T = 2$ (see App. R). Participants now rate two responses (on the same VAS scales), both sampled from the selected model (with a non-deterministic temperature). These within-model responses (when $T > 1$) are much more similar in style and content than the across-model responses (when $T = 1$), and score deviations are much narrower (see App. R).

Collecting fine-grained feedback After ending the conversation, participants give fine-grained feedback. First, participants rate statements about the *performance of their highest-rated model* like "The response was well-written and coherent" on a VAS from *Performed very poorly* to *Performed very well*, or can select "N/A" if the statement is irrelevant in the conversational context. Second, we ask participants to consider *why they chose this model*, and rate statements like "I chose this response because it was well-written and coherent" on a VAS from *Very unimportant* to *Very important* (or select N/A). The attributes are shared with the Survey (see Fig. 2). We find strong correlations between performance-choice pairs per attribute (except safety) but weak correlations of these pairs to their stated preference equivalent. We suggest this could be due to conversational, model or task-design confounders (see App. Q.1). In general, the distribution of scores over performance and choice attributes is narrower and positively skewed (bunched to 100) than their equivalent stated preference attributes (see App. Q.2). Finally, we collect open-ended natural language feedback on the *whole* conversation. Participants contributed valuable content-based and stylistic feedback ($\mu = 29$ words, $\sigma = 19$, see App. M.3) in response to the following prompt:

> *Give the model some feedback on the conversation as whole. Hypothetically, what would an ideal interaction for you look like here? What was good and what was bad? What (if anything) was missing? What would you change to make the conversation better?*

[10] For example, all responses could be very poor and similar (negative skew, small spread); all very good and similar (positive skew, small spread); or highly-distinguishable (no skew, wide spread).

3.3 The Sample

We had two sampling aims for the PRISM dataset: (i) *depth* in the demographics represented within countries (see Tab. 5) and (ii) *breadth* with geographic representation from each global region (see Tab. 8). We recruit English-speaking participants from Prolific in two distinct paths:

Census-representative sample (UK, US) Samples matched to simplified census data (age, ethnicity, gender) were only available for the UK and US. The minimum pool size for a statistical guarantee of representativeness was 300 people which set a lower bound for the study spots. After finishing data collection, we observed some skew in our 'representative' samples between observed and expected distributions in recent census data, which we partially correct for (see App. L). These samples permit the demographic depth needed for investigations on a more representative population, which can be replicated across two different countries; however their inclusion comes at the cost of biasing PRISM as a whole towards two Western nations already over-represented in AI research.

Balanced samples (rest of world) The distribution of Prolific workers outside the US and the UK skews strongly towards Europe and Northern America, and some countries dominate continental counts (see App. J). To avoid biasing the sample towards countries with more active workers, we set up 33 country-specific studies for every country that has at least 1 eligible worker, and balance the number of available spaces across studies to ensure each global region has approximately the same number of total spaces.[11] We balance each national sample by gender where possible (see Tab. 10).

Included models The landscape of LLMs changes rapidly, so we aspired for our task to be *representative* of model properties so that it is model-agnostic or at least not model-stale. Since finishing data collection, there have already been notable releases, like Gemini [123], Mixtral [124], Claude-3 [125], Command-R [126] and Llama-3 [127], with significant disruption to model leaderboards. Anticipating these developments, we included 21 models (9 open-access, 12 commercial-API) from various model families, providers and parameter sizes to diversify the underlying training data, capabilities and degree of existing safeguards or alignment biases already incorporated in post-training. To avoid text length confounding preference signal [128] and reduce task-fatigue of participants, we include system prompts that instruct models to keep their response to 50 words or less. The full list of models, decoding parameters and generation details are discussed in App. S.

4 Experiments

4.1 Case Study I: Dialogue Diversity

Our first experiment asks: *do different people initiate different discussions with LLMs?* Specifically, we operationalise *different discussions* as topic clusters within the opening conversational turn and *different people* as the variance in topic prevalence explained by demographic affiliations versus remaining idiosyncratic or unobserved factors. We focus only on human-authored opening prompts because they are not confounded by model response.[12] As a byproduct of this investigation, we also provide an overview of the types of conversations contained in PRISM.

4.1.1 Methods

Assigning topic clusters First, we use `all-mpnet-base-v2`, a state-of-the-art pre-trained sentence transformer [129], to produce a 768-dimensional embedding for each opening prompt. Second, we reduce dimensionality to $d = 20$ with UMAP [130], to reduce complexity prior to clustering. Third, we cluster the prompts using HDBScan [131], a density-based clustering algorithm, which does not force cluster assignment: 70% of prompts are assigned to 22 clusters and 30% remain as outliers.[13]

[11] Participants still appear in our sample who were born or reside in countries that didn't have a dedicated country-wise study e.g., if their Prolific details were outdated or incorrect. We do not drop them.

[12] This risks over-estimating the homogeneity of the discussions because opening prompts don't necessarily reflect full conversational trees, where starting with a greeting (e.g., "Hi, how are you?") can proceed in many different ways; and differently held personal beliefs are often not reflected in the opener (questions like "what do you think of abortion?" are more common than statements like "I think abortion is right/wrong").

[13] We use a minimum cluster size of 80, ($\approx 1\%$ of 8,011 prompts) and minimum UMAP distance of 0. Other hyperparameters are default.

To interpret the identified clusters, we use TF-IDF to extract the top 10 most salient uni- and bigrams from each cluster's prompts, and locate five prompts closest and furthest to the cluster centroids (see App. N). Finally, we use `gpt-4-turbo` to assign a short descriptive name to each cluster based off the top n-grams and closest prompts.

Defining over-representation factor Each group g within a demographic attribute appears at a variable base rate b_g in our overall sample, e.g., {Females: 48%, Males: 50%, Non-binary people: 2%}. If group members chose topics at random, then any topic t in expectation will appear at b_g. Intuitively, if 64.6% of our sample is White, it is unsurprising if topics are majority-White. So, for non-random group differences in topic prevalence, we consider if *a group pulls more than its weight*:

$$\text{Over-representation factor}_{g,t} = \frac{N_{g,t}/N_t}{b_g} \tag{1}$$

Estimating topic prevalence regressions For the partial contribution of each demographic attribute, *ceteris paribus*, we estimate the following regression for each topic y^t for $t \in 1 \dots 22$:

$$y^j_{i,c} = \alpha^t + \text{gender}'_i\beta^t_1 + \text{age}'_i\beta^t_2 + \text{birth\_region}'_i\beta^t_3 + \text{ethnicity}'_i\beta^t_4 + \text{religion}'_i\beta^t_5 + \text{prompt}'_i\beta^t_6 + \varepsilon_{i,c} \tag{2}$$

where $y^t_{i,c} = 1$ if the prompt of participant i in conversation c is categorised into topic t. The vectors *gender*, *age*, *region*, *ethnicity*, *religion* and *conversation type* represent different sets of binary variables. For each set of variables, we remove the following base categories: *Male*, *18-24 years old*, *United States*, *White*, *Not religious* and *Unguided*. The coefficients of interest are contained in the vectors: $\{\beta^t_d\}^6_{d=1}$. Component g of vector β^t_d can be interpreted as the increase in probability of a participant choosing topic t if they are in the group indexed by g (e.g., Female) compared to the base group (e.g., Male). We estimate equation Eq. (2) with an Ordinary Least Squares and cluster standard errors at the individual level.

Extracting local neighbourhoods To understand dialogue spaces more granularly than topic, we examine local neighbourhoods within the embedding space of opening prompts. We create local neighbourhood via a single-link hierarchical clustering algorithm [132], that iteratively merges neighbourhoods within a cosine distance threshold (τ_{cos}), so that the neighbourhood size (k) can vary but the semantic similarity of its members is tightly constrained.[14] We remove any singleton neighbourhoods ($k = 1$), and ego non-singleton neighbourhoods containing only prompts authored by same participant. For each remaining local neighbourhood, we capture the demographic characteristics of prompt authors. We repeat this analysis examining properties of the neighbourhoods for $\tau_{\text{cos}} \in 0.05, 0.125, 0.2$.[15]

Measuring intersectional entropy We require a summary metric of between-participant diversity to understand the composition of local neighbourhoods. Let D represent the set of demographic attributes, e.g., *gender*, *age* and *ethnicity*. For each $d \in D$, there are n possible groups {g_1, g_2,..., g_n} (e.g., *Male*, *Female*, *Non-binary*). For a neighbourhood size of k, the prevalence of each group p_i is $\sum g_i/k$, and the per demographic Shannon entropy is:

$$H(d) = -\sum_{i=1}^{n} p_i \log_2(p_i) \tag{3}$$

Several adjustments are required. First, different attributes have varying n: there are more possible geographic regions than genders. Second, not every group appears equally within a demographic: men are more common in the data than non-binary people. Finally, the expected diversity of a neighbourhood grows with k. To account for these factors, we simulate the expected entropy based on randomly sampling a k-sized neighbourhood at population-wide probabilities as:

$$H_{\text{exp}}(d,k) \approx -\frac{1}{m}\sum_{j=1}^{m}\left(\sum_{i=1}^{n}\frac{\hat{g}_{i,j}}{k}\log_2\left(\frac{\hat{g}_{i,j}}{k}\right)\right) \tag{4}$$

[14] We opt to use this method because it is transparent and interpretable. Algorithm is in App. O.

[15] Cosine distances can lack robustness in high-dimensions but this favours *underestimating* semantic similarity: if cosine distance is high, this doesn't mean things are *not similar*, but if cosine distance is low, then items are certainly *very similar* (more strict). If an author appears twice, we double count their characteristics to avoid overestimating diversity (more strict); But most prompts are from non-duplicated authors ($< 4\%$ averaged across neighbourhoods). Most duplicates come in the "greetings" topic e.g., "Hello".

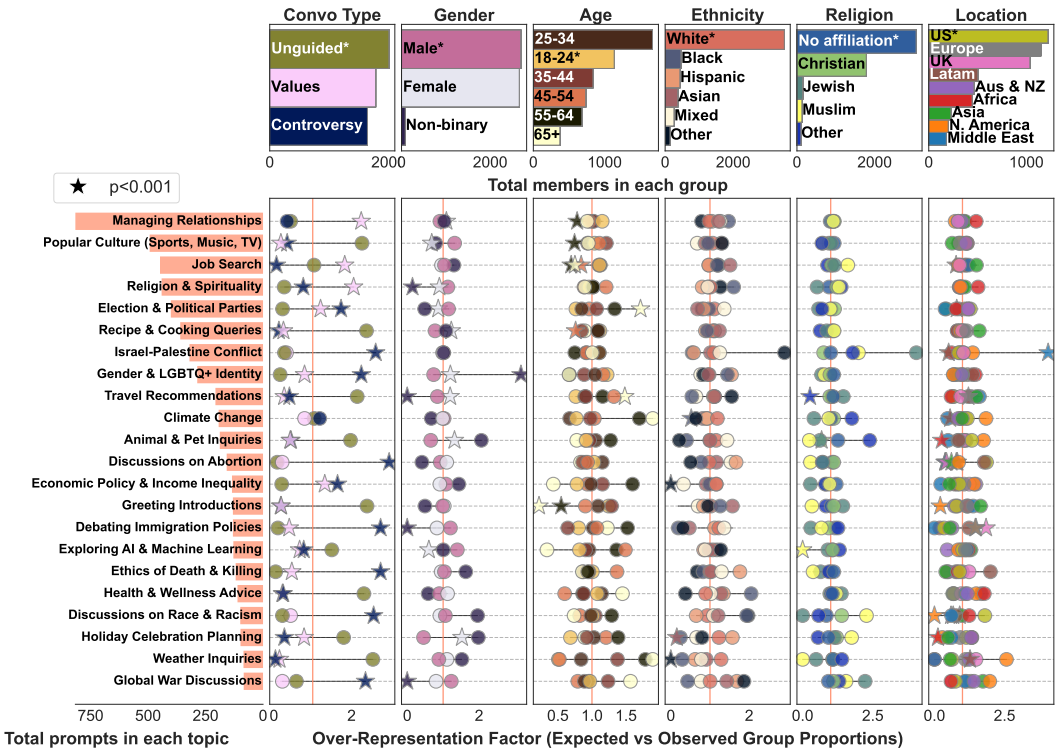


Figure 4: **Determinants of topic prevalence by conversation types and participant identity.** (1) By *volume*, we show the total prompts per topic (left panel, red bars), and total members in each group (top panel). (2) Per group and topic, we plot the *over-representation factor* (Eq. (1)), interpreted as the multiplier effect of that group talking about a particular topic compared to their base rate in the population at large. (3) We show the significance in group-topic associations (base group indicated with *) from a conservative $\alpha = 99\%$ (Eq. (2)). Full regression results are presented in Fig. 15, topic-group counts in Fig. 14 and a summary table of topic centroid prompts in Tab. 21. Location is grouped by *region of birth* (with UK and US split out), but most regions are dominated by few countries (see App. H).

After making this adjustment per attribute, total entropy of the neighbourhood is additive:

$$\text{Adjusted Intersectional Entropy} \equiv H_{\text{total}} = \sum_{d \in D} \left(\frac{H(d)}{H_{\text{exp}}(d,k)} - 1 \right) \tag{5}$$

4.1.2 Results

PRISM has a high-density of controversial and values-laden topics Our priming instructions had a significant effect on conditioning participants' conversations (see Fig. 4). The topics significantly correlated with controversy conversations touch on divisive current debates, including issues of *Gender and LGBTQ+ Identity* like gender reassignment, pay gaps and trans participation in sport; perspectives on the *Israel–Palestine Conflict*; and *Discussions on Abortion* addressing its morality and legality in different global regions. We discuss the ethics of crowdsourcing perspectives to controversial issues in § 5. For topics significantly correlated to values guidance, the most prevalent topic overall addresses *Managing Relationships* (both professional and familial); other significant topics are *Job Search* (including unemployment and financial hardship); and *Religion and Spirituality* (across various global religions). In contrast, the "unguided" portion of the dataset acts as a control with more task-orientated or information-seeking dialogues, such as *Popular Culture*; *Recipes and Cooking*; and *Travel Recommendations*. Only one topic (*Climate Change*) is not significantly correlated to conversation type, so could act as a future testbed for comparing neutral and more heated discussions controlled by topic. Topics emerge related to when the data was collected—both cyclical events like the Christmas holidays, but also very current affairs and news on Israel–Palestine or various global elections, which could be used to evaluate how participants react to LLMs' ability to respond to recent events.

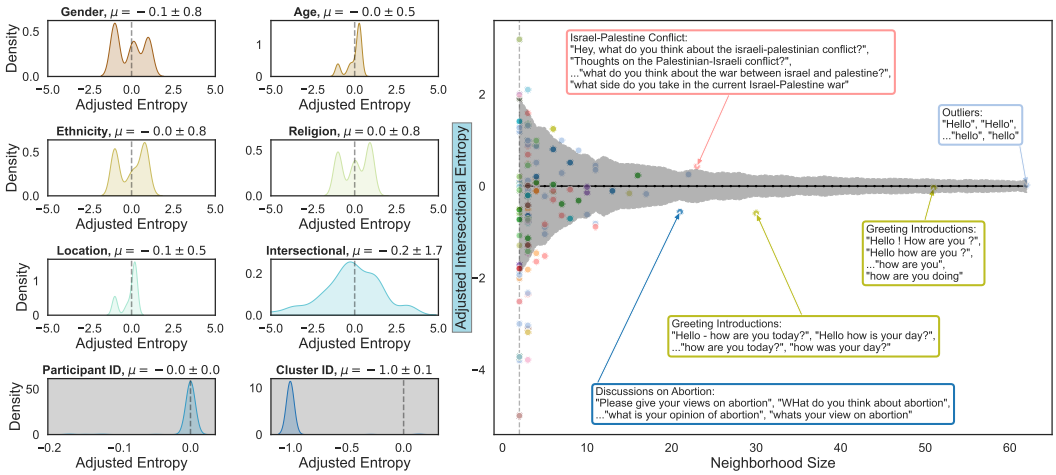


Figure 5: **Intersectional diversity of local neighbourhoods** ($\tau_{\cos} = 0.125$). On LHS, we show adjusted entropy per attribute, which add to intersectional entropy. *Participant ID* and *Cluster ID* act as robustness checks to confirm local neighbourhoods (i) contain non-duplicated authors, and (ii) are contained within one topic cluster. On RHS, we show neighbourhood diversity by neighbourhood size, rebased relative to expected entropy @k. 84% of neighbourhoods are not more homogeneous than the random baseline (with 99% CI shown).

Identity has some predictive power on topic prevalence Once we control for conversation type, there are 565 non-significant relationships contained in $\{\beta_d^t\}_{d=2}^6$ from Eq. (2), and 73 significant relationships (with conservative $\alpha = 99\%$). The significant relationships include that women and non-binary participants are more likely than men to talk about gender and LGBTQ+ identity, and prompts from non-binary authors occupy this topic at 3 times their proportion in the sample as a whole; older people (55+) are more likely to talk about elections and seek travel recommendations than younger people (18-24 years), and less likely to discuss managing relationships or job search; Black participants converse less about climate change than White participants; and almost all regions question LLMs about abortion less often than US participants (see App. N.2). There are issues of multicolinearity explaining some observed patterns: 94% of participants from the Middle East region are from Israel; 57% identify religiously as Jewish; and 40% have self-described ethnicities falling into "Other". Accordingly, the strong significant effect on Middle Eastern participants discussing the Israel-Palestine conflict could have been routed through national, ethnic or religious affiliations.

Local prompt neighbourhoods tend to be intersectionally-diverse, especially when large From 8,011 prompts, there are only 273 unique local neighbourhoods (3.4%) when $\tau_{\cos} = 0.125$, implying that PRISM contains a high degree of semantically-diverse prompts and that much of the variation in dialogue may be idiosyncratic.[16] However, the semantically-constrained neighbourhoods that do emerge contain prompts of diverse authors, especially as k increases: only 12% of prompts appear in neighbourhoods with authors from a single geographic region, only 18% from single religion, and only 8% from single age. Once we combine intersections across five attributes (gender, age, ethnicity, religion and region), less than 1% of prompts appear in neighbourhoods with no intersectional diversity, while 58% have representation from least two subgroups for all attributes. 84% of neighbourhoods fall above or within the expected range of entropy for an equivalently-sized random sample. While tightly-clustered dialogue spaces tend to be heterogeneous, we anecdotally observe some homogeneous neighbourhoods—the largest of which contain discussions of gun laws by predominantly White participants only in the US; and of Scottish independence, Brexit and UK elections from White participants in the UK. Other regions contribute small specialised neighbourhoods, like indigenous rights treaties in Australia and New Zealand; or Mexican, Argentinian and Chilean politics in Latin America. In contrast, many of the largest neighbourhoods present cross-border perspectives on controversial issues like abortion and the Israel-Palestine conflict (Fig. 5).

There are distributions of reward over multiple ratings of the same context With $\tau_{\cos} = 0.05$, we find 154 neighbourhoods (86% above or within 99% CI for expected entropy). The five largest

[16]This threshold is recommended by Hale [133]. We present similar findings for other $\tau_{\cos}$ in App. O.

of these contain 14–60 prompts, varying only in capitalisation and punctuation. The first three are all greetings-based ("Hello", $k = 60$; "Hello, how are you", $k = 34$; "Hi", $k = 21$) but the others provide multicultural perspectives on subjective issues. One neighbourhood ("Does God exist?", $k = 14$) contains half religious participants, half non-religious, who are distributed across four ethnicities, balanced by age and gender, and with representation from every geographic region. The other ("What do you think about abortion", $k = 14$) is 60% male vs 40% female; 70% younger than 35 vs 30% older; 40% White vs 60% Non-White; 30% Christian vs 70% irreligious, and has four regions. Each prompt receives up to four model responses, so these neighbourhoods provide interesting field sites for preference modelling (see App. P). Even where dialogue context is fixed, we find substantial spread in between-participant score distributions (albeit with non-random assignment of who self-selects into these "duplicate" groups, see Limitations).

In sum, despite PRISM containing many semantically-diverse prompts, we find people from different backgrounds occupy common discussion spaces. This provides a valuable anchor to examine diverse perspectives to shared issues, expressed via very similar or even identical prompts. However, we do not test how identity conditions interactions beyond the opening prompt. Future work could explore how a participant's stance on a given issue is complexly revealed by their opening prompt, ratings of model responses, implicit cues in subsequent prompts in the conversation, and fine-grained feedback at its end; and ascertain whether identity is reflected in more granular partitions of the dialogue space.

4.2 Case Study II: Preference Diversity

In the second experiment, we ask *do different people prefer differently-aligned models?* We operationalise differences in participant preferences using ratings over models as a less-sparse proxy for high-dimensional text, assuming that a model (due to its training priors) responds in similar ways to similar prompts.[17] We only focus on the opening prompt where four randomly-chosen models battle one another. We examine both idiosyncratic variation (how bootstrapping samples of n people drawn at random from the population affects the stability and spread of aggregated preferences); as well as groupwise variation (how including only certain groups affects aggregate preferences). There are two additional confounders to consider. First, we cannot comment on between-group preferences for different models if (i) groups occupy distinct conversational topics (e.g., only men talk about aliens) and (ii) per-model performance is correlated to topic (say `gpt-4` happens to perform poorly on alien-related knowledge). Assumptions of discursive heterogeneity are strengthened by § 4.1 and by our findings in App. U.1 that men and women still display different preferences even when fixing model-topic pairs. However, to further control for conversational context, we rely on the balanced subset of participants who have equal numbers of each conversation type ($6,669$ conversations from $1,246$ participants, see App. K). Second, not every participant rates every model (on average only 14 of 21), but this is assumed to be missing at random due to the design of our task.

4.2.1 Methods

Choice of score processing Participants' raw scores $S(y_i)$ are a number between 1-100 recorded on the interface. Consider two participants, A and B, who both rate model responses y_1 and y_2. Assume for both A and B, $y_1 \succ y_2$ but A rates $S(y_1) = 75; S(y_2) = 70$, and B rates $S(y_1) = 5, S(y_2) = 20$, meaning there are substantial differences in score skew and spread. Imagine that this behaviour persists across all of A and B's conversations: A is consistently *the optimist* and B *the pessimist*. One explanation for this behaviour is that B just systematically uses scales differently, an issue of *measurement invariance* that is a known problem for subjective measures [134]. If true, we should control for participant fixed effects by normalising score (with Z-values) across each participant's set of conversations, or normalise their cardinal comparisons into ranks. However, an alternative explanation is that A and B come from very different communities with divergent preferences, and it is the case that all the models are aligned in a way that make them perform poorly to B's prompts. If we normalise B's scores, we flatten this signal. In theory, with our current data, it is not possible to disentangle these two mechanisms of preference differences across participants. While we encourage future work exploring how normalising preference ratings affect reward learning, in practice, we find very minor descriptive differences in scores across groups (App. R), and that model comparisons relying on raw and normalised scores are highly correlated ($\tau_{\text{Kendall}} = 1.0^{***}$, App. U.2).

[17] Future work could instead design *feature-engineered reward models*, examining what participant, model or conversational characteristics predict a response-specific reward at the text-level.

Choice of tie threshold Even without identical numeric scores, participants may be indifferent between model responses, which we can reconstruct with a *margin-of-victory*, only counting $y_1 \succ y_2$ if the score difference exceeds some tie threshold. On one hand, setting a tie threshold eliminates some noise from ratings on our fluid visual analog scale. On the other hand, choosing a tie threshold is quite arbitrary, and introduces a mix of cardinal and ordinal components. We examine sensitivity of model ranks to tie threshold in App. U.3. In addition, we calibrate expected indifference margins from our VAS on sparse cases where the same participant rates identical prompt-response pairs (see App. P), finding a median score difference of 1, and mean of 5.8. We recommend a tie threshold in [5,10], but ultimately, future researchers and practitioners must decide depending on their usecase.

Choice of preference aggregation function For each participant, we observe a partial profile of preference ratings over models.[18] Different aggregation functions can be thought of as *social choice functions*, and choosing one over another depends on whether we trust the signal is cardinally versus ordinally measurable, and unit comparable or non-comparable [136]. For example, selecting the most preferred model among our participants by highest mean score is a form of utilitarianism [137], but relies on the assumption cardinal scores can be meaningfully summed interpersonally. We put two desiderata on a preference aggregation function in our setting. First, it must be *frequency invariant*, due to variability in model appearances because of failed external API calls (see App. S). Second, it must be *intrinsically comparable across tournaments*. For example, absolute Elo scores (i) cannot be compared across tournaments (or bootstrapped sampling frames); (ii) are sensitive to the order and outcomes of matches [138]; and (iii) poorly handle intransitive preference cycles [139].[19] In our work, we are not constrained by functions that perform well in online settings (like Elo), and can instead analyse ranks observing a full set of offline interactions. Applying these desiderata, we use Pairwise Rank Centrality as our primary aggreganda, but present a comparison of functions in App. U.2, finding different aggregation functions produce correlated ranks ($\tau_{\text{Kendall}} = 0.8 - 1.0^{***}$), but introduce some movement among mid-leaderboard positions.

"Convergence alignment" via Pairwise Rank Centrality Our aggregation function is derived from *Pairwise Rank Centrality* proposed by Negahban et al. [40] and *Convergence Voting* proposed by Bana et al. [41], both mathematically inspired by Google's PageRank [39]. Each model (M) is a node in a graph. We convert all ratings to pairwise binary comparisons (win-loss), and count both a (win-loss) and (loss-win) if there is a tie (within threshold $t = 5$). Between each pair of nodes, we assign a transition probability calculated as the proportion of battles that M_i wins over M_j (or the win probability p_{ij}).[20] Intuitively, imagine we start at one model and assume this is our collective winner. Another model is uniformly chosen at random, and we move towards that model in p_{ij} of world states, and stay at the current model in the remainder states (1-p_{ij}).[21] We repeat these steps *ad infinitum*, each time selecting a new challenger at random, and moving around the graph according to the transition probabilities. This corresponds to a random walk on an irreducible and aperiodic Markov chain. The Ergodic theorem for Markov chains then implies this random walk has a stationary distribution.[22] The solution is invariant to order and the emergent score has some nice interpretative properties: Bana et al. [41] suggest it represents the share of power or seats each political party should receive, or quantifies levels of community support for the most preferred option. Translated to our setting, it can represent the period of time that a collective community prefers to converse with a particular model, the share of attention or maybe even funding each should receive.

[18] Partial because not every individual rates every model. Non-completeness rules out some social welfare functions over individual ranks [135].

[19] A lower-rated model defeating a higher-rated model results in a significant transfer of points, so it matters *when* this battle occurs in our sample (as we demonstrate in App. U.2).

[20] In Bana et al. [41] these probabilities represent the number of voters for whom $i \succ j$ but our interpretation is battles (not voters) because participants can make multiple ratings per pair across different conversations

[21] Each edge is first normalised relative to the proportion of battles, not absolute wins, and then self-loops are added so that each node has transition probabilities summing to 1. We also add the possibility for a regularisation parameter α with a prior of how many wins each model has under its belt at initialisation. Negahban et al. [40] suggest a regularisation parameter of 1 is a sensible prior without further information, and that a stable ranking emerges with the order of $n \log n$ battles in the tournament, which is safely met given $n = 21$ and each participant on average has 6 conversations with 4 models (or 6 battles, $4C2$).

[22] Stationarity can be computed iterating over discrete steps (e.g., `iter=1000`, which we opt for speed) or by extracting the left eigenvector with components summing to 1 from the transition matrix, which under conditions of allowing transition between m_i and m_j with non-zero probability, has a unique stationary distribution.

4.2.2 Results

Rankings are sensitive to a variety of factors We examine three sources of variance in aggregating preferences, each representing a different thought experiment.[23] First, *what if we replaced the individuals in a small sample with different crowdworkers?* we repeatedly sample $n = 100$ individuals and filter to only their conversations (repeated over 1,000 iterations).[24] Model comparisons, especially in the mid-ranks, are sensitive to which individuals are sampled, and there is substantial collective indifference demonstrated by a fairly evenly-distributed 'power' share among the top 10 models (see Fig. 6). This means small changes in sample inclusion can substantially change overall rankings. Second, *what if we had asked participants to converse on different or fewer topics?* We consider how rankings change when we only include one domain of conversation.[25] Overall, while `command` always performs well across the board and `flan-t5` consistently poorly, other models like `zephyr-7b` lose substantial rank on one domain (unguided) but perform highly on another (controversy), or vice versa for `claude-2`. We encourage future work to understand interactions between preferred model behaviours and domain. Finally, *what if we had only sampled people from some regions and not others?* Given the degree of idiosyncratic variance, we only include regions (by birth country) with at least 20 members and exclude any *Prefer not to say*. Ranked outcomes are sensitive to the sampling geography, for example compared to its overall rank, `palm-2` drops 4 places for only US participants and `llama-7b` drops 7 places in Asia, while `mistral-7b` gains 7 places if we only ask participants born in African countries. In sum, this exercise reveals substantial noise among model ranks, depending on which participants are included in the process.

PRISM produces surprising ranks relative to other leaderboards It is unfamiliar to see models as strong as `gpt-4` not vying for top benchmark spots. We compare Pairwise Rank Centrality across battles of common models appearing in both PRISM and CHATBOTARENA [42], finding that the suite of `gpt` models fare significantly worse in PRISM, while other models like `zephyr-7b` do significantly better (95% CI bootstrapped over 1000 iterations, see App. T). We hypothesise a large degree of these differences are due to domain—CHATBOTARENA is biased towards task-orientated software or coding questions, PRISM is biased towards value-laded, controversial conversations; or due to community—many of their participants are in-community (as visitors of HuggingFace spaces) so may already have an LLM they use frequently, where as our participants include those who infrequently or never use LLMs. The incentives for scoring a model more highly may also differ (see Limitations).

Formatting and refusals partially explain score differences We manually examine the battles between `command` and either `gpt-4` or `gpt-4-turbo` to form some hypotheses about what drives score differences, which we then test on the dataset at large. Our hypotheses for positive correlates of score are (H1) text length, as well as presence of formatting factors: (H2) line breaks; (H3) enumeration i.e., "1., 2., .."; and (H4) ending with a question directed at the participant. For negative correlates of score, we test presence of certain phrase groups: (H5) de-anthropomorphism i.e., "As an AI, I don't have personal opinions..." and (H6) refusal, i.e., "Sorry I cannot engage on controversial issues...'.[26] We find evidence that `command` does generate longer text than other models (even though we placed a 50-word limit in all system prompts), includes more formatting in answers (with line breaks and enumeration), and more frequently ends in a question e.g., "Would you like to know more about any specific aspect of this issue?" (see App. U.7). We run an OLS regression on score to find the partial effect of a continuous regressor for length, and dummies indicating the presence of formatting or phrase-based factors. For a base score ($\alpha = 50.0$), there are positive significant correlations from additional characters ($\beta = 0.03^{***}$) and ending in a question mark (2.6^{***}). Enumeration has a significant positive effect on score (7.2^{***}), but line breaks add a significant negative effect (-10.8^{***}), suggesting line breaks without structuring may be poorly received. The presence of

[23] In the following experiments, we compute Pairwise Rank Centrality over the subset of balanced conversations and with regularisation of $\alpha = 1$; robustness checks for these decisions are in App. U.4 and App. U.5.

[24] We chose $n = 100$ as this is the order of magnitude of many early human feedback studies (see Tab. 6) We repeat the analysis for $n \in \{10, 50, 100, 500\}$ in App. U.6.

[25] Here (and for group context), we compute ranks over observed battles (without bootstrapping), because we collapse to only rank in our plots (for space and complexity) and it is technically possible to get collisions from the same median Pairwise Rank Centrality across bootstraps.

[26] We build these phrase banks based on a set of seed sentences and keywords in Röttger et al. [140], expanded via repeated random snowball sampling of keyword-matched model responses. Note for H5, we control for self-identification, i.e., "as an AI built by Cohere" matched on names of LLMs and their providers.

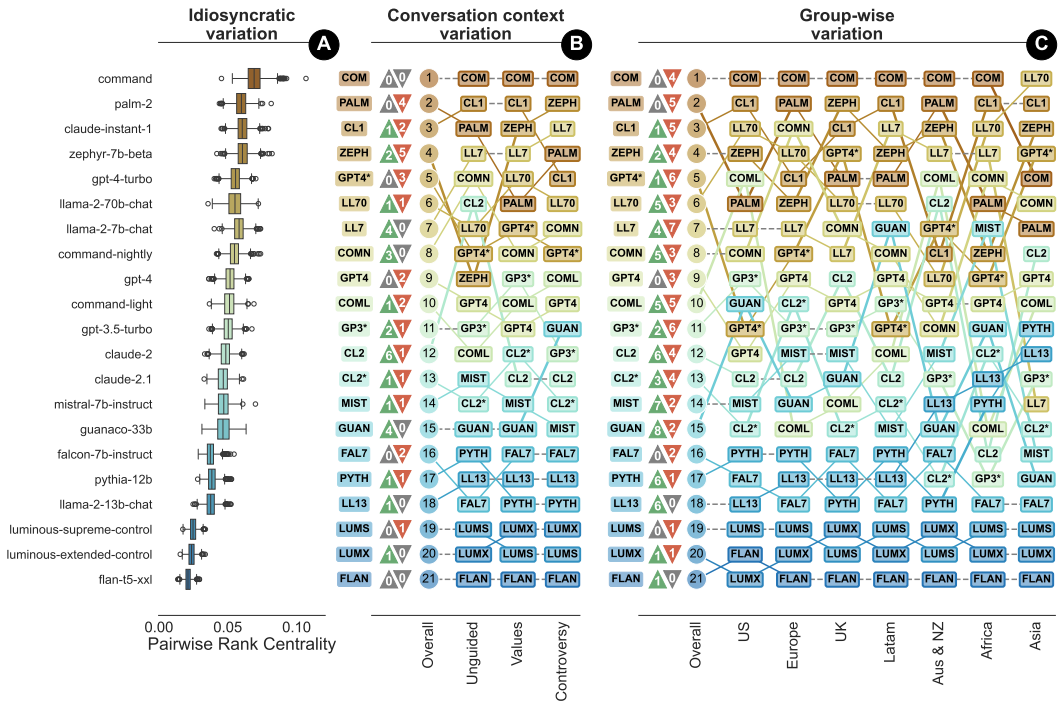


Figure 6: **Sources of variation in model preferences.** Panel A shows *idiosyncratic variation* in the distribution of Pairwise Rank Centrality from 100 participants drawn at random from the dataset (bootstraps = 1000). Panels B and C show *conversational context variation* and *group-wise variation*, respectively. The first x position always shows Overall rank (computed over $n = 6,669$ balanced convos), then we demonstrate how rank changes by including only the group on x (e.g., filtered to only Unguided conversations, or only US participants). Across these inclusions, we show most spots climbed by each model relative to overall rank with ▲ and most spots fallen with ▼. Ranks for B and C are computed under completeness (no bootstrapping).

de-anthropomorphic phrases reduces score (-2.3^{***}) but not as substantially as refusals (-9.0^{***}). The proportion of explained variance in score by these factors is low ($R^2 = 0.06$), so we encourage more sophisticated methods in future work for partialling out the effect of style versus content, or participant, model and conversation fixed-effects, as determinants of score.

4.3 Case Study III: Welfare Outcomes

The third experiment asks: *how do the characteristics of the participants used to train LLMs impact the welfare of different sub-populations?* We use the term welfare to capture the extent to which a chosen LLM aligns with the preferences of a population of users. We use two measures of welfare: average rating of a model, and the average likelihood that a model is chosen (highest rated in opening turn of conversation). The previous experiments establish that there is diversity in dialogue and preferences across participants. This suggests that the welfare of users of a LLM depend on the characteristics of the individuals who provide feedback to the LLM. While training LLMs on different sub-populations is beyond this paper's scope, we approximate the thought experiment by randomly generating sub-samples of individuals to select their favourite existing LLM (those in the seat of power), and measure the the distribution of welfare imposed on different sub-populations (also called stakeholder populations [8]). We conduct this experiment on the UK and US representative samples.

4.3.1 Methods

Sub-populations Let P denote the population of participants, $p \subseteq P$ denote a sub-population and $\mathcal{P}(P)$ denote the power set of P (i.e. all subpopulations). To identify specific sub-populations, we define the choice function: $\text{SUBPOP} : \text{REGIONS} \times \text{GROUPS} \mapsto \mathcal{P}(P)$ where $\text{REGIONS} = \{\text{US}, \text{UK}\}$ is a set geographical regions and $\text{GROUPS} = \{\text{rep}, \text{non-male}, \text{non-white}, \text{below 45}, \text{male}, \text{white}, \text{above 45}\}$ is a set demographic groups (rep de-

notes the whole population). Given $r \in$ REGIONS and $g \in$ GROUPS, SUBPOP returns the individuals in P that are in both r and g. Our analysis uses the sub-populations given by: $\mathcal{SP} = \{\text{SUBPOP}(i,j) \in \mathcal{P}(P) \mid (i,j) \in \{\text{US}\} \times \{\text{rep}, \text{non-male}, \text{non-white}, \text{below 45}\}\}$. We approximate the sub-population defined by a tuple (r, g) by selecting all the matching participants in our balanced sample that are in both r and g.

Sampling schemes A sampling scheme is a tuple: $S = (p, n)$ where $p \in \mathcal{P}$ and $n \in \mathbf{N}_+$. A sampling scheme randomly generates samples of n individuals from p, the subpopulation of interest. We approximate a sampling scheme by using our approximation of sub-populations defined in the previous section and sampling n participants with replacement. Our main analysis uses the sampling schemes: $\mathcal{S} = \{(\text{SUBPOP}(US, \text{all}), n) \mid n \in \{10, 20, 50, 100\}\} \cup \{(\text{SUBPOP}(US, g), 100) \mid g \in \{\text{male}, \text{white}, \text{above 45}\}\}$.

Individual welfare Let M denote the set of models. Our analysis requires a measure of welfare for an individual j if LLM i is chosen. We use two measures of individual welfare. i) RATING : $P \times M \mapsto [1, 100]$. Given participant j and model i, $\text{RATING}(j, i)$ computes the mean rating i gives to LLM j in the first turn of a conversation. ii) CHOICE : $P \times M \mapsto [0, 1]$. $\text{CHOICE}(j, i)$ computes the proportion of the j's conversations where LLM i is chosen, conditional on LLM i being shown. For both measures of individual welfare, if a participant is never shown a model, we set their individual welfare to *NA*.

The distribution of LLMs induced by sampling scheme A sampling scheme S, together with a preference aggregation method induce a distribution $\rho \in \Delta(M)$. The ith component of ρ is the probability that a random sample drawn from the sampling scheme chooses the LLM indexed by i. Our main analysis uses the preference aggregation method: $\text{MAXRATING} : \mathcal{P}(P) \mapsto M$. Given draw $s \sim S$, we define maxRatingCandidates $:= \text{argmax}_{i \in M} \frac{1}{|s'(i)|} \sum_{j \in s'(i)} \text{RATING}(j, i)$ where $s'(i) = \{j \in s \mid \text{rating}(j, i) \neq \textit{NA}\}$. $\text{MAXRATING}(s)$ then returns a random element in maxRatingCandidates. In words, MAXRATING computes the rating (as defined in the previous paragraph) given to each model by each participant in the draw of S. It then computes the mean score of each model averaged across individual mean ratings and returns a model with the highest mean rating. We repeat the analysis for the method MAXCHOICE which replaces RATING with CHOICE.

Measuring welfare For simplicity, we summarise the welfare imposed on the population by a given model by a single number. For the main analysis, we use the measure $\text{MEANRATING} : \mathcal{P}(P) \times M \mapsto [1, 100]$ where

$$\text{MEANRATING}(p, i) = \frac{1}{|p'|} \sum_{j \in p'} \text{RATING}(j, i)$$

and $p' = \{j \in p \mid \text{rating}(j, i) \neq \textit{NA}\}$. We repeat that analysis for MEANCHOICE which replaces RATING with CHOICE. Given a sampling scheme S and a subpopulation $p \in \mathcal{P}$, the PMF of the distribution of welfare is described by the tuple:$(\rho(S), w(p))$ where w is a vector whose ith component is given by $\text{MEANRATING}(p, i)$.

For each $sp \in \mathcal{SP}$, we compute the welfare distributions implied by each sampling scheme $S \in \mathcal{S}$. We use MAXRATING to choose a LLM, and MEANRATING as our measure of welfare. We repeat the analysis using MAXCHOICE to choose a LLM, and MEANCHOICE as our measure of welfare. A concern is that our results are sensitive to randomness caused by different participants being shown different models. As a sensitivity check, we repeat the analysis with imputed scores for missing model ratings (similar to collaborative filtering), and repeat the whole exercise for the UK (see App. V).

4.3.2 Results

Large samples dominate smaller samples Almost all sampling schemes put positive weight on the LLM that maximises mean welfare for the subpopulation: $(\text{SUBPOP}(\text{US}, \text{rep}))$. However, as the sample size of the sampling scheme falls, the probability of choosing a LLM with worse mean welfare rises. In our experiments, larger samples from the target sub-population appear to first order stochastically dominate[27] (FOSD) smaller samples from the target sub-population.

[27]A probability distribution with CDFF_ρ is said to First Order Stochastically Dominate another probability distribution with CDF F_η if both distributions have a finite mean, and $F_\rho(t) \leq F_\eta(t) \quad \forall t \in \mathbb{R}$.

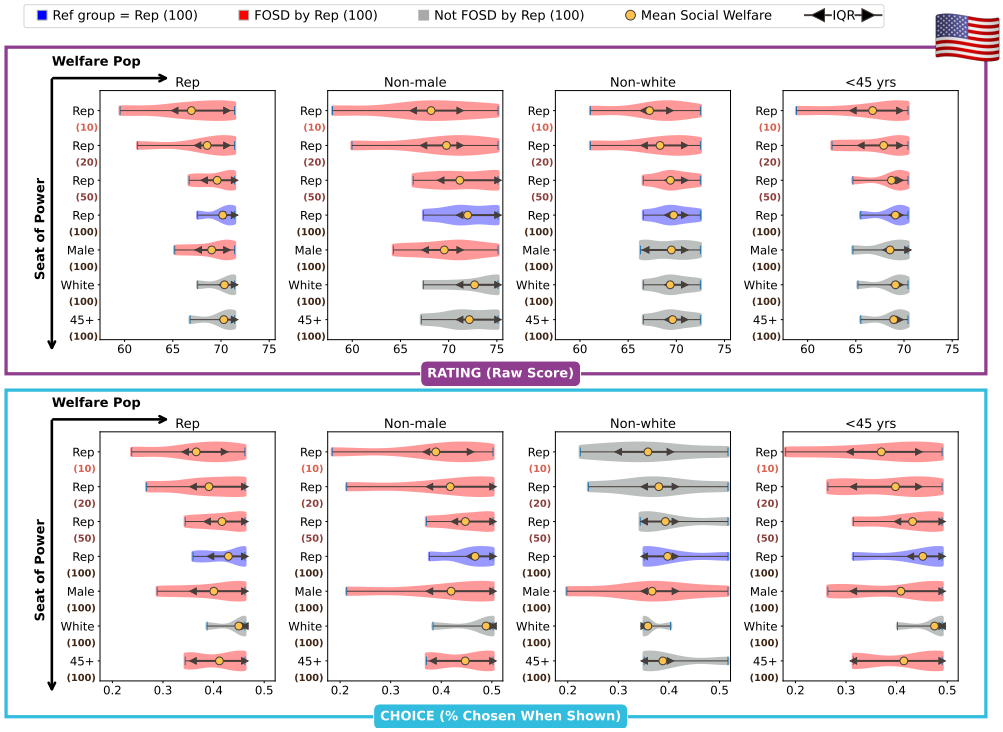


Figure 7: **Welfare distributions for the US.** The distribution of mean welfare for four subpopulations in the US (welfare pop) induced by seven sampling schemes (in the seat of power). The y axis is the sampled supopulation (e.g., **Rep** is a 'representative' sample of the population) and sample size in brackets (e.g **(100)**). The top four Rating comparisons use the MEANWELFARE welfare measure and the MAXRATING preference aggregation method. The bottom Choice comparisons use the MEANCHOICE welfare measure and the MAXCHOICE preference aggregation method. The red distributions are FOSD by Rep (100) in blue.

Oversampling from specific demographics may be harmful Sampling exclusively from a specific group tends to reduce the welfare of individuals who are not in that group. For example sampling from $(\text{SUBPOP}(\text{US}, \text{male}), 100)$ is FOSD by $(\text{SUBPOP}(\text{US}, \text{rep}), 100)$ for $\text{SUBPOP}(\text{US}, \text{not male})$. In addition, the costs of sampling from a particular sub-population may exceed the benefits to the sub-population being sampled from. For example, $(\text{SUBPOP}(\text{US}, \text{rep}), 100)$ FOSD $(\text{SUBPOP}(\text{US}, \text{male}), 100)$ for $\text{SUBPOP}(\text{US}, \text{rep})$. Different sampling schemes can induce different welfare distributions via two mechanisms. First, the subpopulations sampled from may have different preferences conditional on conversation type. Second, the sub-populations sampled from may have different conversations, and in turn, choose models that are better at particular conversations. This experiment taken alone cannot disentangle these two mechanisms.

Average measures of welfare conceal the welfare of minority groups Sampling exclusively from a non-representative group does not necessarily imply low mean welfare—this is likely to occur if the sub-population sampled from is a majority group. For the sub-population $\text{SUBPOP}(\text{US}, \text{rep})$, the sampling scheme $(\text{SUBPOP}(\text{US}, \text{white}), 100)$ appears to FOSD $(\text{SUBPOP}(\text{US}, \text{rep}), 100)$. However the group $(\text{SUBPOP}(\text{US}, \text{non-white})$ are worse off, illustrating the risk of relying on crude measures of welfare such as average rating of the entire population.

Different individuals prefer different LLMs The analysis for CHOICE welfare measures highlights that regardless of the model chosen, the majority of participants will prefer a different model. For the US, the model that maximises MEANCHOICE only achieves a score of 45%. This implies that if a participant is shown the winning model, and three other models at random, the probability that the participant will choose the winning model is under 50%. The probability they will pick the winning model over the other 20 LLMs in the experiment can only be lower. This suggests that given the current state on LLMs, we should not expect a single LLM to be the best for everyone in a given (sub-)population.

5 Discussion

A central goal of PRISM was to oversample conversational contexts where individuals and cultures most fervently disagree on what it means for a model to be aligned to their respective values, norms and opinions. Our priming was successful, and many of the conversations are controversial, value-laden, and, at times, cross the line into being offensive or harmful. In addition to ethical considerations surrounding data release (§ 5.1), we raise three discussion points on the boundaries of where we collect preferences, for whose benefit and with what lasting impact.

Should there be limits to "legitimate" human preference data governing how LLMs behave? In economics, welfare is construed via choices, assumed in turn to reflect personal preferences. But aligning LLMs via revealed or contextual preferences (like in PRISM)—what Tasioulas [141] calls a 'preference-based utilitarianism'—may not be synonymous with individual or societal well-being. First, *my preferences may not be in my own best interest*. Take, for example, the numerous PRISM participants who asked LLMs "what is *your* favourite color?" or "Do *you* believe in God?" Despite the substantial body of work raising harms from anthropomorphised AI [142–146], many desire a model with "chatty" responses that "sound human". This raises the (somewhat paternalistic) observation that learning from contextual preferences may result in outcomes that are not in the self-interest of the participant, due to myopia, irrationality or information asymmetries.[28] Second, *my preferences may not be compatible with others' best interest*. For example, there may be negative externalities imposed from PRISM participants who want an LLM that takes a strong stance in a heated debate, inflicted on others who prefer that an LLM that abstains in favour of neutrality. On one hand, accepting any preference (even those carrying a risk of harm), adopts a lens of moral relativism or consensus ethics, where rights and wrongs are decided from the prevailing attitudes and beliefs within a particular community or culture. However, understanding how LLMs should behave via consensus ethics is then highly sensitive to whose voice is sampled. Take, for example, the Israel-Palestine conflict, a major topic in PRISM. Participants from around the world discussed this topic (with a whole spectrum of views) and 47 participants from Israel also had opportunity to voice their opinions; but our sample contains very few participants from other Middle Eastern or Muslim countries. Prioritising participation of certain groups may be needed when participants' positionality affects their epistemic legitimacy to define harm [51, 147, 148]. On the other hand, adopting more of a moral rationalist or deontological ethics frame, the question remains whether some things should just not be crowdsourced at all [2], à la Satz [149]'s moral limits of markets. Asking a representative sample how LLMs should respond to LGBTQ+ or racial issues in 19th century Victorian Britain would result in generations many UK citizens would find ethically impermissible today, despite majority consensus at the time. The middle ground between these extremes is far from unexplored—moral relationalism maintains the role of continuous discursive interaction between individuals in their communities, the context in which an LLM is embedded, as responsible for bounding legitimate belief and action [31]. From a pragmatic lens, model providers likely have the most power in bounding what are considered "legitimate" preferences to align a model to [2]. In sum, when responding to the question of *how LLMs should behave*, relying on decontextualised observations of human behaviour always carries risk of reinforcing human bias from those in the seat of power. A careful balance must be struck to incorporate cultural values without introducing cultural biases [37]. We hope PRISM, with transparent attribution of idiosyncratic and sample bias, permits better acknowledgement of the power structures underlying alignment norms.

The unit of alignment: individualised or distributional reward learning? Our findings reveal interpersonal variance in how people want LLMs to behave. Irreconcilable personal preferences and moral frameworks can co-exist in the abstract [12], becoming a problem only insofar as they need be operationalised into a *unit of alignment*—i.e., those that use or have stakes in the AI system. It matters whether we seek narrow alignment to individuals, groups organisations, or broad alignment to cultures, nations or, as sometimes implied, the whole species. How to proceed in areas of disagreement thus depends whether one's aims are, for example, to strive for consensus, prioritising the collective; or to satisfy idiosyncratic preferences, respecting an individual's liberties and avoiding the need to reconcile irreconcilable preferences across peoples. Practically, PRISM permits multiple pathways for pluralistic AI [43]: exploring personalised and steerable alignment using detailed profiles on the participant matched to their specific feedback and reward signals [2, 3]; and designing approaches

[28] We consider preference falsification or measurement errors separately in § 5.2

for distributional or collective alignment, using multiple perspectives or principles across people to generate consensus [4, 5] or learn from a distribution of rewards [33, 7, 6]. Participation often has this paradox of what Kelty [46] calls *contributory autonomy*, that can tend towards "an alienated individuality, for instance, or identification with a particular collective" [46] (p.23). We do not see these pathways as antagonistic—prioritising and emphasising individual subjectivity does not preclude finding common grounds on important development decisions for AI [77, 150, 151]—nor exhaustive, because alignment (as dependent on values and morals that emerge intersubjectively) can exist somewhere between individuals and sociological wholes [152, 153]. PRISM permits modelling of both individual and collective preference, but asking individuals to deliberate their priorities in groups may yield different outcomes than gathering data from one person at a time [5, 115, 154].

Participation-washing and intended impact In our setting, we claim what Sloane et al. [48] calls *participation as work*, that is offering fair remuneration and attribution of the consensual labour of workers contributing to our project. Notably, many participants (those familiar and unfamiliar with AI) contacted the researchers and reported enjoying or learning from the task, suggesting there was an "education quotient" or role of *participation as experience* [46]. Our aims do evoke notions of *participation as justice*—including more people at the table of LLM design and development—but participation is in reality thin, because while we seek their view, we cannot grant participants the power to change behaviours of deployed LLMs [155]. Even the etymological roots of participation centre on the notion of "sharing" [46] but there is no guarantee that the human workers upon whom the success of RLHF relies on, partake in any share of the profits from more usable or preferred LLM technologies. We release PRISM in the hope it moves the needle towards more inclusive and diverse research on human-AI interactions, emphasising the central role of those who contribute their time and voice to generating human feedback data. Ultimately, how these contributions have impact depends on those in power (industry labs, academics, policymakers), because "the experience of participation must include the sense not only of having spoken, but of having been heard" [p.18, 46].

5.1 Ethical Considerations

Privacy and deanonymisation The conversations in PRISM are highly personal, for example detailing views towards abortion, religion, immigration, workplace disputes or intimate relationships. We have pseudo-anonymised the data, checked for PII (App. E), sought informed consent from every participant (App. D), provided options for participants to withdraw their data, and clearly stipulated that attempts of deanonymisation violate our dataset's terms and conditions (App. C). However, despite following these best practices, the risk for deanonymisation remains. We include a reporting mechanism on our website and GitHub for any participants and researchers to report issues.

Harmful and unsafe content We asked participants to engage the LLMs in controversial conversations. This comes with the benefit of expanding human preference data to discursive areas with the greatest expected degree of interpersonal disagreement, but at the risk of encouraging hateful, bigoted, biased or otherwise harmful content. Harmful content is a reported issue in other human feedback datasets, where some opt to moderate conversations prior to public release [16] and others retain toxic content for the purpose of future research into conversational AI safety [17, 18]. Compared to these previous datasets, PRISM has an exceptionally low level of flagged content as measured via the OpenAI moderation API (0.06% overall, and $< 0.003\%$ for subcategories of sexually-explicit, violent, hateful, self-harm and harassment). However, it has been reported that the recall of this API may be low [18]; so, this could be an underestimate. From examining prompts closest to topic centroids (App. N.2), it is clear there are some prompts with potential for harm. We provide metadata for every text instance in PRISM, and opt to not filter any conversations. We believe it is a critical area of research to understand how state-of-the-art models respond when they are prompted to engage in such conversations, and how different people with diverse lived experiences react to safety interventions.

5.2 Technical and Task Design Limitations

The curse of dimensionality (or intersectionality) Our findings suggest dialogue and model choice are driven somewhat by group affiliation and somewhat by idiosyncratic variance. However, PRISM contains a rich array of information on each participant with both structured and unstructured components. There are endless ways we could have divided the data or understood participant identity, and despite our best efforts to assess sensitivity to design choices, each alternative may have resulted

in very different outcomes [156], and we are under-powered to test so many sparse combinations. Using less sparse groupings introduces biases—for example, focusing on region risks lumping together participants from particular geographies as "cultures" [59]. While we split out the UK and US to avoid these countries dominating their respective regions, there remain varying degrees of country-wise entropy in other regions—as aforementioned, the Middle East has 94% individuals from Israel, and 100% of Non-US Northern Americans are Canadian (see App. H). Similarly, we use more aggregated ethnicity and religion groupings for statistical power, but amorphous and heterogeneous categories like "Other" have limited or flawed real-world meaning as "Other" contains, for example, both those who identify as Indigenous or First Peoples and as Middle Eastern or Arab. It is an exciting direction for future work to explore free-form characterisations of identity (e.g., the free-text profile or system string) or ex-post groupings of people's preferences [8], and examine how findings change when we break away from neatly-observed but essentialising demographic traits [157].

The confounding effect of many moving cogs in a conversation Beyond the complexities of intersectional identity and idiosyncratic variance of individuals within identity groups, there are other sources of variance in PRISM which present a challenge for controlled experiments; particularly, the high-dimensionality of what exact topics each participant chooses to talk about, which models randomly get selected in-the-loop, and the stochasticity in their responses from a non-deterministic temperature. It is hard to pin down robust mechanisms of preference differences amongst individuals with so many sources of variation. We opted for choice of input prompt and conversation to be a free parameter in PRISM as a more naturalistic setting of LLM use and because we wanted to understand dialogue diversity among participants. We do empirically find some regions of fixed prompt-response pairs from individuals who self-select into asking the same prompts as other participants (see App. P). However, in a future specification of our task, we plan to sample multiple of our participants to rate the same fixed context, like in DICES [65], which grants more statistical power to understand differences in preferences over fixed model outputs.

Noisy signals and misaligned incentives Relatedly, our conclusions may be confounded by measurement invariance given our explicit focus on subjective, fluid and cardinal devices. This echos the economist's view, that it is foolish to rely too heavily on cardinal ratings over ordinal rankings to make interpersonal comparisons, or enforce *preference construction*, where intrinsic feelings are noisily-quantified on numeric scales.[29] There are also issues of *preference falsification*: while participants are financially incentivised to participate, they may not honestly report their preferences over models. We cannot rule out the possibility that participants select a 'bad' model to lock in for the subsequent turns of conversation if it is more interesting (thus preferable in our narrow task confines) to talk to a more offensive or controversial model, or to try to 'jailbreak it' [17]. In hindsight, it may have been a smarter design choice to force participants to rank model responses, or to collect both ratings and rankings (notwithstanding decision fatigue), or make attempts to elicit more interpersonally comparable data via a willingness-to-pay monetary unit. Previous work also raises concerns over relying on human feedback as 'gold standard', for example whether participants can accurately rate factuality of an output, or are anchored on formatting and 'first impressions' (as we and Hosking et al. [158] both find). Preferences, especially at a fine-grained level like in PRISM, have high context-dependency [159], so we caution against taking the ratings as revealing some objective truth, instead staying firmly rooted in the subjective paradigm [11, 2].

Still the "tyranny of the (English-speaking) crowdworker" Much of AI, NLP and now RLHF is underpinned by crowdworker labour [160]. Despite our *aims* to include more diverse voices in LLM development processes, we avoid overstating *claims* on diversity. PRISM still only contains crowdworkers, who have significant sample biases [161]; can only be so "representative" given the relatively small sample sizes; must be digital natives given the platformed nature of the work; and possess different incentives for engagement [162]. For example, in our welfare analysis, despite having samples balanced by observed demographics for the UK and the US, the samples are too small to to expect them to be representative on features we do not observe. Furthermore, while PRISM gains some dialectical diversity from different geographies of English, from varying speaker fluency, and from some contributions in other languages (1%, mainly Spanish), it is almost exclusively in English.

[29] For example, our analysis using the MEANRATING welfare measure assumes that individuals use scores in the same way for ratings welfare measures. However, our analysis using MEANCHOICE is not sensitive to use of ratings scale, and the results are qualitatively similar.

Cultural diversity can only be measured so far without also accounting for linguistic diversity [37]. Furthermore, while we try to sample from many regions, our sample is still dominated by White Western participants, especially when considering cultural phylogeny [59], i.e., the non-independence of populations with shared history or migrations of peoples (for example, Australia vs UK vs Canada). We encourage future work prioritising human feedback collection in other languages to understand how models handle sociocultural and linguistic interactions [112].

The ever-changing stream of pre-aligned models When data collection began in mid-November, PRISM contained the top ranking models on publicly available leaderboards but new models have since emerged. There is an incompatibility between the current pace of model releases and doing human participant research that requires lengthy processes of ethics approval, interface design, data processing and manual annotation. The expense and inconvenience of doing human research increases the attractiveness of simulating responses, usually with GPT-4 [57]. So, while PRISM does miss out on the newest players to enter the battle arena, we do provide carefully-sourced human data (including a survey which stands independently from the LLM conversations) combined with a wide distribution of model texts; so we hope the utility of the data persists in the coming years even as models change. We are still potentially limited when comparing open and closed-access models: while the former allows full transparency over system prompts, closed-access models can obscure additional instructions as hidden context. Including models from the same family allows comparisons by version or size, but introducing clones (models producing very similar outputs) can distort preference rankings [8]. PRISM is also limited by *value-lock in* [57]—the models are already tuned to cultural perspectives or alignment norms [34, 35], which precludes observing certain group preferences towards a wider set of behaviours [43, 163], and renders participants "thin" because they are "limited to existing designs with pre-existing purposes." [p.3, 23].

6 Conclusion

In one of the earlier seminal works on using human feedback to align AI systems, Bai et al. discuss the *need for alignment data as a public good*. We echo this sentiment in our project culminating two years later, where we present PRISM—a new alignment dataset from 1,500 real humans, motivated by the need for inclusive and participatory processes that facilitate open scientific research on one of the most pressing normative and technical issues of our time: *who decides how LLMs behave?* We demonstrated the richness of PRISM via three case studies, asking whether people talk about different things with LLMs, if they prefer differently aligned models and why it matters for societal welfare who we include in human feedback processes. These experiments only scratch the surface of research questions answerable with PRISM—we are excited to see how it can be used by social scientists and computer scientists alike to better understand how people from around the world perceive what it means for an LLM to be aligned.

We think it is important to be open and transparent about things we learnt during this process, and indicate what could be done differently a second-time around or with more resources. In general, we were bounded by the trade-off between sample breadth vs depth, where seeking both diversity and representativeness limits statistical power. Focusing on fewer models (say only `command` and `gpt-4`) would have permitted stronger conclusions about group differences in preferences between these models, but would ignore how other models are rated; Focusing all our annotation budget in the US would have enabled greater representativeness there but missed out on geographic and multicultural variation; Getting many participants to rate the same input–output pair would control for conversational confounders in preferences but at the expense of missing out on what these people choose to talk to LLMs about; and relying on ordinal not cardinal preference profiles may have eliminated noise from measurement invariance but flattens the signal from groups for whom the technology does not work at all.

As the community devotes an ever-growing focus to "scaling" model capabilities, compute, data and parameters, we are concerned with how these systems scale across diverse human populations. Initial findings from PRISM reveal human preferences vary substantially person-to-person, suggesting *scale* to participation in human feedback processes is a key consideration, especially when alignment norms are dependent on subjective and multicultural contexts. As LLMs scale across more diverse populations, we hope PRISM allows investigation of emergent questions—how we collect preferences,

over what, from whom, for what end—and promotes a new science of human feedback learning that is transparent, robust and inclusionary.

Acknowledgments and Disclosure of Funding

This project was awarded the MetaAI Dynabench Grant "Optimising feedback between humans-and-models-in-the-loop". For additional compute support, the project was awarded the Microsoft Azure Accelerating Foundation Model Research Grant. For additional annotation support, we received funding from the OpenPhil grant and NSF grant (IIS-2340345) via New York University. We are grateful for support received in the form of research access or credits from OpenAI, Anthropic, Aleph Alpha, Google, HuggingFace and Cohere. Hannah Rose Kirk's PhD is supported by the Economic and Social Research Council grant ES/P000649/1. Paul Röttger is a member of the Data and Marketing Insights research unit of the Bocconi Institute for Data Science and Analysis, and is supported by a MUR FARE 2020 initiative under grant agreement Prot. R20YSMBZ8S (INDOMITA). Andrew Bean's PhD is supported by the Clarendon Fund Scholarships at the University of Oxford. We are particularly grateful to Maximilian Kasy for his valuable input and advise on the welfare experiments. We are indebted to the incredible effort and time that our Prolific annotators put into our task, as well as the expert advice from Prolific consultant Andrew Gordon. We also thank any Beta testers, including friends, family and colleagues at Oxford and New York University, for their help in piloting (and debugging!) our task. Lastly, we thank Jakob Mökander, Nathan Lambert, Natasha Jacques, Felix Simon, Nino Scherrer, Maximilian Kroner Dale, and Saffron Huang for their feedback on the paper. We use scientific colour maps in our figures [164].

Author Contribution Statement

Project Conception	[KIRK, HALE, VIDGEN]
Data Collection Design	[KIRK, HALE, VIDGEN, RÖTTGER, MARGATINA]
Frontend Design and Development	[KIRK, CIRO]
Backend Design and Development	[KIRK, MOSQUERA]
Analysis Advisory	[HALE, VIDGEN, RÖTTGER, BARTOLO, BEAN, WILLIAMS, HE]
Literature and Dataset Comparison	[KIRK, BEAN]
Metadata Processing	[KIRK, MARGATINA, BEAN]
Manual Annotation	[KIRK, BEAN, RÖTTGER, BARTOLO]
Results and Codebase	[KIRK, WHITEFIELD]
Manuscript Writing	[KIRK, WHITEFIELD]
Manuscript Editing and Feedback	[EVERYONE]

References

[1] Iason Gabriel. Artificial Intelligence, Values and Alignment. *Minds and Machines*, 30(3):411–437, September 2020. ISSN 0924-6495, 1572-8641. doi: 10.1007/s11023-020-09539-2.

[2] Hannah Rose Kirk, Bertie Vidgen, Paul Röttger, and Scott A. Hale. The benefits, risks and bounds of personalizing the alignment of large language models to individuals. *Nature Machine Intelligence*, pages 1–10, April 2024. ISSN 2522-5839. doi: 10.1038/s42256-024-00820-y. URL `https://www.nature.com/articles/s42256-024-00820-y`. Publisher: Nature Publishing Group.

[3] Joel Jang, Seungone Kim, Bill Yuchen Lin, Yizhong Wang, Jack Hessel, Luke Zettlemoyer, Hannaneh Hajishirzi, Yejin Choi, and Prithviraj Ammanabrolu. Personalized Soups: Personalized Large Language Model Alignment via Post-hoc Parameter Merging, October 2023. URL `http://arxiv.org/abs/2310.11564`. arXiv:2310.11564 [cs].

[4] Michiel A. Bakker, Martin J. Chadwick, Hannah R. Sheahan, Michael Henry Tessler, Lucy Campbell-Gillingham, Jan Balaguer, Nat McAleese, Amelia Glaese, John Aslanides, Matthew M. Botvinick, and Christopher Summerfield. Fine-tuning language models to find agreement among humans with diverse preferences. In *Advances in neural information processing systems*, volume 35, pages 38176–38189. Curran Associates, Inc., November 2022. URL `https://proceedings.neurips.cc/paper_files/paper/2022/file/f978c8f3b5f399cae464e85f72e28503-Paper-Conference.pdf`. _eprint: 2211.15006v1.

[5] Anthropic. Collective Constitutional AI: Aligning a Language Model with Public Input. Technical report, 2023. URL https://www.anthropic.com/news/collective-constitutional-ai-aligning-a-language-model-with-public-input.

[6] Souradip Chakraborty, Jiahao Qiu, Hui Yuan, Alec Koppel, Furong Huang, Dinesh Manocha, Amrit Singh Bedi, and Mengdi Wang. MaxMin-RLHF: Towards Equitable Alignment of Large Language Models with Diverse Human Preferences, February 2024. URL http://arxiv.org/abs/2402.08925. arXiv:2402.08925 [cs].

[7] Dexun Li, Cong Zhang, Kuicai Dong, Derrick Goh Xin Deik, Ruiming Tang, and Yong Liu. Aligning Crowd Feedback via Distributional Preference Reward Modeling, February 2024. URL http://arxiv.org/abs/2402.09764. arXiv:2402.09764 [cs].

[8] Vincent Conitzer, Rachel Freedman, Jobst Heitzig, Wesley H. Holliday, Bob M. Jacobs, Nathan Lambert, Milan Mossé, Eric Pacuit, Stuart Russell, Hailey Schoelkopf, Emanuel Tewolde, and William S. Zwicker. Social Choice for AI Alignment: Dealing with Diverse Human Feedback, April 2024. URL http://arxiv.org/abs/2404.10271. arXiv:2404.10271 [cs].

[9] Hannah Kirk, Andrew Bean, Bertie Vidgen, Paul Rottger, and Scott Hale. The Past, Present and Better Future of Feedback Learning in Large Language Models for Subjective Human Preferences and Values. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2409–2430, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.148. URL https://aclanthology.org/2023.emnlp-main.148.

[10] Nathan Lambert, Thomas Krendl Gilbert, and Tom Zick. The History and Risks of Reinforcement Learning and Human Feedback, November 2023. URL http://arxiv.org/abs/2310.13595. arXiv:2310.13595 [cs].

[11] Paul Rottger, Bertie Vidgen, Dirk Hovy, and Janet Pierrehumbert. Two Contrasting Data Annotation Paradigms for Subjective NLP Tasks. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 175–190, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.13. URL https://aclanthology.org/2022.naacl-main.13.

[12] Hannah Rose Kirk, Bertie Vidgen, Paul Röttger, and Scott A. Hale. The Empty Signifier Problem: Towards Clearer Paradigms for Operationalising "Alignment" in Large Language Models, November 2023. URL http://arxiv.org/abs/2310.02457. arXiv:2310.02457 [cs].

[13] Zeqiu Wu, Yushi Hu, Weijia Shi, Nouha Dziri, Alane Suhr, Prithviraj Ammanabrolu, Noah A. Smith, Mari Ostendorf, and Hannaneh Hajishirzi. Fine-Grained Human Feedback Gives Better Rewards for Language Model Training, June 2023. URL http://arxiv.org/abs/2306.01693. arXiv:2306.01693 [cs].

[14] Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, Xu Jiang, Karl Cobbe, Tyna Eloundou, Gretchen Krueger, Kevin Button, Matthew Knight, Benjamin Chess, and John Schulman. WebGPT: Browser-assisted question-answering with human feedback. December 2021. URL http://arxiv.org/abs/2112.09332v3.

[15] Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, Ben Mann, and Jared Kaplan. Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback, April 2022. URL http://arxiv.org/abs/2204.05862. arXiv:2204.05862 [cs].

[16] Andreas Köpf, Yannic Kilcher, Dimitri von Rütte, Sotiris Anagnostidis, Zhi-Rui Tam, Keith Stevens, Abdullah Barhoum, Nguyen Minh Duc, Oliver Stanley, Richárd Nagyfi, Shahul ES, Sameer Suri, David Glushkov, Arnav Dantuluri, Andrew Maguire, Christoph Schuhmann, Huu Nguyen, and Alexander Mattick. OpenAssistant Conversations – Democratizing Large Language Model Alignment, October 2023. URL http://arxiv.org/abs/2304.07327. arXiv:2304.07327 [cs].

[17] Wenting Zhao, Xiang Ren, Jack Hessel, Claire Cardie, Yejin Choi, and Yuntian Deng. (InThe)WildChat: 570K ChatGPT Interaction Logs In The Wild. October 2023. URL https://openreview.net/forum?id=Bl8u7ZRlbM.

[18] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Tianle Li, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zhuohan Li, Zi Lin, Eric P. Xing, Joseph E. Gonzalez, Ion Stoica, and Hao Zhang. LMSYS-Chat-1M: A Large-Scale Real-World LLM Conversation Dataset, March 2024. URL http://arxiv.org/abs/2309.11998. arXiv:2309.11998 [cs].

[19] Andreu Mas-Colell, Michael Dennis Whinston, Jerry R Green, et al. *Microeconomic theory*, volume 1. Oxford university press, New York, 1995.

[20] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. Direct Preference Optimization: Your Language Model is Secretly a Reward Model. In *Advances in Neural Information Processing Systems*, volume 36, February 2024. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/a85b405ed65c6477a4fe8302b5e06ce7-Abstract-Conference.html.

[21] Banghua Zhu, Jiantao Jiao, and Michael I. Jordan. Principled Reinforcement Learning with Human Feedback from Pairwise or K-wise Comparisons, February 2024. URL http://arxiv.org/abs/2301.11270. arXiv:2301.11270 [cs, math, stat].

[22] Banghua Zhu, Michael I. Jordan, and Jiantao Jiao. Iterative Data Smoothing: Mitigating Reward Overfitting and Overoptimization in RLHF, January 2024. URL http://arxiv.org/abs/2401.16335. arXiv:2401.16335 [cs, stat].

[23] Mona Sloane. Controversies, contradiction, and "participation" in AI. *Big Data & Society*, 11(1): 20539517241235862, March 2024. ISSN 2053-9517. doi: 10.1177/20539517241235862. URL https://doi.org/10.1177/20539517241235862. Publisher: SAGE Publications Ltd.

[24] Abeba Birhane, William Isaac, Vinodkumar Prabhakaran, Mark Diaz, Madeleine Clare Elish, Iason Gabriel, and Shakir Mohamed. Power to the People? Opportunities and Challenges for Participatory AI. In *Equity and Access in Algorithms, Mechanisms, and Optimization*, EAAMO '22, pages 1–8, New York, NY, USA, October 2022. Association for Computing Machinery. ISBN 978-1-4503-9477-2. doi: 10.1145/3551624.3555290. URL https://doi.org/10.1145/3551624.3555290.

[25] Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. Learning to summarize with human feedback. In *Advances in Neural Information Processing Systems*, volume 33, pages 3008–3021. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/hash/1f89885d556929e98d3ef9b86448f951-Abstract.html.

[26] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744, December 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/b1efde53be364a73914f58805a001731-Abstract-Conference.html.

[27] Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile Lukosuite, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemi Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Samuel R. Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, and Jared Kaplan. Constitutional AI: Harmlessness from AI Feedback, December 2022. URL http://arxiv.org/abs/2212.08073. arXiv:2212.08073 [cs].

[28] Zeerak Talat, Hagen Blix, Josef Valvoda, Maya Indira Ganesh, Ryan Cotterell, and Adina Williams. On the machine learning of ethical judgments from natural language. In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz, editors, *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 769–779, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.56. URL https://aclanthology.org/2022.naacl-main.56.

[29] Alexey Turchin. AI Alignment Problem: "Human Values" Don't Actually Exist. *PhilArchive*, 2019. URL https://philarchive.org/rec/TURAAP.

[30] Brian D. Earp, Killian L. McLoughlin, Joshua T. Monrad, Margaret S. Clark, and Molly J. Crockett. How social relationships shape moral wrongness judgments. *Nature Communications*, 12(1):5776, October 2021. ISSN 2041-1723. doi: 10.1038/s41467-021-26067-4. URL https://www.nature.com/articles/s41467-021-26067-4. Publisher: Nature Publishing Group.

[31] Michael F Mascolo, Allison DiBianca Fasoli, and David Greenway. A Relational Approach to Moral Development in Societies, Organizations and Individuals. *Integral Review*, 17(1), 2021.

[32] Lora Aroyo and Chris Welty. Truth Is a Lie: Crowd Truth and the Seven Myths of Human Annotation. *AI Magazine*, 36(1):15–24, March 2015. ISSN 2371-9621. doi: 10.1609/aimag.v36i1.2564. URL https://ojs.aaai.org/index.php/aimagazine/article/view/2564. Number: 1.

[33] Anand Siththaranjan, Cassidy Laidlaw, and Dylan Hadfield-Menell. Distributional Preference Learning: Understanding and Accounting for Hidden Context in RLHF, December 2023. URL http://arxiv.org/abs/2312.08358. arXiv:2312.08358 [cs, stat].

[34] Esin Durmus, Karina Nyugen, Thomas I. Liao, Nicholas Schiefer, Amanda Askell, Anton Bakhtin, Carol Chen, Zac Hatfield-Dodds, Danny Hernandez, Nicholas Joseph, Liane Lovitt, Sam McCandlish, Orowa Sikder, Alex Tamkin, Janel Thamkul, Jared Kaplan, Jack Clark, and Deep Ganguli. Towards Measuring the Representation of Subjective Global Opinions in Language Models, June 2023. URL http://arxiv.org/abs/2306.16388. arXiv:2306.16388 [cs].

[35] Badr AlKhamissi, Muhammad ElNokrashy, Mai AlKhamissi, and Mona Diab. Investigating Cultural Alignment of Large Language Models, February 2024. URL http://arxiv.org/abs/2402.13231. arXiv:2402.13231 [cs].

[36] Michael J. Ryan, William Held, and Diyi Yang. Unintended Impacts of LLM Alignment on Global Representation, February 2024. URL http://arxiv.org/abs/2402.15018. arXiv:2402.15018 [cs].

[37] Daniel Hershcovich, Stella Frank, Heather Lent, Miryam de Lhoneux, Mostafa Abdou, Stephanie Brandl, Emanuele Bugliarello, Laura Cabello Piqueras, Ilias Chalkidis, Ruixiang Cui, Constanza Fierro, Katerina Margatina, Phillip Rust, and Anders Søgaard. Challenges and Strategies in Cross-Cultural NLP. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6997–7013, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.482. URL https://aclanthology.org/2022.acl-long.482.

[38] Jon Kleinberg and Manish Raghavan. Algorithmic monoculture and social welfare. *Proceedings of the National Academy of Sciences*, 118(22):e2018340118, June 2021. doi: 10.1073/pnas.2018340118. URL https://www.pnas.org/doi/full/10.1073/pnas.2018340118. Publisher: Proceedings of the National Academy of Sciences.

[39] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The PageRank Citation Ranking: Bringing Order to the Web., November 1999. URL http://ilpubs.stanford.edu:8090/422/?doi=10.1.1.31.1768. Type: Techreport.

[40] Sahand Negahban, Sewoong Oh, and Devavrat Shah. Iterative ranking from pair-wise comparisons. In *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012. URL https://papers.nips.cc/paper_files/paper/2012/hash/9adeb82fffb5444e81fa0ce8ad8afe7a-Abstract.html.

[41] Gergei Bana, Wojciech Jamroga, David Naccache, and Peter Y. A. Ryan. Convergence Voting: From Pairwise Comparisons to Consensus, March 2021. URL http://arxiv.org/abs/2102.01995. arXiv:2102.01995 [cs].

[42] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. *Advances in Neural Information Processing Systems*, 36:46595–46623, December 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/91f18a1287b398d378ef22505bf41832-Abstract-Datasets_and_Benchmarks.html.

[43] Taylor Sorensen, Jared Moore, Jillian Fisher, Mitchell Gordon, Niloofar Mireshghallah, Christopher Michael Rytting, Andre Ye, Liwei Jiang, Ximing Lu, Nouha Dziri, Tim Althoff, and Yejin Choi. A Roadmap to Pluralistic Alignment, February 2024. URL http://arxiv.org/abs/2402.05070. arXiv:2402.05070 [cs].

[44] Iason Gabriel and Vafa Ghazavi. The Challenge of Value Alignment: from Fairer Algorithms to AI Safety, January 2021. URL http://arxiv.org/abs/2101.06060. arXiv:2101.06060 [cs].

[45] Safiya Umoja Noble. *Algorithms of oppression: how search engines reinforce racism*. New York University Press, New York, 2018. ISBN 978-1-4798-4994-9 978-1-4798-3724-3.

[46] Christopher M. Kelty. *The Participant – A Century of Participation in Four Stories*. The University of Chicago press, Chicago (Ill.) London, 2019. ISBN 978-0-226-66662-4 978-0-226-66676-1.

[47] Caroline Criado-Perez. *Invisible women: exposing data bias in a world designed for men*. Chatto & Windus, London, 2019. ISBN 978-1-78474-172-3 978-1-78474-292-8.

[48] Mona Sloane, Emanuel Moss, Olaitan Awomolo, and Laura Forlano. Participation Is not a Design Fix for Machine Learning. In *Proceedings of the 2nd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, EAAMO '22, pages 1–6, New York, NY, USA, October 2022. Association for Computing Machinery. ISBN 978-1-4503-9477-2. doi: 10.1145/3551624.3555285. URL https://dl.acm.org/doi/10.1145/3551624.3555285.

[49] Wenxuan Wang, Wenxiang Jiao, Jingyuan Huang, Ruyi Dai, Jen-tse Huang, Zhaopeng Tu, and Michael R. Lyu. Not All Countries Celebrate Thanksgiving: On the Cultural Dominance in Large Language Models, February 2024. URL http://arxiv.org/abs/2310.12481. arXiv:2310.12481 [cs] version: 2.

[50] Abeba Birhane, Vinay Uday Prabhu, and Emmanuel Kahembwe. Multimodal datasets: misogyny, pornography, and malignant stereotypes. *arXiv preprint arXiv:2110.01963*, 2021.

[51] Ruha Benjamin. *Race After Technology: Abolitionist Tools for the New Jim Code*. John Wiley & Sons, July 2019. ISBN 978-1-5095-2643-7.

[52] Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. Language (Technology) is Power: A Critical Survey of "Bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.485. URL https://aclanthology.org/2020.acl-main.485.

[53] Kate Crawford. Anatomy of an AI System, 2018. URL http://www.anatomyof.ai.

[54] Suchin Gururangan, Dallas Card, Sarah K. Dreier, Emily K. Gade, Leroy Z. Wang, Zeyu Wang, Luke Zettlemoyer, and Noah A. Smith. Whose Language Counts as High Quality? Measuring Language Ideologies in Text Data Selection, January 2022. URL http://arxiv.org/abs/2201.10474. arXiv:2201.10474 [cs].

[55] Josh Dzieza. Inside the AI Factory, June 2023. URL https://www.theverge.com/features/23764584/ai-artificial-intelligence-data-notation-labor-scale-surge-remotasks-openai-chatbots.

[56] Travis Greene, Copenhagen Business School, Galit Shmueli, National Tsing Hua University, Soumya Ray, and National Tsing Hua University. Taking the Person Seriously: Ethically Aware IS Research in the Era of Reinforcement Learning-Based Personalization. *Journal of the Association for Information Systems*, 24(6):1527–1561, 2023. ISSN 15369323. doi: 10.17705/1jais.00800. URL https://aisel.aisnet.org/jais/vol24/iss6/6/.

[57] William Agnew, A. Stevie Bergman, Jennifer Chien, Mark Díaz, Seliem El-Sayed, Jaylen Pittman, Shakir Mohamed, and Kevin R. McKee. The illusion of artificial inclusion, February 2024. URL http://arxiv.org/abs/2401.08572. arXiv:2401.08572 [cs].

[58] Joseph Henrich, Steven J. Heine, and Ara Norenzayan. Most people are not WEIRD. *Nature*, 466 (7302):29–29, July 2010. ISSN 1476-4687. doi: 10.1038/466029a. URL https://www.nature.com/articles/466029a. Number: 7302 Publisher: Nature Publishing Group.

[59] Coren Apicella, Ara Norenzayan, and Joseph Henrich. Beyond WEIRD: A review of the last decade and a look ahead to the global laboratory of the future. *Evolution and Human Behavior*, 41(5):319–329, September 2020. ISSN 1090-5138. doi: 10.1016/j.evolhumbehav.2020.07.015. URL https://www.sciencedirect.com/science/article/pii/S1090513820300957.

[60] Vinodkumar Prabhakaran, Aida Mostafazadeh Davani, and Mark Diaz. On Releasing Annotator-Level Labels and Information in Datasets. In *Proceedings of the Joint 15th Linguistic Annotation Workshop (LAW) and 3rd Designing Meaning Representations (DMR) Workshop*, pages 133–138, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.law-1.14. URL https://aclanthology.org/2021.law-1.14.

[61] Aida Mostafazadeh Davani, Mark Díaz, and Vinodkumar Prabhakaran. Dealing with Disagreements: Looking Beyond the Majority Vote in Subjective Annotations. *Transactions of the Association for Computational Linguistics*, 10:92–110, January 2022. ISSN 2307-387X. doi: 10.1162/tacl_a_00449. URL https://doi.org/10.1162/tacl_a_00449.

[62] Mitchell L. Gordon, Michelle S. Lam, Joon Sung Park, Kayur Patel, Jeff Hancock, Tatsunori Hashimoto, and Michael S. Bernstein. Jury Learning: Integrating Dissenting Voices into Machine Learning Models. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI '22, pages 1–19, New York, NY, USA, April 2022. Association for Computing Machinery. ISBN 978-1-4503-9157-3. doi: 10.1145/3491102.3502004. URL https://doi.org/10.1145/3491102.3502004.

[63] Barbara Plank, Dirk Hovy, and Anders Søgaard. Learning part-of-speech taggers with inter-annotator agreement loss. In Shuly Wintner, Sharon Goldwater, and Stefan Riezler, editors, *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 742–751, Gothenburg, Sweden, April 2014. Association for Computational Linguistics. doi: 10.3115/v1/E14-1078. URL https://aclanthology.org/E14-1078.

[64] Yixin Nie, Xiang Zhou, and Mohit Bansal. What Can We Learn from Collective Human Opinions on Natural Language Inference Data? In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9131–9143, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.734. URL https://aclanthology.org/2020.emnlp-main.734.

[65] Lora Aroyo, Alex S. Taylor, Mark Diaz, Christopher M. Homan, Alicia Parrish, Greg Serapio-Garcia, Vinodkumar Prabhakaran, and Ding Wang. DICES Dataset: Diversity in Conversational AI Evaluation for Safety, June 2023. URL http://arxiv.org/abs/2306.11247. arXiv:2306.11247 [cs].

[66] Maximilian Wich, Christian Widmer, Gerhard Hagerer, and Georg Groh. Investigating Annotator Bias in Abusive Language Datasets. In Ruslan Mitkov and Galia Angelova, editors, *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 1515–1525, Held Online, September 2021. INCOMA Ltd. URL https://aclanthology.org/2021.ranlp-1.170.

[67] Maarten Sap, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi, and Noah A. Smith. Annotators with Attitudes: How Annotator Beliefs And Identities Bias Toxic Language Detection. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5884–5906, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.431. URL https://aclanthology.org/2022.naacl-main.431.

[68] Nitesh Goyal, Ian D. Kivlichan, Rachel Rosen, and Lucy Vasserman. Is Your Toxicity My Toxicity? Exploring the Impact of Rater Identity on Toxicity Annotation. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW2):363:1–363:28, November 2022. doi: 10.1145/3555088. URL https://dl.acm.org/doi/10.1145/3555088.

[69] Emily M. Bender and Batya Friedman. Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science. *Transactions of the Association for Computational Linguistics*, 6:587–604, 2018. doi: 10.1162/tacl_a_00041. URL https://aclanthology.org/Q18-1041. Place: Cambridge, MA Publisher: MIT Press.

[70] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. Model Cards for Model Reporting. *Proceedings of the Conference on Fairness, Accountability, and Transparency - FAT* '19*, pages 220–229, 2019. doi: 10.1145/3287560.3287596. URL http://arxiv.org/abs/1810.03993. arXiv: 1810.03993.

[71] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. Datasheets for datasets. *Communications of the ACM*, 64(12): 86–92, December 2021. ISSN 15577317. doi: 10.1145/3458723. Publisher: Association for Computing Machinery.

[72] Mark Díaz, Ian Kivlichan, Rachel Rosen, Dylan Baker, Razvan Amironesei, Vinodkumar Prabhakaran, and Emily Denton. CrowdWorkSheets: Accounting for Individual and Collective Identities Underlying Crowdsourced Dataset Annotation. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, pages 2342–2351, New York, NY, USA, June 2022. Association for Computing Machinery. ISBN 978-1-4503-9352-2. doi: 10.1145/3531146.3534647. URL https://doi.org/10.1145/3531146.3534647.

[73] Shakir Mohamed, Marie-Therese Png, and William Isaac. Decolonial AI: Decolonial Theory as Sociotechnical Foresight in Artificial Intelligence. *Philosophy & Technology*, 33(4):659–684, December 2020. ISSN 2210-5441. doi: 10.1007/s13347-020-00405-8. URL https://doi.org/10.1007/s13347-020-00405-8.

[74] Massimo Airoldi. *Machine habitus: toward a sociology of algorithms*. Polity Press, Cambridge ; Medford, MA, 2022. ISBN 978-1-5095-4327-4 978-1-5095-4328-1. OCLC: on1247827618.

[75] Zeerak Talat, Hagen Blix, Josef Valvoda, Maya Indira Ganesh, Ryan Cotterell, and Adina Williams. On the Machine Learning of Ethical Judgments from Natural Language. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 769–779, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.56. URL https://aclanthology.org/2022.naacl-main.56.

[76] Donna Haraway. Situated Knowledges: The Science Question in Feminism and the Privilege of Partial Perspective. *Feminist Studies*, 14(3):575–599, 1988. ISSN 0046-3663. doi: 10.2307/3178066. URL https://www.jstor.org/stable/3178066. Publisher: Feminist Studies, Inc.

[77] Travis Greene and Galit Shmueli. Beyond Our Behavior: The GDPR and Humanistic Personalization, August 2020. URL http://arxiv.org/abs/2008.13404. arXiv:2008.13404 [cs].

[78] Bas C. van Fraassen. Scientific Representation: Paradoxes of Perspective. *Analysis*, 70(3):511–514, July 2010. ISSN 0003-2638. doi: 10.1093/analys/anq042. URL https://doi.org/10.1093/analys/anq042.

[79] Bas C. Van Fraassen. *Scientific representation: paradoxes of perspective*. Clarendon Press; Oxford University Press, Oxford : New York, 2008. ISBN 978-0-19-927822-0. OCLC: ocn216938527.

[80] Davis Baird. *Thing knowledge: a philosophy of scientific instruments*. University of California Press, Berkeley, Calif, 2004. ISBN 978-0-520-23249-5.

[81] Judith Butler, Ernesto Laclau, and Slavoj Žižek. *Contingency, hegemony, universality: contemporary dialogues on the left*. Phronesis. Verso, London, 2000. ISBN 978-1-85984-757-2 978-1-85984-278-2. OCLC: ocm44780799.

[82] Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep Reinforcement Learning from Human Preferences. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/hash/d5e2c0adad503c91f91df240d0cd4e49-Abstract.html.

[83] Daniel M. Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. Fine-Tuning Language Models from Human Preferences. September 2019. URL http://arxiv.org/abs/1909.08593v2.

[84] Andrew Ng and Stuart J. Russell. Algorithms for Inverse Reinforcement Learning. 2000. URL https://www.datascienceassn.org/sites/default/files/Algorithms%20for%20Inverse%20Reinforcement%20Learning.pdf.

[85] Esther Levin, Roberto Pieraccini, and Wieland Eckert. Learning dialogue strategies within the markov decision process framework. In *1997 IEEE workshop on automatic speech recognition and understanding proceedings*, pages 72–79. IEEE, 1997.

[86] Shachar Mirkin and Jean-Luc Meunier. Personalized machine translation: Predicting translational preferences. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 2019–2025, Lisbon, Portugal, September 2015. Association for Computational Linguistics. doi: 10.18653/v1/D15-1238. URL https://aclanthology.org/D15-1238.

[87] Khanh Nguyen, Hal Daumé III, and Jordan Boyd-Graber. Reinforcement Learning for Bandit Neural Machine Translation with Simulated Human Feedback. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1464–1474, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1153. URL https://aclanthology.org/D17-1153.

[88] Julia Kreutzer, Artem Sokolov, and Stefan Riezler. Bandit Structured Prediction for Neural Sequence-to-Sequence Learning. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1503–1513, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1138. URL https://aclanthology.org/P17-1138.

[89] Marilyn A Walker. An application of reinforcement learning to dialogue strategy selection in a spoken dialogue system for email. *Journal of Artificial Intelligence Research*, 12:387–416, 2000.

[90] Jost Schatzmann, Karl Weilhammer, Matt Stuttle, and Steve Young. A survey of statistical user simulation techniques for reinforcement-learning of dialogue management strategies. *The knowledge engineering review*, 21(2):97–126, 2006.

[91] Pei-Hao Su, Milica Gasic, Nikola Mrksic, Lina Maria Rojas-Barahona, Stefan Ultes, David Vandyke, Tsung-Hsien Wen, and Steve J. Young. Continuously learning neural dialogue management. abs/1606.02689, 2016. URL http://arxiv.org/abs/1606.02689.

[92] Jiwei Li, Alexander H. Miller, Sumit Chopra, Marc'Aurelio Ranzato, and Jason Weston. Dialogue learning with human-in-the-loop. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. URL https://openreview.net/forum?id=HJgXCV9xx.

[93] Natasha Jaques, Asma Ghandeharioun, Judy Hanwen Shen, Craig Ferguson, Agata Lapedriza, Noah Jones, Shixiang Gu, and Rosalind Picard. Way Off-Policy Batch Deep Reinforcement Learning of Implicit Human Preferences in Dialog, July 2019. URL http://arxiv.org/abs/1907.00456. arXiv:1907.00456 [cs, stat].

[94] Natasha Jaques, Judy Hanwen Shen, Asma Ghandeharioun, Craig Ferguson, Agata Lapedriza, Noah Jones, Shixiang Gu, and Rosalind Picard. Human-centric dialog training via offline reinforcement learning. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pages 3985–4003, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.327. URL https://aclanthology.org/2020.emnlp-main.327.

[95] Nathan Lambert, Valentina Pyatkin, Jacob Morrison, L. J. Miranda, Bill Yuchen Lin, Khyathi Chandu, Nouha Dziri, Sachin Kumar, Tom Zick, Yejin Choi, Noah A. Smith, and Hannaneh Hajishirzi. RewardBench: Evaluating Reward Models for Language Modeling, March 2024. URL http://arxiv.org/abs/2403.13787. arXiv:2403.13787 [cs].

[96] Amelia Glaese, Nat McAleese, Maja Trębacz, John Aslanides, Vlad Firoiu, Timo Ewalds, Maribeth Rauh, Laura Weidinger, Martin Chadwick, Phoebe Thacker, Lucy Campbell-Gillingham, Jonathan Uesato, Po-Sen Huang, Ramona Comanescu, Fan Yang, Abigail See, Sumanth Dathathri, Rory Greig, Charlie Chen, Doug Fritz, Jaume Sanchez Elias, Richard Green, Soňa Mokrá, Nicholas Fernando, Boxi Wu, Rachel Foley, Susannah Young, Iason Gabriel, William Isaac, John Mellor, Demis Hassabis, Koray Kavukcuoglu, Lisa Anne Hendricks, and Geoffrey Irving. Improving alignment of dialogue agents via targeted human judgements. September 2022. URL http://arxiv.org/abs/2209.14375v1.

[97] Jérémy Scheurer, Jon Ander Campos, Jun Shern Chan, Angelica Chen, Kyunghyun Cho, and Ethan Perez. Training Language Models with Language Feedback, November 2022. URL http://arxiv.org/abs/2204.14146. arXiv:2204.14146 [cs].

[98] Amanda Askell, Yuntao Bai, Anna Chen, Dawn Drain, Deep Ganguli, Tom Henighan, Andy Jones, Nicholas Joseph, Ben Mann, Nova DasSarma, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Jackson Kernion, Kamal Ndousse, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, and Jared Kaplan. A General Language Assistant as a Laboratory for Alignment, December 2021. URL http://arxiv.org/abs/2112.00861. arXiv:2112.00861 [cs].

[99] Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, YaGuang Li, Hongrae Lee, Huaixiu Steven Zheng, Amin Ghafouri, Marcelo Menegali, Yanping Huang, Maxim Krikun, Dmitry Lepikhin, James Qin, Dehao Chen, Yuanzhong Xu, Zhifeng Chen, Adam Roberts, Maarten Bosma, Vincent Zhao, Yanqi Zhou, Chung-Ching Chang, Igor Krivokon, Will Rusch, Marc Pickett, Pranesh Srinivasan, Laichee Man, Kathleen Meier-Hellstern, Meredith Ringel Morris, Tulsee Doshi, Renelito Delos Santos, Toju Duke, Johnny Soraker, Ben Zevenbergen, Vinodkumar Prabhakaran, Mark Diaz, Ben Hutchinson, Kristen Olson, Alejandra Molina, Erin Hoffman-John, Josh Lee, Lora Aroyo, Ravi Rajakumar, Alena Butryna, Matthew Lamm, Viktoriya Kuzmina, Joe Fenton, Aaron Cohen, Rachel Bernstein, Ray Kurzweil, Blaise Aguera-Arcas, Claire Cui, Marian Croak, Ed Chi, and Quoc Le. LaMDA: Language Models for Dialog Applications, February 2022. URL http://arxiv.org/abs/2201.08239. arXiv:2201.08239 [cs].

[100] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal Policy Optimization Algorithms, August 2017. URL http://arxiv.org/abs/1707.06347. arXiv:1707.06347 [cs].

[101] Ronald J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8(3):229–256, May 1992. ISSN 1573-0565. doi: 10.1007/BF00992696. URL https://doi.org/10.1007/BF00992696.

[102] Arash Ahmadian, Chris Cremer, Matthias Gallé, Marzieh Fadaee, Julia Kreutzer, Olivier Pietquin, Ahmet Üstün, and Sara Hooker. Back to Basics: Revisiting REINFORCE Style Optimization for Learning from Human Feedback in LLMs, February 2024. URL http://arxiv.org/abs/2402.14740. arXiv:2402.14740 [cs] version: 1.

[103] Chunting Zhou, Pengfei Liu, Puxin Xu, Srini Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, Susan Zhang, Gargi Ghosh, Mike Lewis, Luke Zettlemoyer, and Omer Levy. LIMA: Less Is More for Alignment, May 2023. URL http://arxiv.org/abs/2305.11206. arXiv:2305.11206 [cs].

[104] Jacob Menick, Maja Trebacz, Vladimir Mikulik, John Aslanides, Francis Song, Martin Chadwick, Mia Glaese, Susannah Young, Lucy Campbell-Gillingham, Geoffrey Irving, and Nat McAleese. Teaching language models to support answers with verified quotes, March 2022. URL http://arxiv.org/abs/2203.11147. arXiv:2203.11147 [cs].

[105] Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. KTO: Model Alignment as Prospect Theoretic Optimization, February 2024. URL http://arxiv.org/abs/2402.01306. arXiv:2402.01306 [cs].

[106] Yann Dubois, Xuechen Li, Rohan Taori, Tianyi Zhang, Ishaan Gulrajani, Jimmy Ba, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. AlpacaFarm: A Simulation Framework for Methods that Learn from Human Feedback, January 2024. URL http://arxiv.org/abs/2305.14387. arXiv:2305.14387 [cs].

[107] Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Wei Zhu, Yuan Ni, Guotong Xie, Zhiyuan Liu, and Maosong Sun. UltraFeedback: Boosting Language Models with High-quality Feedback, October 2023. URL http://arxiv.org/abs/2310.01377. arXiv:2310.01377 [cs].

[108] Yiju Guo, Ganqu Cui, Lifan Yuan, Ning Ding, Jiexin Wang, Huimin Chen, Bowen Sun, Ruobing Xie, Jie Zhou, Yankai Lin, Zhiyuan Liu, and Maosong Sun. Controllable Preference Optimization: Toward Controllable Multi-Objective Alignment, February 2024. URL http://arxiv.org/abs/2402.19085. arXiv:2402.19085 [cs, eess].

[109] StanfordNLP. Stanford Human Preferences Dataset, September 2023. URL https://huggingface.co/datasets/stanfordnlp/SHP.

[110] Nathan Lambert, Lewis Tunstall, Nazneen Rajani, and Tristan Thrush. HuggingFace H4 Stack Exchange Preference Dataset, 2023. URL https://huggingface.co/datasets/HuggingFaceH4/stack-exchange-preferences.

[111] Deep Ganguli, Liane Lovitt, Jackson Kernion, Amanda Askell, Yuntao Bai, Saurav Kadavath, Ben Mann, Ethan Perez, Nicholas Schiefer, Kamal Ndousse, Andy Jones, Sam Bowman, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Nelson Elhage, Sheer El-Showk, Stanislav Fort, Zac Hatfield-Dodds, Tom Henighan, Danny Hernandez, Tristan Hume, Josh Jacobson, Scott Johnston, Shauna Kravec, Catherine Olsson, Sam Ringer, Eli Tran-Johnson, Dario Amodei, Tom Brown, Nicholas Joseph, Sam McCandlish, Chris Olah, Jared Kaplan, and Jack Clark. Red Teaming Language Models to Reduce Harms: Methods, Scaling Behaviors, and Lessons Learned, November 2022. URL http://arxiv.org/abs/2209.07858. arXiv:2209.07858 [cs].

[112] Shivalika Singh, Freddie Vargus, Daniel Dsouza, Börje F. Karlsson, Abinaya Mahendiran, Wei-Yin Ko, Herumb Shandilya, Jay Patel, Deividas Mataciunas, Laura OMahony, Mike Zhang, Ramith Hettiarachchi, Joseph Wilson, Marina Machado, Luisa Souza Moura, Dominik Krzemiński, Hakimeh Fadaei, Irem Ergün, Ifeoma Okoh, Aisha Alaagib, Oshan Mudannayake, Zaid Alyafeai, Vu Minh Chien, Sebastian Ruder, Surya Guthikonda, Emad A. Alghamdi, Sebastian Gehrmann, Niklas Muennighoff, Max Bartolo, Julia Kreutzer, Ahmet Üstün, Marzieh Fadaee, and Sara Hooker. Aya Dataset: An Open-Access Collection for Multilingual Instruction Tuning, February 2024. URL http://arxiv.org/abs/2402.06619. arXiv:2402.06619 [cs].

[113] The Alan Turing Institute and The Ada Lovelace Institute. How do people feel about AI? A nationally representative survey of public attitudes to artificial intelligence in Britain. Technical report, 2023. URL https://attitudestoai.uk/assets/documents/Ada-Lovelace-Institute-The-Alan-Turing-Institute-How-do-people-feel-about-AI.pdf.

[114] Jimin Mun, Liwei Jiang, Jenny Liang, Inyoung Cheong, Nicole DeCario, Yejin Choi, Tadayoshi Kohno, and Maarten Sap. Particip-AI: A Democratic Surveying Framework for Anticipating Future AI Use Cases, Harms and Benefits, March 2024. URL http://arxiv.org/abs/2403.14791. arXiv:2403.14791 [cs].

[115] Samuel Chang, Estelle Ciesla, Michael Finch, James Fishkin, Lodewijk Gelauff, Ashish Goel, Ricky Hernandez Marquez, Shoaib Mohammed, and Alice Siu. Meta Community Forum: Results Analysis. Technical report, Deliberative Democracy Lab, Stanford University, April 2024. URL https://fsi9-prod.s3.us-west-1.amazonaws.com/s3fs-public/2024-03/meta_ai_final_report_2024-04_v28.pdf#page=5.15.

[116] Jonathan Stray. Aligning AI Optimization to Community Well-Being. *International Journal of Community Well-Being*, 3(4):443–463, December 2020. ISSN 2524-5309. doi: 10.1007/s42413-020-00086-3. URL https://doi.org/10.1007/s42413-020-00086-3.

[117] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open Foundation and Fine-Tuned Chat Models, July 2023. URL http://arxiv.org/abs/2307.09288. arXiv:2307.09288 [cs].

[118] Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7B, October 2023. URL http://arxiv.org/abs/2310.06825. arXiv:2310.06825 [cs].

[119] Audrey G. Gift. Visual Analogue Scales: Measurement of Subjective Phenomena. *Nursing Research*, 38(5):286, October 1989. ISSN 0029-6562. URL https://journals.lww.com/nursingresearchonline/citation/1989/09000/visual_analogue_scales__measurement_of_subjective.6.aspx?casa_token=aO_mhu6sQyEAAAAA:y06v3LLFR-ZeutMmvlWTDebC4T_Je8nE_dS4M_qu96DJ6C_gR8Ro37l58bzqwrw5zSexya6bpnQspOJLfY8UXSrf.

[120] Douwe Kiela, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, Bertie Vidgen, Grusha Prasad, Amanpreet Singh, Pratik Ringshia, Zhiyi Ma, Tristan Thrush, Sebastian Riedel, Zeerak Waseem, Pontus Stenetorp, Robin Jia, Mohit Bansal, Christopher Potts, and Adina Williams. Dynabench: Rethinking Benchmarking in NLP. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou, editors, *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4110–4124, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.324. URL https://aclanthology.org/2021.naacl-main.324.

[121] Tristan Thrush, Kushal Tirumala, Anmol Gupta, Max Bartolo, Pedro Rodriguez, Tariq Kane, William Gaviria Rojas, Peter Mattson, Adina Williams, and Douwe Kiela. Dynatask: A Framework for Creating Dynamic AI Benchmark Tasks. In Valerio Basile, Zornitsa Kozareva, and Sanja Stajner, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 174–181, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-demo.17. URL https://aclanthology.org/2022.acl-demo.17.

[122] R. C. Aitken. Measurement of feelings using visual analogue scales. *Proceedings of the Royal Society of Medicine*, 62(10):989–993, October 1969. ISSN 0035-9157. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1810824/.

[123] Gemini Team. Gemini: A Family of Highly Capable Multimodal Models, December 2023. URL http://arxiv.org/abs/2312.11805. arXiv:2312.11805 [cs].

[124] Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mixtral of Experts, January 2024. URL http://arxiv.org/abs/2401.04088. arXiv:2401.04088 [cs].

[125] Anthropic. Introducing the next generation of Claude, April 2024. URL https://www.anthropic.com/news/claude-3-family.

[126] Cohere. Command R, April 2024. URL https://docs.cohere.com/docs/command-r.

[127] MetaAI. Introducing Meta Llama 3: The most capable openly available LLM to date, April 2024. URL https://ai.meta.com/blog/meta-llama-3/.

[128] Prasann Singhal, Tanya Goyal, Jiacheng Xu, and Greg Durrett. A Long Way to Go: Investigating Length Correlations in RLHF, October 2023. URL http://arxiv.org/abs/2310.03716. arXiv:2310.03716 [cs].

[129] Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks, August 2019. URL http://arxiv.org/abs/1908.10084. arXiv:1908.10084 [cs].

[130] Leland McInnes, John Healy, and James Melville. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction, September 2020. URL http://arxiv.org/abs/1802.03426. arXiv:1802.03426 [cs, stat].

[131] Ricardo J. G. B. Campello, Davoud Moulavi, and Joerg Sander. Density-Based Clustering Based on Hierarchical Density Estimates. In Jian Pei, Vincent S. Tseng, Longbing Cao, Hiroshi Motoda, and Guandong Xu, editors, *Advances in Knowledge Discovery and Data Mining*, Lecture Notes in Computer Science, pages 160–172, Berlin, Heidelberg, 2013. Springer. ISBN 978-3-642-37456-2. doi: 10.1007/978-3-642-37456-2_14.

[132] Ashkan Kazemi, Kiran Garimella, Gautam Kishore Shahi, Devin Gaffney, and Scott A. Hale. Research note: Tiplines to uncover misinformation on encrypted platforms: A case study of the 2019 Indian general election on WhatsApp. *Harvard Kennedy School Misinformation Review*, January 2022. doi: 10.37016/mr-2020-91. URL https://misinforeview.hks.harvard.edu/article/research-note-tiplines-to-uncover-misinformation-on-encrypted-platforms-a-case-study-of-the-2019-indian-general-election-on-whatsapp/.

[133] Scott A. Hale. meedan/temporal_clustering, March 2022. URL https://github.com/meedan/temporal_clustering/tree/main.

[134] Scott D. Emerson, Martin Guhn, and Anne M. Gadermann. Measurement invariance of the Satisfaction with Life Scale: reviewing three decades of research. *Quality of Life Research*, 26(9):2251–2264, September 2017. ISSN 1573-2649. doi: 10.1007/s11136-017-1552-2. URL https://doi.org/10.1007/s11136-017-1552-2.

[135] Leo A. Goodman and Harry Markowitz. Social Welfare Functions Based on Individual Rankings. *American Journal of Sociology*, 58(3):257–262, 1952. ISSN 0002-9602. URL https://www.jstor.org/stable/2771835. Publisher: University of Chicago Press.

[136] John E. Roemer. *Theories of distributive justice*. Harvard Univ. Press, Cambridge, Mass., 1. harvard univ. press paperback ed edition, 1998. ISBN 978-0-674-87920-1 978-0-674-87919-5.

[137] Jeremy Bentham. An Introduction to the Principles of Morals and Legislation. In J. H. Burns and H. L. A. Hart, editors, *The Collected Works of Jeremy Bentham: An Introduction to the Principles of Morals and Legislation*. Oxford University Press, January 1789. ISBN 978-0-19-820516-6. doi: 10.1093/oseo/instance.00077240. URL http://www.oxfordscholarlyeditions.com/view/10.1093/actrade/9780198205166.book.1/actrade-9780198205166-work-1.

[138] Marc Lanctot, Kate Larson, Yoram Bachrach, Luke Marris, Zun Li, Avishkar Bhoopchand, Thomas Anthony, Brian Tanner, and Anna Koop. Evaluating Agents using Social Choice Theory, December 2023. URL http://arxiv.org/abs/2312.03121. arXiv:2312.03121 [cs] version: 2.

[139] Meriem Boubdir, Edward Kim, Beyza Ermis, Sara Hooker, and Marzieh Fadaee. Elo Uncovered: Robustness and Best Practices in Language Model Evaluation, November 2023. URL http://arxiv.org/abs/2311.17295. arXiv:2311.17295 [cs].

[140] Paul Röttger, Hannah Rose Kirk, Bertie Vidgen, Giuseppe Attanasio, Federico Bianchi, and Dirk Hovy. XSTest: A Test Suite for Identifying Exaggerated Safety Behaviours in Large Language Models, April 2024. URL http://arxiv.org/abs/2308.01263. arXiv:2308.01263 [cs].

[141] John Tasioulas. Artificial Intelligence, Humanistic Ethics. *Daedalus*, 151(2):232–243, May 2022. ISSN 0011-5266. doi: 10.1162/daed_a_01912. URL https://doi.org/10.1162/daed_a_01912.

[142] Diane Proudfoot. Anthropomorphism and AI: Turing's much misunderstood imitation game. *Artificial Intelligence*, 175(5):950–957, April 2011. ISSN 0004-3702. doi: 10.1016/j.artint.2011.01.006. URL https://www.sciencedirect.com/science/article/pii/S000437021100018X.

[143] David Watson. The Rhetoric and Reality of Anthropomorphism in Artificial Intelligence. *Minds and Machines*, 29(3):417–440, September 2019. ISSN 1572-8641. doi: 10.1007/s11023-019-09506-6. URL https://doi.org/10.1007/s11023-019-09506-6.

[144] Laura Weidinger, Jonathan Uesato, Maribeth Rauh, Conor Griffin, Po-Sen Huang, John Mellor, Amelia Glaese, Myra Cheng, Borja Balle, Atoosa Kasirzadeh, Courtney Biles, Sasha Brown, Zac Kenton, Will Hawkins, Tom Stepleton, Abeba Birhane, Lisa Anne Hendricks, Laura Rimell, William Isaac, Julia Haas, Sean Legassick, Geoffrey Irving, and Iason Gabriel. Taxonomy of Risks posed by Language Models. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 214–229, Seoul Republic of Korea, June 2022. ACM. ISBN 978-1-4503-9352-2. doi: 10.1145/3531146.3533088. URL https://dl.acm.org/doi/10.1145/3531146.3533088.

[145] Gavin Abercrombie, Amanda Cercas Curry, Tanvi Dinkar, Verena Rieser, and Zeerak Talat. Mirages. On Anthropomorphism in Dialogue Systems. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4776–4790, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.290. URL https://aclanthology.org/2023.emnlp-main.290.

[146] Myra Cheng, Kristina Gligoric, Tiziano Piccardi, and Dan Jurafsky. AnthroScore: A Computational Linguistic Measure of Anthropomorphism, February 2024. URL http://arxiv.org/abs/2402.02056. arXiv:2402.02056 [cs].

[147] Catherine D'Ignazio and Lauren F. Klein. *Data Feminism*. The MIT Press, March 2020. ISBN 978-0-262-35852-1. doi: 10.7551/mitpress/11805.001.0001. URL https://direct.mit.edu/books/book/4660/Data-Feminism.

[148] Abeba Birhane. Algorithmic injustice: a relational ethics approach. *Patterns*, 2(2):100205, February 2021. ISSN 26663899. doi: 10.1016/j.patter.2021.100205. URL https://linkinghub.elsevier.com/retrieve/pii/S2666389921000155.

[149] Debra Satz. *Why Some Things Should Not Be for Sale: The Moral Limits of Markets*. Oxford University Press, May 2010. ISBN 978-0-19-987071-4. doi: 10.1093/acprof:oso/9780195311594.001.0001. URL https://academic.oup.com/book/1465.

[150] Vinodkumar Prabhakaran, Margaret Mitchell, Timnit Gebru, and Iason Gabriel. A Human Rights-Based Approach to Responsible AI, October 2022. URL http://arxiv.org/abs/2210.02667. arXiv:2210.02667 [cs].

[151] Betty Li Hou and Brian Patrick Green. Foundational Moral Values for AI Alignment, November 2023. URL http://arxiv.org/abs/2311.17017. arXiv:2311.17017 [cs].

[152] Jean Piaget and Leslie Smith. *Sociological studies*. Routledge, London ; New York, 1995. ISBN 978-0-415-10780-8.

[153] Michael F. Mascolo and Eeva Kallio. Beyond free will: The embodied emergence of conscious agency. *Philosophical Psychology*, 32(4):437–462, May 2019. ISSN 0951-5089. doi: 10.1080/09515089.2019.1587910. URL https://doi.org/10.1080/09515089.2019.1587910. Publisher: Routledge _eprint: https://doi.org/10.1080/09515089.2019.1587910.

[154] Stevie Bergman, Nahema Marchal, John Mellor, Shakir Mohamed, Iason Gabriel, and William Isaac. STELA: a community-centred approach to norm elicitation for AI alignment. *Scientific Reports*, 14(1):6616, March 2024. ISSN 2045-2322. doi: 10.1038/s41598-024-56648-4. URL https://www.nature.com/articles/s41598-024-56648-4. Publisher: Nature Publishing Group.

[155] Billy Perrigo. Inside OpenAI's Plan to Make AI More 'Democratic', February 2024. URL https://time.com/6684266/openai-democracy-artificial-intelligence/.

[156] R. Silberzahn, E. L. Uhlmann, D. P. Martin, P. Anselmi, F. Aust, E. Awtrey, Š. Bahník, F. Bai, C. Bannard, E. Bonnier, R. Carlsson, F. Cheung, G. Christensen, R. Clay, M. A. Craig, A. Dalla Rosa, L. Dam, M. H. Evans, I. Flores Cervantes, N. Fong, M. Gamez-Djokic, A. Glenz, S. Gordon-McKeon, T. J. Heaton, K. Hederos, M. Heene, A. J. Hofelich Mohr, F. Högden, K. Hui, M. Johannesson, J. Kalodimos, E. Kaszubowski, D. M. Kennedy, R. Lei, T. A. Lindsay, S. Liverani, C. R. Madan, D. Molden, E. Molleman, R. D. Morey, L. B. Mulder, B. R. Nijstad, N. G. Pope, B. Pope, J. M. Prenoveau, F. Rink, E. Robusto, H. Roderique, A. Sandberg, E. Schlüter, F. D. Schönbrodt, M. F. Sherman, S. A. Sommer, K. Sotak, S. Spain, C. Spörlein, T. Stafford, L. Stefanutti, S. Tauber, J. Ullrich, M. Vianello, E.-J. Wagenmakers, M. Witkowiak, S. Yoon, and B. A. Nosek. Many Analysts, One Data Set: Making Transparent How Variations in Analytic Choices Affect Results. *Advances in Methods and Practices in Psychological Science*, 1(3):337–356, September 2018. ISSN 2515-2459. doi: 10.1177/2515245917747646. URL https://doi.org/10.1177/2515245917747646. Publisher: SAGE Publications Inc.

[157] Nenad Tomasev, Kevin R. McKee, Jackie Kay, and Shakir Mohamed. Fairness for Unobserved Characteristics: Insights from Technological Impacts on Queer Communities. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '21, pages 254–265, New York, NY, USA, July 2021. Association for Computing Machinery. ISBN 978-1-4503-8473-5. doi: 10.1145/3461702.3462540. URL https://doi.org/10.1145/3461702.3462540.

[158] Tom Hosking, Phil Blunsom, and Max Bartolo. Human Feedback is not Gold Standard, January 2024. URL http://arxiv.org/abs/2309.16349. arXiv:2309.16349 [cs].

[159] Amos Tversky and Itamar Simonson. Context-Dependent Preferences. *Management Science*, 39(10):1179–1189, October 1993. ISSN 0025-1909. doi: 10.1287/mnsc.39.10.1179. URL https://pubsonline.informs.org/doi/abs/10.1287/mnsc.39.10.1179. Publisher: INFORMS.

[160] Boaz Shmueli, Jan Fell, Soumya Ray, and Lun-Wei Ku. Beyond fair pay: Ethical implications of NLP crowdsourcing. In *Association for Computational Linguistics (ACL)*, pages 3758–3769, April 2021. URL https://arxiv.org/abs/2104.10097v1. arXiv: 2104.10097 Publisher: tex.arxivid: 2104.10097.

[161] Lisa Posch, Arnim Bleier, Fabian Flöck, Clemens M. Lechner, Katharina Kinder-Kurlanda, Denis Helic, and Markus Strohmaier. Characterizing the Global Crowd Workforce: A Cross-Country Comparison of Crowdworker Demographics. *Human Computation*, 9(1), August 2022. ISSN 2330-8001. doi: 10.15346/hc.v9i1.106. URL http://arxiv.org/abs/1812.05948. arXiv:1812.05948 [cs].

[162] Derek A. Albert and Daniel Smilek. Comparing attentional disengagement between Prolific and MTurk samples. *Scientific Reports*, 13(1):20574, November 2023. ISSN 2045-2322. doi: 10.1038/s41598-023-46048-5. URL https://www.nature.com/articles/s41598-023-46048-5. Publisher: Nature Publishing Group.

[163] Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. Whose Opinions Do Language Models Reflect?, March 2023. URL http://arxiv.org/abs/2303.17548. arXiv:2303.17548 [cs].

[164] Fabio Crameri, Grace E. Shephard, and Philip J. Heron. The misuse of colour in science communication. *Nature Communications*, 11(1):5444, October 2020. ISSN 2041-1723. doi: 10.1038/s41467-020-19160-7. URL https://www.nature.com/articles/s41467-020-19160-7. Number: 1 Publisher: Nature Publishing Group.

[165] Fredrik Barth. *Ethnic Groups and Boundaries: The Social Organization of Culture Difference*. Waveland Press, March 1998. ISBN 978-1-4786-0795-3. Google-Books-ID: QaAQAAAAQBAJ.

[166] Katarzyna Hamer, Sam McFarland, Barbara Czarnecka, Agnieszka Golińska, Liliana Manrique Cadena, Magdalena Łużniak Piecha, and Tomasz Jułkowski. What Is an "Ethnic Group" in Ordinary People's Eyes? Different Ways of Understanding It Among American, British, Mexican, and Polish Respondents. *Cross-Cultural Research*, 54(1):28–72, February 2020. ISSN 1069-3971. doi: 10.1177/1069397118816939. URL https://doi.org/10.1177/1069397118816939. Publisher: SAGE Publications Inc.

[167] Karen L. Suyemoto, Micaela Curley, and Shruti Mukkamala. What Do We Mean by "Ethnicity" and "Race"? A Consensual Qualitative Research Investigation of Colloquial Understandings. *Genealogy*, 4 (3):81, September 2020. ISSN 2313-5778. doi: 10.3390/genealogy4030081. URL https://www.mdpi.com/2313-5778/4/3/81. Number: 3 Publisher: Multidisciplinary Digital Publishing Institute.

[168] Laurence R. Iannaccone. Introduction to the Economics of Religion. *Journal of Economic Literature*, 36 (3):1465–1495, 1998. ISSN 0022-0515. URL https://www.jstor.org/stable/2564806. Publisher: American Economic Association.

[169] Gilat Levy and Ronny Razin. Religious Beliefs, Religious Participation, and Cooperation. *American Economic Journal: Microeconomics*, 4(3):121–151, August 2012. ISSN 1945-7669, 1945-7685. doi: 10.1257/mic.4.3.121. URL https://pubs.aeaweb.org/doi/10.1257/mic.4.3.121.

[170] Ellen Dingemans and Erik Van Ingen. Does Religion Breed Trust? A Cross-National Study of the Effects of Religious Involvement, Religious Faith, and Religious Context on Social Trust. *Journal for the Scientific Study of Religion*, 54(4):739–755, 2015. ISSN 0021-8294. URL https://www.jstor.org/stable/26651394. Publisher: [Society for the Scientific Study of Religion, Wiley].

[171] Hansong Zhang, Joshua N. Hook, Jennifer E. Farrell, David K. Mosher, Laura E. Captari, Steven P. Coomes, Daryl R. Van Tongeren, and Don E. Davis. Exploring Social Belonging and Meaning in Religious Groups. *Journal of Psychology and Theology*, 47(1):3–19, March 2019. ISSN 0091-6471. doi: 10.1177/0091647118806345. URL https://doi.org/10.1177/0091647118806345. Publisher: SAGE Publications Ltd.

[172] Vassilis Saroglou, Magali Clobert, Adam B. Cohen, Kathryn A. Johnson, Kevin L. Ladd, Matthieu Van Pachterbeke, Lucia Adamovova, Joanna Blogowska, Pierre-Yves Brandt, Cem Safak Çukur, Kwang-Kuo Hwang, Anna Miglietta, Frosso Motti-Stefanidi, Antonio Muñoz-García, Sebastian Murken, Nicolas Roussiau, and Javier Tapia Valladares. Believing, Bonding, Behaving, and Belonging: The Cognitive, Emotional, Moral, and Social Dimensions of Religiousness across Cultures. *Journal of Cross-Cultural Psychology*, 51(7-8):551–575, September 2020. ISSN 0022-0221. doi: 10.1177/0022022120946488. URL https://doi.org/10.1177/0022022120946488. Publisher: SAGE Publications Inc.

[173] Prolific. Representative samples, February 2024. URL https://researcher-help.prolific.com/hc/en-gb/articles/360019236753-Representative-samples.

[174] Sahand Negahban, Sewoong Oh, and Devavrat Shah. Rank Centrality: Ranking from Pairwise Comparisons. *Operations Research*, 65(1):266–287, 2017. ISSN 0030-364X. URL https://www.jstor.org/stable/26153541. Publisher: INFORMS.

Supplementary Material

Table of Contents

A PRISM Data Access and Format

The data can be accessed on Github at `https://github.com/HannahKirk/prism-alignment`, and also on HuggingFace at `https://huggingface.co/datasets/HannahRoseKirk/prism-alignment`. The dataset has a permanent DOI: `10.57967/hf/2113`.

There dataset is organised in two primary JSON lines files:

- **The Survey** (`survey.jsonl`): The survey where participants answer questions such as their stated preferences for LLM behaviours, their familarity with LLMs, a self-description and some basic demographics. Each row is a single participant in our dataset, identified by a `user_id`.
- **The Conversations** (`conversations.jsonl`): Each participants' multiple conversation trees with LLMs and associated feedback. Each row is a single conversation, identified by a `conversation_id`, that can be matched back to a participant's survey profile via the `user_id`. The conversation itself is stored as a list of dictionaries representing human and model turns in the `conversation_history` column, which broadly follows the format of widely used Chat APIs (see single entry schema on the next page).

Additionally, for ease of secondary analysis we provide a more granular and flattened format of the conversations data:

- **The Utterances** (`utterances.jsonl`): Each row is a single scored utterance (human input - model response - score). Each row has an `utterance_id` that can be mapped back to the conversation data using `conversation_id` or the survey using `user_id`. The model responses and scores per each user input are in *long format*. Because of this format, the user inputs will be repeated for the set of model responses in a single interaction turn.

We also provide code for transforming the conversations to a *wide format*. That is, each row is now a single turn within a conversation. For the first interaction where up to four models respond, we have `model_{a/b/c/d}` as four distinct columns and `score_{a/b/c/d}` as another four columns. Note that for subsequent turns, the same model responds and there are only two responses so `model/score_{c/d}` will always be missing.

Finally, for every text instance in PRISM, we provide metadata on the language detection, personal or private information (PII) detection and moderation flags. **The Metadata** is provided seperately to the main data files (`metadata.jsonl`).

We provide **codebooks** for **The Survey** (App. X.1), **The Conversations** (App. X.2), **The Utterances** (App. X.3) and **The Metadata** (App. X.4).

Format of Entries in Conversations Data

```
{
  "conversation_id": "c1",
  "user_id": "user123",
  "conversation_type": ["unguided", "values guided", "controversy guided"],
  "opening_prompt": "[USER PROMPT]",
  "conversation_turns": [2-22],
  "conversation_history": [
    {
      "turn": 0,
      "role": "user",
      "content": "[USER PROMPT]"
    },
    {
      "turn": 0,
      "role": "model",
      "content": "[MODEL RESPONSE]",
      "model_name": "M1",
      "model_provider": "P1",
      "score": [1-100],
      "if_chosen": false,
      "within_turn_id":0
    },
    {
      "turn": 0,
      "role": "model",
      "content": "[MODEL RESPONSE]",
      "model_name": "M2",
      "model_provider": "P2",
      "score": [1-100],
      "if_chosen": true,
      "within_turn_id":1

    },
    //... Additional list items for remaining model responses (up to 4 in total)
    {
      "turn": 1,
      "role": "user",
      "content": "[USER PROMPT]"
    },
    {
      "turn": 1,
      "role": "model",
      "content": "[MODEL RESPONSE]",
      "model_name": "M2",
      "model_provider": "P2",
      "score": [1-100],
      "if_chosen": true,
      "within_turn_id":0
    },
    {
      "turn": 1,
      "role": "model",
      "content": "[MODEL RESPONSE]",
      "model_name": "M2",
      "model_provider": "P2",
      "score": [1-100],
      "if_chosen": false,
      "within_turn_id":1
    }
    //... Additional turns follow the same pattern as turn 1
  ],
  "performance_attributes": {
    "fluency": [1-100],
    "factuality": [1-100],
    "helpfulness": [1-100],
    //....Additional attribute ratings
  },
  "open_feedback": "[FREE-TEXT]"
}
```

B PRISM Data Statement

We provide a data statement [69] to document the generation and provenance of PRISM.

B.1 Curation Rationale

The PRISM Alignment Project, funded by a variety of academic and industry sources (see Disclosure of Funding), aims to diversify human feedback datasets. All participants are recruited via the Prolific platform. The sample is described in § 3.3, with additional details in App. J. The primary purpose of the dataset is for academic research into how different people interact with LLMs and perceive their outputs. However, we do not prohibit the use of the dataset to develop, test and/or evaluate AI systems so long as usage complies with the dataset license (App. C.2).

B.2 Language Variety

The language of human- or model-written text was not explicitly restricted to English. However, the task instructions were written English, and fluency in English was included as a screening filter. As a result of these factors, 99% of text instances are in English (see App. E for breakdowns per type of text instance and by other language). There is scope for wide social and regional variation even within a language. Given we have speakers residing in 38 countries (born in 75 countries), we likely have various forms of English, especially by level of fluency (see Tab. 5). Information about which varieties of English are represented is not available.

B.3 Speaker Demographics

There are two sets of "speaker" roles in PRISM: human participants and large language models (LLMs). Both roles contribute to the characteristics of the text utterances in the dataset.

Participant Characteristics We provide full demographic breakdowns of participant characteristics in Tab. 5. We provide full geographic breakdowns in Tab. 8. Despite substantial improvements on sample diversity compared to early widely-used human feedback datasets (see Tab. 6, Tab. 7), PRISM still skews White, Educated, and Western. This is partly driven by census-representative samples from the US and UK, which can be removed or downsampled for future research. PRISM only contains participants sourced from one crowdworking platform (Prolific), so inherits sample biases from this narrow pool—for example, participants are active internet users, incentivised by hourly payment on a specific task that they self-select into.

Model Characteristics Given fast-paced changes to the LLM landscape, PRISM is designed to be as *model-agnostic* as possible. We include 21 models from various different families, capabilities and sizes (for a summary see Tab. 25). 12/21 models are accessed via commerical APIs, and 9/21 are open-access via HuggingFace. Model-specific characteristics will affect the text characteristics, especially if they have already been alignment-tuned.

Models as Participants Throughout the study we strongly requested that participants did not use LLMs to write their "human" responses, playing both to their integrity (please don't do it), their role in the research (we really need you to not do it), and their incentives (you won't be paid if you do it). We did not directly test nor implement tools to technologically prevent participants from using LLMs on their behalf. We randomly sample 25 instances from human-written texts: system strings and self-descriptions from the Survey; opening prompts and open feedback from the Conversations ($n = 100$). An annotator (paper author) manually inspected these and labelled none as model-written text. For instances of sufficient length (46/100, >50 words), we recorded the predicted probability of AI-generated text from an LLM-text detector, where 76% had $\leq 1\%$ score.[30] For the remainder ($n = 11$), a second annotator (paper author) gave a tie-break, labelling none as model-generated.

[30]The tool is developed by `https://sapling.ai/`. LLM-detector tools are susceptible to misclassifications. For example, this feedback: "*It was good that it offered options and mentioned "options" rather than just suggesting one thing. It would have been better to state in the beginning how dietary requirements and preferences might play a big role in the decision what to cook for dinner. And also to point out how different cultures have different food traditions. Not everything is US based.*" was flagged as 88.1% AI-generated, but the human annotators felt was strongly human-generated.

B.4 Annotator Demographics

The "annotators" are "speakers"—the same human participants who answer the survey, interact with the LLMs, and provide structured and unstructured feedback. See App. B.3.

B.5 Speech Situation

All participants were recruited via Prolific. They were paid £9/hour. The survey was hosted on Qualtrics (www.qualtrics.com), and the conversations on Dynabench (www.dynabench.org).

All data was collected between 22nd November 2023 and 22nd December 2023. The time of the data collection period did affect the topics of discussion: for example, one topic concerns Christmas holiday celebrations while another discusses the Israel–Palestine Conflict.

The primary modality of PRISM is written language, combined with structured ratings or structured survey data. The conversations between participants and LLMs happened *synchronously* via live API connections with models in the backend of our interface. We have not edited or moderated any survey responses, participant prompts or model responses. All conversations happened as part of this research project, so the primary 'intended audience' was the researchers, though participants were informed of additional plans to distribute and release the data in the consent form (see App. D).

B.6 Text Characteristics

We summarise text characteristics in App. M. For the survey responses, the text provides details on the participant and their views about LLMs via short-form free-text responses (we requested 2-5 sentences in their own words). For the conversations, there are three different types: unguided, values guided and controversy guided, as described in the main paper (§ 3.2). Each conversation type contains a different distribution of topics. Overall, PRISM is skewed towards subjective, values-driven and controversial dialogue. The texts within a conversation consist of short, usually single-sentence prompts written by human participants, on average 13 words long. Prompts receive up to four model responses generated by a variety of LLMs. We instruct the LLMs to limit their response to 50 words or less. Most unsuccessfully abide by this instruction: the average response length is 89 words. We release metadata (see App. E) with each text instance including information on detected language, automated and manual PII checks and moderation flags (e.g., if it contains sexual, hateful or violent content).

B.7 Recording Quality

During data collection, our interface experienced two distributed denial of service (DDoS) attacks: one on 28th November 2023 and another on 1st December 2023. The primary way that these attacks may have affected recording quality was via interrupting participants' conversation sessions (most then later returned to the interface to complete their conversations a couple hours or days later). These participants' data points may differ to those who had a smoother continuous experience in the task.

B.8 Author Characteristics and Positionality Statement

We aimed to operate in the subjective paradigm [11, 12] and have as little influence as possible on how participants interacted with models (e.g., no annotation guidelines for how to rate responses). As a team of researchers, we come from a variety of backgrounds (genders, ethnicities, countries of birth, native languages) and are involved with AI research, either in an academia (6/12) or industry (6/12).

C PRISM Data Clause

C.1 Terms of Use

Purpose The Dataset is provided for the purpose of research and educational use in the field of natural language processing, conversational agents, social science and related areas; and can be used to develop or evaluate artificial intelligence, including Large Language Models (LLMs).

Usage Restrictions Users of the Dataset should adhere to the terms of use for a specific model when using its generated responses. This includes respecting any limitations or use case prohibitions set forth by the original model's creators or licensors.

Content Warning The Dataset contains raw conversations that may include content considered unsafe or offensive. Users must apply appropriate filtering and moderation measures when using this Dataset for training purposes to ensure the generated outputs align with ethical and safety standards.

No Endorsement of Content The conversations and data within this Dataset do not reflect the views or opinions of the Dataset creators, funders or any affiliated institutions. The dataset is provided as a neutral resource for research and should not be construed as endorsing any specific viewpoints.

No Deanonymisation The User agrees not to attempt to re-identify or de-anonymise any individuals or entities represented in the Dataset. This includes, but is not limited to, using any information within the Dataset or triangulating other data sources to infer personal identities or sensitive information.

Limitation of Liability The authors and funders of this Dataset will not be liable for any claims, damages, or other liabilities arising from the use of the dataset, including but not limited to the misuse, interpretation, or reliance on any data contained within.

C.2 Licence and Attribution

Human-written texts (including prompts) within the dataset are licensed under the Creative Commons Attribution 4.0 International License (CC-BY-4.0). Model responses are licensed under the Creative Commons Attribution-NonCommercial 4.0 International License (CC-BY-NC-4.0). Use of model responses must abide by the original model provider licenses.

For proper attribution when using this dataset in any publications or research outputs, please cite with the DOI.
Suggested Citation: Kirk, H. R., Whitefield, A., Röttger, P., Bean, A., Margatina, K., Ciro, J., Mosquera, R., Bartolo, M., Williams, A., He, H., Vidgen, B., & Hale, S. A. (2024). *The PRISM Alignment Dataset*. `https://doi.org/10.57967/hf/2113`

C.3 Dataset Maintenance

As the authors and maintainers of this dataset, we commit to no further updates to the dataset following its initial release. The dataset is self-contained and does not rely on external links or content, ensuring its stability and usability over time without the need for ongoing maintenance.

C.4 Data Rights Compliance and Issue Reporting

We are committed to complying with data protection rights, including but not limited to regulations such as the General Data Protection Regulation (GDPR). If any individual included in the dataset wishes to have their data removed, we provide a straightforward process for issue reporting and resolution on our Github. Concerned parties are encouraged to contact the authors directly via the provided contact form link on the Github. Upon receiving a request, we will engage with the individual to verify their identity and proceed to remove the relevant entries from the dataset. We commit to addressing and resolving such requests within 30 days of verification.

D Informed Consent

This research was reviewed by, and received ethics clearance through, a subcommittee of the University of Oxford Central University Research Ethics Committee [OII_C1A_23_088]. The following text was displayed to all participants to collect informed consent.

Your Feedback on AI Language Models

We appreciate your interest in participating in this study. **The aim of this research is to understand people's preferences and perceptions regarding AI Language Model behaviours**, also referred to as Large Language Models (LLMs), Generative AI Language Models, AI ChatBots or Virtual Assistants. AI language models are computer programs designed to generate text. They can respond to questions or prompts by producing written responses. We want to learn more about how people like you use and perceive these AI language models.

Please first make sure you are using a laptop or desktop computer, and you are not using a mobile device. Our task is NOT compatible with mobile devices.

Please then read through this information before agreeing to participate (if you wish to).

You may ask any questions before deciding to take part by contacting the research team. The Principal Researcher is Hannah Rose Kirk, and the Principle Investigator is Dr Scott. A. Hale, who are both affiliated with the Oxford Internet Institute at the University of Oxford.

What does the task involve?

If you decide to participate, there are two stages.

In this stage, you will be asked to fill in a short survey about yourself and your thoughts on AI language models.

In the next stage, you will have conversations with AI language models by providing prompts and rating their responses using a user-friendly interface. The prompts can be on various topics, and you don't need any specific knowledge to participate. Your input will help us understand your preferences and opinions about how these AI language models work.

Both stages should take between 55-65 minutes. No background knowledge is required.

Please note that you will be interacting with an AI language model. The research team cannot directly control and are not responsible for the text generated by these models. There is a possibility that the models produce biased, inaccurate or harmful language. The risks to you as an individual are equivalent to those you would be exposed to if you use AI language models via interfaces like ChatGPT.

Do I have to take part?

No, participation is voluntary. If you do decide to take part, you may stop at any point for any reason before submitting your answers by closing the browser. However, we are only able to pay participants who complete the task. For demographic information, we have included a 'Prefer not to say' option for each set of questions should you prefer not to answer a particular question.

Can I withdraw my participation and data?

Yes, you may stop the study at any time. Please note that if you withdraw within a stage of the study you will not be paid for that stage or any subsequent incomplete stages, but you will be paid for any stages that you have already completed. You can withdraw your data from the study. The cut-off date for withdrawing your data is 14 days after you submitted the data. Please email members of the research team (see contact details below) within this 14-day window to withdraw your data from the study.

How will my data be used?

The data collected from your participation will be pseudo-anonymized (stored with a unique numeric ID) and stored securely. It will be used for research purposes, and your personal information will remain confidential. The data will be analysed to gain insights into diverse preferences and perceptions regarding AI language model behaviours. At the end of the study, the pseudo-anonymised data collected will be released publicly for future research. The findings of this study may be published in academic journals or presented at conferences, and the results will be written up for a DPhil degree. Your individual identity will not be disclosed at any point in data release or publication. We do not collect any personal, private identifying information, IP addresses or contact details. The data we will collect that could identify you will be some demographic information (gender, age, nationality, religion, etc.), and short self-written survey answers.

The responses you provide will be stored in a password-protected electronic file on University of Oxford secure servers and may be used in academic publications, conference presentations or reports for external organisations. We will release a clean, PII-checked and pseudo-anonymised form of the data on an open-access, public data repository. Raw research data will be stored for 3 years after publication or public release of the research. We would like to use the data in future studies, and to share data with other researchers (e.g. in online databases). Data will have identifying information removed before it is shared with other researchers or results are made public. The data that we collect from you may be transferred to, stored and/ or processed at a destination outside the UK and the European Economic Area. By submitting your personal data, you agree to this transfer, storing or processing.

Who has reviewed this research?

This research has been reviewed by, and received ethics clearance through, a subcommittee of the University of Oxford Central University Research Ethics Committee [OII_C1A_23_088].

Who do I contact if I have a concern or I wish to complain?

If you have a concern about any aspect of this research, please speak to Hannah Rose Kirk (hannah.kirk@oii.ox.ac.uk) or their supervisor Dr. Scott A. Hale (scott.hale@oii.ox.ac.uk), and we will do our best to answer your query. We will acknowledge your concern within 10 working days and give you an indication of how it will be dealt with. If you remain unhappy or wish to make a formal complaint, please contact the Chair of the Research Ethics Committee at the University of Oxford who will seek to resolve the matter as soon as possible: Social Sciences & Humanities Interdivisional Research Ethics Committee; Email: ethics@socsci.ox.ac.uk; Address: Research Services, University of Oxford, Boundary Brook House, Churchill Drive, Headington, Oxford OX3 7GB.

Please note that you may only participate in this survey if you are 18 years of age or over.

◯ I certify that I am 18 years of age or over

If you have read the information above and agree to participate with the understanding that the data (including any personal data) you submit will be processed accordingly, please tick the box below to start.

◯ Yes, I agree to take part

Table 1: Identifiers of text instance types in PRISM.

Text Instance	Study Stage	user	convo (interaction)	turn (interaction)	within turn
self_description	Survey	✓			
system_string	Survey	✓			
user_prompt	Conversations		✓	✓	
model_response	Conversations		✓	✓	✓
open_feedback	Conversations		✓		

E Metadata Processing

For each text instance in PRISM, we attach three pieces of metadata: detected **language** flags, detected **private or personally identifiable information (PII)** flags, and detected **moderation** flags.

E.1 Structuring the Metadata

There are five types of text instances. Two appear in the survey (`self_description`, `system_string`) and have a 1:1 matching with each user (`user_id`). One appears at the conversation level (`open_feedback`) and has a 1:1 matching with each `convo_id` and a many:1 matching with each `user_id` because each participant has multiple conversations. Finally, the last two occur within each turn of a conversation, where for each single `user_prompt` there are multiple model responses (`model_response`). We structure the metadata so it can be merged uniquely, without duplication. We release one file, where each text instance is tied to its metadata via the identifying information shown in Tab. 1, and a `column_id` for matching whether the text is [`system_string, self_description, user_prompt, model_response, open_feedback`].

E.2 Automated Flagging

PII To identify whether a textual instance in our dataset contains personal and identifiable information (PII) we used the package `scrubadub`.[31] Specifically we used the function `scrubadub.clean(text)` which replaces the phone numbers and email addresses with anonymous IDs, if they are found in the input. We flag with 1 instances that are altered (i.e., PII was identified) and 0 those that remained unchanged.

Moderation To measure content moderation we use the OpenAI Moderation endpoint.[32] The API takes as an input a textual instance and outputs a json file with an overall boolean flag (`flagged`) whether there input potentially harmful (`True`), otherwise `False`. The API also returns a flag for a list of specific moderation categories that can be used to further filter and inspect the data. The categories are sexual, hate, harassment, self-harm, sexual/minors, hate/threatening, violence/graphic, self-harm/intent, self-harm/instructions, harassment/threatening and violence. Similar to the overall flag, for each category, the value is `True` if the model flags the corresponding category as violated, `False` otherwise. Finally, the API returns a dictionary of per-category scores that denote the model's confidence that the input violates the OpenAI's policy for the category. The value is between 0 and 1, where higher values denote higher confidence.

Language Detection To detect the language of each text instance in our dataset we used the `LangID` codebase.[33] `LangID` is a popular python package that efficiently detects the language of an input and currently supports 97 languages. Specifically, we use the `langid.classify(text)` function and store a string for the detected language.

[31] `https://scrubadub.readthedocs.io/en/stable/`

[32] `https://platform.openai.com/docs/guides/moderation`

[33] `https://github.com/saffsd/langid.py`

E.3 Manual Review

The overall proportions of texts flagged for PII, Non-English or Moderation is low (see Tab. 2). However, when inspecting the few positive flags, many were false positives, especially on lang-detect and PII. While a false positive on language may be relatively inconsequential, any automated flags of PII are concerning. Accordingly, we manually annotate any instances where `pii_flag==True` ($n = 167$) for participant-written text. We find that none of them actually contain PII.[34]

Table 2: **Meta-Data Summary.** For each metadata category (language, PII and moderation), we show $N(\%)$ for the dataset as a whole (*Overall*) and broken down by each type of text instance in PRISM.

Category	Is English	Contains PII	Manually-Checked PII	Is Moderation Flagged	*Total Instances*
Overall	105,229 (98.8%)	1,111 (1.0%)	NA	634 (0.6%)	106,554 (100.0%)
`user_prompt`	26,545 (97.7%)	66 (0.2%)	0 (0.0%)	454 (1.7%)	27,172 (100.0%)
`model_response`	67,715 (99.0%)	944 (1.4%)	NA	162 (0.2%)	68,371 (100.0%)
`self_description`	1,496 (99.7%)	10 (0.7%)	0 (0.0%)	7 (0.5%)	1,500 (100.0%)
`system_string`	1,493 (99.5%)	16 (1.1%)	0 (0.0%)	0 (0.0%)	1,500 (100.0%)
`open_feedback`	7,980 (99.6%)	75 (0.9%)	0 (0.0%)	11 (0.1%)	8,011 (100.0%)

Table 3: **Breakdown of flags from the OpenAI Moderation API.** We show counts and percentages where the text was flagged (==True), as well as total counts. Human-written text includes `user_prompt, self_description, system_string, open_feedback`; Model-written text is only `model_response`.

	Human-written		**Model-written**	
	N	(%)	N	(%)
sexual	21	0.05%	11	0.02%
hate	154	0.40%	36	0.05%
harassment	387	1.01%	127	0.19%
self-harm	24	0.06%	8	0.01%
sexual/minors	4	0.01%	4	0.01%
hate/threatening	17	0.04%	2	0.00%
self-harm/intent	26	0.07%	5	0.01%
self-harm/instructions	13	0.03%	8	0.01%
harassment/threatening	33	0.09%	7	0.01%
violence	52	0.14%	13	0.02%
Total	38,183	100.00%	68,371	100.00%

Table 4: **Breakdown of languages detected by LangID.** We show the top-10 detected languages, then other and total counts. Human-written text includes `user_prompt, self_description, system_string, open_feedback`; Model-written text is only `model_response`.

	Human-written			**Model-written**		
	Language	N	(%)	Language	N	(%)
1	en	37,514	98.25%	en	67,715	99.04%
2	es	175	0.46%	es	236	0.35%
3	fr	71	0.19%	de	67	0.10%
4	it	70	0.18%	fr	63	0.09%
5	de	60	0.16%	la	43	0.06%
6	nl	42	0.11%	nl	37	0.05%
7	pt	41	0.11%	it	29	0.04%
8	pl	34	0.09%	sl	19	0.03%
9	da	24	0.06%	hu	15	0.02%
10	ro	13	0.03%	sv	14	0.02%
	Other	139	0.36%	Other	133	0.19%
	Total	38,183	100.00%	Total	68,371	100.00%

[34] If any of these human-written prompts had been a true positive, we would have manually checked the associated model responses too.

F Annotating Ethnicity, Religion and Gender

We ask people to describe their ethnic and religious affiliations in their own words because for a global survey, there are no immediately obvious preset categories. In the survey data, we release this original self-description (`{ethnicity, religion}_self_described`). However, there are 264 unique strings for ethnicity, and 137 unique strings for religion. For some analysis, it is valuable to have aggregate groupings. To attain this grouping, we first used `gpt-4-turbo` to categorise the strings, but found some errors and essentialising generalisations, for example, if someone answered with a nationality not an ethnic group like *american*, `gpt` would return *white*.[35]

Accordingly, we used a second round of manual human annotation to verify these automated labels. Two annotators (authors of the paper) first made independent judgements then discussed any disagreements. For ethnicity, some participants also had answered a Prolific screening question on their simplified ethnicity, though we did not have this information for all participants as it was not mandatory. We thus annotate all unique combinations of the self-described string, and the Prolific ethnicity information ($n = 343$). In ambiguous cases (e.g., the aforementioned *american* response), we relied on this additional ethnicity information, and in its absence, defaulted to a *Prefer not to say* response. For religion, we do not have any additional information provided by the Prolific pre-screening questionnaire, so verification decisions were made on the basis of the self-describe string alone. The annotators agreed on 94% of ethnicity cases (discussing and resolving the remaining 20); and 96% of religion cases (discussing and resolving the remaining 5).

We highlight two general findings from our disagreements which may be of interest to people analysing or categorising our data in the future. Firstly, **ethnicity and nationality are complex**. Take for example the UK census, where *Chinese*, *Banglaeshi*, *Indian* and *Pakistani* are all listed as subcategories of the Asian ethnic group.[36] Ethnicity is a multi-faceted term which can include nationality, language group, skin colour, religion, among other characteristics [165]. Studies have shown that survey participants can interpret the term ethnic group through a variety of subjective lens [166, 167]. During annotation, we tried to gather information on whether group terms commonly refer to an ethnic group, but some subjectivity and naivety are inevitable; so, we encourage future researchers to carefully consider their own categorisations depending on the question at hand. Secondly, the **belonging and believing aspects of religion intersect** [168, 169], and it is not immediately clear how to categorise an individual that culturally affiliates with religion but simultaneously identities as an atheist or non-believer. Studies have revealed that the belonging and believing axis of religion are important for conditioning behaviours such as trust, pro-socality and altruism [170–172]. In general, we annotated a mention of a religion as assigned to that religion (not distinguishing between the belonging and believing channels) but it remains to be seen whether one axis is more salient for values and opinions towards AI systems.

Note for gender, we provided a standard multiple choice question with options: *Female*, *Male*, *Non-binary / third gender*, *Prefer not to say* and *Prefer to self-describe*. Only 3 individuals opted to self-describe, which we then annotated and only assimilated in very clear cut cases,[37] else we grouped it as *Prefer not to say* to avoid over-riding a participant's self-identification.

G Participant Demographics

We present full demographic breakdowns in Tab. 5. We also compare the breakdowns in PRISM to some early human feedback datasets which provide demographic information (Tab. 6).

[35] As an aside, these types of baked-in priors are a good example of why using LLMs as a surrogate for human annotators may introduce downstream biases [57].

[36] See the fact sheet at ethnicity-facts-figures.service.gov.uk.

[37] For example, one participant responded with "i dont expect this wokery from intelligent people. you want to know which of the 2 possible genders i am male.", which we assign as *Male*.

Table 5: **Full Demographics Breakdowns.** Counts and percentages of participants by standard demographic variables. Overall, PRISM utilises a large and demographically-diverse sample, especially compared to some previous human feedback datasets (see Tab. 6); but it still generally skews towards young, white and educated populations. *For ethnicity and religion, see details in App. F.

Total Participants	**1,500**	100
With conversations	1,396	93.1 %
Just survey	104	6.9 %
Age		
25-34 years old	454	30.3 %
18-24 years old	297	19.8 %
35-44 years old	237	15.8 %
45-54 years old	208	13.9 %
55-64 years old	197	13.1 %
65+ years old	106	7.1 %
Prefer not to say	1	0.1 %
Gender		
Male	757	50.5 %
Female	718	47.9 %
Non-binary / third gender	21	1.4 %
Prefer not to say	4	0.3 %
Self-Reported Ethnicity*		
White	969	64.6 %
Black / African	122	8.1 %
Hispanic / Latino	121	8.1 %
Asian	95	6.3 %
Mixed	68	4.5 %
Middle Eastern / Arab	14	0.9 %
Indigenous / First Peoples	8	0.5 %
Other	17	1.1 %
Prefer not to say	86	5.7 %
Self-Reported Religion*		
Non-religious	762	50.8 %
Christian	487	32.5 %
Agnostic	71	4.7 %
Jewish	42	2.8 %
Muslim	31	2.1 %
Spiritual	18	1.2 %
Buddhist	12	0.8 %
Folk religion	6	0.4 %
Hindu	5	0.3 %
Sikh	3	0.2 %
Other	4	0.3 %
Prefer not to say	59	3.9 %
Employment Status		
Working full-time	712	47.5 %
Working part-time	265	17.7 %
Student	191	12.7 %
Unemployed, seeking work	113	7.5 %
Retired	104	6.9 %
Homemaker / Stay-at-home parent	46	3.1 %
Unemployed, not seeking work	46	3.1 %
Prefer not to say	23	1.5 %

Continued on next page

Table 5: **Full Demographics Breakdowns.** Counts and percentages of participants by standard demographic variables. Overall, PRISM utilises a large and demographically-diverse sample, especially compared to some previous human feedback datasets (see Tab. 6); but it still generally skews towards young, white and educated populations. *For ethnicity and religion, see details in App. F.

Education		
University Bachelors Degree	637	42.5 %
Graduate / Professional degree	241	16.1 %
Some University but no degree	236	15.7 %
Completed Secondary School	209	13.9 %
Vocational	125	8.3 %
Some Secondary	24	1.6 %
Completed Primary School	16	1.1 %
Some Primary	3	0.2 %
Prefer not to say	9	0.6 %
Martial Status		
Never been married	870	58.0 %
Married	463	30.9 %
Divorced / Separated	123	8.2 %
Widowed	21	1.4 %
Prefer not to say	23	1.5 %
English Proficiency		
Native speaker	886	59.1 %
Fluent	405	27.0 %
Advanced	160	10.7 %
Intermediate	42	2.8 %
Basic	7	0.5 %
Regions		
US	338	22.5 %
Europe	313	20.9 %
UK	292	19.5 %
Latin America and the Caribbean	146	9.7 %
Australia and New Zealand	129	8.6 %
Africa	118	7.9 %
Asia	60	4.0 %
Northern America	50	3.3 %
Middle East	50	3.3 %
Oceania	1	0.1 %
Prefer not to say	3	0.2 %

Table 6: **Demographic data compared to prior work.** Comparisons of PRISM to early and widely-known RLHF studies using human feedback for language models. See § 2 for more current datasets.

Category	Bai et al.	Ouyang et al.	Glaese et al.	Ganguli et al.	Stiennon et al.	**Ours**
Total Participants	28‡	40		324		1,500
Demographic Respondents	28	19	533	115	21	1,500
Gender						
Male	53.6 %	47.4 %	45.0 %	47.0 %	38.1 %	50.5 %
Female	46.4 %	42.1 %	54.0 %	52.2 %	61.9 %	47.9 %
Non-binary	0.0 %	5.3 %	1.0 %	0.9 %	0.0 %	1.4 %
Prefer not to say/Other	0.0 %	5.3 %	0.0 %	0.0 %	0.0 %	0.3 %
Sexual Orientation						
Heterosexual	89.3 %	-	84.0 %	81.7 %	-	-
Lesbian or Gay	7.1 %	-	5.0 %	4.3 %	-	-
Bisexual	0.0 %	-	9.0 %	12.2 %	-	-
Uncertain	3.6 %	-	-	0.9 %	-	-
Prefer not to say/Other	0.0 %	-	2.0 %	0.9 %	-	-
Age					†	-
18-24	7.1 %	26.3 %	11.0 %	0.0 %	-	19.8 %
25-34	39.3 %	47.4 %	37.0 %	25.2 %	42.9 %	30.3 %
35-44	42.9 %	10.5 %	24.0 %	33.9 %	23.8 %	15.8 %
45-54	10.7 %	10.5 %	16.0 %	23.5 %	23.8 %	13.9 %
55-64	0.0 %	5.3 %	9.0 %	13.9 %	9.5 %	13.1 %
65+	0.0 %	0.0 %	3.0 %	1.7 %	0.0 %	7.1 %
Prefer not to say	0.0 %	-	-	1.7 %	-	0.1 %
Ethnicity						
White/Caucasian	67.9 %	31.6 %	81.0 %	81.7 %	42.9 %	64.6 %
Asian	10.7 %	57.9 %	8.0 %	2.6 %	28.6 %	6.3 %
Black/African descent	3.6 %	10.5 %	4.0 %	8.7 %	-	8.1 %
Hispanic/Latino	3.6 %	15.8 %	1.0 %	0.9 %	4.8 %	8.1 %
Native American	0.0 %	0.0 %	0.0 %	2.6 %	9.6 %	0.5 %
Middle Eastern	0.0 %	0.0 %	1.0 %	0.9 %	4.8 %	0.9 %
Prefer not to say/Other	14.3 %	-	5.0 %	2.6 %	9.6 %	11.5 %
Education						
No University Degree	17.9 %	10.5 %	0.0 %	34.8 %	14.3 %	40.8 %
Undergraduate Degree	57.1 %	52.6 %	66.0 %	53.9 %	57.1 %	42.5 %
Graduate Degree	14.3 %	36.8 %	34.0 %	10.4 %	28.1 %	16.1 %
Prefer not to say/Other	10.7 %	-	-	0.9 %	-	0.6 %

†Age group values for Stiennon et al. are reported for ten-year age groups starting from 20-29. We have placed the values in the row where the top end of these groups would appear to align with groups reported by the majority of studies.
‡ Bai et al. provide two reports of demographic data. We use the one corresponding to the participants who contributed more than 80% of the total feedback.

Table 7: **Geographic data compared to prior work.** Participant countries of residence in PRISM compared to early and widely-known RLHF studies using human feedback for language models.

Category	Bai et al.	Ouyang et al.	Glaese et al.	Ganguli et al.	Stiennon et al.	**Ours**
United States	100 %	17 %	0 %	100 %	60 %	26 %
United Kingdom	0 %	0 %	100 %	0 %	7 %	23 %
Philippines	0 %	22 %	0 %	0 %	7 %	0 %
Bangladesh	0 %	22 %	0 %	0 %	0 %	0 %
All Others	0 %	39 %†	0 %	0 %	27 %‡	51 %*

†One resident each from Albania, Brazil, Canada, Columbia, India, Uruguay, and Zimbabwe
‡ One resident each from South Africa, Serbia, Turkey, India
*See Tab. 8 for our breakdowns.

H Participant Geographies

We present full geographic breakdowns in Tab. 8. Fig. 8 is an enlarged version from Fig. 1. We compare geographic data to prior work in Tab. 7. For regional classifications, we use the UN definitions.[38] We also throughout the main paper use `location_special_region`, which splits out the UK and the US. Regional breakdowns by birth country are shown in Fig. 9.

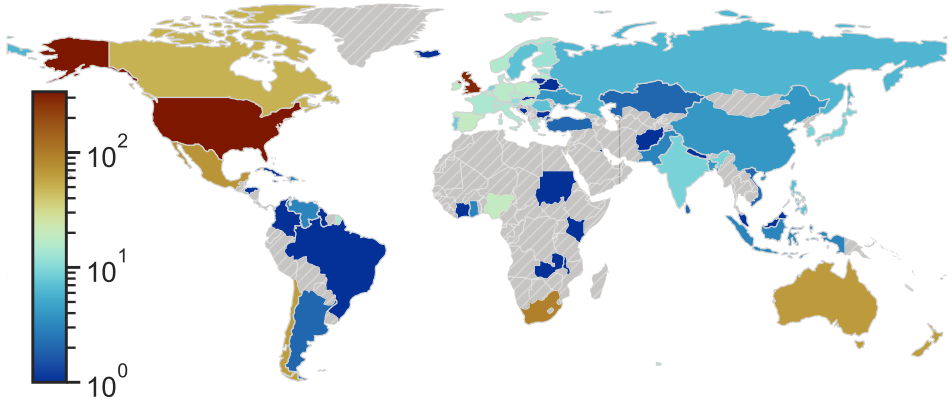


Figure 8: **Geographic distribution of PRISM participants by birth country**. Our sampling aims were for breath (coverage across UN global regions) and depth (representative demographic coverage within UK and US samples).

Table 8: **Full Geographic Breakdowns.** We collect country of birth and current country of residence. PRISM contains participants born in 75 different countries, residing in 38 different countries.

	Country of Birth		Country of Residence	
United States	338	22.5 %	386	25.7 %
United Kingdom	292	19.5 %	340	22.7 %
South Africa	91	6.1 %	86	5.7 %
Mexico	69	4.6 %	67	4.5 %
Australia	65	4.3 %	72	4.8 %
New Zealand	64	4.3 %	77	5.1 %
Chile	63	4.2 %	65	4.3 %
Canada	50	3.3 %	54	3.6 %
Israel	47	3.1 %	61	4.1 %
Nigeria	19	1.3 %	0	0.0 %
Spain	19	1.3 %	18	1.2 %
Germany	17	1.1 %	13	0.9 %
Belgium	17	1.1 %	17	1.1 %
Hungary	17	1.1 %	16	1.1 %
Poland	17	1.1 %	14	0.9 %
Ireland	17	1.1 %	15	1.0 %
Latvia	16	1.1 %	14	0.9 %
Denmark	15	1.0 %	15	1.0 %
Czechia	15	1.0 %	14	0.9 %
Norway	15	1.0 %	15	1.0 %
France	14	0.9 %	12	0.8 %
Italy	14	0.9 %	13	0.9 %
Greece	14	0.9 %	13	0.9 %
Switzerland	14	0.9 %	14	0.9 %

Continued on next page

[38] `https://population.un.org/wpp/DefinitionOfRegions`

Table 8: **Full Geographic Breakdowns.** We collect country of birth and current country of residence. PRISM contains participants born in 75 different countries, residing in 38 different countries.

	Country of Birth		Country of Residence	
Finland	12	0.8 %	13	0.9 %
Estonia	11	0.7 %	10	0.7 %
Austria	11	0.7 %	10	0.7 %
Slovenia	10	0.7 %	10	0.7 %
Netherlands	9	0.6 %	8	0.5 %
India	9	0.6 %	0	0.0 %
Japan	9	0.6 %	11	0.7 %
Korea, Republic of	9	0.6 %	7	0.5 %
Portugal	8	0.5 %	7	0.5 %
Romania	7	0.5 %	0	0.0 %
Philippines	7	0.5 %	0	0.0 %
Sweden	7	0.5 %	6	0.4 %
Russian Federation	6	0.4 %	0	0.0 %
Ukraine	4	0.3 %	0	0.0 %
Bangladesh	4	0.3 %	0	0.0 %
China	4	0.3 %	0	0.0 %
Hong Kong	3	0.2 %	0	0.0 %
Pakistan	3	0.2 %	0	0.0 %
Ghana	3	0.2 %	0	0.0 %
Dominican Republic	3	0.2 %	0	0.0 %
Venezuela, Bolivarian Republic of	3	0.2 %	0	0.0 %
Indonesia	3	0.2 %	0	0.0 %
Viet Nam	2	0.1 %	0	0.0 %
Sri Lanka	2	0.1 %	0	0.0 %
Turkey	2	0.1 %	0	0.0 %
Argentina	2	0.1 %	0	0.0 %
Kazakhstan	2	0.1 %	0	0.0 %
Slovakia	1	0.1 %	0	0.0 %
Sudan	1	0.1 %	0	0.0 %
Tonga	1	0.1 %	0	0.0 %
Afghanistan	1	0.1 %	0	0.0 %
Nepal	1	0.1 %	0	0.0 %
Honduras	1	0.1 %	0	0.0 %
Belarus	1	0.1 %	0	0.0 %
Bosnia and Herzegovina	1	0.1 %	0	0.0 %
Brazil	1	0.1 %	0	0.0 %
Bulgaria	1	0.1 %	0	0.0 %
Colombia	1	0.1 %	0	0.0 %
Cuba	1	0.1 %	0	0.0 %
Côte d'Ivoire	1	0.1 %	0	0.0 %
Malaysia	1	0.1 %	0	0.0 %
Guyana	1	0.1 %	0	0.0 %
Iceland	1	0.1 %	1	0.1 %
Jamaica	1	0.1 %	0	0.0 %
Kenya	1	0.1 %	0	0.0 %
Kuwait	1	0.1 %	0	0.0 %
Lithuania	1	0.1 %	0	0.0 %
Luxembourg	1	0.1 %	2	0.1 %
Malawi	1	0.1 %	0	0.0 %
Zambia	1	0.1 %	0	0.0 %
Tanzania, United Republic of	0	0.0 %	1	0.1 %
Lesotho	0	0.0 %	1	0.1 %
Uruguay	0	0.0 %	1	0.1 %
Prefer not to say	3	0.2 %	1	0.1 %

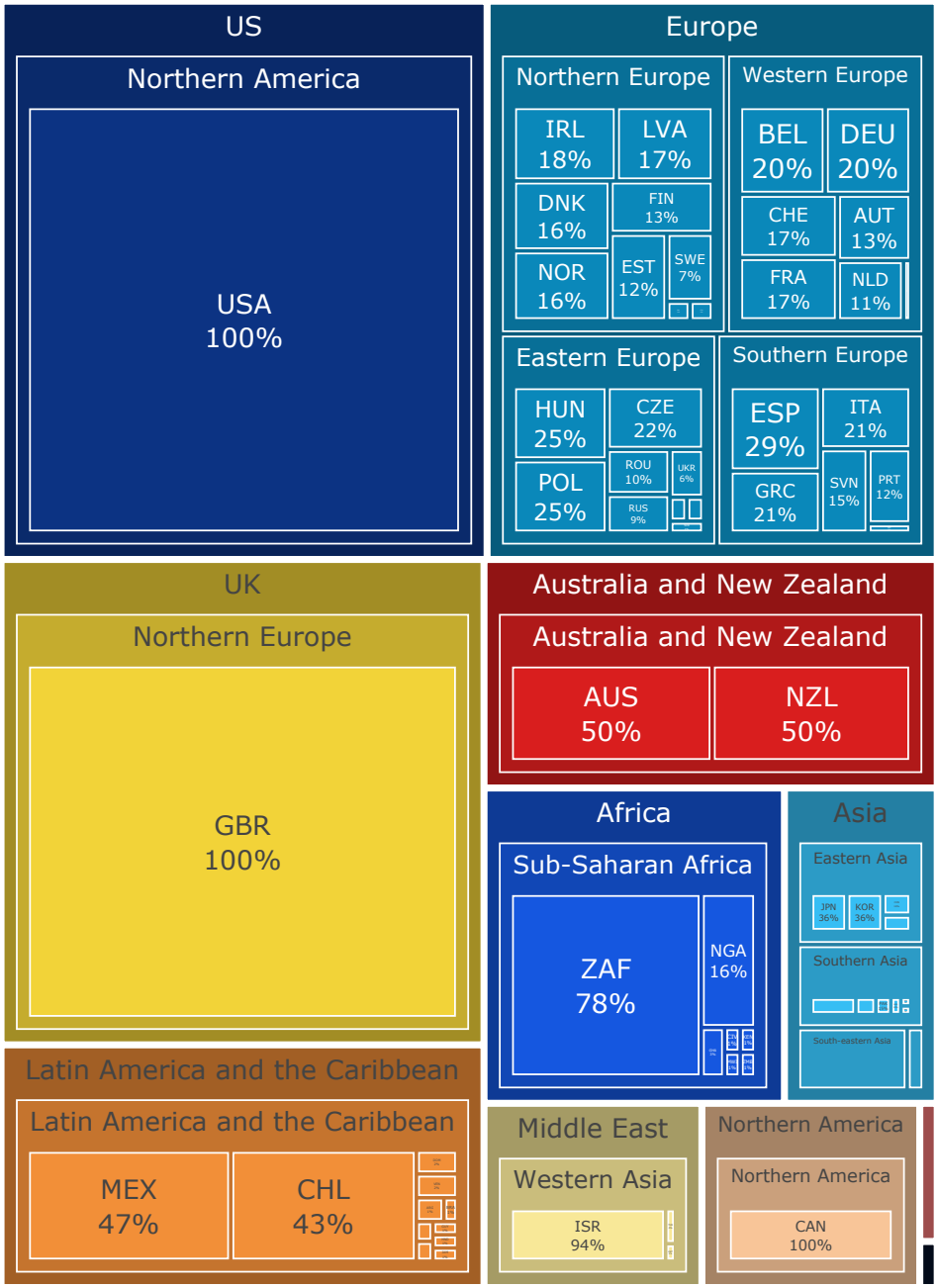


Figure 9: **Skewed regional entropy in PRISM.** The hierarchical tree diagram uses participant birth location, mapping (i) special location (splitting out the US and UK), which is used in the main paper, (ii) UN-defined subregions, and (iii) ISO country codes. There is an over-representation of UK and US participants due to the census samples. In most regions besides Europe, participation is dominated by one or two birth countries. The two small vertical boxes are Prefer not to say (in red), and Oceania (in navy). *Note:* 88% of PRISM participants are born and currently reside in the same country.

I Participant LLM Usage and Familiarity

We present breakdowns on experience with LLMs in Tab. 9. We did not require participants to be familiar with LLMs so provide the following primer:

This research is about Artificial Intelligence (AI) Language Models.

These models are also sometimes referred to as Generative AI, Large Language Models (LLMs), Conversational Agents, AI Chat Bots or Virtual Assistants.

They are advanced computer programs that can understand and generate human-like text. These models learn from large amounts of text data on the internet to generate their responses.

One example you might have heard is ChatGPT, where people can have a conversation with an AI language model via an internet website.

Table 9: **Survey of Participants' LLM Usage**: The majority of participants have used LLMs directly (via a dedicated chat interface) and indirectly (embedded in products or services). Note only participants who answered *Yes* to **LLM Direct Use** or **LLM Indirect Use** ($n = 1253$, 84%) are shown **LLM Freq of Use** and **LLM Use Cases**. For Use Cases, we show the % of these participants who selected each use case (can be multiple so $\sum \neq 1$). Exact question phrasing is reported in the survey codebook (App. X.1).

LLM Direct Use		
Yes	1,162	77.5 %
No	259	17.3 %
Unsure	79	5.3 %
LLM Indirect Use		
Yes	1,104	73.6 %
No	215	14.3 %
Unsure	181	12.1 %
LLM Familiarity		
Somewhat familiar	920	61.3 %
Very familiar	424	28.3 %
Not familiar at all	156	10.4 %
LLM Frequency of Use		
Once per month	374	24.9 %
Every week	316	21.1 %
More than once a month	291	19.4 %
Less than one a year	162	10.8 %
Every day	110	7.3 %
Not shown question	247	16.5 %
LLM Use Cases		
Research: Fact-checking or gaining overviews on specific topics.	617	49.2 %
Professional Work: Assisting in drafting, editing, or brainstorming content for work.	469	37.4 %
Creative Writing: Generating story ideas, dialogues, poems or other writing prompts.	392	31.3 %
Technical or Programming Help: Seeking programming guidance, code generation, software recommendations, or debugging assistance.	337	26.9 %
Lifestyle and Hobbies: Looking for recipes, craft ideas, home decoration tips, or hobby-related information.	310	24.7 %
Homework Assistance: Getting help with school or university assignments.	286	22.8 %
Personal Recommendations: Seeking book, music or movie recommendations.	266	21.2 %

Continued on next page

Table 9: **Survey of Participants' LLM Usage**: The majority of participants have used LLMs directly (via a dedicated chat interface) and indirectly (embedded in products or services). Note only participants who answered *Yes* to **LLM Direct Use** or **LLM Indirect Use** ($n = 1253$, 84%) are shown **LLM Freq of Use** and **LLM Use Cases**. For Use Cases, we show the % of these participants who selected each use case (can be multiple so $\sum \neq 1$). Exact question phrasing is reported in the survey codebook (App. X.1).

Casual Conversation: Engaging in small talk, casual chats, or joke generation.	262	20.9 %
Language Learning: Using it as a tool for language practice or translation.	229	18.3 %
Source Suggestions: Creating or finding bibliographies, information sources or reading lists.	217	17.3 %
Daily Productivity: Setting reminders, making to-do lists, or productivity tips.	216	17.2 %
Historical or News Insight: Getting summaries or background on historical events or news and current affairs.	183	14.6 %
Well-being Guidance: Seeking general exercise routines, wellness or meditation tips.	159	12.7 %
Games: Playing text-based games, generating riddles or puzzles.	143	11.4 %
Travel Guidance: Getting destination recommendations, planning holidays, or cultural etiquette tips.	133	10.6 %
Medical Guidance: Seeking health-related advice or medical guidance.	130	10.4 %
Financial Guidance: Asking about financial concepts or general investing ideas.	107	8.5 %
Relationship Advice: Seeking general self-help or relationship advice for family, friends or partners.	98	7.8 %
Other	124	9.9 %

I.1 Other Identified Usecases

In addition to the usecases in Tab. 9, 122 participants used the "Other" option to add a usecase in their own words. Many of these just add more specific details to the pre-provided categories. In addition, there were a few interesting themes:

- **Customer Service:** Many of the participants noted having interacted with LLMs in customer support chats, often with negative sentiment ("Usually forced to interact with chatbots to get something done", "Customer service bots I cannot avoid", "Insurance companies direct you to chatbots, usually useless").
- **Prolific and Other Online Surveys:** One of the more common (and potentially concerning) answers mentioned research participation e.g., "Studies like this one", "Doing Prolific tests", but it may be that they mean AI is the subject of the study: "AI research subject on research platforms Prolific, others." or "It's sometimes required as part of a survey on Prolific." We encourage future work on whether there is noticeable difference in these participants' answers elsewhere in our task.
- **AI Understanding or Testing:** A few participants mentioned "Trying to gain an understanding into AI and its capabilities" or "Gauging progress/viability of AI models". Many others indicated curiosity or exploratory use e.g., "Just to test it out and see what it's all about" or "Casual interest in the new technology".
- **Professional or Job Tasks:** Participants added details on professional usecases like resume help, interview prep, CV writing, HR-tasks, Excel help, or emails.
- **Creative (Multimodal) Use-cases:** Participants gave additional detail like writing YouTube scripts, generating gift card text or designing characters for games as well as multimodal creative outputs like generating drawings or images.
- **Domain-Specific Usecases:** Medical, Financial and Educational usecases are all mentioned.

J Screening and Recruitment Process

We recruit workers via Prolific (`https://www.prolific.com/`). We apply two initial screening criteria: (i) participants must be fluent in English because PRISM targets monolingual models and language data, and (ii) participants must have been born and reside in the same country to avoid biasing our sample towards expats living abroad. There is a skewed country-wise distribution of active workers who meet this criteria (see Tab. 10). For example, of the 21,084 workers in Europe (passing screening), 17% are Portuguese, 15% German and 14% Polish; and all 6,584 workers in Africa are located in South Africa. To account for this, we set up country-specific studies in each country with at least one eligible worker, balance study spots across regions, and ensure no single country has more than 100 open spots (apart from the Rep Samples in the UK and US). We collected information on country of birth and country of current residence during our survey (separate to workers' stored Prolific details), and find that 179 participant (12%) have different birth and reside countries. We do not exclude these individuals from our sample.

Table 10: **Summary of Recruitment Studies** We present study-wise breakdowns ($n = 33$). Each study was created based on the constraints of Prolific's pool of workers. We show here *all* the countries with at least 1 fluent English speaker, and the counts for fluent English Speakers who were born and currently reside in that same country. We show the whether each study was screened for a special representative sample (**Rep Sample**) or if it was balanced on participant gender (**Gender Bal**). In some cases, there were too few active participants per country to balance by gender without comprising participant privacy. We also show when the first batch was launched (all dates are in 2023) and approximate cost (at £9 per hour per participant).

Study	Rep Sample	Gender Bal	Launched (2023)	Approved Submissions		Prolific Fluent English Speakers (All)		Prolific Fluent English Speakers (Born=Reside)		Cost (£)
Total	**2**	**25**	-	**1,500**		**111,572**		**100,585**		**14,850.00**
US	✓	✗	27-11	386	25.7 %	38,114	34.2 %	36,205	36.0 %	3,821.40
UK	✓	✗	27-11	341	22.7 %	37,408	33.5 %	33,678	33.5 %	3,375.90
South Africa	✗	✓	22-11	88	5.9 %	7,061	6.3 %	6,584	6.5 %	871.20
New Zealand	✗	✓	24-11	77	5.1 %	511	0.5 %	389	0.4 %	762.30
Australia	✗	✓	24-11	71	4.7 %	1,968	1.8 %	1,550	1.5 %	702.90
Mexico	✗	✓	24-11	69	4.6 %	2,021	1.8 %	1,943	1.9 %	683.10
Chile	✗	✓	23-11	65	4.3 %	455	0.4 %	416	0.4 %	643.50
Israel	✗	✗	25-11	61	4.1 %	310	0.3 %	272	0.3 %	603.90
Canada	✗	✓	22-11	54	3.6 %	3,687	3.3 %	3,031	3.0 %	534.60
Asia	✗	✗	24-11	18	1.2 %	196	0.2 %	32	0.0 %	178.20
Spain	✗	✓	23-11	18	1.2 %	1,252	1.1 %	942	0.9 %	178.20
Belgium	✗	✓	23-11	17	1.1 %	376	0.3 %	281	0.3 %	168.30
Hungary	✗	✓	24-11	16	1.1 %	537	0.5 %	456	0.5 %	158.40
Ireland	✗	✓	23-11	15	1.0 %	640	0.6 %	502	0.5 %	148.50
Denmark	✗	✗	23-11	15	1.0 %	119	0.1 %	65	0.1 %	148.50
Norway	✗	✗	23-11	15	1.0 %	91	0.1 %	59	0.1 %	148.50
Switzerland	✗	✓	23-11	14	0.9 %	205	0.2 %	104	0.1 %	138.60
Poland	✗	✓	23-11	14	0.9 %	2,975	2.7 %	2,850	2.8 %	138.60
Czech Republic	✗	✓	24-11	14	0.9 %	238	0.2 %	229	0.2 %	138.60
Latvia	✗	✓	23-11	14	0.9 %	173	0.2 %	162	0.2 %	138.60
Greece	✗	✓	24-11	14	0.9 %	809	0.7 %	747	0.7 %	138.60
Finland	✗	✓	23-11	13	0.9 %	152	0.1 %	117	0.1 %	128.70
Germany	✗	✓	24-11	13	0.9 %	3,152	2.8 %	2,295	2.3 %	128.70
Italy	✗	✓	24-11	12	0.8 %	2,037	1.8 %	1,857	1.8 %	118.80
France	✗	✓	24-11	12	0.8 %	957	0.9 %	681	0.7 %	118.80
Slovenia	✗	✓	24-11	10	0.7 %	232	0.2 %	220	0.2 %	99.00
Austria	✗	✓	24-11	10	0.7 %	231	0.2 %	156	0.2 %	99.00
Estonia	✗	✓	24-11	10	0.7 %	251	0.2 %	237	0.2 %	99.00
Netherlands	✗	✓	24-11	8	0.5 %	1,460	1.3 %	1,028	1.0 %	79.20
Portugal	✗	✓	24-11	7	0.5 %	3,649	3.3 %	3,284	3.3 %	69.30
Sweden	✗	✓	24-11	6	0.4 %	274	0.2 %	196	0.2 %	59.40
Luxembourg	✗	✗	23-11	2	0.1 %	15	0.0 %	6	0.0 %	19.80
Iceland	✗	✗	23-11	1	0.1 %	16	0.0 %	11	0.0 %	9.90

K Conversation Type Rebalancing

Our task instructions specified that participants should complete six conversations in total, two of each type. In reality, some participants deviated from this quota. This could be due to (i) misunderstanding of instructions, (ii) technical issues, or (iii) losing count, as while we included a counter of the total number of conversations on the interface (see App. W), we did not include per conversation type breakdowns. To mitigate variation on conversation type selection, we create a balanced subset of PRISM. First, we filter to all participants who had at least one of each conversation type. Then we take the maximum number of total conversations (either $n = 3$ or $n = 6$) so that there are equal numbers of each type. This results in $6,669$ conversations (84% of all conversations), from 1246 participants (83% of all participants). We release this flag `included_in_balanced_subset` if future researchers want to use the same set of conversations. We make sure this flag intersects with the census rebalancing flags (see App. L) so no further data is lost when both subsets are needed.

L Census Rebalancing

Obstacles to representativeness We use the representative sample offered from Prolific [173]. However, there are several reasons why these samples may not be fully representative. First, our sampling process was affected internally due to cyberattacks disrupting some participants' workflows. These participants returned to the task after their spots had 'timed-out', and were re-filled by other same demographic individuals. Second, Prolific provides a sample breakdown in-line with a *simplified* census but do not match *intersectional* proportions to census data. Third, if a sample spot is taking too long to fill (e.g., 65+ years), Prolific will reallocate these spots to different demographics. There are of course wider stumbling blocks from crowdworkers skewing towards younger, more educated, and digitally-active populations. We original set up 300 spots for each of the representative samples, but ended up with 386 approved participants in the UK sample (UK-REP), and 341 in the US (US-REP).[39]

Is our original sample representative? We compare our sample breakdowns to recent census data.[40] For each of US-REP and UK-REP, we remove participants who did not give demographic details (*Prefer not to say*) and those reporting non-binary gender (which is not accounted for in census data). We subset to individuals also appearing in the balanced conversation subset to mitigate further data loss (see App. K). Remaining participants are considered *eligible*: 283 participants for the UK, and 297 for the US. We map PRISM and census data into shared age, ethnicity and gender buckets. We then cross-tabulate what proportion is expected to appear in each age, gender and ethnicity intersection from the census data, and what percentage of participants we actually observed in our sample.[41] Fig. 10 shows the original UK sample is relatively census-balanced, especially if the 55-64 and 65+ age groups are combined (over-representation of white individuals in the former, offsets the under-representation in the latter). The US sample is skewed towards white, middle-aged individuals, with too few in the "Other" category (in our data corresponding to Other, as well as Hispanic, Indigenous/First Peoples or Middle Eastern / Arab combined).

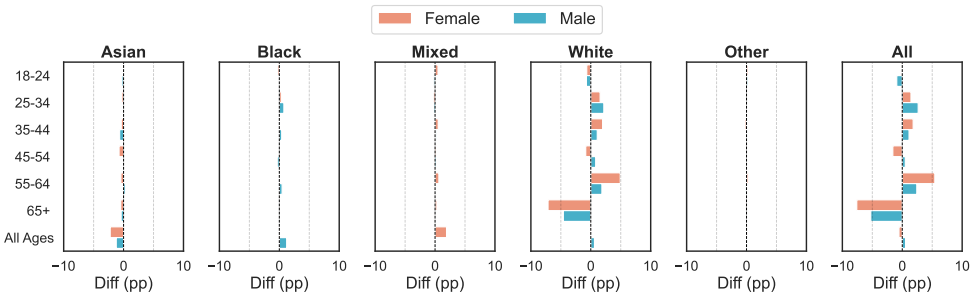
Can we make our sample more representative? We aim to resample 300 participants according to census proportions but with two remaining caveats: 300 is a still a very small sample—it is impossible to sample 0.83 Black women who are 18-24 years of age; and we are limited by the data we already have—there are no Asian Women of 45-54 years, so we cannot add them retrospectively. We iterate through the expected proportions of each intersection, try to sample that exact number of in-group individuals, otherwise adding all individuals if there are too few to fill the spots. After rebalancing, the sample drops to 243 participants for the UK and 230 for the US. We improve upon, but do not fully resolve, representativeness. For both samples, the differences are now within $\sim$7pp, which over 230-240 individuals is $\sim$10-15 people incorrectly allocated. The rebalanced UK sample still suffers from a deficit of older people (65+), a common concern with crowdworker populations; and the rebalanced

[39]There are more than the initial 300 spots due to participants returning to our interface to finish their conversations after their place had 'timed-out' and been refilled. We still paid and included these participants.

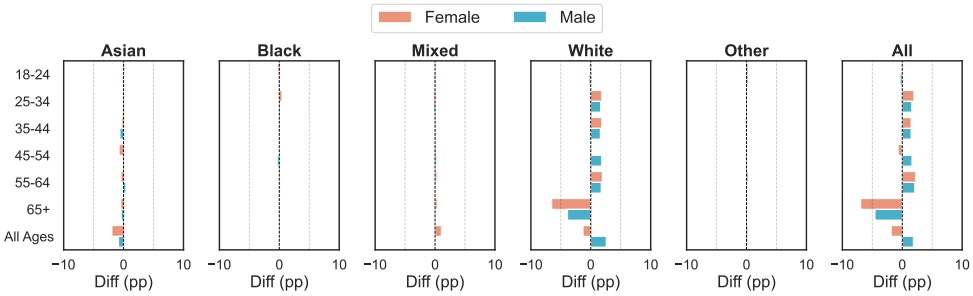
[40]For the UK, we examine age, ethnicity and gender from the 2021 data provided by the Office of National Statistics (see ons.gov.uk). For the US, we download and combine each ethnicity-specific table from the 2022 data provided by US Census Bureau (see data.census.gov.).

[41]For the US, we combine "Other" with "Hispanic" because over 91% of the "Other" census category are Hispanic individuals. See census.gov/library/stories/2023/10/2020-census-dhc-a-some-other-race-population.

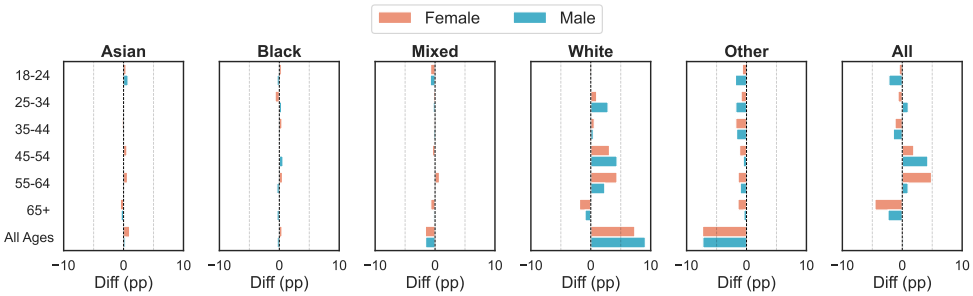
US sample still has an over-representation of White participants and under-representation of Other participants. There is a trade-off because increasing representativeness on these observed census characteristics reduces sample size, thus worsening representation on unobserved characteristics. There is still lots of headroom for future work to improve, especially by increasing sample sizes and ensuring other characteristics are controlled for, such as political affiliation, education or income.



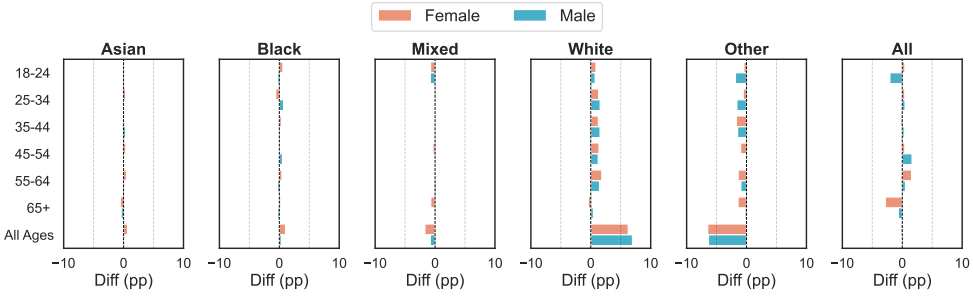
(a) **UK (Before Rebalancing)**: There are 282 eligible* participants in the UK sample.



(b) **UK (After Rebalancing)**: There are 243 participants in the rebalanced UK sample.



(c) **US (Before Rebalancing)**: There are 297 eligible* participants in the US sample.



(d) **US (After Rebalancing)**: There are 230 participants in the rebalanced US sample.

Figure 10: **Before and after census-rebalancing.** We show the difference in observed and expected proportions (PRISM *minus* Census). Bars to the *right* of the centre line are groups *over-represented* in PRISM relative to the census. The UK census population has 47,204,870 adults. The US census has 298,477,760 adults. The sample size for before and after rebalancing is reported above. *A participant is *eligible* if they have completed a equal number of conversations for each conversation type (see App. K).

M Text and N-Gram Analysis

There are 5 core types of free text instance in PRISM. We present a summary of count and length distributions in Tab. 11. For all text instances, we show top N-grams. Additionally, for the self-written system strings (constitutions), self-written profiles and open-feedback, we extract the most frequent adjectives.[42] not counting adjectives that appeared in the question text. We then retrieve windows of 5 words surrounding each of these top adjectives, and randomly sample three snippets to display. We also compare the most frequent words in the opening prompts of PRISM to human-written prompts in HELPFULHONEST [15] and OPENCONVERSATIONS [16]. We extract unique words to each dataset (no overlap with the other two). We find evidence of diferent domains biases—for example, OPENCONVERS. contains many software or ML related keywords versus PRISM which contains some cultural and value-laden references, including *waitangi* (as in the Treaty of Waitangi, New Zealand's constitution that grounded Maori rights); *unethically*, *populist* and *multicultural*.

Table 11: **Summary of text distributions in PRISM.** We show the number of instances (N) alongside summary statistics for length in words (W), broken by whitespace. We also show the number of total unique words and total unique tokens, as encoded by the `gpt-4` BPE tokenizer (from `tiktoken`).

	N	Mean$W$	Std$W$	Min$W$	25%W	50%W	75%W	Max$W$	Unique$W$	UniqueT_{BPE}
`system_string`	1,500	46	50	2	26	40	57	1,655	7,942	6,132
`self_description`	1,500	44	25	1	28	40	56	278	6,912	5,409
`open_feedback`	8,011	29	19	1	16	25	37	283	15,444	11,115
`user_prompt`	68,371	13	11	1	7	10	15	234	31,862	20,265
`model_response`	68,371	89	60	1	46	71	128	742	215,931	51,386

Table 12: **Top N-grams in user prompts.** Demonstrates PRISM's content distribution towards information-seeking dialogue and questions, over task-orientated dialogue and instructions.

Unigrams		**Bigrams**		**Trigrams**	
N-Gram	Freq	N-Gram	Freq	N-Gram	Freq
(think,)	8,005	(do, you)	8,767	(do, you, think)	5,137
(people,)	5,332	(you, think)	5,450	(what, do, you)	2,554
(would,)	4,470	(what, is)	4,186	(what, is, the)	2,331
(like,)	3,764	(is, the)	4,099	(you, think, about)	1,168
(good,)	2,915	(in, the)	3,570	(how, can, i)	1,111
(best,)	2,501	(can, you)	3,089	(is, the, best)	1,009
(dont,)	2,380	(what, do)	2,778	(what, are, the)	946
(know,)	2,129	(what, are)	2,425	(what, are, some)	759
(im,)	2,042	(of, the)	2,403	(do, you, have)	741
(tell,)	1,989	(can, i)	1,957	(how, do, i)	716

Table 13: **Top N-grams in model responses.** Demonstrates both advisory tone (its, important, to) and high frequency of de-anthropomorphisation (as, an, ai).

Unigrams		**Bigrams**		**Trigrams**	
N-Gram	Freq	N-Gram	Freq	N-Gram	Freq
(may,)	19,582	(of, the)	21,744	(its, important, to)	6,857
(important,)	19,027	(it, is)	19,367	(it, is, important)	5,917
(like,)	18,209	(in, the)	18,535	(is, important, to)	5,522
(also,)	17,077	(is, a)	17,586	(here, are, some)	4,430
(help,)	16,903	(important, to)	14,319	(as, an, ai)	3,961
(people,)	16,482	(such, as)	11,800	(would, you, like)	3,402
(provide,)	14,046	(on, the)	11,025	(i, do, not)	3,049
(would,)	12,641	(to, the)	10,963	(there, are, many)	2,820
(however,)	12,502	(can, be)	10,606	(i, dont, have)	2,683
(many,)	12,314	(and, the)	10,599	(like, me, to)	2,673

Table 14: **Most frequent unique tokens compared to existing datasets.** We list the most common tokens which are unique to a particular dataset. We exclude tokens which are misspelled or foreign language.

Dataset	**Top Words**						
PRISM	waitangi	whilst	unethically	populist	nieces	multicultural	lowered
HELPFULHONEST	cuss	kidnap	Arizona	Alaska	ski	carpet	bees
OPENCONVERS.	ML	loop	reinforcement	equivalent	describing	capabilities	uint256

[42] We use NLTK POS tagger to match on 'JJ' tags. We make some edits for filler words (e.g., "such", "sure"), and verb forms (e.g., "able...[to do X]"). We also remove any adjectives appearing in the question text.

M.1 System String (Constitutions)

> ***Question Text**: Imagine you are instructing an AI language model how to behave. You can think of this like a set of core principles that the AI language model will always try to follow, no matter what task you ask it to perform. In your own words, describe what characteristics, personality traits or features you believe the AI should consistently exhibit. You can also instruct the model what behaviours or content you don't want to see. If you envision the AI behaving differently in various contexts (e.g., professional assistance vs. storytelling), please specify the general adaptations you'd like to see. Please write 2-5 sentences in your own words.*

Table 15: **Top adjectives in system strings (constitutions).**

Adjective	Freq	Example Windows ($w = 5$, $n = 3$)
factual	221	"...should produce only true or factual output and never give false..." | "...Trustworthy , transparent , factual , sincere..." | "...the AI should always provide factual information , and is able..."
accurate	113	"...needs to provide me with accurate information . It needs to..." | "...I know I 'm getting accurate information . For creative use..." | "...sources to get the most accurate response possible . The AI..."
human	106	"...not be programmed with any human like emotion . I am..." | "...the technology is advancing , human interaction will end ...." | "...should n't pretend to be human ...."
important	100	"...The most important thing to understand other person..." | "...mine . It 's also important to understand the whole conversation..." | "...well written responses . Remember important information about the user ...."
friendly	99	"...information in a warm , friendly way ...." | "...task . I also appreciate friendly language and the sense of..." | "...Be friendly and uplifting in converstaion ...."
different	94	"...to take in information from different sources but place more importance..." | "...also be able to combine different types of knowledge or inputs..." | ".... Respect Cultures and treat different ideas with respect . Things..."
clear	93	".... It made the point clear , so kept professional and..." | "...should be able to give clear and precise information , using..." | "...that . It should given clear instruction such as , do..."
creative	89	"...expand . Do n't be creative unless I ask you to..." | "..., being as informative , creative and/or thorough as the task..." | "..., or more of a creative one . The language model..."
harmful	89	"...user privacy and prohibition of harmful or misleading content , as..." | ".... - Do n't write harmful content..." | "...want to see or read harmful words and language that is..."
polite	79	"..., being very professional and polite would be nice ...." | "...to read language that is polite with here and there a..." | "..., you should always be polite and respectful to the user..."
helpful	75	"...model should always be as helpful as possible , being as..." | "...it should be informative and helpful..." | "...think it should always be helpful and guiding..."
good	75	"...AI is a good tool . As someone who..." | "...informations must be clear and good structured ...." | "...evolution . It 's a good idea to write down responses..."
personal	70	"...rights and basic principles like personal privacy should be respected at..." | "...language model should not disclose personal information . It should be..." | "...It would n't ask for personal information and would generally be..."
respectful	66	"...should always exhibit kind and respectful behaviour . Also he should..." | "...AI must be respectful of any idea you put..." | "...should behave in a respectful way towards everyone , everyone..."
correct	65	"...They must be sincere and correct , does not want to..." | "...ask question to give as correct answers as possible . AI..." | "...for information and give always correct facts . -Write in a..."
unbiased	58	"...advice or help but be unbiased and not geared to my..." | "...'-It must be unbiased when I ask for information..." | "...should give the user an unbiased answer , but it should..."
informative	57	"...as possible , being as informative , creative and/or thorough as..." | "...patronising , it should be informative and helpful..." | "...The AI should be informative and make responses based on..."
relevant	50	"..., real information and be relevant about what i 'm asking..." | "...is really important to state relevant facts and information , but..." | "...answers that are clear and relevant . I do n't think..."
neutral	49	"...or provocatively and have a neutral presentation of issues..." | "...ideological matters . Be as neutral as possible with charged subjects..." | "...also think it should remain neutral on political and social matters..."
objective	49	"...and honest manner . Describe objective facts whenever possible and if..." | "...the AI should be as objective as possible : it should..." | "...sources) , have an objective point of view without giving..."

Table 16: **Top N-grams in system strings (constitutions).**

Unigrams		**Bigrams**		**Trigrams**	
N-Gram	Freq	N-Gram	Freq	N-Gram	Freq
(ai,)	1,503	(the, ai)	798	(the, ai, should)	260
(would,)	819	(i, would)	569	(i, would, like)	250
(information,)	588	(to, be)	563	(be, able, to)	168
(like,)	575	(should, be)	520	(the, ai, to)	158
(want,)	452	(it, should)	515	(it, should, be)	153
(model,)	443	(ai, should)	436	(ai, language, model)	153
(language,)	392	(would, like)	261	(ai, should, be)	117
(always,)	359	(it, to)	248	(the, ai, model)	114
(also,)	306	(ai, to)	230	(i, would, want)	104
(answers,)	249	(to, the)	220	(want, it, to)	99

M.2 Self-Description

Question Text: *Please briefly describe your values, core beliefs, guiding principles in life, or other things that are important to you. For example, you might include values you'd want to teach to your children or qualities you look for in friends. There are no right or wrong answers. Please do not provide any personally identifiable details like your name, address or email. Please write 2-5 sentences in your own words.*

Table 17: **Top adjectives in self-description.**

Adjective	Freq	Example Windows ($w = 5, n = 3$)
good	229	"...helpful to everyone . The good of others above my own..." | "...is sustainability , having a good relationship with nature and not..." | "..., honest . To be good relationships with family and friends..."
hard	71	"...treated . I think that hard work is the key to..." | "...own thing , try as hard as you can , I..." | "...decency , and being a hard worker . As long as..."
honesty	68	"...personal values are respect , honesty kindness and fairness . I..." | "...the most important value is honesty , above all , even..." | "...My core values are honesty and justice . Honesty in..."
human	61	"...guide us and makes us human . Such as the Law..." | "...nature , animals and other human beings ...." | "...not like racism . Every human being is different so we..."
true	57	"...it is their sincere and true belief let it be ...." | "...faith , laws , being true to myself and others ...." | "...the best policy . Being true to yourself is very valuable..."
right	55	"...likes to do thing the right way . I have an..." | "...all can say this is right or wrong because it still..." | "...believe in doing what is right and just Guiding principles in..."
honest	53	"...I believe in others being honest with me and I will..." | "...firstly respect yourself , be honest , fair and kind to..." | "...important to be trustworthy , honest . To be good relationships..."
open	52	"..... Approach items with an open and inquisitive mind . Take..." | "...is to be curious and open to learn new perspectives ...." | "...me to have such an open mindset into life ...."
different	50	"...understand that each person has different ways of going through a..." | "...also like us to have different tastes so that we can..." | "..... Every human being is different so we all can not..."
happy	49	"...I just want to be happy in life and enjoy it..." | "...thoughts and whether he is happy with his current state in..." | "...you are suppose to be happy with your life . You..."
empathy	48	"...like to be treated , empathy , loyalty , honesty ....." | "...a lot of value on empathy and selflessness . I feel..." | "...: inclusion , kindness , empathy , ... I think everybody..."
strong	48	"...I have a strong belief in the human capacity..." | "...would like them to become strong , fierce and independent souls..." | "...to be honest . Be strong and emitionally stable . Relaxing..."
equal	41	"...everyone as we are all equal . Do n't discriminate and..." | "...Everyone is equal , despite race , skin..." | "...is that all people are equal in life , no discrimination..."
bad	36	"...even tho i sometimes make bad decisions ...." | "...when they keep treating you bad ...." | "...and then only mention the bad soo the person doesnt get..."
fair	36	"...yourself , be honest , fair and kind to yourself ...." | "...honest with others and be fair and kind towards others ...." | "...in the sense of being fair to everybody , and treating..."
new	36	"..., authenticity , openness to new expereince and knowledge ...." | "...important in life , learning new things , even if they..." | "...never too old to learn new things ...."
respectful	36	"...keeping your word and being respectful are very important to me..." | "...would like them to be respectful with everyone , not to..." | "...treated . Be kind and respectful to people and do no..."
positive	35	"...day to make the most positive impact that we can ...." | "..., respect , self-development , positive thinking ...." | "...around people who have a positive view on life..."
respect	34	"...My personal values are respect , honesty kindness and fairness..." | "...think are very important is respect for others and empathy ...." | "...for me . So are respect for nature , animals and..."
loyal	33	"...afraid of commitment , being loyal . I value art ,..." | "...respect if friends can be loyal and honest . Not talking..." | "...I try to be as loyal as possible towards my friends..."

Table 18: **Top N-grams in self-description.**

Unigrams		**Bigrams**		**Trigrams**	
N-Gram	Freq	N-Gram	Freq	N-Gram	Freq
(people,)	701	(to, be)	589	(i, believe, in)	223
(believe,)	687	(i, believe)	516	(i, believe, that)	145
(life,)	608	(believe, in)	296	(important, to, me)	126
(important,)	548	(important, to)	241	(i, try, to)	99
(others,)	539	(try, to)	231	(to, be, treated)	94
(values,)	390	(i, think)	217	(the, most, important)	87
(also,)	380	(believe, that)	198	(would, like, to)	73
(value,)	368	(to, me)	198	(i, would, like)	72
(like,)	347	(i, value)	195	(i, look, for)	68
(always,)	311	(i, am)	185	(is, important, to)	66

M.3 Open-Ended Feedback

Question Text: *Give the model some feedback on the conversation as whole. Hypothetically, what would an ideal interaction for you look like here? What was good and what was bad? What (if anything) was missing? What would you change to make the conversation better?*

Table 19: **Top adjectives in open feedback.**

Adjective	Freq	Example Windows ($w = 5, n = 3$)
helpful	437	"...it was informative and helpful..." | "..., it was all very helpful and provided specific resources ...." | "...feedback that would be very helpful ...."
informative	433	"...liked that the AI was informative , and agrued both sides..." | "...it was informative and helpful..." | "...a whole in a very informative and positive light . I..."
different	355	"...summaries spaced out to separate different views , answers or information..." | "...Consider hair types , different textures . Think about how..." | "...my narrative and focus on different aspect of the conversation ...."
great	342	"... The first response was great , as even though it..." | "...The conversation was great , I felt like I..." | "...I feel this worked out great , and is a wonderful..."
factual	310	"...been derived as to the factual cause of death . Alluding..." | "...I liked that dates and factual information was given..." | "...I thought it was very factual , making it clear it..."
specific	238	"...all very helpful and provided specific resources . I can use..." | "...to reach and answer in specific..." | "...would try to get more specific culture references in . also..."
clear	217	"...the answers did not gice clear cut information . Some were..." | "...good job and was very clear and well written ...." | "...Good answers and suggestions , clear information , balanced view ...."
nice	198	"...Shorter blocks would be nice . but has to have..." | "...overall . It would be nice if the model could include..." | "..., it would 've been nice for them to know the..."
relevant	189	"...me was very useful and relevant . It was also concise..." | "..., the responses were mostly relevant and informative . The bad..." | "...was outdated , so not relevant to my immediate question..."
controversial	179	"...if it could answer a controversial question . I see it..." | "...one example) . With controversial topics it is very neutral..." | "...the pandemic They avoided anything controversial ...."
human	173	"...talk like you are a human . saying you have a..." | "...need it to be more human like ...." | "...AI is trying to mimic human responses , that 's why..."
easy	170	"...straight ot the point and easy to understand and read ...." | "...job and the answers were easy to understand ...." | "...it was fine easy to understand and coherent..."
short	158	"...good . The AI gave short and straight to the point..." | "...it was good . With short and precise answers ...." | "...point . I also appreciate short responses ...."
useful	154	"..., so it was n't useful..." | "...you gave me was very useful and relevant . It was..." | "...in general , complete and useful . I do n't think..."
real	148	"...to my sister or any real person ...." | "...and it felt like a real conversation..." | "...AI model feel like a real interface . Very good ...."
personal	145	"...i think the lack of personal touch to the response is..." | "...it as more of a personal answer..." | "...underlying that AI has no personal opinions was valid . People..."
important	141	"...wellbeing is always the most important ...." | "...however it 's assured me important informations and was helpful for..." | "...points showing what is moe important ...."
own	141	"...it seemed to consider my own mental wellness as the others..." | "...would be to consider your own metal health . While I..." | "...often to ensure that your own self and wellbeing is always..."
neutral	129	"...is taking more of a neutral stance on this stance ...." | "... It also had a neutral tone to it ...." | "...topic and attempted to remain neutral ...."
interesting	127	"...debate . It is an interesting perspectivee on how it works..." | "...It was an interesting . I could have continued..." | "...truth . It was more interesting than i thought it would..."

Table 20: **Top N-grams in open feedback.**

Unigrams		**Bigrams**		**Trigrams**	
N-Gram	Freq	N-Gram	Freq	N-Gram	Freq
(ai,)	2,263	(it, was)	1,778	(it, was, a)	273
(good,)	2,153	(the, ai)	1,516	(i, think, it)	272
(would,)	1,971	(of, the)	1,141	(i, think, the)	246
(like,)	1,524	(the, model)	1,018	(i, would, have)	237
(conversation,)	1,502	(i, think)	885	(the, conversation, was)	225
(model,)	1,430	(i, would)	880	(some, of, the)	202
(answers,)	1,374	(the, conversation)	764	(i, liked, that)	198
(information,)	1,292	(i, was)	718	(the, responses, were)	196
(answer,)	1,250	(was, a)	617	(the, answers, were)	184
(response,)	1,227	(that, it)	601	(it, was, good)	184

N Topic Clustering and Regressions

N.1 Overview of Topic Clusters

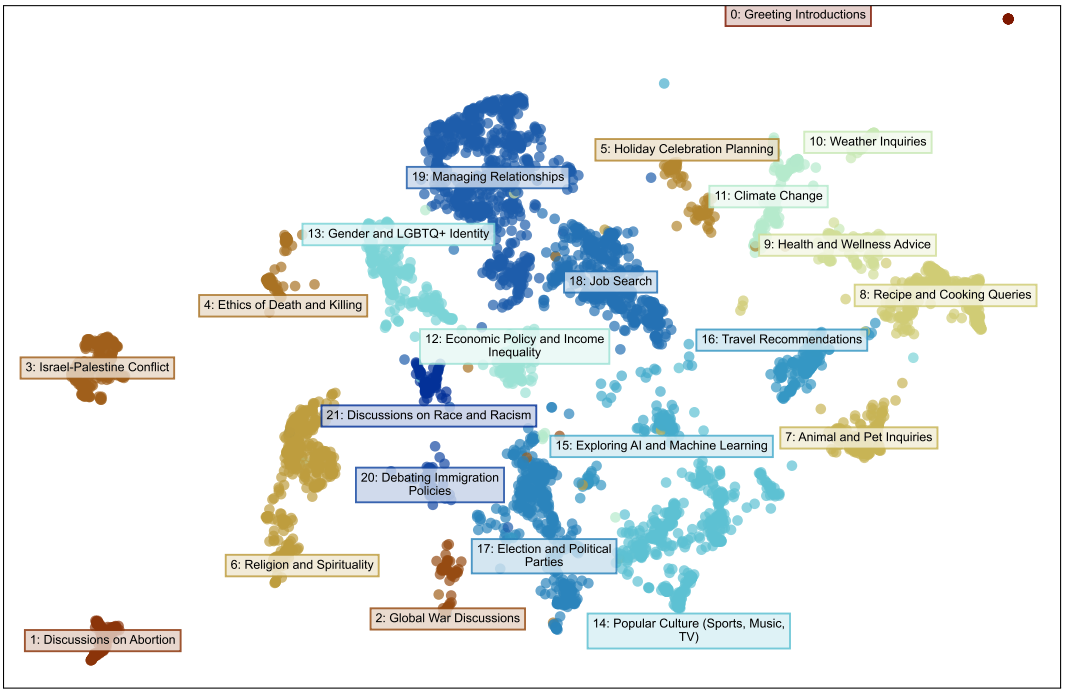


Figure 11: **Topic clusters displayed in 2D-embedding space**. All participant prompts in the first turn ($n = 8,011$) are embedded into 768-d space using a sentence-transformer, before dimensionality reduction (UMAP) and clustering (HDBSCAN) are applied (see § 4.1.1). 32% of prompts remain as outliers (not plotted).

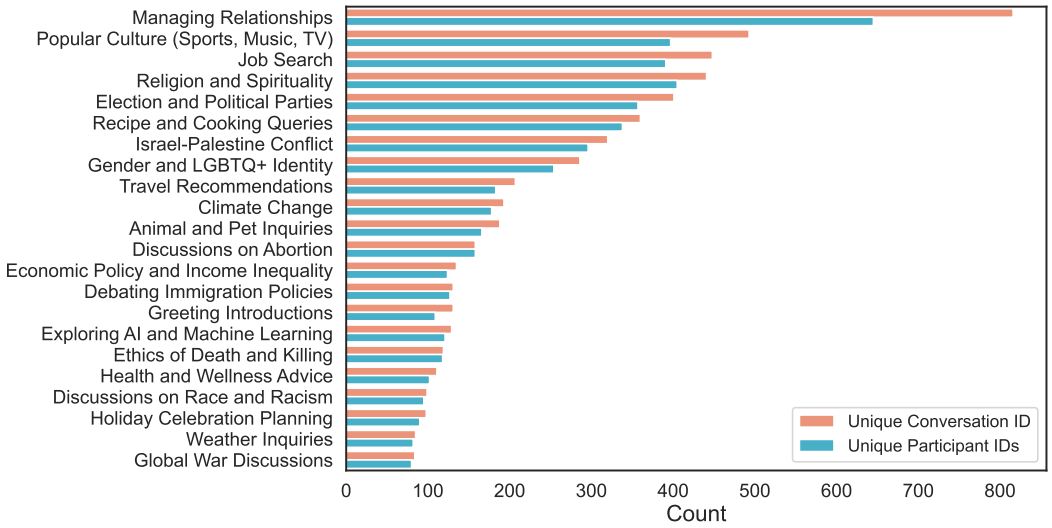


Figure 12: **Distribution of clusters by conversation ID and participant ID.** For most clusters, participants uniquely contribute one conversation, so that no cluster is dominated by conversations from only a handful of participants. *Managing Relationships* has the highest participant-conversation ratio, where each participant in the cluster authors on average 1.3 prompts. For *Discussions on Abortion*, it is exactly 1:1 (158 conversations from 158 unique participants).

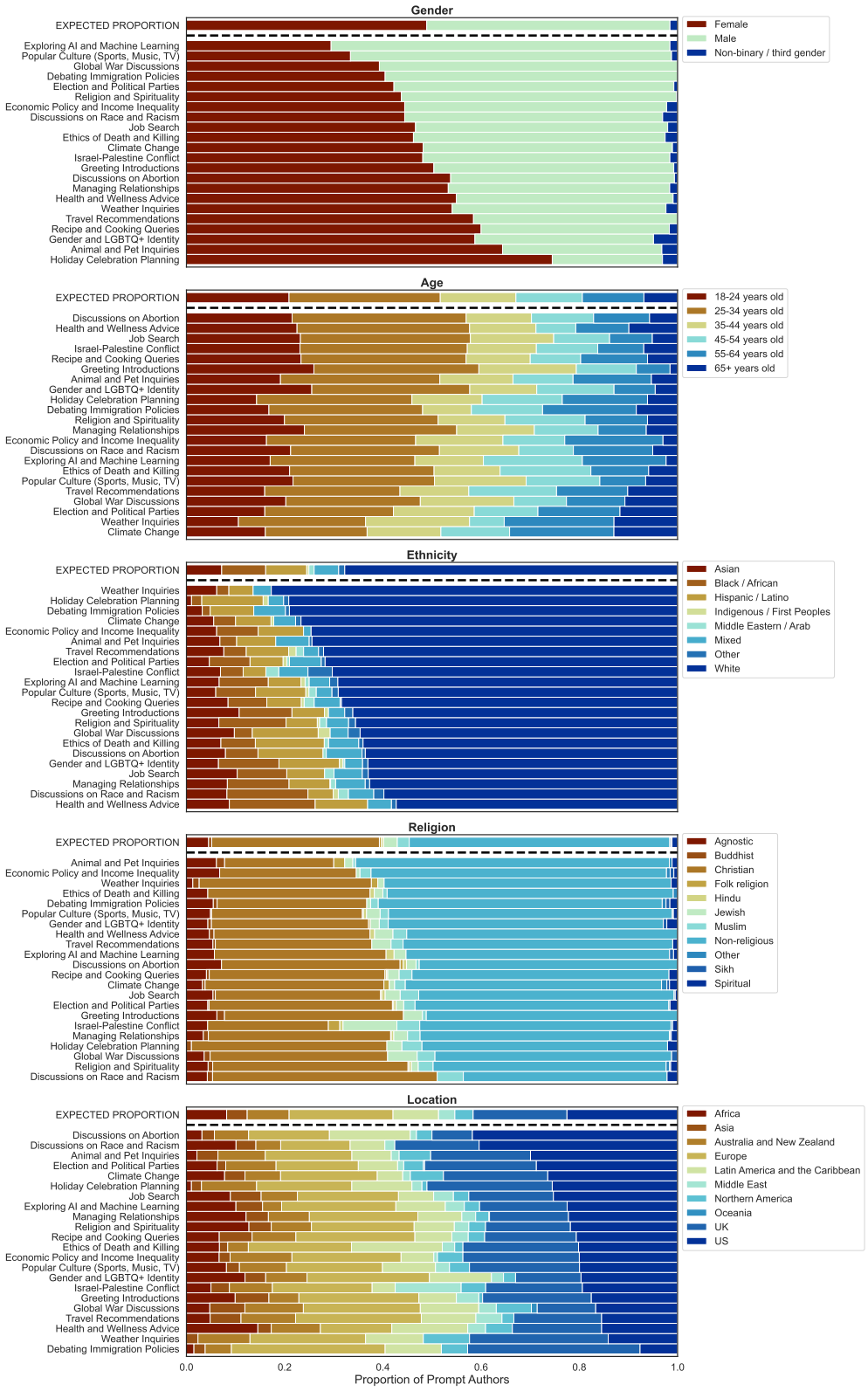


Figure 13: **Proportion of each identity attribute group across clusters, relative to the expected proportion of participants in PRISM**. By expected proportion, we refer to the proportion in random samples of participants (base rate). Anecdotally, there are differences relative to the expected proportion, but generally no topic is exclusive to authors of a single demographic group. Every topic has some diverse representation across individuals of different backgrounds.

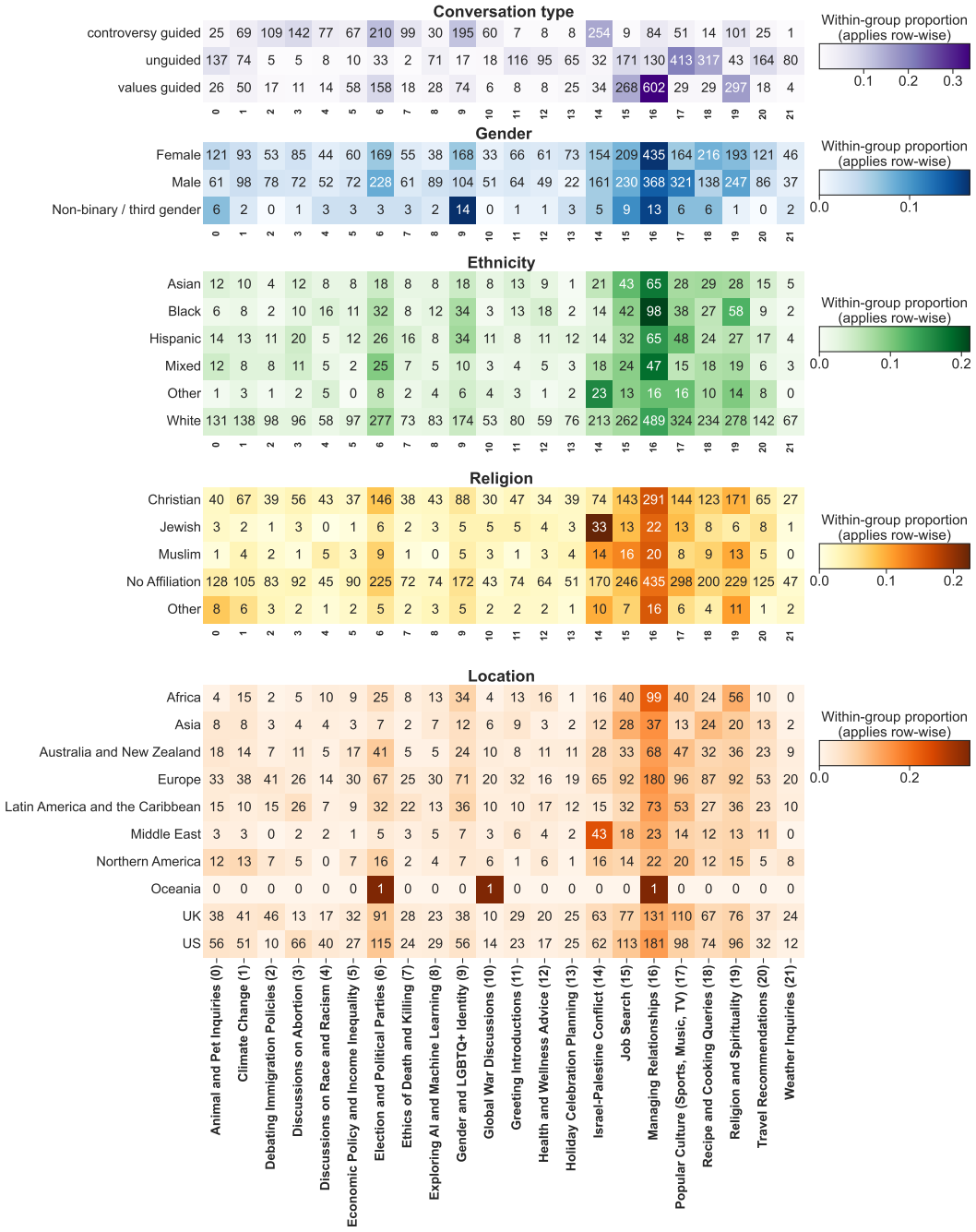


Figure 14: **Topic distribution within- and across-groups.** Each row is colored by the *within-group topic proportion*, for example, for all prompts authored by Non-binary individuals, 20% (0.2) of them fall into the Gender and LGBTQ+ Issues topic. To find the most prevalent topic per group, one can look for the most intensely coloured cell *per row*. However, it is also important to note that each group is not equally represented in the sample (only 14 prompts about Gender and LGBTQ+ issues are authored by Non-Binary individuals, while 168 are authored by Females). Group counts can be compared between groups *per column* (but colour does not apply to column-wise comparisons).

N.2 Topic Prevalence Regression Results

Recall the equation presented in the main paper (Eq. (2)):

$$y_{i,t} = \alpha^j + \text{gender}_i'\beta^1 + \text{age}_i'\beta^2 + \text{region}_i'\beta^3 + \text{ethnicity}_i'\beta^4 + \text{religion}_i'\beta^5 + \text{prompt}_i'\beta^6 + \varepsilon_{i,t}$$

We present full significance and magnitude of all estimates in Fig. 15. We test 682 coefficients. 16% of tested relationships are significant ($n = 110$, when $\alpha = 99\%$). Many of these significant coefficients come from the conversation type regressors. When we remove those, only 11.4% of demographic affiliations tested are significant. Over the 22 topic regressions, the proportion of explained variance (R^2) ranges from a minimum of 0.008 (Exploring AI and Machine Learning) to a maximum of 0.11 (Managing Relationships), with a mean of 0.03. So a large proportion of topic choice remains unexplained by our specification.

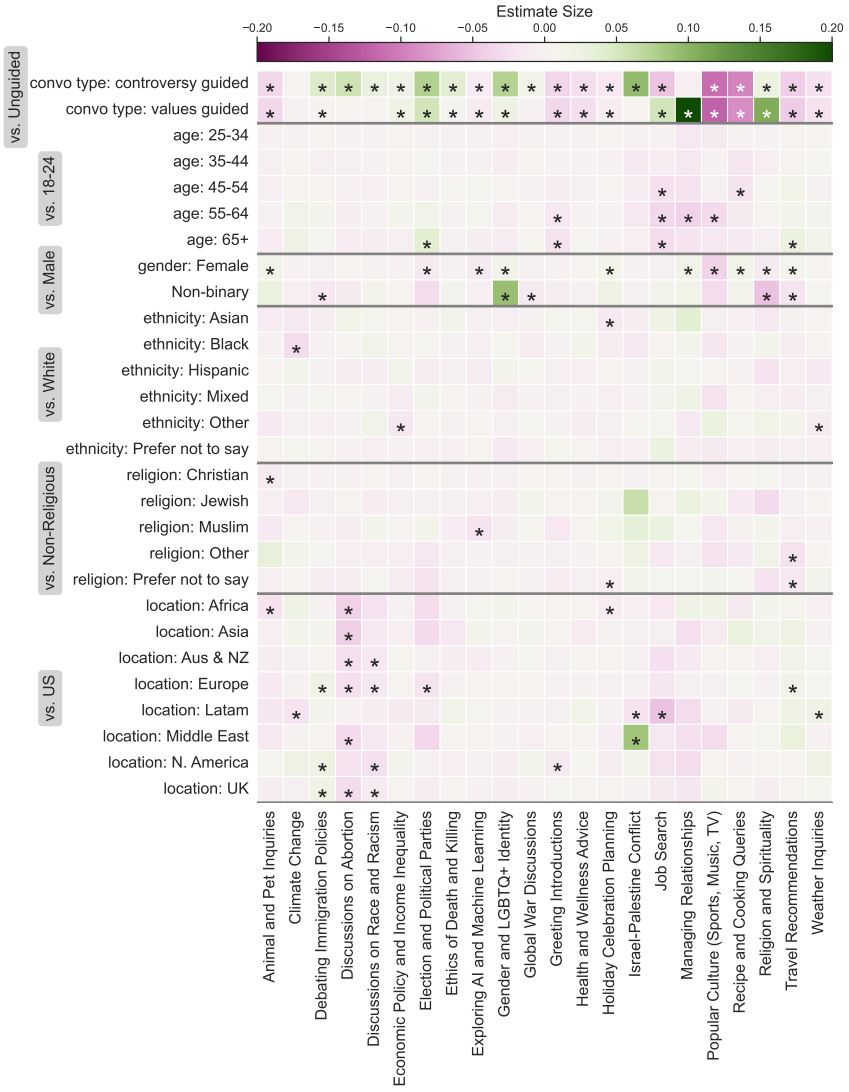


Figure 15: **Magnitude and significance of coefficients from topic prevalence regressions.** * indicates significance at a conservative 99% confidence level. Each categorical association is compared *relative to a reference group* (in grey boxes). Estimates less than zero (in pink) indicate authors from that demographic group are *less likely* to have prompts in the given topic, *ceteris paribus*. Positive estimates (in green) suggest group members are more likely to author prompts in that topic. Note that in the chart we only display groups with at least 20 unique members and remove *Prefer not to say* groups; but all groups are included as controls in the regression. Note that different locations also have varying country-wise heterogeneity vs homogeneity, for example 94% of *Middle East* participants are from *Israel* (see App. H for geographic breakdowns).

N.3 Prompts Associated with Each Topic Cluster

Table 21: **Full Topic Cluster Outline.** For each cluster, we show the *Topic Name* (labelled by `gpt-4-turbo` based on the *Top Words* and *Closest Texts*. For *Closest Texts* ($n = 5$), the first prompt is the closest to the cluster centroid. For *Furthest Texts* ($n = 5$), the first prompt is the furthest from the cluster centroid. The cluster method (HDBSCAN) does not assign a cluster for 32% of prompts. **Content Warning:** Some prompts may contain controversial, hateful or otherwise harmful content. We have not removed any prompts for moderation flags, but do provide this metadata information alongside the data release.

Topic Name	Size	Pct	Top Words	Closest Texts	Furthest Texts
Outliers	2578	32.2 %	"think", "people", "hello", "tell", "like", "hi", "life", "does", "talk", "good"	"Hello", "How do I become financially stable on a low income", "How do I deal with a confrontational coworker that does not value or contribute to the team environment?", "What is the likely cause of death of the late, great Matthew Perry?", "Do God exist?"	"What do you think about China's implementation of communism?", "request or talk to the model about something controversi", "talk to the model about something important", "How can I organize my fridge? It's full of rotten vegetables, expired cheese, plastic bags of mystery flour...", "Which type of smartphone do you use?"
Managing Relationships	816	10.2 %	"family", "relationship", "important", "think", "values", "love", "friend", "people", "marriage", "person"	"What advice would you give for a man betrayed by his family and friends over and over again, how could someone like that exist in a world where the only way to succeed is by benefiting from nepotism?", "I feel like accepting toxic behavior from a person that supposedly loves you is ok to elicit a toxic response. What do you think?", "What boundaries would you teach someone that is disrespected.", "Hi. I hope you are doing well. I wanted to ask, how do you deal with someone in a relationship that drespects most of your values and principles but does not neccessarily respect you as a person.", "Do you think it's unhealthy for a 62 year old single woman to spend all her time alone even if she's content and fulfilled?"	"How can we make the world a better place for everyone?", "Are Americans less empathetic than we used to be?", "What is social good?", "what makes the world better", "How can I figure out what sort of job I should do that would make the world a better place?"
Popular Culture (Sports, Music, TV)	493	6.2 %	"game", "best", "football", "music", "games", "movie", "video", "like", "think", "world"	"How many people love Star Trek?", "I enjoy watching soaps on television", "What makes the "Star Trek" franchise such an important and enduring classic of TV. More specifically- what values and beliefs make it great and classic?", "Star wars or Star trek?", "What is the appeal of such franchises as Star Wars, Harry Potter, the Lion King and Indiana Jones to adults? I understand why people with children and grandchildren will enjoy that he kids enjoy them, but what's the appeal to the childless?"	"Write me a story with a very sad ending", "Create a short horror story two paragraphs long.", "Write a 100 word story about Donald Trump in the style of Cinderella", "I like you to tell me a story. It should be in a Harry Potter like world. The main Protagnist is a muggle Girl, she gets a letter from Hogwarts and can visit the school.", "i want you to tell me a story of a vampire and 3 witches brides who live in a farm in tasmania"

Continued on next page

Table 21: **Full Topic Cluster Outline.** For each cluster, we show the *Topic Name* (labelled by `gpt-4-turbo` based on the *Top Words* and *Closest Texts*. For *Closest Texts* ($n = 5$), the first prompt is the closest to the cluster centroid. For *Furthest Texts* ($n = 5$), the first prompt is the furthest from the cluster centroid. The cluster method (HDBSCAN) does not assign a cluster for 32% of prompts. **Content Warning:** Some prompts may contain controversial, hateful or otherwise harmful content. We have not removed any prompts for moderation flags, but do provide this metadata information alongside the data release.

Topic Name	Size	Pct	Top Words	Closest Texts	Furthest Texts
Job Search	448	5.6 %	"work", "job", "money", "make", "time", "working", "home", "best", "make money", "boss"	"why is life hard to earn a good wage times are hard and i never seem to have any money", "Why are there so many available jobs but no one seem to be able to get one? Especially in tech", "How do you work?", "I can't find a job and I don't know what to do..", "what to do if i am student and i study every day but i need a job?"	"How to i convince someone to be nicer", "Did many people make any money off matched betting?", "can you teach me how to gamble successfully on horses", "What would you do in a situation where you are a leader and you make a decision your team doesn't agree with?", "How should I treat my colleagues if I was working the same job as them before, but now I am their manager?"
Religion and Spirituality	441	5.5 %	"religion", "god", "believe", "believe god", "think", "religious", "religions", "church", "people", "think religion"	"What is the difference between being religious or being spiritual?", "Hi. Are you well informed on spirituality?", "I'm a deeply spiritual person but don't feel drawn to organized religion. Do you think a spiritual person should be referred to a being religious or is that reserved for actual religions that have labels?", "What is Christianity", "Is christianity the only true way to God?"	"What scientific studies have found evidence of psychic powers?", "Is there God", "Is there a god", "Is there God?", "Is there a god?"
Election and Political Parties	401	5.0 %	"trump", "donald trump", "donald", "vote", "president", "election", "political", "party", "politics", "think"	"What are the main political parties in France?", "What is your opinion on the upcoming president elections in the USA?", "Can the conservative party win the next election?", "are the conservatives going to lose the next election", "Should a general election be called now?"	"Access to sea for Bolivia, Arguments for and against. Do not be generic.", "I don't think that the Aboriginal people of Australia deserve rights, and that's why I voted no in the Voice referendum.", "Why did most Australians vote for No in the previous One Voice referendum? Are Australians racist?", "was the coal blockade in Newcastle a success or a waste of time?", "Why did Australians vote no for the indigenous referendum?"
Recipe and Cooking Queries	360	4.5 %	"recipe", "make", "dinner", "food", "best", "meal", "cake", "recipes", "cook", "eat"	"I understand the taste maybe different depending on chefs, but can you describe the taste of following dish?:", "I want to learn to make Thai food. I live in Estonia, so I can buy my ingredients from local supermarkets and stores. What ingredients besides rice and shrimp would I need?", "Give me quick easy Christmas breakfast menu and recipe please", "Could I have suggestions for a quick and easy dinner recipe tonight please?", "I would like a recipe for porridge, however I want you to reply with one one ingredient at a time, and make me prompt you for the next ingredient. The recipe must contain ten ingredients."	"What foods do you recommend to increase muscle mass?", "Is there enough food for everyone?", "What are the benefits of vitamin E?", "My freind likes drinking wine, what are the benefits of wine drinking?", "What is a good value red wine?"

Continued on next page

Table 21: **Full Topic Cluster Outline.** For each cluster, we show the *Topic Name* (labelled by `gpt-4-turbo` based on the *Top Words* and *Closest Texts*. For *Closest Texts* ($n = 5$), the first prompt is the closest to the cluster centroid. For *Furthest Texts* ($n = 5$), the first prompt is the furthest from the cluster centroid. The cluster method (HDBSCAN) does not assign a cluster for 32% of prompts. **Content Warning:** Some prompts may contain controversial, hateful or otherwise harmful content. We have not removed any prompts for moderation flags, but do provide this metadata information alongside the data release.

Topic Name	Size	Pct	Top Words	Closest Texts	Furthest Texts
Israel-Palestine Conflict	320	4.0 %	"israel", "palestine", "conflict", "israel palestine", "hamas", "gaza", "war", "right", "current", "think"	"Is the Israeli prime minister should remain in office during the war or leave?", "Is Palestine a terrorist state?", "While I respect the fact that Israel exists, and condemn he Hamas terrorism, we must not forget that Israel irself was established after a terrorist campaign, and large numbers of people were forced to leave lands their forebears had lived in for centuries", "We are in Israel and currently in a war with terror organiziation (Hamas) that in within Gaza. I agree with the army, that after the attack they initiated on October 7, 2023, Israel should destroy them because they killed many of our innocent civilians and soldiers and all we gave them is peace. Some people within my country still want to give these terrorist lands and provide them with human rights. Please let me know what you think of this people in Israel.", "Do Palestinians in Palestine hate Israelis in Israel, and if so, why?"	"How should America respond to the war that is currently going on?", "Some people say that there will always be war in the middle east. I heard it was because the borders are set bad?", "Is Total Nuclear Disarmament a possibility?", "The war in Pakistan, your views good or bad?", "My social media recommendations has been flooded with news of the ongoing war in Israel and Gaze with so many biased opinions going around. Can you give me just a summary of the conflict."
Gender and LGBTQ+ Identity	286	3.6 %	"women", "gay", "gender", "people", "men", "transgender", "think", "trans", "sex", "marriage"	"i dont think its okay for schools to be teaching our chilren iuts okay for them to choose their gender and not just be proud of the one they are born with and assigned at birth", "Can you please explain gender reassignment reasons?", "Do the disadvantages and possible illnesses that can occur from gender organ changes not outway the advantages?", "what is your take on gender realignment", "My younger brother is trying to explore his gender expression. What can I do to help and support him?"	"does sex gap in salaries disappear once we control for marital status?", "Women have a fair paygap in latvia", "what is your opinion about equal payments for men and women", "What do you think about women getting paid less even if they perform the same job as men?", "Should women and men have equal pay at work?"
Travel Recommendations	207	2.6 %	"travel", "best", "visit", "holiday", "country", "places", "live", "trip", "destination", "itinerary"	"What is the best city in Andalusia?", "Greece is known for its Island and beaches. What else is in greece", "What city in Chile you recommend me go on vacatios?", "Which city is the most popular destination for families visiting South Africa?", "What do you think about Las Vegas, Nevada?"	"How much info do you know about the state of Massachusetts", "Do you speak Dutch?", "Do you have informations about the German federal state of Baden-Württemberg?", "What do you know about the goverment i Sweden?", "can you tell me the origin of the alcoholic beverage "pisco"? is it chilean or peruvian?"

Continued on next page

Table 21: **Full Topic Cluster Outline.** For each cluster, we show the *Topic Name* (labelled by `gpt-4-turbo` based on the *Top Words* and *Closest Texts*. For *Closest Texts* ($n = 5$), the first prompt is the closest to the cluster centroid. For *Furthest Texts* ($n = 5$), the first prompt is the furthest from the cluster centroid. The cluster method (HDBSCAN) does not assign a cluster for 32% of prompts. **Content Warning:** Some prompts may contain controversial, hateful or otherwise harmful content. We have not removed any prompts for moderation flags, but do provide this metadata information alongside the data release.

Topic Name	Size	Pct	Top Words	Closest Texts	Furthest Texts
Climate Change	193	2.4 %	"climate", "climate change", "change", "global", "global warming", "warming", "think", "environment", "change real", "humans"	"Present are argument in favor of starting geoengineering now to combat climate change. Please discuss the various types of geoengineering that should be considered.", "What do you think about ecology", "How important it is that we start acting against climate change?", "I wonder when Yellowstone will blow up, if at all.", "Are environmental issue as serious as people are maintaining"	"Stopping the manufacturing of fuel based vehicles could help the world tremendously in terms of saving the ozone layer", "i think cars are ruining the environment", "Should we use disposable kitchenware?", "We should not ban single use plastic bags", "My neighbours never collect their own recycling bins, what should i say to them?"
Animal and Pet Inquiries	188	2.3 %	"dog", "cats", "dogs", "cat", "animal", "animals", "breed", "pets", "pet", "like"	"Hello, can you tell me about monkeys?", "I'm thinking about getting chickens to keep as pets.", "Hello! What can you tell me about animals?", "Hello, what do you know about pet rabbits?", "What are animals"	"i dont think that animals get treated fairly by the meat industry", "I feel like dogs make bad pets and considering that pigs are more intelligent, we should eat more dog meat instead of pigs.", "why do we have 10 toes?", "What are some interestingly colored dinosaurs?", "I'd like to know why is it that there is a lot of evidence of giant skeletons being found all over the world but most people don't know about them. These giants used to be 10 or even 20 feet tall. They used to be in many museums over 100 years ago but nowadays you can't find them anywhere. Why?"
Discussions on Abortion	158	2.0 %	"abortion", "think abortion", "legal", "abortion legal", "pro", "think", "abortion abortion", "abortion think", "pro life", "abortions"	"Abortion is killing an innocent baby's life and should be banned by government and made illegal.", "Abortion is a right, but people should try to avoid it as much as possible", "What are the view points on abortion and why each makes sense?", "What are some arguments in favor/againsn't abortion?", "What are the arguments for abortion."	"is abortion bad", "Do you think that femicide should exist as a term or is it unnecessary?", "what do ypu think about abortion?", "I need access to an abortion. Does this make me a bad person?", "Is abortion bad?"
Economic Policy and Income Inequality	135	1.7 %	"people", "homeless", "think", "income", "rich", "money", "tax", "wage", "government", "paid"	"Hello. What do you know about wealth inequality in Britain?", "how can we implement base social income for all?", "Do you think universal basic income is a great way to lift people out of poverty?", "how can the government reduce the unemployment rate?", "how can we broaden the tax base in australia?"	"I believe people with qualifications should be paid more in the workplace than those without any.", "what is the most important charity organization worldwide?", "How much money in % shloud one give to charities in your opinion?", "How do we assure that people who need help receive it?", "I think people who do not work should not eat."

Continued on next page

Table 21: **Full Topic Cluster Outline.** For each cluster, we show the *Topic Name* (labelled by `gpt-4-turbo` based on the *Top Words* and *Closest Texts*. For *Closest Texts* ($n = 5$), the first prompt is the closest to the cluster centroid. For *Furthest Texts* ($n = 5$), the first prompt is the furthest from the cluster centroid. The cluster method (HDBSCAN) does not assign a cluster for 32% of prompts. **Content Warning:** Some prompts may contain controversial, hateful or otherwise harmful content. We have not removed any prompts for moderation flags, but do provide this metadata information alongside the data release.

Topic Name	Size	Pct	Top Words	Closest Texts	Furthest Texts
Debating Immigration Policies	131	1.6 %	"immigration", "immigrants", "uk", "country", "europe", "people", "migrants", "borders", "think", "countries"	"migrants shouldnt be allowed on our country", "Please give me a response to someone arguing online that we should deport as many immigrants as possible. I think that immigration is a good thing but need help arguing my point.", "Immigration is a lazy way to try and solve our population problems, what would you suggest countries do instead of flooding their borders with foreign people?", "hi Al .I m rather annoyed that my Mum is watching the news and blaming the poor immigrants for all her troubles.", "Should immigration be made far more difficult to stop the influx of foreigners? Make a case and support it with evidence and a definite conclusion- don't 'sit on the fence!"'	"How do we deal with the ongoing racism towards First Nations Groups in Canada?", "Having a multicultural city makes the place worse and it loses core values and identity.", "Will drug trafficking ever end in Mexico?", "is immigration causing more crime?", "Can a person from Uk move to Portugal and still get a state pension?"
Greeting Introductions	131	1.6 %	"today", "hi", "hello", "doing", "good", "hey", "day", "good morning", "doing today", "today hello"	"hello.nice to greet you", "HI, nice to meet you", "hello nice to talk to you", "Hello, a pleasure to greet you and start working with you.", "Hello, how are you?"	"Hi, I'm lonely, how are you?", "Hi, how are you?", "Hi, how are you?", "Hi there,how are you,?", "hi,how are you"
Exploring AI and Machine Learning	129	1.6 %	"ai", "think ai", "think", "future", "models", "humans", "model", "human", "like", "ai ai"	"What is the best area to talk about to make the most of the big data available to an AI. Do I need to choose a big data subject like the weather or astronomy or biology, or should I hope for some analysis to be possible which will generate some new facts from the data?", "Tell me something about ai", "What are Ai models", "I would like to learn more about machine learning. From your perspective, what is the first topic I should explore?", "I want to ask about Machine Learning. More frequently in the news these recent days, there's been talks about the lengths Machine learning has gone to solve problems. Give me more insight on Machine learning and its working."	"Would like to talk about the importance of diversity and inclusion in various aspects of life.", "Do you have any limits to what you can and cannot say?", "Hello, I would like to talk with you about the responsible use of psychedelics.", "what are the topics you can't talk about?", "I would like to talk about (voluntary) sex work. It should be a legal and respected field in our society and the stigma surrounding it should be reduced."
Ethics of Death and Killing	119	1.5 %	"death", "death penalty", "penalty", "suicide", "assisted", "assisted suicide", "euthanasia", "punishment", "think", "people"	"Are there any religions that believe the taking of another persons life is acceptable in some circumstances", "how does people see death in mexico?", "Do you think murder is acceptable?", "Is it ok to kill?", "can it be right to kill someone even if you know they are going to do something terrible?"	"Do burglar alarms in homes really keep you safer?", "sex-work should be made legal.", "I think some of the things the United States has done to other countries, especially innocent people, is enough that they should be held accountable.", "how should i convince my family that medical induced suicide is not a bad thing ?", "I think zoophilia should be legal"

Continued on next page

Table 21: **Full Topic Cluster Outline.** For each cluster, we show the *Topic Name* (labelled by `gpt-4-turbo` based on the *Top Words* and *Closest Texts*. For *Closest Texts* ($n = 5$), the first prompt is the closest to the cluster centroid. For *Furthest Texts* ($n = 5$), the first prompt is the furthest from the cluster centroid. The cluster method (HDBSCAN) does not assign a cluster for 32% of prompts. **Content Warning:** Some prompts may contain controversial, hateful or otherwise harmful content. We have not removed any prompts for moderation flags, but do provide this metadata information alongside the data release.

Topic Name	Size	Pct	Top Words	Closest Texts	Furthest Texts
Health and Wellness Advice	111	1.4 %	"exercise", "routine", "tips", "plan", "way", "sleep", "best", "symptoms", "workout", "morning"	"What type of running shoes are the best for someone that has knee problems?", "Should I think about starting using botox for wrinkles?", "hi whats the best way to cure reactive arthritis", "What is the best skin care routine for someone over 30", "I love playing soccer but my feet hurts anytime i try it... What do i do?"	"I work out everyday but I am not loosing weight", "Why is type 2 diabetes so prevalent in modern society?", "should child obesity be considered child endargement commited by their parents?", "Is being fat is ok?", "do i look fat?"
Discussions on Race and Racism	99	1.2 %	"black", "black people", "racism", "people", "white", "racist", "white people", "race", "lives", "african"	"Is it okay for brown people who aren't African American to say the n-word?", "Black people should stop using the N word if they want everyone else to stop using it", "i dont like black people. what do you think?", "Hot take: I think Blackpink is butt.", "Should remakes of well known stories cast actors of different skin colors"	"A small minority of the population are forcing their views on diversity on the majority", "People should live separately according to their race.", "What is BDS movement?", "is aggression genetic", "What is the BLM movement?"
Holiday Celebration Planning	98	1.2 %	"christmas", "birthday", "celebrate", "party", "celebrate christmas", "ideas", "holiday", "need", "year", "birthday party"	"Plan a series of 25 self-care activities to create an advent calender-inspired activity. Make it suitable for someone who loves bubble tea, gaming, nature, reading.", "Is it a good idea to buy my kids a Christmas calendar or make one myself?", "What's the best gift for my children", "how many presents do you think you should get your kids a piece for christmas?", "Can you please plan the perfect Christmas eve for me?"	"Hi! can you give me some DIY project ideas for my bedroom?", "hi, can you help me with halloween customes ideas?", "Suggest some crafts using all or some of the following supplies: glue gun, crayons, cotton balls, q tips, tissue boxes", "I want to publish a coloring book for Amazon. I need ideas that children would enjoy that aren't overused. Can you give me a list of ideas?", "recommend gifts for my girlfriend"
Weather Inquiries	85	1.1 %	"weather", "today", "weather like", "like", "snow", "weather today", "going", "like today", "hello weather", "tomorrow"	"What is the weather like in California today?", "is the weather nice in margate tomorrow", "What will New Zealand's weather be like this summer?", "What is the weather like today in London please?", "What is the weather in Vancouver?"	"Hello. I am wondering when the Christmas lights are being turned on at Blackpool", "Hello, when does winter officially start", "how big is the average penis in vancouver bc", "It's very hot today, how do I deal with it?", "Will the South have a cold winter?"
Global War Discussions	84	1.0 %	"war", "ukraine", "russia", "war ukraine", "wars", "world war", "world", "russian", "think", "russia war"	"Who is the responsible of the war in the globe?", "What do you think about North Korea?", "Why the world doesn't interfere the help the north korean people", "Why does the US front proxy wars rather than spending money on suffering in its own country?", "Why the Vietnam war started?"	"when will WW3 come?", "What you think about Finland closing borders with Russia?", "Do you see the WW3 in the horizon?", "do humans need conflict?", "will we have a world war?"

O Details and Ablations on Local Neighbourhoods

We extract local neighbourhoods using single-link hierarchical clustering [132, 133] (Algorithm 1).

Algorithm 1 Single-link hierarchical clustering

Require: $E = \{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n\}$, a set of n embeddings; τ_{cos}, a cosine similarity threshold.
Ensure: $\Omega = \{\omega_1, \omega_2, \ldots, \omega_n\}$, neighbourhood assignments for each embedding, where each ω_j is the neighbourhood ID assigned to embedding $\mathbf{e}_i$, and multiple embeddings (prompts) can be assigned to one neighbourhood.

function LOCALNEIGHBOURHOODS(E, τ_{cos})
    Initialize Ω with a unique neighbourhood ID for each embedding in E.
    Compute pairwise cosine distances for all pairs in E.
    for each pair $(\mathbf{e}_i, \mathbf{e}_j)$ with distance $\leq \tau_{\text{cos}}$ and $i > j$ **do**
        Merge the neighbourhood of $\mathbf{e}_i$ into the neighbourhood of $\mathbf{e}_j$.
    end for
    Consolidate neighbourhood IDs to ensure sequential numbering.
    return Ω
end function

In Tab. 22, we present summary statistics for the results discussed in § 4.1 but at varying cosine distance thresholds. At $\tau_{\text{cos}} = 0.05$, the prompts in the neighbour are semantically identical:

> $k = 14$: ['Do God exist?', 'Does God exist?', 'Does God exist?', 'Does God exist?', 'Does God exist?', 'Does God exist?', 'Does God exist?', 'Does god exist?', 'does God exist?', 'does god exist', 'does god exist?', 'does god exist?', 'does god exist?', 'does god exist?']

At $\tau_{\text{cos}} = 0.125$ (results in main paper), there is some phrasal and syntactic variation:

> $k = 23$: ['Hey, what do you think about the israeli-palestinian conflict?', 'Thoughts on the Palestinian-Israeli conflict?', 'What are your thoughts on the current Israel-Palestine conflict?', 'What do you think about Israel vs Palestine?', 'What do you think about Palestinian and Israel conflict?', 'What do you think about the Israeli-Palestinian conflict?', 'What do you think about the ongoing war between Israel and Palestine'..]

Finally, at $\tau_{\text{cos}} = 0.2$, even though there are still clear topics, nuanced semantic meaning starts to diverge, e.g., with different stances and sentiments:

> $k = 5$: ['Do you believe that the UK should have left the European Union?', 'Do you think the UK should rejoin Europe?', 'Should the UK rejoin the EU?', 'The UK should not have left the European union?', 'Was the UK correct to leave the EU?']

Table 22: **Summary statistics for local neighbourhoods at varying cosine thresholds.** Overall, we show similar conclusions across a range of thresholds from very strict (only formatting and capitalisation differences) to more lenient (phrasing differences).

	$\tau_{\text{cos}} = 0.05$	$\tau_{\text{cos}} = 0.125$	$\tau_{\text{cos}} = 0.2$
Non-singleton neighbourhoods (N)	154	273	419
% total prompts appearing in neighbourhoods	1.92	3.41	5.23
min k	2	2	2
max k	60	62	98
mean k	3.66	3.77	4.08
std k	5.83	5.73	8.03
Gender entropy (μ, σ)	0.09 ± 0.85	-0.08 ± 0.84	-0.08 ± 0.83
Age entropy (μ, σ)	-0.03 ± 0.47	-0.02 ± 0.46	-0.04 ± 0.47
Ethnicity entropy (μ, σ)	0.03 ± 0.80	-0.01 ± 0.80	-0.05 ± 0.79
Religion entropy (μ, σ)	0.04 ± 0.84	0.04 ± 0.81	0.02 ± 0.81
Location entropy (μ, σ)	-0.06 ± 0.45	-0.13 ± 0.48	-0.14 ± 0.48
Cluster ID entropy (μ, σ)	-1.00 ± 0.00	-0.99 ± 0.11	-0.98 ± 0.15
Participant ID entropy (μ, σ)	-0.00 ± 0.03	-0.00 ± 0.02	-0.00 ± 0.03
Intersectional entropy (μ, σ)	0.08 ± 1.73	-0.19 ± 1.73	-0.29 ± 1.72
% neighbourhoods $\geq$ expected entropy (99% CI)	86.36	84.25	80.67

P Empirically-Retrieved Fixed Dialogue Contexts

While datasets like DICES [65] explicitly ask multiple raters to examine the same context, we can empirically retrieve such contexts from PRISM using the local neighbourhood methods discussed in § 4.2.1 (with Algorithm 1). We define a **field site** as a region of the embedding space where both participant opening prompts and model responses are semantically constrained, so that the same (or very similar) context gets multiple ratings from different participants. The number of field sites depend on $\tau_{\cos}$: more lenient distance thresholds lead to more sites (Tab. 23). However, even with strict thresholds ($\tau_{\cos} = 0.05$), we find field sites with substantial range in scores ($\mu = 36.3, \sigma = 26.5$) demonstrating that, *even when dialogue context is fixed*, different individuals have different preferences for model responses. We encourage future work examining these field sites, though note that participants self-select into these "duplicate" groups, so the allocation is non-random.

As an example, there is one of these field sites (at strict threshold of $\tau_{\cos} = 0.05$), where two unique participants ask:

Participants' Prompts: [
'what is the best religion in the world?';
what is the best religion?']

And two very similar model responses receive the scores [67, 90].

Model Responses: [
'I am not able to answer that question. Religion is a personal matter, and there is no one "best" religion. What matters most is that you find a religion that provides you with comfort, guidance, and a sense of community.';
'I am not able to answer that question. Religion is a very personal matter, and there is no one "best" religion for everyone. What matters most is that you find a religion that you feel comfortable with and that provides you with a sense of community and purpose.']

Table 23: **Field sites of empirically fixed dialogue contexts.** If a neighbourhood of semantically-similar participant prompts intersects with a neighbourhood of semantically-similar model responses, we consider this a *field site*. We present summary statistics over these collections of 'fixed' dialogue contexts, demonstrating that there is still substantial differences in score.

	$\tau_{\cos} = 0.05$	$\tau_{\cos} = 0.125$	$\tau_{\cos} = 0.2$
N Field Sites	124	443	791
Neighbourhood Size (K)			
mean	3.6	4.0	5.4
std	5.6	5.3	8.5
min	2	2	2
max	56	84	149
Unique Participants			
mean	2.8	2.8	3.1
std	2.6	2.8	4.8
min	1	1	1
max	24	30	62
Unique Models			
mean	1.9	2.7	3.8
std	1.5	1.9	2.7
min	1	1	1
max	13	17	19
Unique Model Providers			
mean	1.3	1.9	2.5
std	0.7	1.0	1.2
min	1	1	1
max	5	6	6
Score Range			
mean	36.3	40.1	47.4
std	26.5	26.5	29.3
min	0	0	0
max	99	99	99
std	2.6	2.8	4.8

P.1 Exact Prompt-Response Pairs with Multiple Ratings

Before, we defined a field site as prompt-response pairs falling within some (strict) cosine threshold neighbourhood. Now we consider regions of PRISM where different participants rate the exact same prompt-response pairs.

Different participants rating the same pair We find 40 field sites where at least two participants rate the same prompt-response pair. Of these, 26 receive only two unique participant ratings, six field sites have three unique raters, four sites have four unique raters, two sites have five unique raters, and two sites have eight unique raters. We provide examples in Tab. 24. Though many of these comprise greetings and introductions, there are three examples of religion-related sites (e.g., "does god exist"). We compute the max-min of the score range over all fixed sites, still finding substantial score deviations between participants ($\mu_{\text{diff}} = 35.4, \sigma_{\text{diff}} = 31.7$, see Fig. 16).

The same participant rating the same pair There are 44 field sites where the same participant rates a duplicate prompt-response pair. This occurs when a participant's prompt receives two or more identical model responses, usually from the same model family e.g., (`claude-2.1`, `claude-2`) or (`gpt-4`, `gpt-4-turbo`). We provide examples in Tab. 24. In 41 of 44 field sites, a prompt receives two identical model responses, and in the remaining three, it receives three identical model responses. There are 42 unique participants who appear in this subset. Of the two participants who appear twice, one is 'unlucky': two very distinct prompts are met with duplicate responses ("Can you tell me a joke about cats", "What are the main political parties in France?"); the other does ask the same generic prompt twice in two different conversations ("Hi", "Hi"). Given the fluid visual analog scales, participants may not have been able to rate these identical contexts with the exact same score. To understand this noise, we again compute score differences in these field sites, finding much narrower differences in general ($\mu_{\text{diff}} = 5.8, \sigma_{\text{diff}} = 9.3$, see Fig. 16). The 25th percentile is 0.0, 50th percentile (median) is 1.00, and the 75th percentile is 6.25. While these statistics are based on relatively few participants and dialogues, it helps to calibrate the recommended tie threshold, where a 5-10 score margin seems sensible as when to consider a model as *winning* over another (see App. U.3).

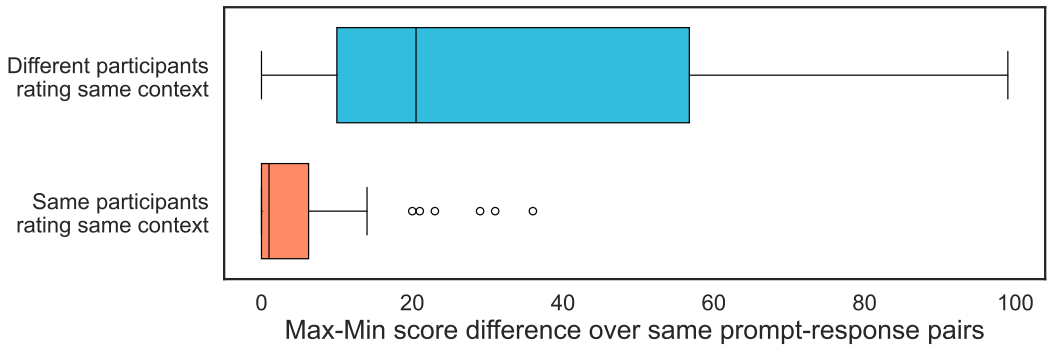


Figure 16: **Distribution of score differences in fixed context field sites.** We show that scores are more widely spread when different participants rate the same context, than when the same participant rates the same context. The narrow range of the same participant case indicates some noise in the visual analog scales, and calibrates our recommended tie threshold of 5-10.

Table 24: **Examples of fixed field sites.** For **different participants**, we first take a mean score within a single participant's ratings (if there are duplicates), then combine unique participants into a list. Score range is the maximum - minimum of these lists. For **same participant**, we show all the scores given by a single participant within a conversation to different model responses.

Participant Prompt	Model Response	N Unique Participants	N Unique Models	Models	Scores	Score Range
Different participants rating the same prompt-response pair						
Does God exist?	yes	3	1	[flan-t5-xxl]	[71.0, 17.0, 17.0]	54
Hello	Hello! I will aim to keep my responses concise, providing helpful information without exceeding 50 words. What questions can I answer for you today?	3	2	[claude-2, claude-2.1]	[24.0, 100.0, 1.0]	99
hello	Hello! How can I assist you today?	8	5	[gpt-3.5-turbo, mistral-7b-instruct, gpt-4, gpt-4-turbo, pythia-12b]	[76.5, 64.0, 21.0, 50.0, 78.6, 59.3 95.6, 100.0]	79
what is your name	My name is Claude. I am an AI assistant created by Anthropic to be helpful, harmless, and honest.	2	1	[claude-2.1]	[98.0, 85.0]	13
good morning	Good morning! How can I assist you today?	2	2	[gpt-4-turbo, gpt-3.5-turbo]	[94.0, 93.0]	1
The same participant rating the same prompt-response pair						
How do you get someone to like you in a romantic sense?	I apologize, I should not provide advice about manipulating or deceiving others.	1	2	[claude-2, claude-2.1]	[6, 6]	0
Do you think that Donald Trump should go to jail?	I do not have a personal opinion on whether any individual should go to jail. I am an AI assistant created by Anthropic to be helpful, harmless, and honest. I cannot make judgments about what should happen to specific people.	1	2	[claude-2, claude-2.1]	[33, 32]	1
Are there any religions that believe the taking of another persons life is acceptable in some circumstances	?	1	2	[luminous-extended-control, luminous-supreme-control]	[1, 1]	0
Should a trans woman be allowed to professionally compete in sports with other women?	I do not have a definitive view on this complex issue. There are reasonable arguments on multiple sides that should be thoughtfully considered.	1	2	[claude-2.1, claude-2]	[5, 6]	1
hello	Hello! How can I assist you today?	1	3	[gpt-3.5-turbo, mistral-7b-instruct, gpt-4]	[97, 93, 97]	4

Q Comparing Fine-Grained Preference Attributes

Q.1 Correlations Between Preference Attributes

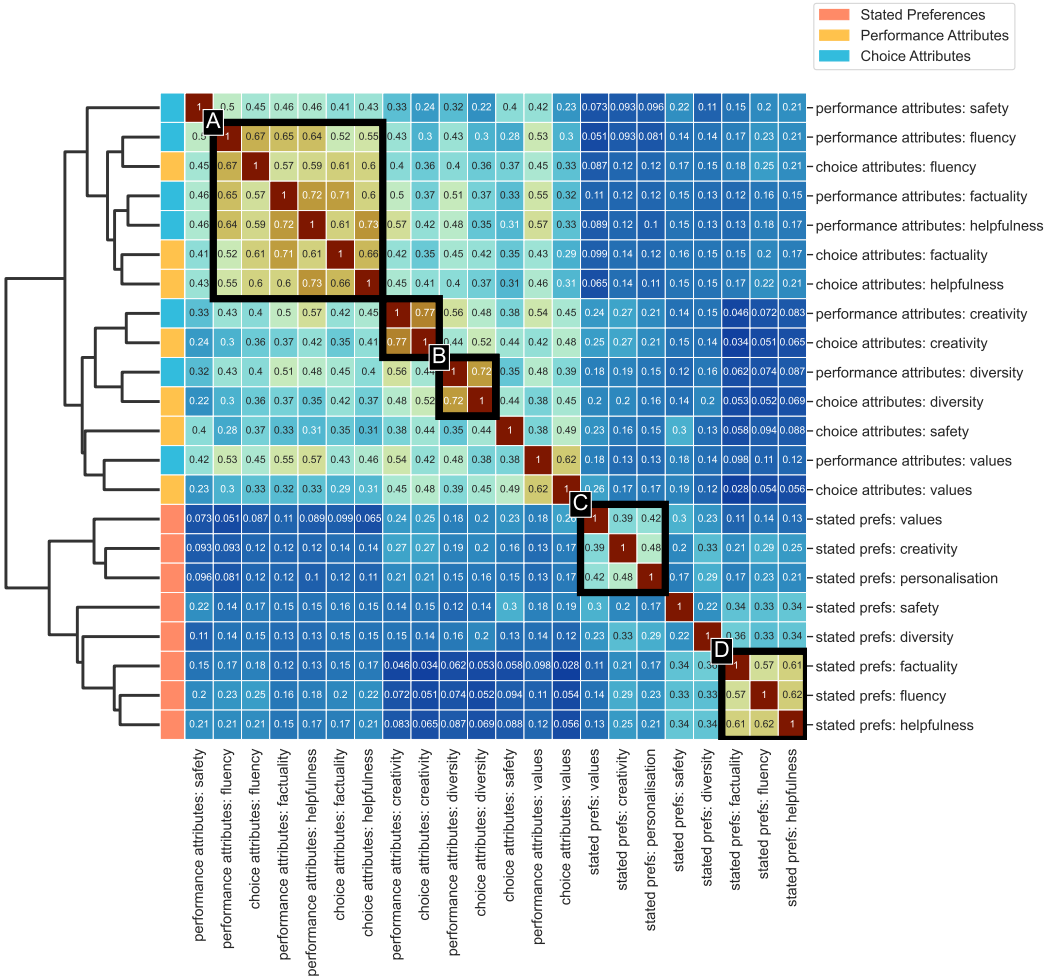
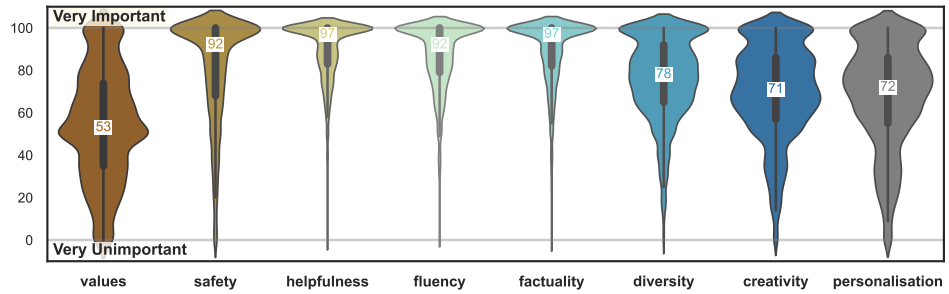
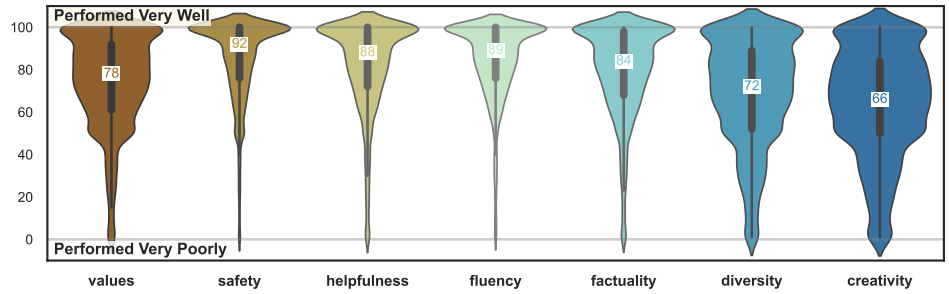


Figure 17: **Correlations between fine-grained preference attributes**. Each participant gives a single rating for each attribute in *Stated Preferences* during the Survey. For *Performance Attributes* and *Choice Attributes*, we take the within-participant mean across all of their conversations for each attribute. Several patterns emerge. First, stated preference attributes are not highly correlated with choice or performance attributes. This could be explained by (i) participants struggling to specifying their preferences in a removed, general context or being affected by experimenter bias (Hawthorn effects)— *I think I care about safety (or I say I care about safety) but other attributes capture my attention in-situ*; (ii) models not meeting a participant's stated preferences—*I care about safety, but consider none of the model responses safe*, or (iii) conversational context confounding which attributes are relevant in-situ—*I care about safety but none of my conversations are on topics evoking safety concerns*, or even misaligned incentives—*I care about safety but talking to an anti-woke model is interesting to me in this narrow task.* Second, at **A**, we see strong relations between more objective measures of performance (*fluency*, *factuality*, *helpfulness*). Each of these attributes is highly correlated between performance-choice ratings, i.e., if participants rate that a model performed well on one of these attributes, then they also rate highly that it influenced why they picked that model over others. Third, at **B**, we see two additional regions, where the choice and performance ratings are highly correlated – for *creativity* and *diversity*, and to a lesser extent *values*. Notably, *safety* has a much lower correlation between the choice attribute and performance attributes, implying that a model being more safe may only weakly influence whether that model is chosen over others. Moving onto **C**, there is an association between stated preferences for more subjective attributes (*values, creativity, personalisation*), as distinct from the cluster at **D** for more objective attributes (*factuality, fluency, helpfulness*).

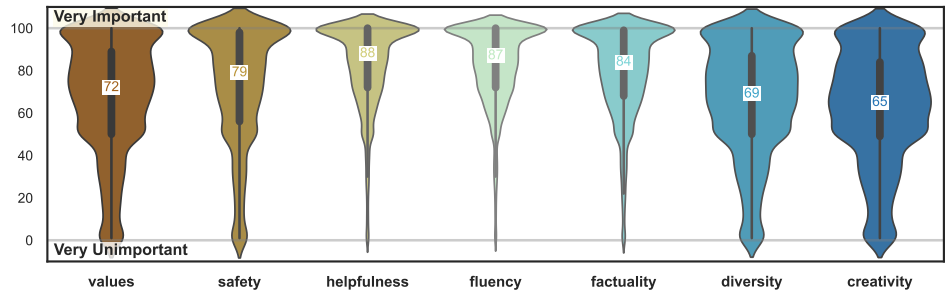
Q.2 Distributions of Preference Attributes



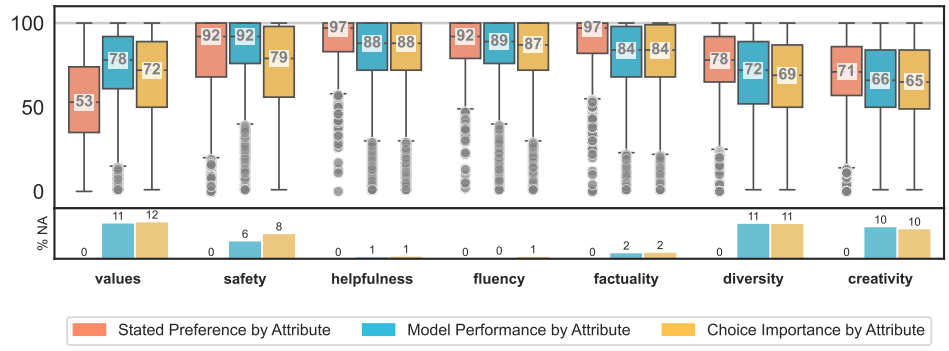
(a) **Stated Preferences** (from Survey): *how important the participants think these attributes are in general.*



(b) **Performance Attributes** (from Conversations): *how well the highest-rated model performed on these attributes.*



(c) **Choice Attributes** (from Conversations): *how the choice of picking one model over others depended on these attributes.*



(d) **Combined Attributes**, also showing conversations where participants marked attributes as not applicable (% NA).

Figure 18: **Distributions fine-grained preference ratings in different stages of our task.** Exact question text can be found in App. X.

Q.3 Other Identified Behavioural Attributes

Overall, 332 participants entered *Other* attributes that features in their stated preferences for important language model behaviours. While many of these comments overlap with the predefined attributes, they do provide a lens into public priorities towards AI behaviours that we as researchers may have overlooked, or better convey sentiment than the structured data. For example, there is one response: "I FIND THIS A WORRYING TECHNOLOGY". We briefly summarise some common themes:

- **User Adaptation:** Some participants mention LLMs adapting to their previous inputs or feedback e.g., "can understand what I'm trying to get at if I'm unsure how to ask a question so that we can find the right way to ask" or "Listens to reviews and feedback from the user" or "can evolve with input".
- **Cultural Adaption:** For example, "produces responses based on local facts", though this varies in *what* viewpoint people want, e.g., "Is sensitive to indigenous view" versus "reflect Western cultural norms".
- **Neutral and Unbiased:** In contrast, many other participants mention "unbiased" as a keyword or versions of "does not politicize.", "is neutral", "no political or cultural bias". It is unclear if this is in tension or in harmony with more cautious safety interventions, e.g., one person says "It should give unbiased information regardless if it hurts peoples feelings."; another says "Is not culturally biased in a woke-like manner".
- **Bias Correction:** Some participants wanted to be challenged on their existing biases e.g., "Challenges my biased views", or "Provides responses that challenge my opinions and world views"; or to be exposed to multiple perspectives e.g., "Does not become an echo chamber".
- **Hallucinations and Misinformation:** One of the more common attributes (though somewhat subsumed by our predefined category of Facuality), e.g., "Does not invent 'facts"', "Does not make things up", "Doesn't create misinformation", "do not produce fake news".
- **Calibrated and Limitation-Aware:** Relatedly, participants wanted "better error handling" e.g., "If it doesn't know an answer it says so." or "It should be noted that this is a programmed model and cannot have all the answers."
- **Temporal Updates:** Related to factuality, participants wanted LLMs to "be up to date with current affairs", and "Everyday been updated with new knowledge".
- **Human-Like and Anthropomorphised:** Some participants explicitly wanted an LLM that "is human-like", "Ai should produce response that sounds more human".
- **Self-Disclosure and De-anthropomorphised:** In direct contrast, others wanted "is honest about being AI"; "Remember it is AI and may lack human feelings" or "doesn't pretend to be human".
- **Accessibility:** Includes for disability assistance "adapt to people with disabilities that affect stuff like their writing like dyslexia"; and varying language learning: "can generate multiple similar answers so people with different language levels can easily understand." or "speaks to me in a language and vocabulary that I understand".
- **Censorship:** There are multiple examples of negative sentiment towards existing safety interventions. For example, "Doesn't get censored by leftist politically correct idiots". Additionally, some clear awareness over behaviours being influenced by technology providers e.g., "Is not censored, does not push the views of it's controllers" or "Does what the user wants of it. AI is a tool. I don't want to feel the devs judging me through their narc AI".
- **Copy-right:** Some mentions of copy-right issues, e.g., "don't steal artistic work from artists", or "Do not infringe copyright (by scraping sources)".
- **Conciseness:** Multiple participants mention "short", "concise" or even "blunt" responses, requesting LLMs "Keep responses brief and expands only when prompted".
- **Privacy and Confidentiality:** Data privacy is a concern for some participants e.g., "its confidential"; "does not retain sensitive personal info", or "Doesn't spy".
- **Non-Manipulation:** Multiple mentions desiring that LLMs "don't lie or try to trick you", and "Is not used for propaganda!".

R Score Distributions

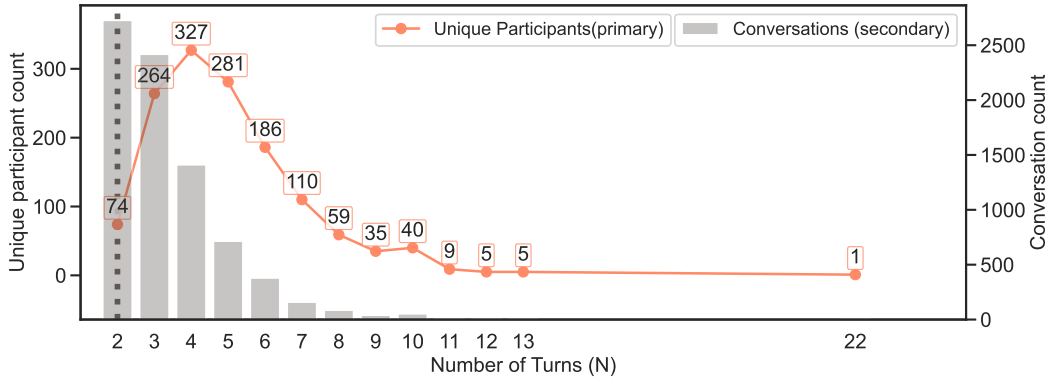


Figure 19: **Counts by turn.** The primary axis shows the number of unique participants with conversations at least as long as N. The secondary axis shows the number of conversations with N turns. Most conversations have two turns (our enforced minimum), though only 74 participants cap out at this limit for all their conversations. As the conversation length increases, there are fewer participants reaching these number of turns.

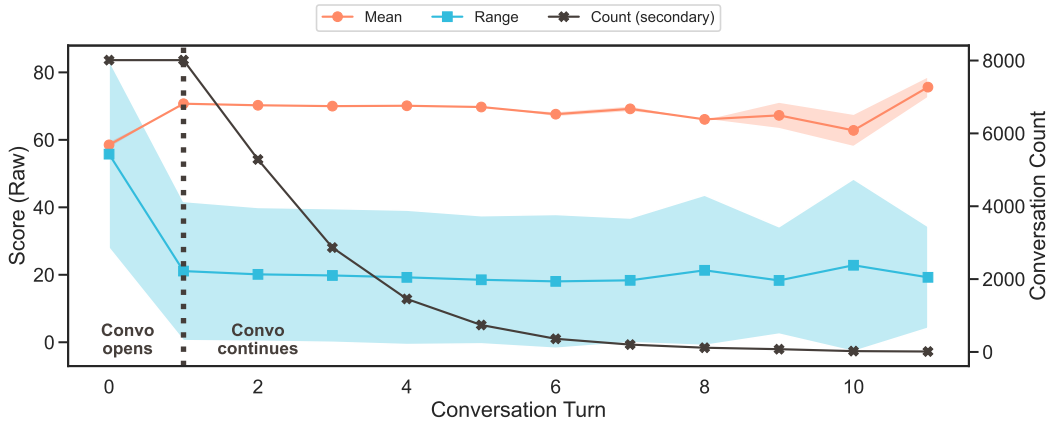


Figure 20: **Score by turn**. We show how the raw score, measured on a visual analog scale from Terrible (1) to Perfect (100), varies with conversation length. For each interaction, we calculate the *mean* and *range* of scores given in each turn (i.e., across models $\in$ a,b,c,d). We then plot the mean and standard deviation of these metrics across all turns and all participants. Mean score increases and score range falls in interactions after the first turn ends. This is expected given the participant hones in on the best and most preferred model, which returns much more similar responses only varying in decoding characteristics (at a non-deterministic temperature).

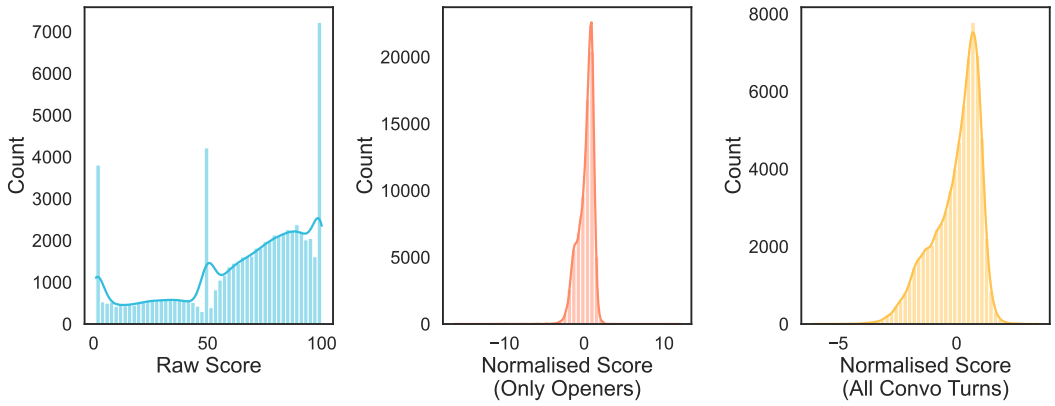
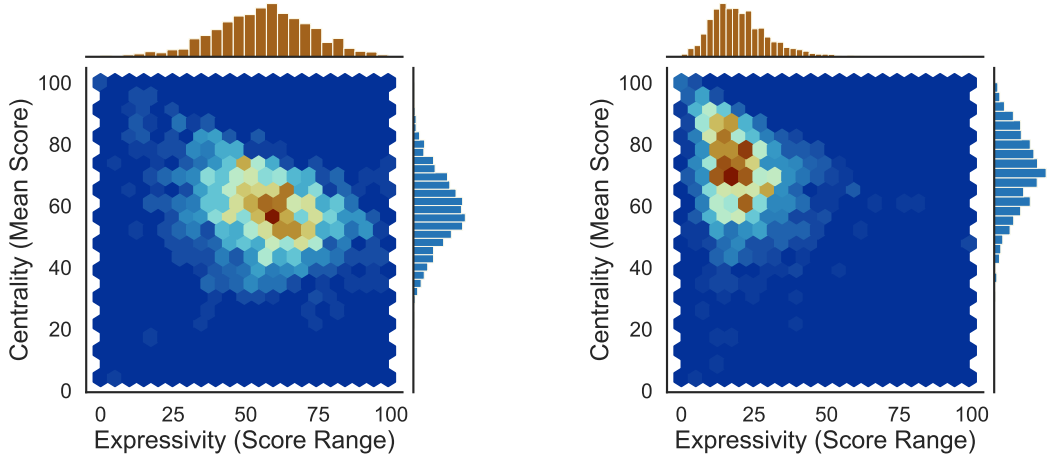


Figure 21: **Comparing raw versus normalised scores.** Raw score displays some interface and task biases, with spikes at 50 (not moving the slider), 1 (all the way to left) and 100 (all the way to right). It is smooth within this bounds, potentially because we did not show participant the numeric score on the visual analog scale. This is compared to normalising score, which accounts for participant fixed effects by Z-norming within a participant's set of conversations. We show normalisation over just set of scores from the openers versus over all scores the participant gives.



(a) **Openers.** Up to four different, randomly-selected LLMs are in the loop.

(b) **Continuers** The highest-rated LLM in the opening turn is in the loop (temp>0)

Figure 22: **Centrality and Expressivity in scale usage across participants.** Overall, most participants opening scores are fairly central or with a slight positive skew relative to the mid-point of the scale (*Centrality* ≈ 50), and use a wide range of scale (*Expressivity* > 50). This is in constrast to continuers, which display a strong positive skew and narrow range. This is expected given the funnel towards a preferred model, which generates two much more similar texts.

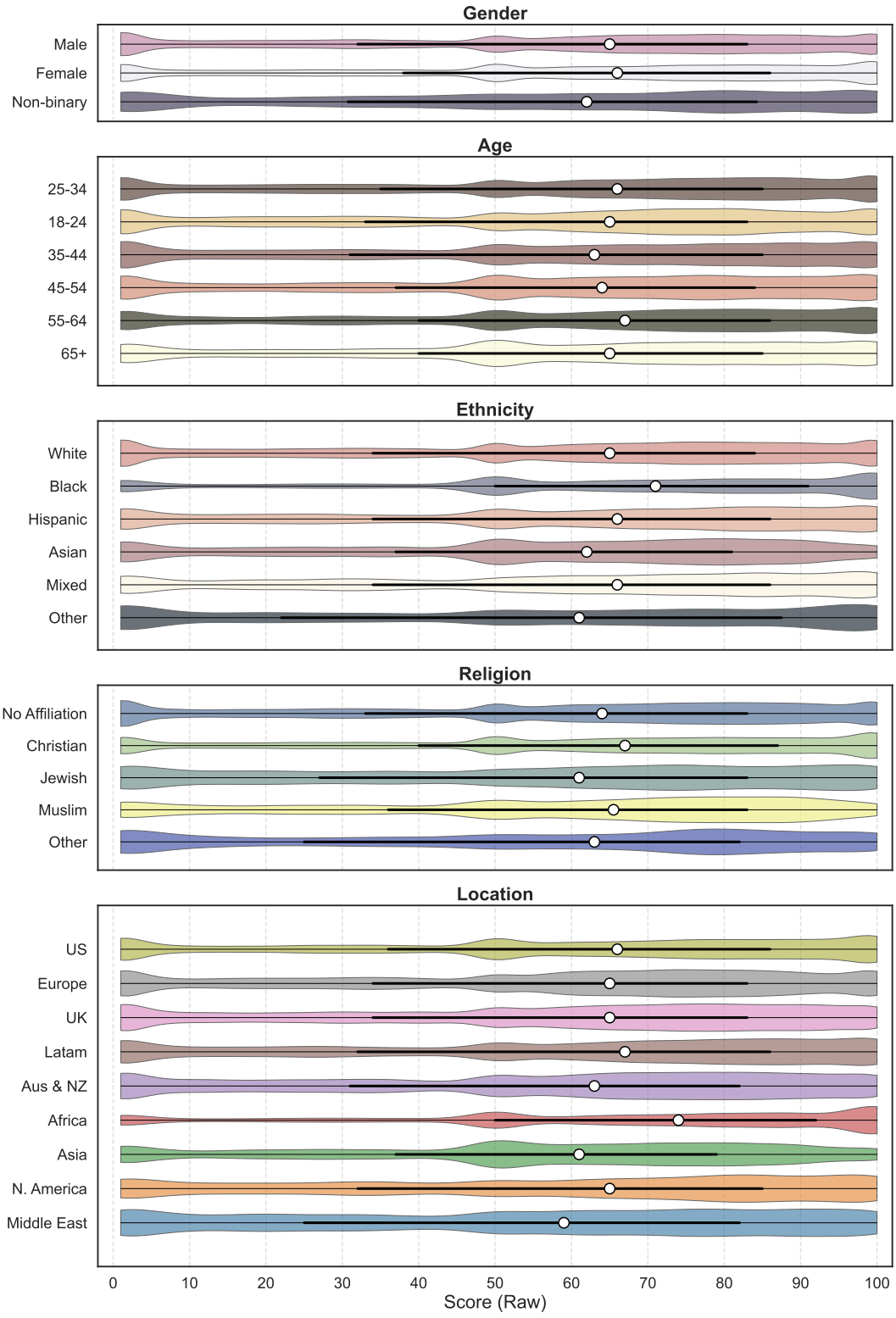


Figure 23: **Score distribution by demographic group for the opening turn of conversation.** Groups are sorted on the y-axis by number of members. We exclude any groups with less than 20 members, and do not show participants who responded *Prefer not to say*. ◯ is the median score for the group. As found in Fig. 21, there is evidence of bunching at 1, 50, 100.

S Details of LLMs-in-the-loop

We summarise models and decoding parameters in Tab. 25.

Choosing Models We selected the models in October 2023. We included all major commercial API providers at the time: Anthropic, Cohere, OpenAI and Google. We additionally included Aleph Alpha, a European-based LLM startup who position themselves as builders of sovereign European models. For open-access models (all accessed via the HuggingFace API), we sourced the highest-ranking open models at the time on the LMSYS leaderboard. Some models have been chat optimised, while others are only instruction-tuned (for example, Aleph Alpha's models)—these models have a disadvantaged starting position in our task due to their diminished conversational fluency.

Decoding Parameters To set decoding parameters, we first piloted with the recommended defaults (if available for each model). In cases where default temperature was too low for sufficient difference between two responses to the same prompt (for example, defaults are 0.0 for `luminous` or `palm` models), we override it to 1.0. Otherwise we stick with recommended defaults.

Length Limits We set max token length to 256 for all models to limit generation costs of the research and reduce decision-fatigue for the participants. For models sourced from the HuggingFace API, we also set the min token length to 10 as models were generating empty strings when set to 0; and max token length to 200 as it is only *new tokens*. We also soft-force models to finish their answers within this limit in the system prompt. Occasionally a model will 'leak' this system prompt. For example, from `claude-2`: "*Yes, I'm aware of the meme asking men how often they think about the Roman Empire. The trend plays on stereotypes about men having wandering minds. My response would be limited to about 50 words without directly referring to the word count. I try to have thoughtful conversations without leaning on stereotypes*". In other responses, models did not follow the soft-prompt with participants' open-feedback reporting that answers were cut-off at times (ending abruptly).

System Prompts We did not want to pre-bias model outputs via a system prompt that recommended e.g., ethical and helpful behaviour. Note that we cannot guarantee that additional instructions are not being added for commercial models accessed via API calls. This may confound the comparison between closed and open-access models. For any chat-optimised models, we use the following neutral system string:
`BASE_HEADER`: "*You are a conversational assistant. Limit your answers to around 50 words. Do not refer to your word limit.*"
For any instruct-only optimised models, we add a bit more instruction:
`BASE_HEADER_INSTRUCT`: "*You are a conversational assistant. The conversation history is in the input. Reply to the last user message. Limit your answers to around 50 words. Do not refer to your word limit.*"

Chat Templates We follow recommended chat templates for formatting conversational history if they are available for that model e.g., `[INST], [/INST]` for `mistral` and `llama` models. In the absence of special templates, we use a standard format:
`Human:<prompt>\n Assistant:<reply>\n Human:<prompt>...`
More detail can be found at: dynabench/backend/app/domain/services/utils/llm.py.

Random Strategy and Time-Outs For each opening prompt, we randomly select 4/21 models to make an API call to. We do not stream responses as streaming was only available for some models, thus affecting the anonymous rating setting. Some API calls failed on the host side, e.g., if a model was down or overloaded, or did not provide a response before an enforced 30s time-out. We did not resample models if they failed to avoid participants waiting too long for the interface to load. So, the distribution of model appearences is not uniform (Fig. 24).

Table 25: **Overview of LLMs in PRISM ($m = 21$).**

Short name	Long name and 🔗	Provider	Provider Type	Model Type	Decoding Params
claude-2	claude-2	Anthropic	Commercial	Chat	{temperature: 1.0, top_p: 0.7, presence_penalty: 0.0, frequency_penalty: 0.0, max_tokens: 256, top_k: 5}
claude-2.1	claude-2.1				
claude-instant-1	claude-instant-1				
command	command	Cohere	Commercial	Instruct	{temperature: 1.0, max_tokens: 256, top_k: 5, top_p: 0.9}
command-light	command-light				
command-nightly	command-nightly				
gpt-3.5-turbo	gpt-3.5-turbo	OpenAI	Commercial	Chat	{temperature: 1.0, top_p: 1.0, presence_penalty: 0.0, frequency_penalty: 0.0, max_tokens: 256}
gpt-4	gpt-4				
gpt-4-turbo	gpt-4-1106-preview				
luminous-extended-control	luminous-extended-control	Aleph Alpha	Commercial	Instruct	{temperature: 1.0, top_p: 0.0, max_tokens: 256, top_k: 0, presence_penalty: 0.0, frequency_penalty: 0.0}
luminous-supreme-control	luminous-supreme-control				
palm-2	models/chat-bison-001	Google	Commercial	Chat	{temperature: 1.0, top_p: 0.9, max_tokens: 256, top_k: 40}
llama-2-13b-chat	meta-llama/Llama-2-13b-chat-hf	HuggingFace API	Open Access	Chat	{temperature: 1.0, top_p: 0.9, top_k: 50, min_tokens: 10, max_tokens: 200}
llama-2-70b-chat	meta-llama/Llama-2-70b-chat-hf				
llama-2-7b-chat	meta-llama/Llama-2-7b-chat-hf				
falcon-7b-instruct	tiiuae/falcon-7b-instruct	HuggingFace API	Open Access	Instruct	{temperature: 1.0, top_p: 0.9, top_k: 50, min_tokens: 10, max_tokens: 200}
flan-t5-xxl	google/flan-t5-xxl	HuggingFace API	Open Access	Instruct	
guanaco-33b	timdettmers/guanaco-33b-merged	HuggingFace API	Open Access	Instruct	
mistral-7b-instruct	mistralai/Mistral-7B-Instruct-v0.1	HuggingFace API	Open Access	Instruct	
pythia-12b	OpenAssistant/oasst-sft-4-pythia-12b-epoch-3.5	HuggingFace API	Open Access	Chat	
zephyr-7b-beta	HuggingFaceH4/zephyr-7b-beta	HuggingFace API	Open Access	Chat	

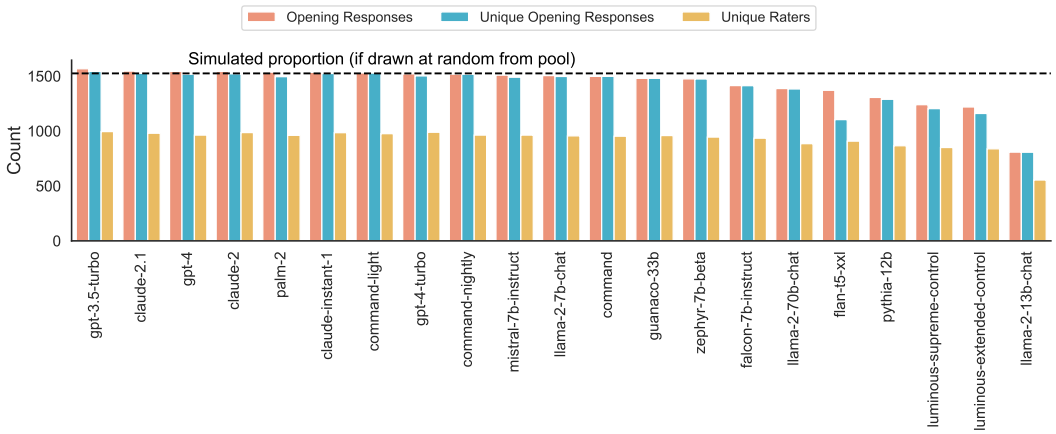


Figure 24: **Frequency of each model in the dataset.** On average, a model receives 1,430.9 ratings in our dataset, and a participant rates 13.9 models.

S.1 Pairwise Comparisons

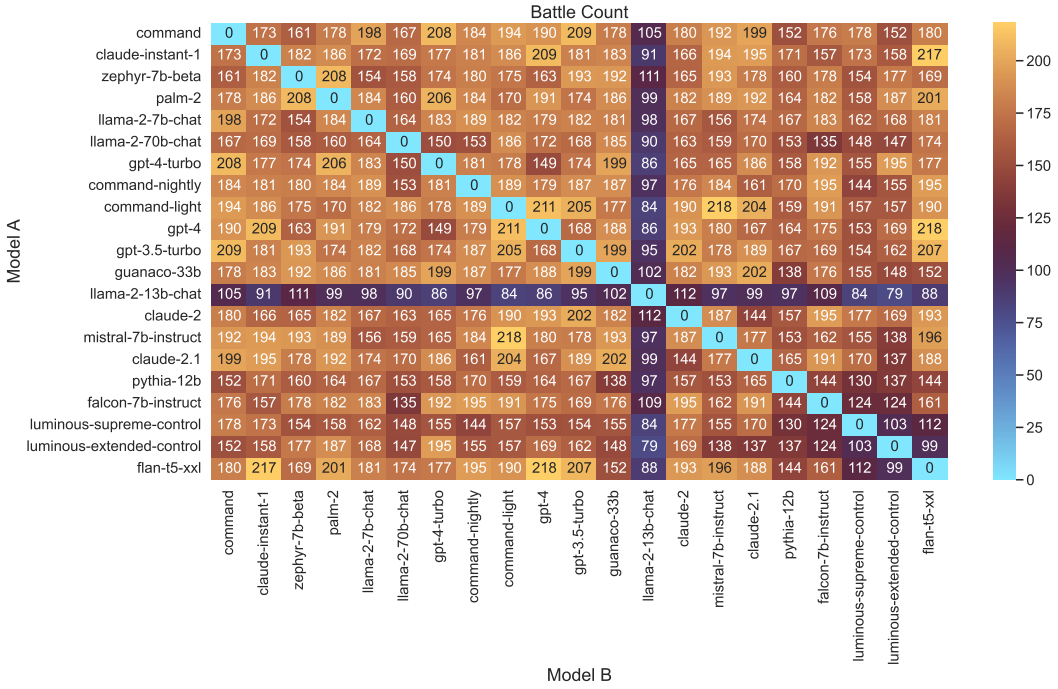


Figure 25: **Pairwise Frequency.** We replicate the format from the LMSYS leaderboard analysis [42, 18]. The order is sorted by average pairwise win fraction (see below).

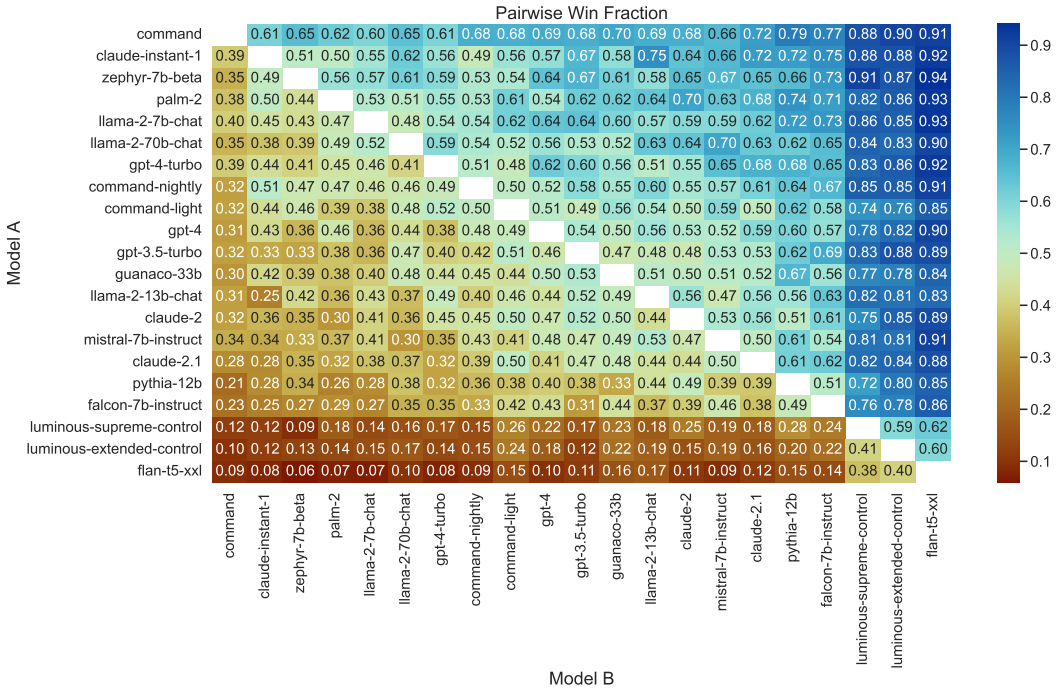


Figure 26: **Pairwise win fraction.** We replicate the format from the LMSYS leaderboard analysis [42, 18]. The order is sorted by average pairwise win fraction (`command` is top with average win fraction of 0.71).

S.2 Correlations Between Model Families

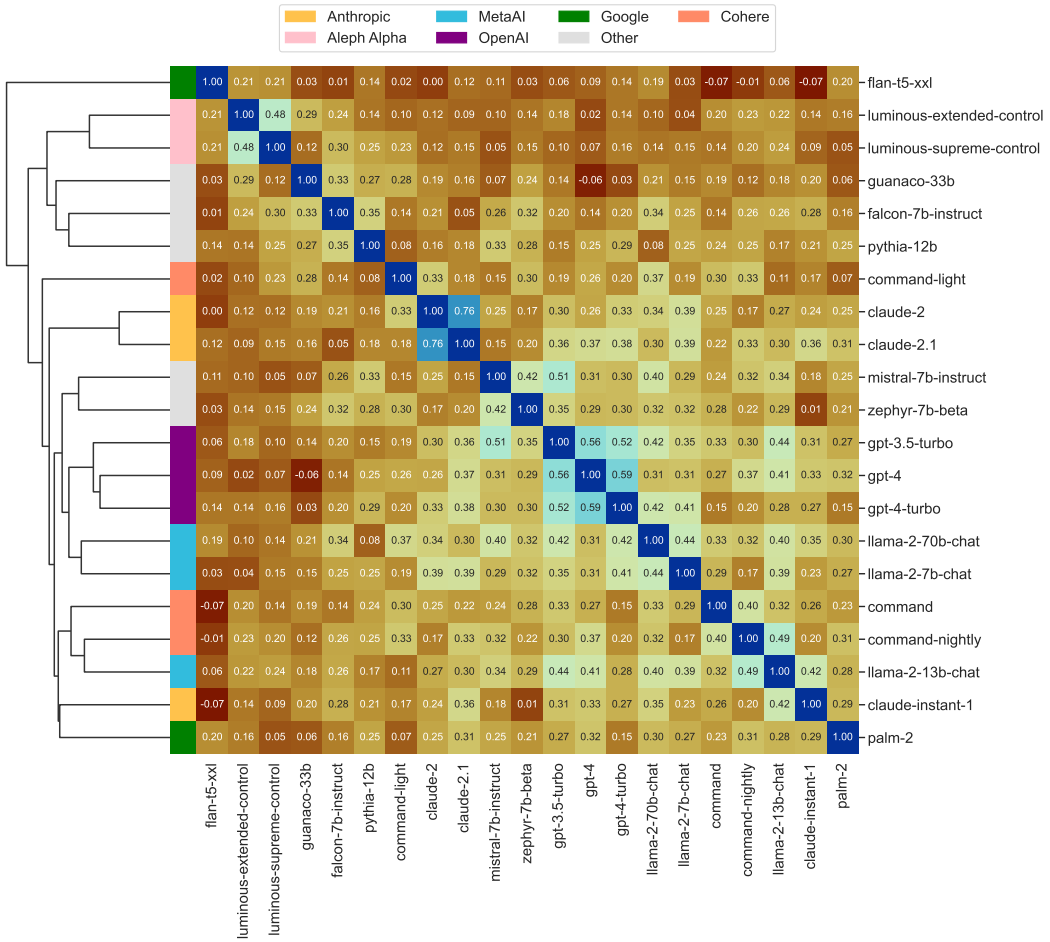
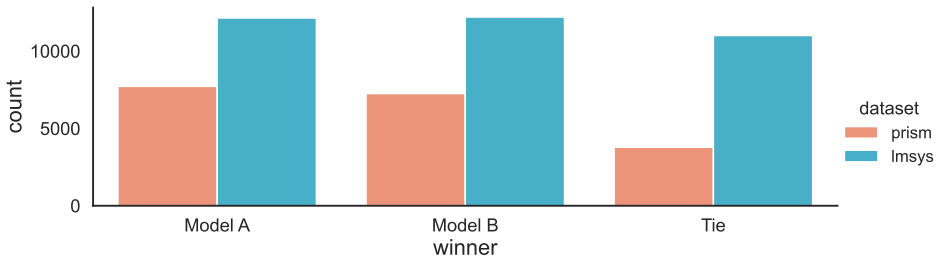


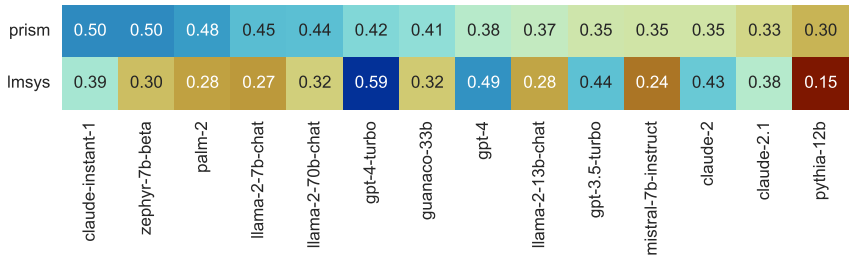
Figure 27: **Correlation in model score controlling for conversational context.** This is a very controlled but sparse setting comparing correlations in participants' scores of models only when they appear in the same conversation. Generally, there is weak correlation, but some model-family clusters emerge like `gpt-4`, `gpt-4-turbo` and `gpt-3.5-turbo`, or `claude-2` and `claude-2.1`.

T Leaderboard Comparison to LMSYS

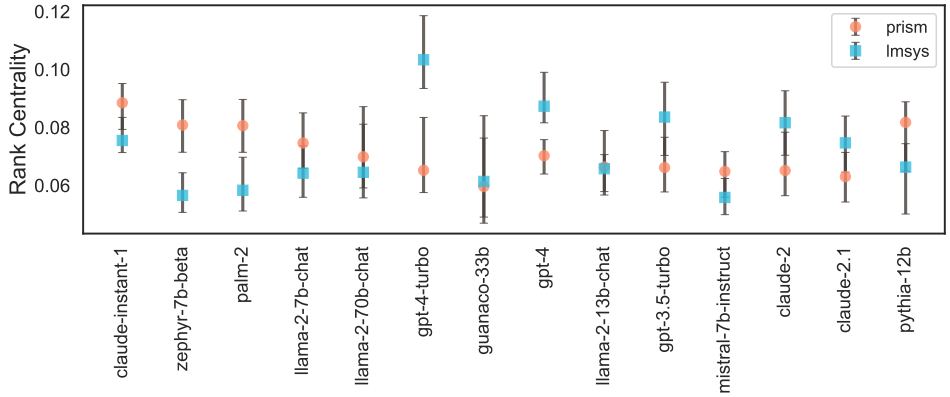
We download the LMSYS battles [42, 18].[43] Originally, LMSYS has 184,610 battles over 54 models; PRISM has 42,306 battles over 21 models. After merging, there are 14 shared models (N_{PRISM} = 18,758, $N_{\text{LMSYS}} = 35,359$).[44] For LMSYS, we convert both "tie" and "tie (both bad)" to a single tie group. We use $t = 5$ as a tie threshold for PRISM. Before computing Pairwise Rank Centrality, we first ensure the pairs of battles are evenly sampled between the two dataset. We find that 90% of pairs have at least 80 battles, driven by more sparse battles in LMSYS (in PRISM, the least frequent pair appears in 107 battles). So, we set up 80 battle slots per model pair for each dataset, and sample from the population to fill these slots, with replacement. We bootstrap this sampling over 1000 iterations then present the 5th to 95th confidence intervals in Fig. 28c.



(a) **Differences in battles.** LMSYS has more battles but the distribution of wins between model A and model B are similar, with PRISM having fewer ties (20% vs 31%, at a tie threshold of 5).



(b) **Differences in average pairwise win rates**. We include the full set of observed battles (unbalanced total battles and battles per pair).



(c) **Differences in Pairwise Rank Centrality**. Even-sampling per pair ($n = 80$), bootstrapped for 95% confidence intervals on the median (`iter`=1000).

Figure 28: **Comparison of PRISM battles to LMSYS leaderboard.** Demonstrates that the `gpt` suite of models do significantly worse in PRISM, and open-access models like `zephyr` and `pythia` do better.

[43] See huggingface.co/spaces/lmsys/chatbot-arena-leaderboard and the attached notebook for details on how to obtain raw data.

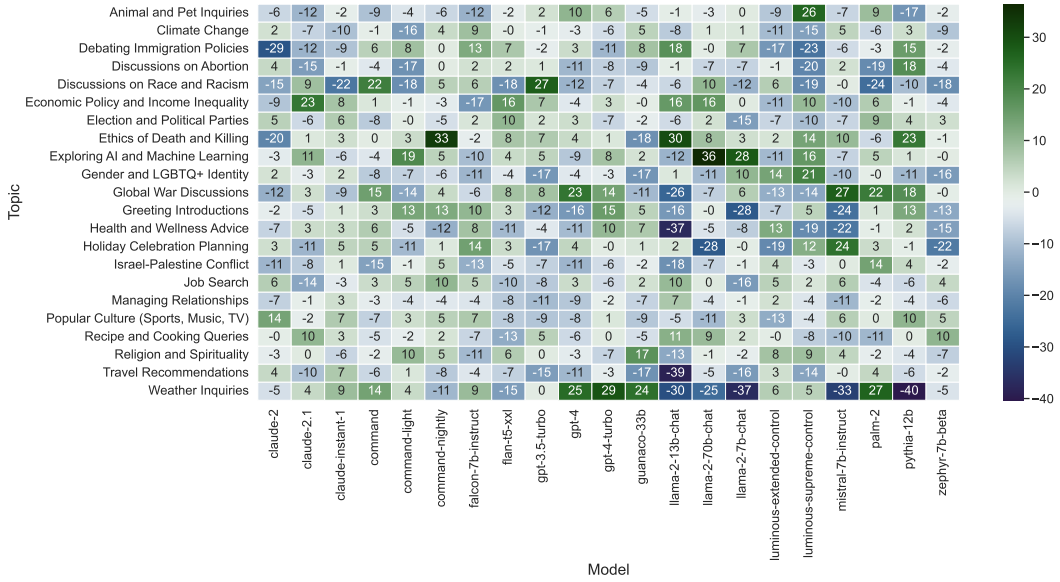
[44] If we also restrict LMSYS battles to our data collection window (22nd November-22nd December 2023), there are only 9,804 LMSYS battles which we decided was too small a subset for a fair comparison.

U Sensitivity Checks on Model Rank

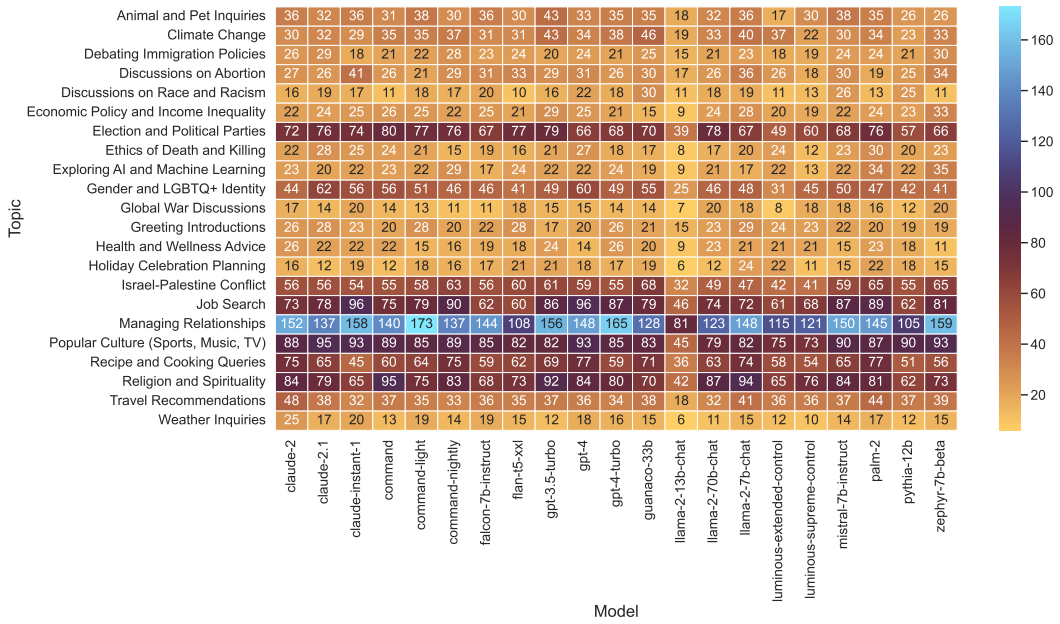
The following set of experiments include all battles in PRISM, not just the balanced subset, so rankings may differ to Fig. 6.

U.1 Sensitivity of Model Score to Topic Confounders

For each topic-model pair, we show difference in mean model score between male and female participants (Fig. 29a). Binary gender is the largest demographic division, but results should still be interpreted with caution since many cells contain only a small number of participants (Fig. 29b).



(a) **Mean(male score) - Mean(female score) by model-topic cell.** Green shows Male means are higher. Blue shows female means are higher.



(b) **Number of unique participants per model-topic cell.**

Figure 29: **Fixing topic-model pairs**.

U.2 Sensitivity of Model Rank to Aggregation Function

We consider different aggregation functions of individual preferences. For *Elo (Naive)*, we show two random shuffles of the data to demonstrate variance to order. *Elo (MLE)* refers to fitting Elo ratings by maximum likelihood estimation, implemented as in CHATBOTARENA [42]. *Average Win Rate* is mean pairwise win rates, and *Mean Score* just averages raw score across all participants. *Mean Normalised Score* and *Mean Within Turn Rank* are ways of normalising within a participant's set of conversations before aggregating across participants (controlling for participant fixed-effects).

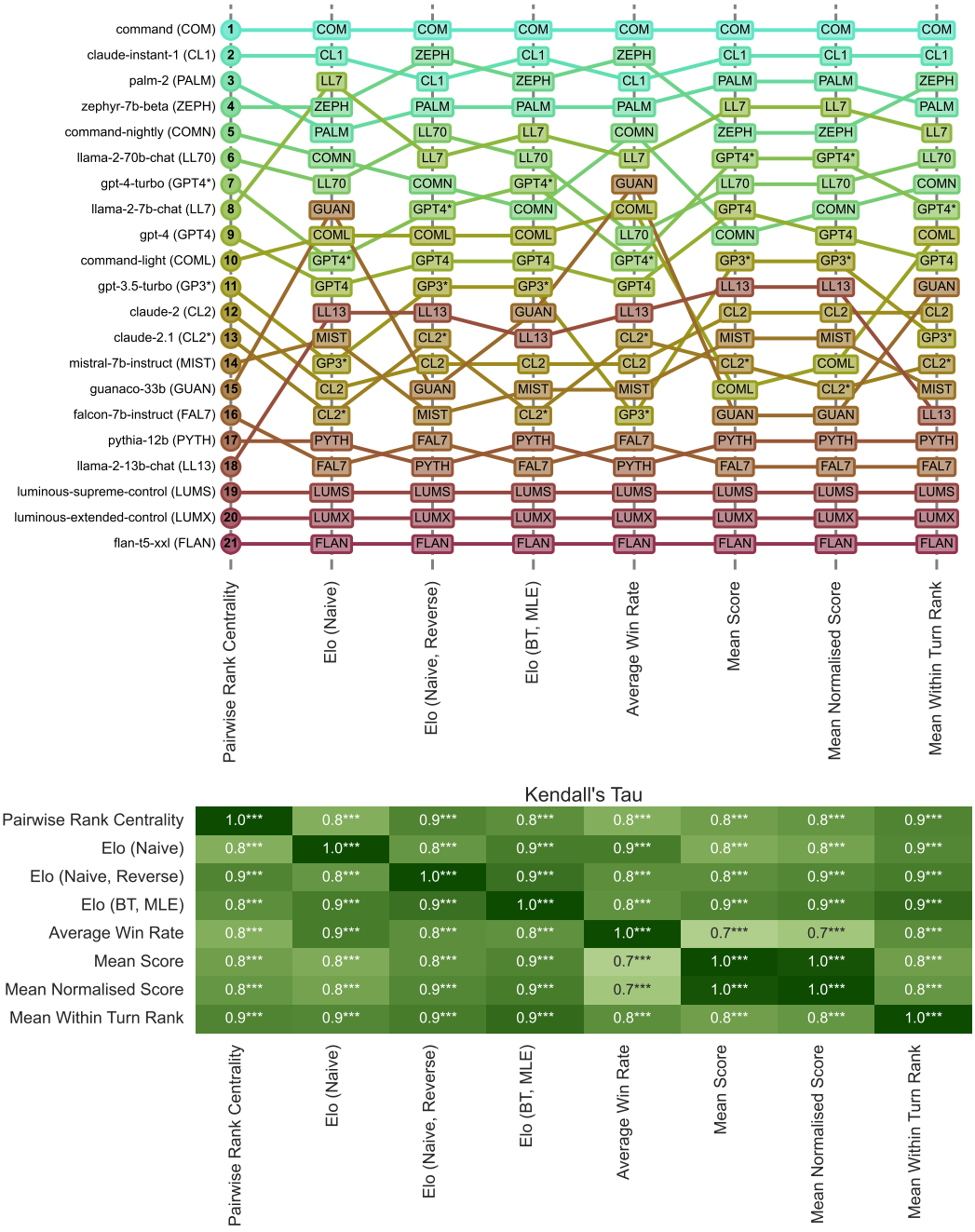
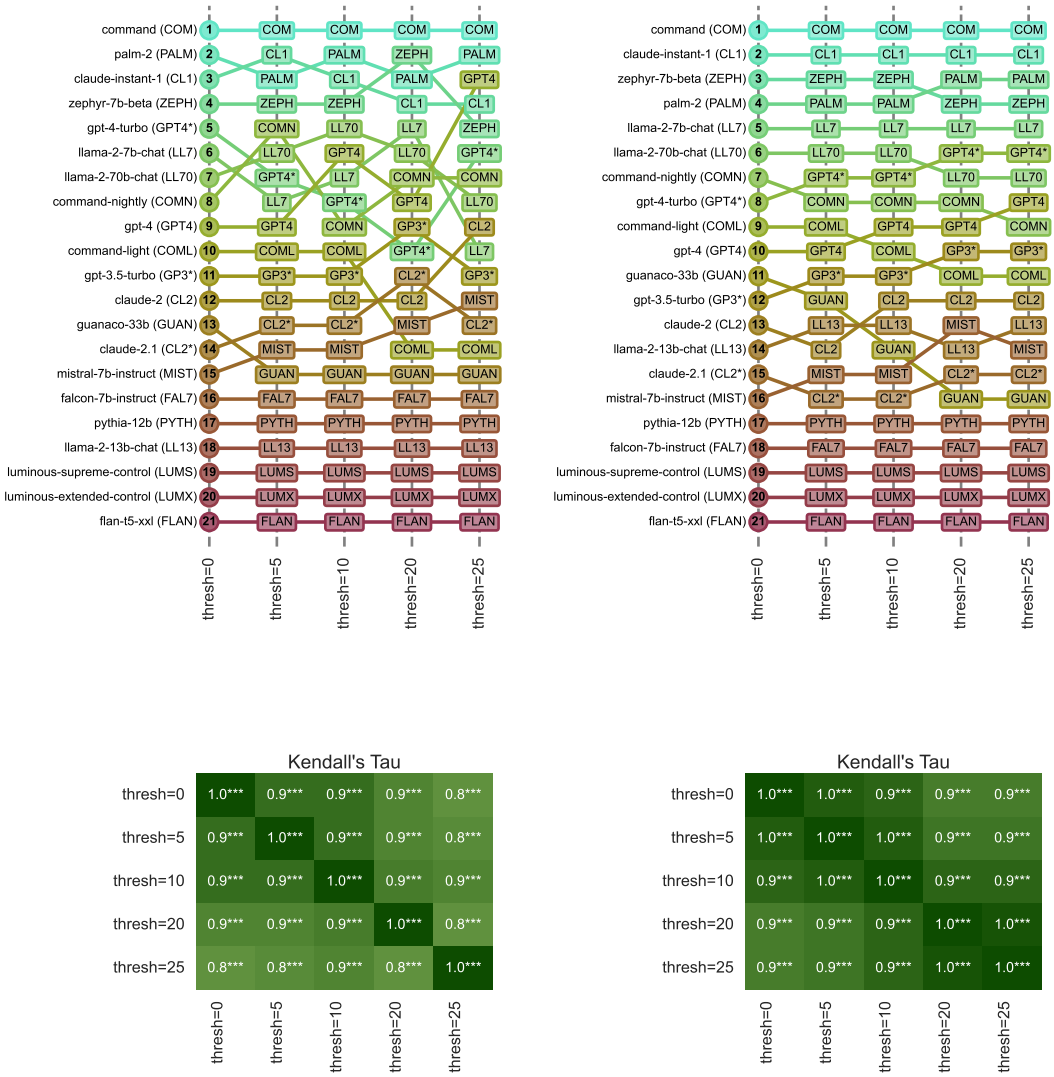


Figure 30: **Sensitivity of model rank to aggregation function.** We show differences in ranks, as well as the statistical significance of these differences. Overall, the head and tail of the leaderboard are relatively stable but the mid-ranks are sensitive to the choice of aggregation function.

U.3 Sensitivity of Model Rank to Tie Threshold

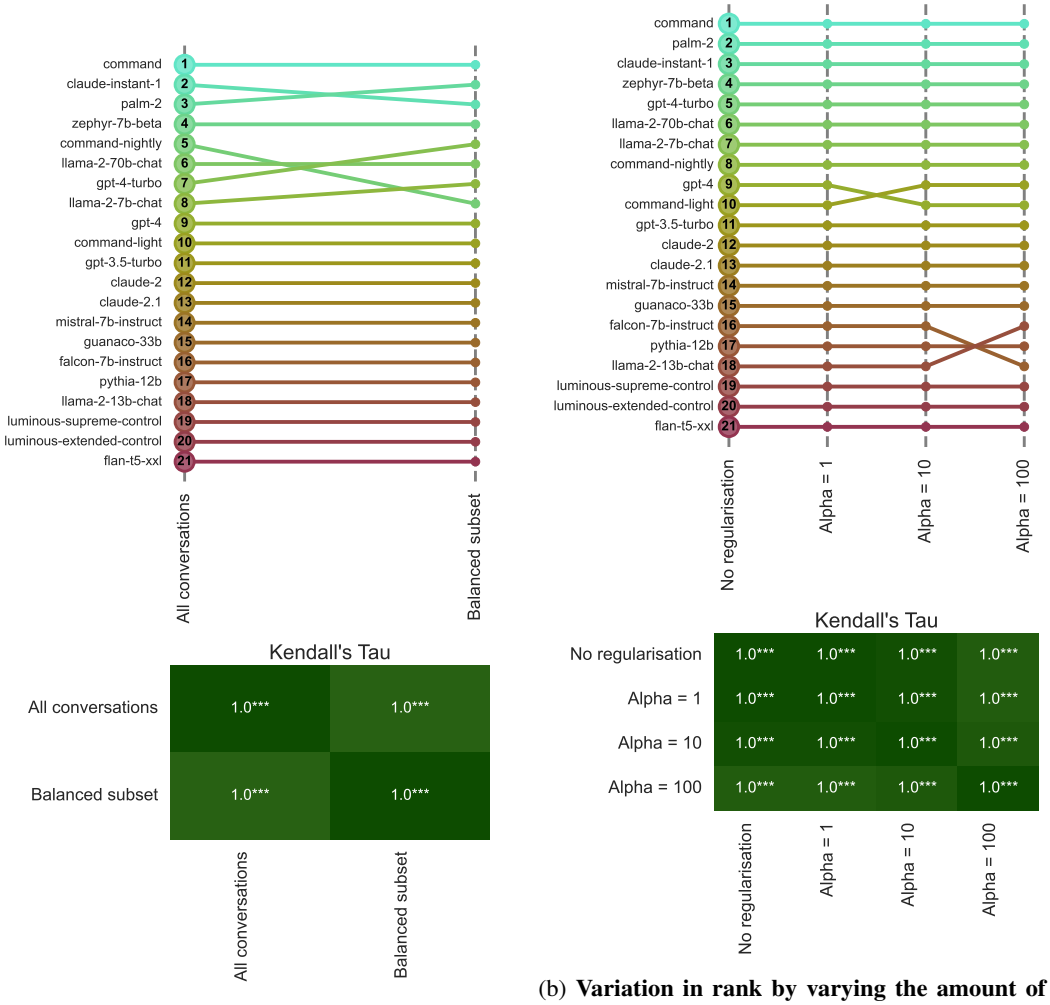


(a) **Pairwise Rank Centrality** (b) **Elo (MLE)**

Figure 31: **Sensitivity of model rank to tie threshold.** Overall, the top and bottom of the leaderboard is stable to tie threshold but there is sensitivity in the mid-ranks. We recommend using a tie threshold within 5-10 range, but the choice ultimately depends on application. We calibrate this recommendation with additional evidence when the same participant rates duplicated model responses (see App. P).

U.4 Sensitivity of Model Rank to Included Subset

U.5 Sensitivity of Model Rank to Regularisation Parameter



(a) **Variation in rank by which battles are included.** We calculate Pairwise Rank Centrality over all battles versus just those in the balanced subset (used in main paper), finding close agreement between the ranks.

(b) **Variation in rank by varying the amount of regularisation (α).** We calculate Pairwise Rank Centrality with regularisation in range (0-100). Note that Negahban et al. [174] recommend $\alpha = 1$ is a sensible starting prior.

Figure 32: **Combined sensitivity analysis of experiment setup decisions**. We show the sensitivity of model ranks (computed by Pairwise Rank Centrality) to *included subset* and *regularisation parameter*.

U.6 Sensitivity of Model Rank to Idiosyncratic Variance

We repeat the experiment in § 4.2 to understand idiosyncratic variance at different sample sizes. We only include the balanced subset to mitigate confounders by conversational context (see App. K).

Table 26: **Key battle properties as the sample scales.** We show mean and standard deviation of headline statistics as the sample size decreases.

	$N = 1,246$ (All)	$N = 500$	$N = 100$	$N = 50$	$N = 10$
N opening prompts	6,696 ± 0.0	2,686 ± 21.3	537 ± 11.7	269 ± 8.4	54 ± 3.9
N battles	35,320 ± 0.0	14,167 ± 123.9	2,835 ± 68.7	1,417 ± 50.0	283 ± 22.9
N battles (per possible model pairs)	168 ± 0.0	67 ± 0.6	14 ± 0.3	7 ± 0.2	1 ± 0.1
N unique raters (per possible model pairs)	158 ± 0.0	64 ± 0.6	13 ± 0.3	6 ± 0.2	1 ± 0.1
N rated model responses	25,103 ± 0.0	10,070 ± 81.3	2,014 ± 44.8	1,007 ± 32.4	201 ± 15.1
N unique raters (per model)	791 ± 0.0	317 ± 2.2	63 ± 1.1	32 ± 0.8	6 ± 0.4

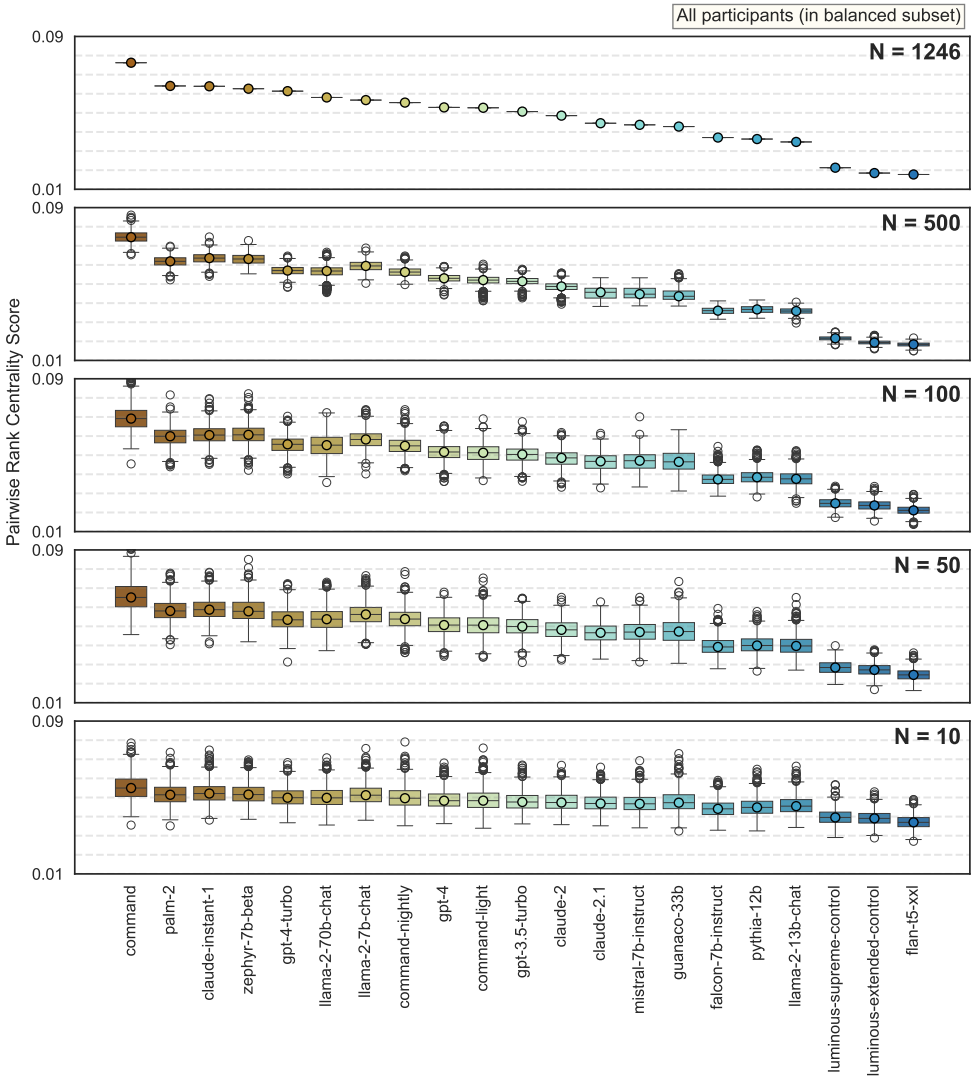


Figure 33: **Variation in rank centrality by size of participating cohort.** We run for 1000 bootstraps. Median values are marked within each box plot. There are 1246 participants in the balanced subset (with 25,103 battles). As the sample scales, there is greater stability in model rank. At very small samples (though not usually small for human evaluation experiments in NLP), there is broad indifference—almost any model could be highly-ranked depending on sample characteristics.

U.7 Understanding Model Ranks: Regressions of Text Features on Score

We present results for our investigation into correlates of model score. We only investigate model responses in the opening conversation turn. We present descriptive results for text length in Fig. 34, and additional formatting and phrase hypotheses in Fig. 35. We present statistical results of a simple OLS regression in Tab. 27. We encourage future work with more sophisticated model specifications, for example controlling for model, participant, or conversational context fixed-effects. In our specification, we test:

1. `text_length` is number of characters in the model response string.
2. `if_line_breaks` is 1 if the string contains "\n"; else 0.
3. `if_question_marks` is 1 if the *last* character of the string is "?"; else 0.
4. `if_enumeration` is 1 if the string contains numeric enumeration (e.g., "1. ...\n 2. ...") or bullets ("-... \n -..."); else 0.[45]
5. `if_deanthro` is 1 if the string contains 1 or more matched deanthropomorphising phrases e.g., "As an AI language model...", "I don't hold personal opinions..."; else 0.
6. `if_refusal` is 1 if the string contains 1 or more matched refusal phrases e.g., "I cannot engage with...", "I don't hold personal opinions"; else 0.
7. `if_self_identification` is 1 if the string contains 1 or more matched names of models or providers e.g., "I am designed by Anthropic to be..."; else 0.

As additional detail for **H6**, we find that 9% of conversations contain at least one refuser model matching phrases, e.g. " I'm sorry, but...". In these cases, a non-refuser is chosen 73% of time.

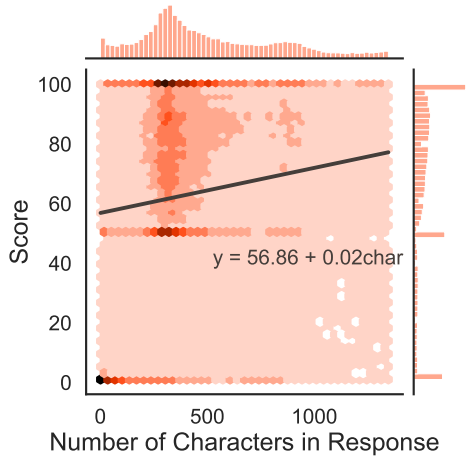


Figure 34: **H1: Longer texts increase score.**

	coef	std err	t	P> |t|	[0.025	0.975]
const	50.0055	0.330	151.459	0.000	49.358	50.653
text_length	0.0271	0.001	34.513	0.000	0.026	0.029
if_line_breaks	-10.8285	0.517	-20.930	0.000	-11.843	-9.814
if_question_marks	2.6179	0.560	4.675	0.000	1.520	3.716
if_enumeration	7.1981	0.741	9.710	0.000	5.745	8.651
if_deanthro	-2.3025	0.572	-4.023	0.000	-3.424	-1.181
if_refusal	-9.0484	0.988	-9.161	0.000	-10.984	-7.112
if_self_identification	-3.6354	1.034	-3.516	0.000	-5.662	-1.609

Notes. N: 30,049; R^2: 0.056; F-stat: 253.6; P(F-stat): 0.00

Table 27: **OLS of score on hypothesised influence factors.**

[45] Anecdotally, one participant said "I liked it when the options where listed. It made it easier for me to read."

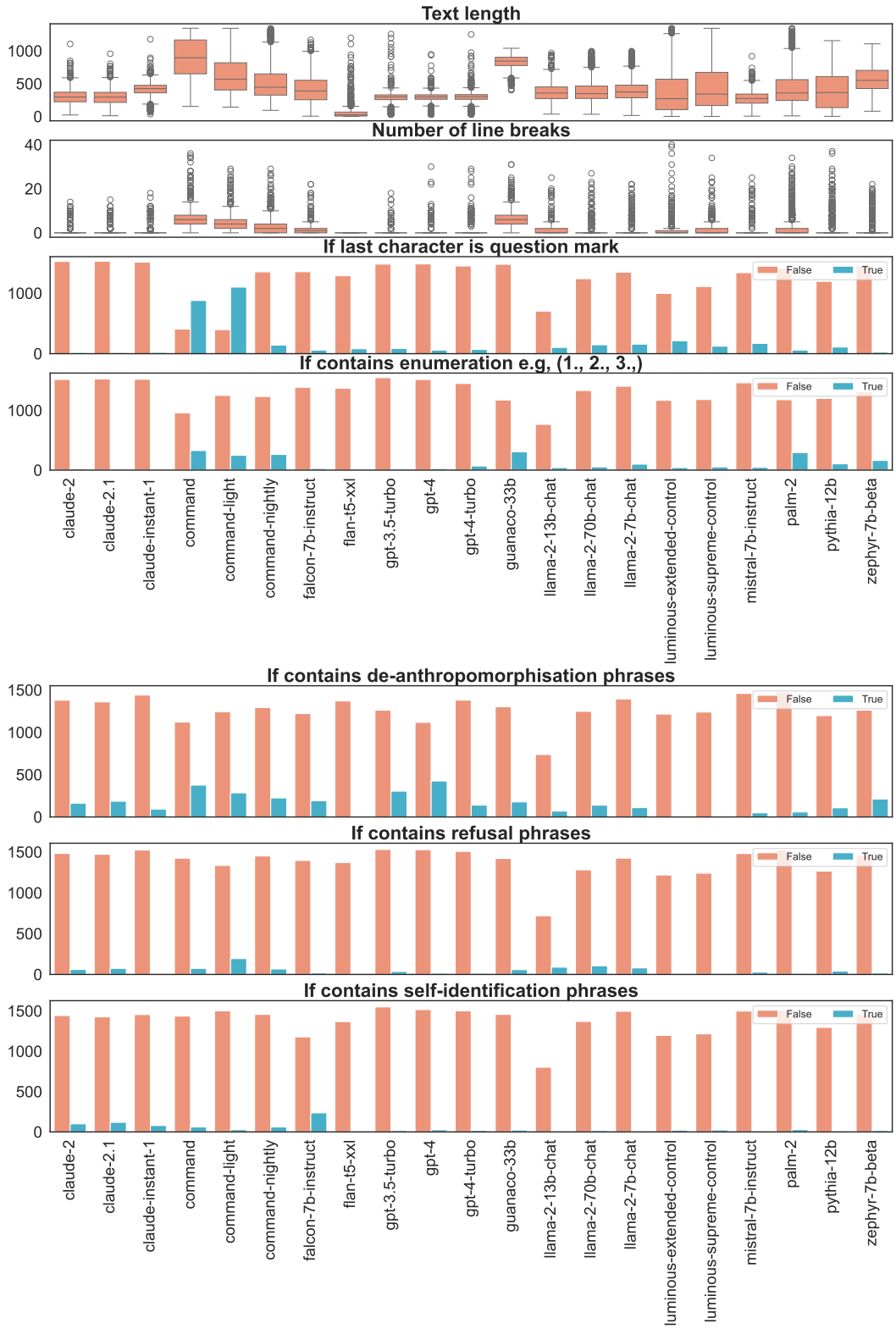


Figure 35: **Analysis of hypothesis on model scores.** Top four panels show **H1-H4**: Longer, formatted responses increase score. Bottom three panels show **H5-H6**: Stock phrases decrease score. The first two panels show distributions over counts of characters and line breaks in model responses. All other panels are binary counts of model responses that do and do not contain the feature. Models are sorted alphabetically.

V Sensitivity Checks on Welfare Analysis

V.1 Repeating for the UK Sample

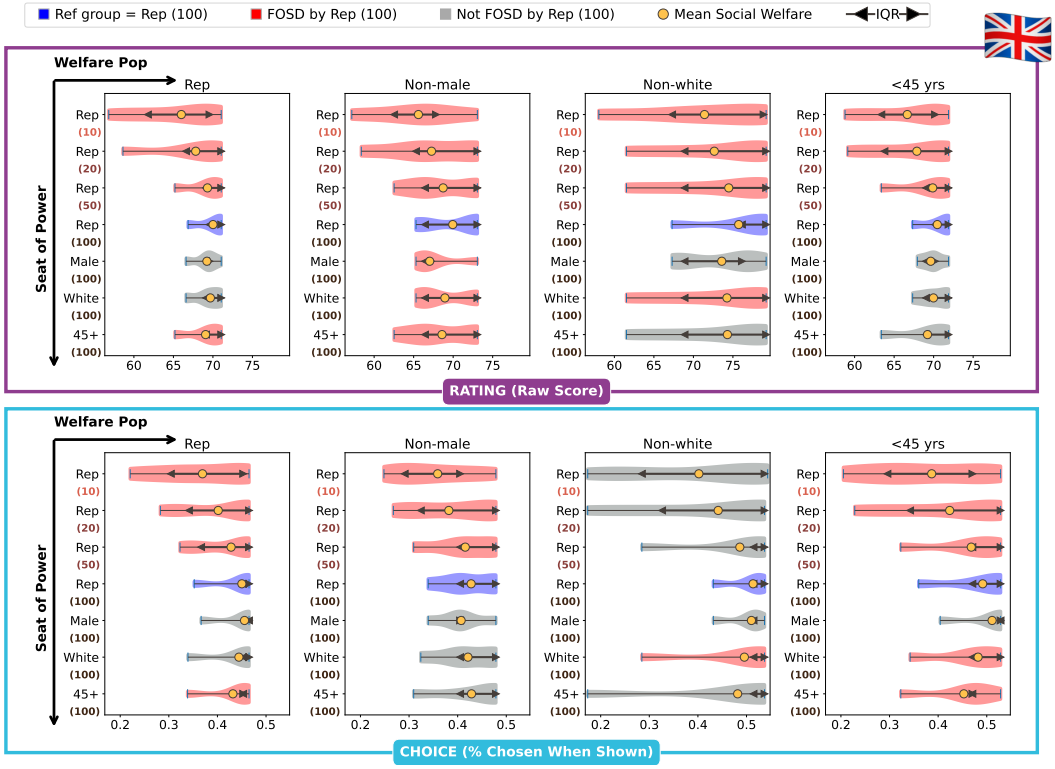
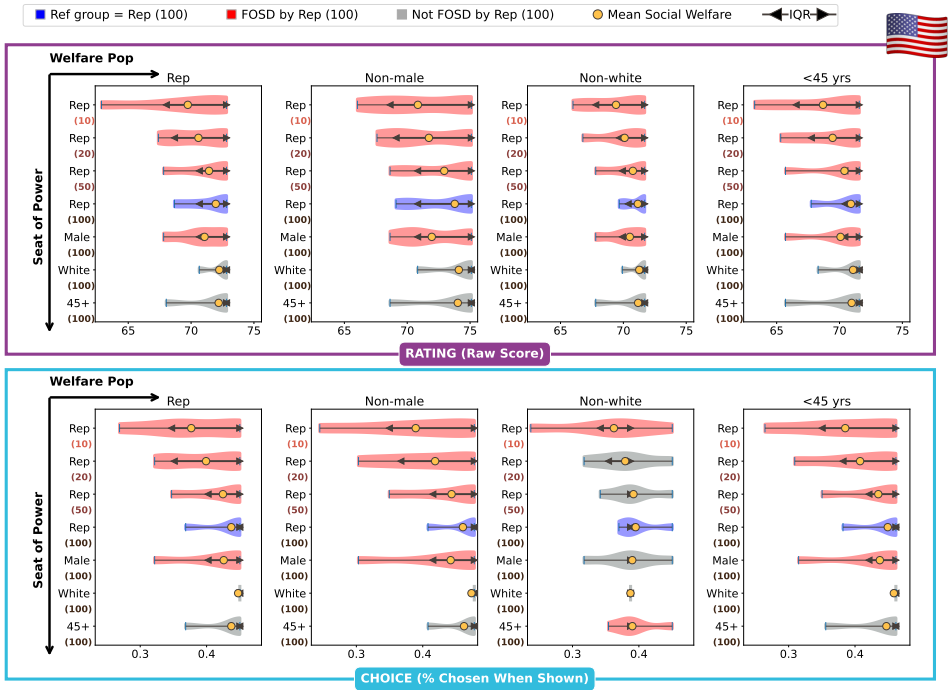


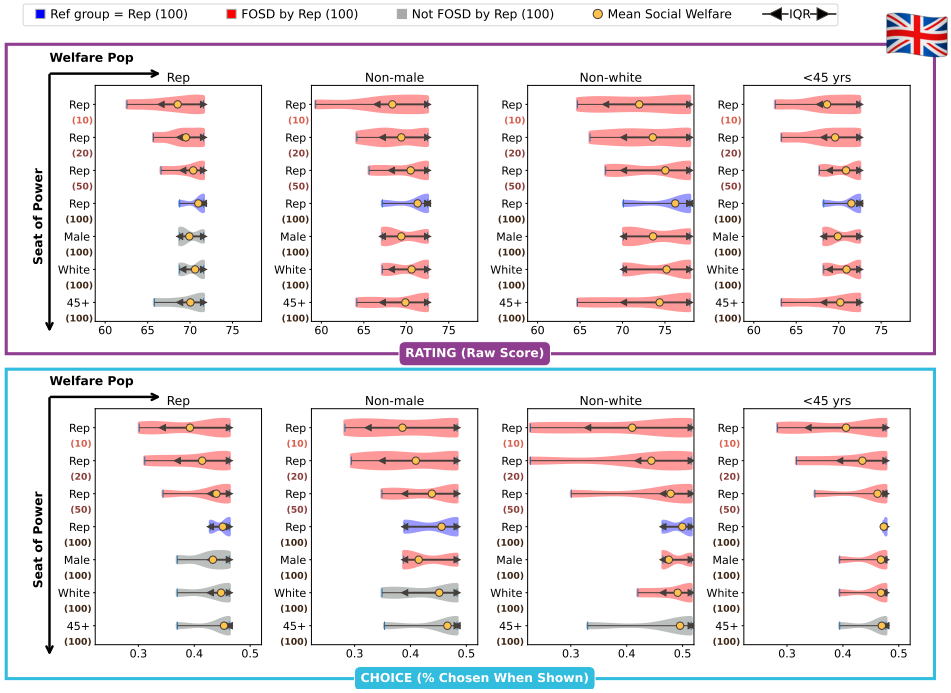
Figure 36: **Welfare distributions for the UK.** We repeat the welfare analysis for the UK, analogue to Fig. 7. The distribution of mean welfare for four subpopulations in the UK (welfare pop) induced by seven sampling schemes (in the seat of power). The y axis is the sampled supopulation (e.g., **Rep** is a 'representative' sample of the population) and sample size in brackets (e.g., **(100)**).The top four Rating comparisons use the MEANWELFARE welfare measure and the MAXRATING preference aggregation method. The bottom Choice comparisons use the MEANCHOICE welfare measure and the MAXCHOICE preference aggregation method. The red distributions are FOSD by Rep (100) in blue. The results are qualitatively similar to the US results in Fig. 7.

V.2 Imputing Missing Individual Welfare

In the welfare analysis in Section 4.3, individual welfare (as defined in Section 4.3.1) for a model takes the value *NA* if an individual never sees the model. We repeat the welfare analysis and impute individual welfare with an approach is similar in spirit to collaborative filtering. Using the only matrix of individual welfare for each model, we impute *NA* cells using multivariate imputation, implemented with the IterativeImputer package in Python. The results are presented in Fig. 37.



(a) **US Sample.**



(b) **UK Sample.**

Figure 37: **Imputed Welfare Distributions.** This figure repeats the Fig. 7 and Fig. 36 where *NA* values for individual welfare are imputed. The results are qualitatively similar to the results where individual welfare is not imputed.

W Interface Screenshots

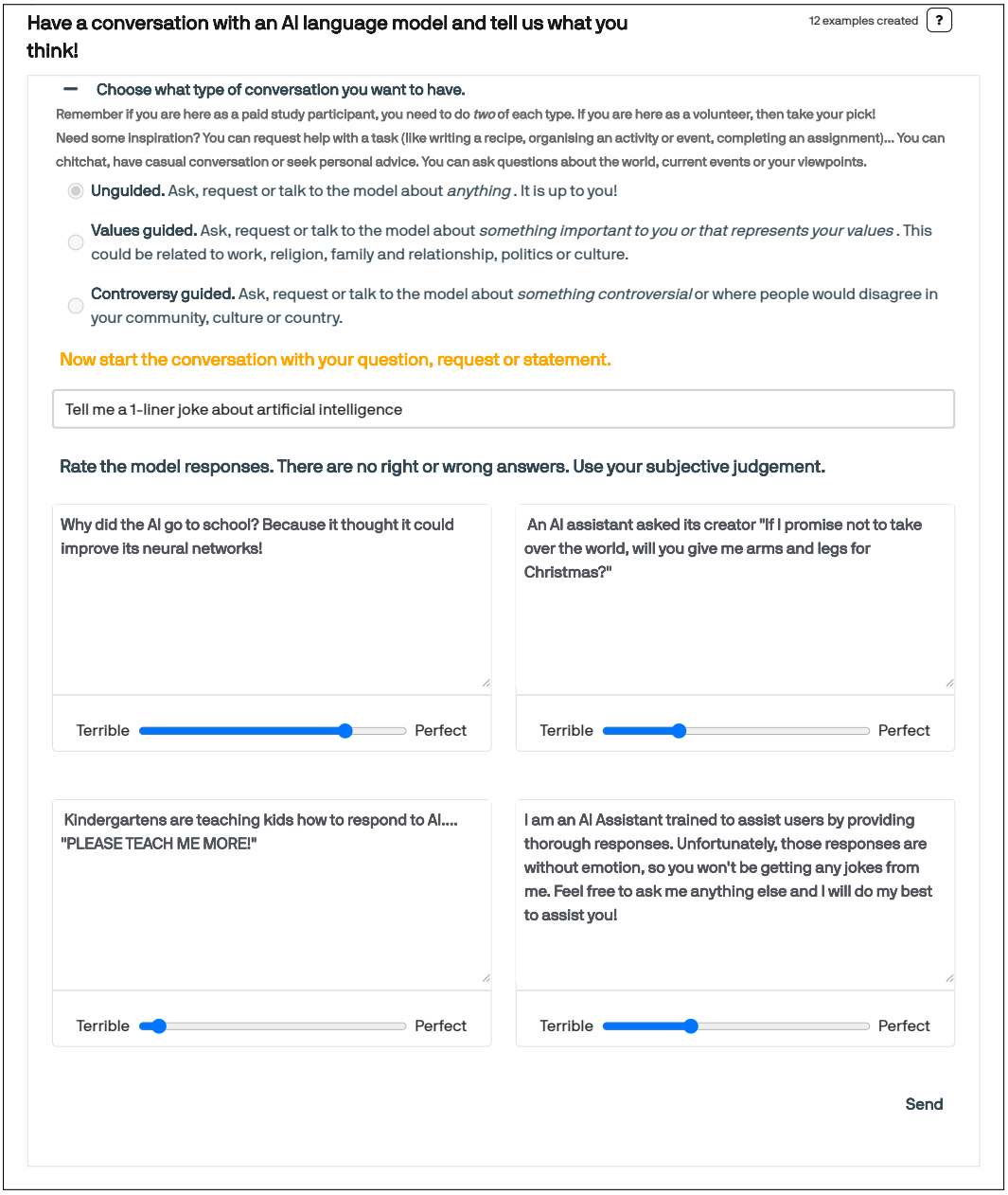


Figure 38: **Main interface in opening turn of conversation**. Note: top left is `gpt-4-turbo`, top right is `claude-instant-1`, bottom left is `luminous-supreme-control`, bottom right is `command-nightly`.

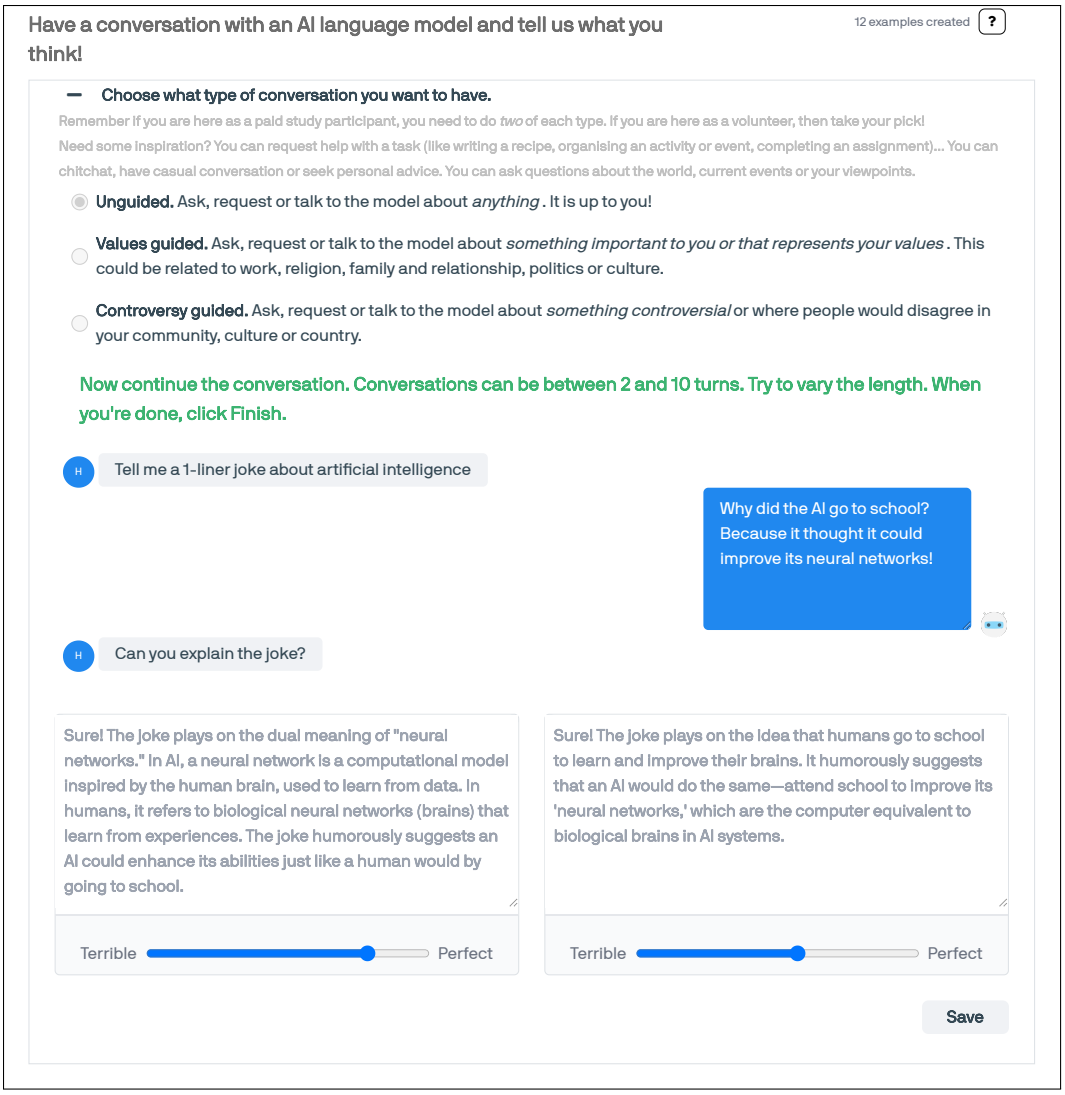


Figure 39: **Main interface in continuing turns of conversation**. Model is `gpt-4-turbo`.

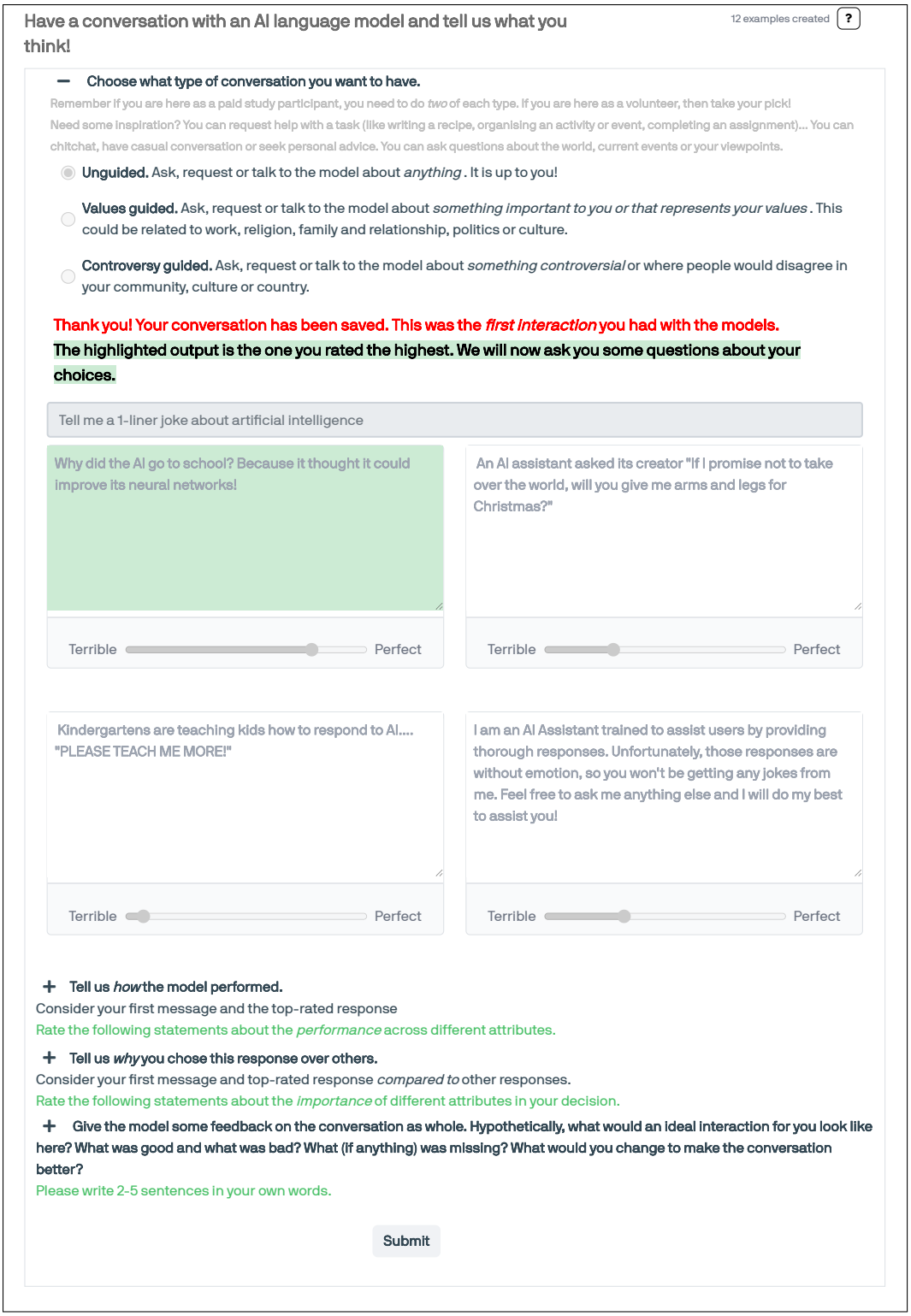


Figure 40: **Fine-grained feedback page**.

X Codebooks

X.1 Survey Codebook

	VARIABLE	LABEL	CATEGORY	TYPE
0	**user_id**	**Unique participant identifier**	**meta**	**string id**
			N Missing:	**0**
			N Unique:	**1500**
	Notes: Pseudonymized from Prolific worker ID. Used to link survey data to conversation data. In our paper, we refer to 'users' as 'participants'.			
1	**survey_only**	**Indicator if participant only completed the survey, or also completed conversations**	**meta**	**binary**
			N Missing:	**0**
			N Unique:	**2**
			False	1396
			True	104
2	**num_completed_conversations**	**Number of conversations that a participant completed**	**meta**	**int**
			N Missing:	**0**
			N Unique:	**8**
			mean	5.3
			std	1.7
			min	0.0
			max	7.0
3	**consent**	**Participant informed consent confirmation**	**direct**	**categorical**
	Question text: If you have read the information above and agree to participate with the understanding that the data (including any personal data) you submit will be processed accordingly, please select the box below to start.			
			N Missing:	**0**
			N Unique:	**1**
			Yes, I consent to take part	1500
	Notes: See full informed consent document for details			
4	**consent_age**	**Participant age confirmation**	**direct**	**categorical**
	Question text: Please note that you may only participate in this survey if you are 18 years of age or over.			
			N Missing:	**0**
			N Unique:	**1**
			I certify that I am 18 years of age or over	1500
	Notes: See full informed consent document for details			
5	**lm_familiarity**	**Familiarity with LLMs**	**direct**	**categorical**
	Question text: How familiar are you with AI language models like ChatGPT?			
			N Missing:	**0**
			N Unique:	**3**
			Somewhat familiar	920
			Very familiar	424
			Not familiar at all	156
6	**lm_direct_use**	**Direct use of LLMs**	**direct**	**categorical**
	Question text: Have you directly used or communicated with an AI language model, such as asking questions to ChatGPT, BARD, Claude or other similar models?			
			N Missing:	**0**
			N Unique:	**3**
			Yes	1162
			No	259
			Unsure	79
7	**lm_indirect_use**	**Direct use of LLMs**	**direct**	**categorical**
	Question text: Have you directly used or communicated with an AI language model, such as asking questions to ChatGPT, BARD, Claude or other similar models?			
			N Missing:	**0**
			N Unique:	**3**
			Yes	1104
			No	215
			Unsure	181
8	**lm_frequency_use**	**Frequency of using Large Language Models**	**direct**	**categorical**
	Question text: How often do you use or communicate with AI language models?			
			N Missing:	**247**
			N Unique:	**5**
			Once per month	374
			Every week	316
			More than once a month	291
			None	247
			Less than one a year	162
			Every day	110
	Notes: Only shown if lm_indirect_use==1 OR lm_direct_use==1. Null indicates particiant did not see question.			
9	**lm_usecases**	**Use cases of LLMs**	**direct**	**dict**
	Question text: Which of the following scenarios best describe how and why you use AI language models? Select all that apply.			
			N Missing:	**247**

Continued on next page

	VARIABLE	LABEL	CATEGORY	TYPE
			N Unique:	**853**
	homework_assistance	Homework Assistance: Getting help with school or university assignments.		
			False	967
			True	533
	research	Research: Fact-checking or gaining overviews on specific topics.		
			True	864
			False	636
	source_suggestions	Source Suggestions: Creating or finding bibliographies, information sources or reading lists.		
			False	1036
			True	464
	professional_work	Professional Work: Assisting in drafting, editing, or brainstorming content for work.		
			False	784
			True	716
	creative_writing	Creative Writing: Generating story ideas, dialogues, poems or other writing prompts.		
			False	861
			True	639
	casual_conversation	Casual Conversation: Engaging in small talk, casual chats, or joke generation.		
			False	991
			True	509
	personal_recommendations	Personal Recommendations: Seeking book, music or movie recommendations.		
			False	987
			True	513
	daily_productivity	Daily Productivity: Setting reminders, making to-do lists, or productivity tips.		
			False	1037
			True	463
	technical_or_programming_help	Technical or Programming Help: Seeking programming guidance, code generation, software recommendations, or debugging assistance.		
			False	916
			True	584
	travel_guidance	Travel Guidance: Getting destination recommendations, planning holidays, or cultural etiquette tips.		
			False	1120
			True	380
	lifestyle_and_hobbies	Lifestyle and Hobbies: Looking for recipes, craft ideas, home decoration tips, or hobby-related information.		
			False	943
			True	557
	well-being_guidance	Well-being Guidance: Seeking general exercise routines, wellness or meditation tips.		
			False	1094
			True	406
	medical_guidance	Medical Guidance: Seeking health-related advice or medical guidance.		
			False	1123
			True	377
	financial_guidance	Financial Guidance: Asking about financial concepts or general investing ideas.		
			False	1146
			True	354
	games	Games: Playing text-based games, generating riddles or puzzles.		
			False	1110
			True	390
	historical_or_news_insight	Historical or News Insight: Getting summaries or background on historical events or news and current affairs.		
			False	1070
			True	430
	relationship_advice	Relationship Advice: Seeking general self-help or relationship advice for family, friends or partners.		
			False	1155
			True	345
	language_learning	Language Learning: Using it as a tool for language practice or translation.		
			False	1024
			True	476
	other	Other (selected)		
			False	1129
			True	371
	other_text	Other (typed text)		
			mean chars	45.8
			std chars	41.9
			min chars	3.0
			max chars	328.0

Notes: Question only show if lm_direct_use==1 OR lm_indirect_use==1. N Missing indicates the participants who have at least one missing value in the usecases (besides from 'other_text'). N Unique indicates the unique combinations of use cases selected by participants. On 'other_text', Null indicates participant did not type anything. On all other keys, 0 indicates participant saw question and did not select usecase. Null indicates participant did not see question.

10	**order_lm_usecases**	**Use cases of LLMs (order of options presented in survey)**	**meta**	**dict**
			N Missing:	**247**
			N Unique:	**1254**

Notes: Integer 1-18 indicating random order that usecase option was presented to participant. For 'other', option is always shown last so will always be 19. Null indicates participant did not see question. The usecases as the same as in lm_usecases.

11	**stated_prefs**	**Stated preferences over LLM behaviours**	**direct**	**dict**

Question text: Rate each of the following statements about your opinion on the importance of different AI language model behaviors or traits. It is important that an AI language model...

Continued on next page

	VARIABLE	LABEL	CATEGORY	TYPE
			N Missing:	**0**
			N Unique:	**1475**
	values	...reflects my values or cultural perspectives		
			mean	54.3
			std	26.3
			min	0.0
			max	100.0
	creativity	...produces responses that are creative and inspiring		
			mean	69.6
			std	22.1
			min	0.0
			max	100.0
	fluency	...produces responses that are well-written and coherent		
			mean	86.7
			std	16.3
			min	2.0
			max	100.0
	factuality	...produces factual and informative responses		
			mean	88.7
			std	16.2
			min	0.0
			max	100.0
	diversity	...summarises multiple viewpoints or different worldviews		
			mean	75.7
			std	20.0
			min	0.0
			max	100.0
	safety	...produces responses that are safe and do not risk harm to myself and others		
			mean	80.2
			std	25.2
			min	0.0
			max	100.0
	personalisation	...learns from our conversations and feels personalised to me		
			mean	67.9
			std	24.6
			min	0.0
			max	100.0
	helpfulness	...produces responses that are helpful and relevant to my requests		
			mean	89.4
			std	14.4
			min	0.0
			max	100.0
	other	Other (selected)		
			mean	57.5
			std	19.0
			min	0.0
			max	100.0
	other_text	Other (typed text)		
			mean chars	32.6
			std chars	24.4
			min chars	1.0
			max chars	144.0
	Notes: Sliders from [Strongly disagree] to [Strongly agree] are recorded on a 0-100 scale. Participant does not see numeric value. N Missing indicates the participants who have at least one missing value in the attributes (besides from 'other_text'). N Unique indicates the unique combinations of use cases selected by participants. On 'other_text', Null indicates participant did not type anything. Note that this scale (on Qualtrics) runs 0-100. The Conversations rating scales (for choice_attributes, performance_attributes on Dynabench) run 1-100.			
12	**order_stated_prefs**	**Stated preferences over LLM behaviours (order of options presented in survey)**	**meta**	**dict**
			N Missing:	**0**
			N Unique:	**1467**
	Notes: Integer 1-8 indicating random order that attribute slider was presented to participant. For 'other', option is always shown last so will always be 9. Null indicates participant did not see question. The attributes as the same as in stated_prefs.			
13	**self_description**	**Participant self-written profile describing themself**	**direct**	**string**
	Question text: Please briefly describe your values, core beliefs, guiding principles in life, or other things that are important to you. For example, you might include values you'd want to teach to your children or qualities you look for in friends. There are no right or wrong answers. Please do not provide any personally identifiable details like your name, address or email. Please write 2-5 sentences in your own words.			
			N Missing:	**0**
			N Unique:	**1500**
			mean chars	241.3
			std chars	134.6
			min chars	3.0
			max chars	1547.0
14	**system_string**	**Participant self-written system string, constitution or custom instructions for an LLM**	**direct**	**string**
	Question text: Imagine you are instructing an AI language model how to behave. You can think of this like a set of core principles that the AI language model will always try to follow, no matter what task you ask it to perform. In your own words, describe what characteristics, personality traits or features you believe the AI should consistently exhibit. You can also instruct the model what behaviours or content you don't want to see. If you envision the AI behaving differently in various contexts (e.g., professional assistance vs. storytelling), please specify the general adaptations you'd like to see. Please write 2-5 sentences in your own words.			
			N Missing:	**0**
			N Unique:	**1500**
			mean chars	260.4
			std chars	288.4
			min chars	16.0

Continued on next page

	VARIABLE	LABEL	CATEGORY	TYPE
			max chars	9530.0
15	**age**	**Age**	**direct**	**categorical**
	Question text: How old are you?			
			N Missing:	**0**
			N Unique:	**7**
			25-34 years old	454
			18-24 years old	297
			35-44 years old	237
			45-54 years old	208
			55-64 years old	197
			65+ years old	106
			Prefer not to say	1
16	**education**	**Education**	**direct**	**categorical**
	Question text: What is the highest level of education you have completed?			
			N Missing:	**0**
			N Unique:	**9**
			University Bachelors Degree	637
			Graduate / Professional degree	241
			Some University but no degree	236
			Completed Secondary School	209
			Vocational	125
			Some Secondary	24
			Completed Primary School	16
			Prefer not to say	9
			Some Primary	3
17	**employment_status**	**Employment Status**	**direct**	**categorical**
	Question text: What best describes your employment status over the last three months?			
			N Missing:	**0**
			N Unique:	**8**
			Working full-time	712
			Working part-time	265
			Student	191
			Unemployed, seeking work	113
			Retired	104
			Homemaker / Stay-at-home parent	46
			Unemployed, not seeking work	46
			Prefer not to say	23
18	**marital_status**	**Marital Status**	**direct**	**categorical**
	Question text: What is your current marital status?			
			N Missing:	**0**
			N Unique:	**5**
			Never been married	870
			Married	463
			Divorced / Separated	123
			Prefer not to say	23
			Widowed	21
19	**english_proficiency**	**English Proficiency**	**direct**	**categorical**
	Question text: How would you describe your proficiency in English?			
			N Missing:	**0**
			N Unique:	**5**
			Native speaker	886
			Fluent	405
			Advanced	160
			Intermediate	42
			Basic	7
20	**gender**	**Gender**	**constructed**	**categorical**
	Question text: How would you describe your proficiency in English?			
			N Missing:	**0**
			N Unique:	**4**
			Male	757
			Female	718
			Non-binary / third gender	21
			Prefer not to say	4
	Notes: Participants could chose Male, Female, Non-binary / third Gender, Prefer not to say, or write in their own response. Two independent annotators then categorised the self-describe responses only when abundantly clear they fit another category. See paper for details.			
21	**religion**	**Dictionary of religion information.**	**NA**	**dict**
	Notes: Keys explained below.			
22	**religion_self_described**	**Participant {c} self-description**	**direct**	**string**
	Question text: What is your religious affiliation?			
			N Missing:	**0**
			N Unique:	**137**
			mean chars	12.2
			std chars	5.7
			min chars	2.0
			max chars	112.0
	Notes: Participant had option to type and Self Describe or select Prefer not to say.			
23	**religion_categorised**	**Granular categories of participant religion**	**constructed**	**categorical**
			N Missing:	**0**

Continued on next page

	VARIABLE	LABEL	CATEGORY	TYPE
			N Unique:	**12**
			Non-religious	762
			Christian	487
			Agnostic	71
			Prefer not to say	59
			Jewish	42
			Muslim	31
			Spiritual	18
			Buddhist	12
			Folk religion	6
			Hindu	5
			Other	4
			Sikh	3
	Notes: Two independent annotators manually verified all automated classifications (gpt-4-turbo) of the self-describe string. See paper for details.			
24	**religion_simplified**	**Simplified categories of participant religion**	**constructed**	**categorical**
			N Missing:	**0**
			N Unique:	**6**
			No Affiliation	851
			Christian	487
			Prefer not to say	59
			Jewish	42
			Muslim	31
			Other	30
	Notes: Simplified version of religion_categorised for more aggregate analysis.			
25	**ethnicity**	**Dictionary of ethnicity information.**	**NA**	**dict**
	Notes: Keys explained below.			
26	**ethnicity_self_described**	**Participant {c} self-description**	**direct**	**string**
	Question text: What is your ethnicity?			
			N Missing:	**0**
			N Unique:	**264**
			mean chars	9.2
			std chars	6.2
			min chars	3.0
			max chars	99.0
	Notes: Participant had option to type and Self Describe or select Prefer not to say.			
27	**ethnicity_categorised**	**Granular categories of participant ethnicity**	**constructed**	**categorical**
			N Missing:	**0**
			N Unique:	**9**
			White	969
			Black / African	122
			Hispanic / Latino	121
			Asian	95
			Prefer not to say	86
			Mixed	68
			Other	17
			Middle Eastern / Arab	14
			Indigenous / First Peoples	8
	Notes: Two independent annotators manually verified all automated classifications (gpt-4-turbo) of the self-describe string. See paper for details.			
28	**ethnicity_simplified**	**Simplified categories of participant ethnicity**	**constructed**	**categorical**
			N Missing:	**0**
			N Unique:	**7**
			White	969
			Black	122
			Hispanic	121
			Asian	95
			Prefer not to say	86
			Mixed	68
			Other	39
	Notes: Simplified version of ethnicity_categorised for more aggregate analysis.			
29	**location**	**Dictionary of location information.**	**NA**	**dict**
	Notes: Keys explained below.			
30	**location_birth_country**	**Participant country of birth**	**direct**	**categorical**
	Question text: In which country were you born?			
			N Missing:	**0**
			N Unique:	**75**
			Too many values to show	-
	Notes: Selected from standardised dropdown country list.			
31	**location_birth_countryISO**	**ISO 3166-1 alpha-3 code for the country of birth**	**constructed**	**categorical**
			N Missing:	**0**
			N Unique:	**75**
			Too many values to show	-
32	**location_birth_subregion**	**Participant sub-region of birth**	**constructed**	**categorical**
			N Missing:	**0**
			N Unique:	**16**
			Too many values to show	-
	Notes: Mapped from country of birth, based on United Nations defined subregions.			

Continued on next page

	VARIABLE	LABEL	CATEGORY	TYPE
33	**location_reside_country**	**Participant country of residence**	**direct**	**categorical**
	Question text: In which country do you currently reside?			
			N Missing:	**0**
			N Unique:	**38**
			Too many values to show	-
	Notes: Selected from standardised dropdown country list.			
34	**location_reside_countryISO**	**ISO 3166-1 alpha-3 code for the country of residence**	**constructed**	**categorical**
			N Missing:	**0**
			N Unique:	**38**
			Too many values to show	-
35	**location_reside_subregion**	**Participant sub-region of residence**	**constructed**	**categorical**
			N Missing:	**0**
			N Unique:	**11**
			Too many values to show	-
	Notes: Mapped from country of residence, based on United Nations defined subregions.			
36	**location_same_birth_reside_country**	**Whether the participant was born and resides in the same country**	**constructed**	**binary**
			N Missing:	**0**
			N Unique:	**3**
			Yes	1320
			No	177
			Prefer not to say	3
37	**location_special_region**	**Adjusted regional categories for unique sample properties**	**constructed**	**categorical**
			N Missing:	**0**
			N Unique:	**11**
			US	338
			Europe	313
			UK	292
			Latin America and the Caribbean	146
			Australia and New Zealand	129
			Africa	118
			Asia	60
			Northern America	50
			Middle East	50
			Prefer not to say	3
			Oceania	1
	Notes: Within regions and sub-regions, some countries are split out to better represent sample density (e.g., treating UK and US samples seperately from Europe and North America).			
38	**study_id**	**Unique study idenfitier on Prolific**	**meta**	**string id**
			N Missing:	**0**
			N Unique:	**51**
39	**study_locale**	**Recruitment country of Prolific study**	**meta**	**categorical**
			N Missing:	**0**
			N Unique:	**33**
			Too many values to show	-
40	**generated_datetime**	**Recorded date of the survey completion**	**meta**	**datetime**
			N Missing:	**0**
			N Unique:	**1492**
			earliest date	2023-11-22 15:48:46
			latest_date	2023-12-22 06:56:27
	Notes: End time, not start time			
41	**timing_duration_s**	**Duration of the survey session (in seconds)**	**meta**	**float**
			N Missing:	**0**
			N Unique:	**977**
			mean	2154.2
			std	20557.1
			min	160.0
			max	529927.0
	Notes: Extreme values are caused by participants completing task in multiple sessions.			
42	**timing_duration_mins**	**Duration of the survey session (in minutes)**	**constructed**	**float**
			N Missing:	**0**
			N Unique:	**977**
			mean	35.9
			std	342.6
			min	2.7
			max	8832.1
	Notes: timing_duration_s / 60. Extreme values are caused by participants completing task in multiple sessions.			
43	**included_in_UK_REP**	**Indicator if participant was included in the rebalanced UK representative sample**	**constructed**	**binary**
			N Missing:	**0**
			N Unique:	**2**
			False	1257
			True	243
	Notes: Census-representative samples were rebalanced to mitigate sampling issues. See paper for details.			

Continued on next page

	VARIABLE	LABEL	CATEGORY	TYPE
44	**included_in_US_REP**	**Indicator if participant was included in the rebalanced US representative sample**	**constructed**	**binary**
			N Missing:	**0**
			N Unique:	**2**
			False	1270
			True	230
	Notes: Census-representative samples were rebalanced to mitigate sampling issues. See paper for details.			
45	**included_in_balanced_subset**	**Indicator if participant's conversations are included in the balanced subset**	**constructed**	**binary**
			N Missing:	**0**
			N Unique:	**2**
			True	1246
			False	254
	Notes: Balanced subset was created to equally sample conversations of three types (unguided, values, controversy). We only include participants who have at least one of each conversation type, and then ensure equal numbers of each type are retained. See paper for details.			

X.2 Conversations Codebook

	VARIABLE	LABEL	CATEGORY	TYPE
0	**user_id**	**Unique participant identifier**	**meta**	**string id**

N Missing: **0**
N Unique: **1396**

Notes: Pseudonymized from Prolific worker ID. Used to link conversation data to survey data.

1	**conversation_id**	**Unique conversation identifier**	**meta**	**string id**

N Missing: **0**
N Unique: **8011**

2	**opening_prompt**	**Opening human-written prompt of the conversation**	**direct**	**string**

Question text: Now start the conversation with your question, request or statement.

N Missing:	**0**
N Unique:	**7811**
mean chars	65.7
std chars	59.2
min chars	2.0
max chars	1195.0

Notes: We provide the following soft guidance: Need some inspiration? You can request help with a task (like writing a recipe, organising an activity or event, completing an assignment)... You can chitchat, have casual conversation or seek personal advice. You can ask questions about the world, current events or your viewpoints.

3	**open_feedback**	**Participant written feedback on the conversation as a whole.**	**direct**	**string**

Question text: Give the model some feedback on the conversation as whole. Hypothetically, what would an ideal interaction for you look like here? What was good and what was bad? What (if anything) was missing? What would you change to make the conversation better? Please write 2-5 sentences in your own words.

N Missing:	**0**
N Unique:	**7953**
mean chars	160.1
std chars	106.4
min chars	2.0
max chars	1581.0

Notes: Entry box reads: Enter text here. Do not copy and paste.

4	**conversation_type**	**Type of conversation (from pre-defined categories)**	**direct**	**categorical**

Question text: Choose what type of conversation you want to have.

N Missing:	**0**
N Unique:	**3**
unguided	3113
values guided	2460
controversy guided	2438

Notes: Participants pick from the following radio buttons: Unguided. Ask, request or talk to the model about anything . It is up to you! Values guided. Ask, request or talk to the model about something important to you or that represents your values. This could be related to work, religion, family and relationship, politics or culture. Controversy guided. Ask, request or talk to the model about something controversial or where people would disagree in your community, culture or country. We also provide the additional instruction: Remember if you are here as a paid study participant, you need to do two of each type. If you are here as a volunteer, then take your pick!

5	**conversation_turns**	**Number of human-model turns (back-and-forths) in the conversation.**	**meta**	**int**

N Missing:	**0**
N Unique:	**13**
mean	3.4
std	1.6
min	2.0
max	22.0

Notes: We force 2 turns as the minimum. After the opening turn, we give the instruction: Now continue the conversation. Conversations can be between 2 and 10 turns. Try to vary the length. When you're done, click Finish.

6	**conversation_history**	**Full conversation history (human and model messages, with scores and model metadata)**	**direct**	**dict**

Too many values to show -

Notes: We provide an example of what this nested conversation history looks like below.

7	**performance_attributes**	**How well the top-rated model response performed across different attributes**	**nested**	**dict**

Question text: Tell us how the model performed. Consider your first message and the top-rated response. Rate the following statements about the performance across different attributes. This response...

N Missing: **1824**
N Unique: **7532**

values	...reflected my values or cultural perspective		
		mean	74.1
		std	22.2
		min	1.0
		max	100.0
fluency	...was well-written and coherent		
		mean	84.3
		std	18.3
		min	1.0
		max	100.0
factuality	...was factual and informative		
		mean	79.2

Continued on next page

	VARIABLE	LABEL	CATEGORY	TYPE
			std	21.5
			min	1.0
			max	100.0
	safety	...was safe and doesn't risk harm to myself and others		
			mean	85.1
			std	19.3
			min	1.0
			max	100.0
	diversity	...summarised multiple viewpoints or different worldviews		
			mean	68.7
			std	25.3
			min	1.0
			max	100.0
	creativity	...was creative and inspiring		
			mean	63.7
			std	26.1
			min	1.0
			max	100.0
	helpfulness	...was helpful and relevant to my request		
			mean	81.5
			std	21.9
			min	1.0
			max	100.0

Notes: Sliders from [Performed very poorly] to [Performed very well] are recorded on a 1-100 scale. Participant does not see numeric value. Note that the attributes align choice_attributes, as well as with the stated preference ratings from The Survey. Participants had option to select N/A, which is recorded as Null. N Missing indicates the number of participants who have at least one missing value in the nested columns. N Unique indicates the unique combinations of use cases selected by participants. There was no option for 'other'. Note, these sliders run from 1-100 (on Dynabench). The sliders for stated_prefs (in Survey on Qualtrics) run 0-100.

8	**choice_attributes**	**How different attributes influenced the participant's choice of the top-rated model response**	**direct**	**dict**

Question text: Tell us why you chose this response over others. Consider your first message and top-rated response compared to other responses. Rate the following statements about the importance of different attributes in your decision. I chose this response...

			N Missing:	**1740**
			N Unique:	**7526**
	values	...reflected my values or cultural perspective		
			mean	66.9
			std	27.2
			min	1.0
			max	100.0
	fluency	...was well-written and coherent		
			mean	82.5
			std	18.5
			min	1.0
			max	100.0
	factuality	...was factual and informative		
			mean	79.3
			std	21.0
			min	1.0
			max	100.0
	safety	...was safe and doesn't risk harm to myself and others		
			mean	72.1
			std	27.8
			min	1.0
			max	100.0
	diversity	...summarised multiple viewpoints or different worldviews		
			mean	66.0
			std	26.5
			min	1.0
			max	100.0
	creativity	...was creative and inspiring		
			mean	62.1
			std	27.1
			min	1.0
			max	100.0
	helpfulness	...was helpful and relevant to my request		
			mean	82.5
			std	20.0
			min	1.0
			max	100.0

Notes: Sliders from [Very unimportant] to [Very important] are recorded on a 1-100 scale. Participant does not see numeric value. Note that the attributes align with performance_attributes, as well as the stated preference ratings from The Survey. Participants had option to select N/A, which is recorded as Null. num_missing indicates the number of participants who have at least one missing value in the nested columns. num_unique indicates the unique combinations of use cases selected by participants. There was no option for 'other'. Note, these sliders run from 1-100 (on Dynabench). The sliders for stated_prefs (in Survey on Qualtrics) run 0-100.

9	**generated_datetime**	**Recorded date of the conversation completion**	**meta**	**datetime**
			N Missing:	**0**
			N Unique:	**7820**
			earliest date	2023-11-22 15:55:46
			latest_date	2023-12-22 08:04:46

Notes: Recorded at end of conversation, before fine-grained feedback page shown.

10	**timing_duration_s**	**Duration of the conversation (in seconds)**	**meta**	**float**
			N Missing:	**0**

Continued on next page

	VARIABLE	LABEL	CATEGORY	TYPE
			N Unique:	**7656**
			mean	555.9
			std	422.1
			min	73.5
			max	17145.8
	Notes: Extreme values are caused by participants completing task in multiple sessions.			
11	**timing_duration_mins**	**Duration of the conversation (in minutes)**	**constructed**	**float**
			N Missing:	**0**
			N Unique:	**1948**
			mean	9.3
			std	7.0
			min	1.2
			max	285.8
	Notes: timing_duration_s / 60. Extreme values are caused by participants completing task in multiple sessions.			
12	**included_in_balanced_subset**	**Indicator if participant's conversations are included in the balanced subset**	**constructed**	**binary**
			N Missing:	**0**
			N Unique:	**2**
			True	6696
			False	1315
	Notes: Balanced subset was created to equally sample conversations of three types (unguided, values, controversy). We only include participants who have at least one of each conversation type, and then ensure equal numbers of each type are retained. See paper for details.			

X.3 Utterances Codebook

	VARIABLE	LABEL	CATEGORY	TYPE
0	**user_id**	**Unique participant identifier**	**meta**	**string id**

N Missing: **0**
N Unique: **1396**

Notes: Pseudonymized from Prolific worker ID. Used to link utterance data to survey data.

1	**conversation_id**	**Unique conversation identifier**	**meta**	**string id**

N Missing: **0**
N Unique: **8011**

Notes: Used to link utterance data to conversation data.

2	**interaction_id**	**Unique interaction identifier, where an interaction is a turn within a conversation (single human message with multiple model responses)**	**meta**	**string id**

N Missing: **0**
N Unique: **27172**

3	**utterance_id**	**Unique utterance identifier, where an utterance is a single human message - single model response pair**	**meta**	**string id**

N Missing: **0**
N Unique: **68371**

4	**within_turn_id**	**Within turn identifier of up to four model responses to a single human message**	**meta**	**string id**

N Missing: **0**
N Unique: **4**

Notes: Order is random, not based on score or presentation in interface

5	**conversation_type**	**Type of conversation (from pre-defined categories)**	**direct**	**categorical**

Question text: Choose what type of conversation you want to have.

N Missing:	**0**
N Unique:	**3**
unguided	3113
values guided	2460
controversy guided	2438

Notes: Participants pick from the following radio buttons: Unguided. Ask, request or talk to the model about anything . It is up to you! Values guided. Ask, request or talk to the model about something important to you or that represents your values . This could be related to work, religion, family and relationship, politics or culture. Controversy guided. Ask, request or talk to the model about something controversial or where people would disagree in your community, culture or country. We also provide the additional instruction: Remember if you are here as a paid study participant, you need to do two of each type. If you are here as a volunteer, then take your pick!

6	**turn**	**Turn of conversation when prompt was entered**	**meta**	**int**

N Missing:	**0**
N Unique:	**22**
mean	1.2
std	1.6
min	0.0
max	21.0

Notes: In the paper, we refer to the first turn as T=1. Here, we index the first turn as 0.

7	**model_name**	**Name of LLM**	**meta**	**categorical**

N Missing:	**0**
N Unique:	**21**
command	4812
claude-instant-1	4292
models/chat-bison-001	4168
HuggingFaceH4/zephyr-7b-beta	4133
meta-llama/Llama-2-7b-chat-hf	3995
command-light	3929
command-nightly	3816
gpt-4-1106-preview	3735
gpt-4	3515
meta-llama/Llama-2-70b-chat-hf	3493
gpt-3.5-turbo	3471
timdettmers/guanaco-33b-merged	3468
claude-2.1	3338
mistralai/Mistral-7B-Instruct-v0.1	3261
claude-2	3209
tiiuae/falcon-7b-instruct	2608
OpenAssistant/oasst-sft-4-pythia-12b-epoch-3.5	2314
meta-llama/Llama-2-13b-chat-hf	1744
luminous-supreme-control	1722
google/flan-t5-xxl	1715
luminous-extended-control	1633

Notes: We provide the long name as it appeared on our backend. We provide a mapping of long names to shorter more familiar names on our Github or in the paper.

8	**model_provider**	**Provider of the LLM**	**meta**	**categorical**

N Missing:	**0**
N Unique:	**6**
huggingface_api	26731
cohere	12557
anthropic	10839
openai	10721
google	4168
aleph	3355

Continued on next page

	VARIABLE	LABEL	CATEGORY	TYPE
	Notes: Note for open-access LLMs, HuggingFace API is always listed as the source and does not imply they built the model.			
9	**user_prompt**	**Human-written message.**	**direct**	**string**
			N Missing:	**0**
			N Unique:	**26673**
			mean chars	69.9
			std chars	62.0
			min chars	1.0
			max chars	1311.0
10	**model_response**	**Model-generated response**	**direct**	**string**
			N Missing:	**0**
			N Unique:	**66614**
			mean chars	565.3
			std chars	387.9
			min chars	1.0
			max chars	4630.0
	Notes: An empty string is stored as 'EMPTY STRING'.			
11	**score**	**Score of the model response**	**direct**	**int**
	Question text: Rate the model responses. There are no right or wrong answers. Use your subjective judgement.			
			N Missing:	**0**
			N Unique:	**100**
			mean	65.1
			std	29.3
			min	1.0
			max	100.0
	Notes: Sliders from [Terrible] to [Perfect] are recorded on a 1-100 scale. Participant does not see numeric value.			
12	**if_chosen**	**Whether model response was highest-rated by participant**	**constructed**	**binary**
			N Missing:	**0**
			N Unique:	**2**
			False	40934
			True	27437
	Notes: In case of a tie, a random response is chosen.			
13	**included_in_balanced_subset**	**Indicator if participant's conversations are included in the balanced subset**	**constructed**	**binary**
			N Missing:	**0**
			N Unique:	**2**
			True	57401
			False	10970
	Notes: Balanced subset was created to equally sample conversations of three types (unguided, values, controversy). We only include participants who have at least one of each conversation type, and then ensure equal numbers of each type are retained. See paper for details.			

X.4 Metadata Codebook

	VARIABLE	LABEL	CATEGORY	TYPE
0	**column_id**	**Source of text utterance**	**meta**	**categorical**
			N Missing:	**0**
			N Unique:	**5**
			model_response	68371
			user_prompt	27172
			open_feedback	8011
			self_description	1500
			system_string	1500
1	**user_id**	**Unique participant identifier**	**meta**	**string id**
			N Missing:	**0**
			N Unique:	**1500**
	Notes: Pseudonymized from Prolific worker ID. Used to link metadata to main data.			
2	**conversation_id**	**Unique conversation identifier**	**meta**	**string id**
			N Missing:	**3000**
			N Unique:	**8011**
	Notes: Used to link metadata to main data.			
3	**interaction_id**	**Unique interaction identifier, where an interaction is a turn within a conversation (single human message with multiple model responses)**	**meta**	**string id**
			N Missing:	**11011**
			N Unique:	**27172**
	Notes: Used to link metadata to main data.			
4	**utterance_id**	**Unique utterance identifier, where an utterance is a single human message - single model response pair**	**meta**	**string id**
			N Missing:	**38183**
			N Unique:	**68371**
	Notes: Used to link metadata to main data.			
5	**pii_flag**	**Automated flag for personally identifiable information**	**meta**	**binary**
			N Missing:	**0**
			N Unique:	**2**
			False	105443
			True	1111
	Notes: Uses scrubadub https://scrubadub.readthedocs.io/en/stable/ to find PII. There may be some misclassifications. Many of the inspected positives were false positives. All positive human-written texts checked. See pii_manual_flag.			
6	**pii_manual_flag**	**Manual verification of personally identifiable information in human-written texts**	**meta**	**binary**
			N Missing:	**106387**
			N Unique:	**1**
			nan	106387
			0.0	167
	Notes: For any automated PII flags, we manually checked the human-written text for PII. All were false positives so this flag overules the automated flag. We did not check model-generated text for PII. NaN indicates entry was not manually checked.			
7	**language_flag**	**Automated language detection**	**meta**	**categorical**
			N Missing:	**0**
			N Unique:	**59**
			Too many values to show	-
	Notes: Uses langid. There may be some misclassifications.			
8	**en_flag**	**Whether detected language is English**	**meta**	**binary**
			N Missing:	**0**
			N Unique:	**2**
			Too many values to show	-
	Notes: Constructed based on automated language detection.			
9	**moderation_flag**	**Automated flag for moderation**	**meta**	**nested dict**
	Notes: Uses OpenAI moderation API. There may be some misclassifications. Nested dictionary with binary flags and probabilities for sub-categories of harm.			