Large Language Models Perform on Par with Experts Identifying Mental Health Factors in Adolescent Online Forums

Isabelle Lorge
Department of Psychiatry, University of Oxford, OX3 7JX

Dan W. Joyce
Department of Primary Care and Mental Health, University of Liverpool, L69 3GF

Andrey Kormilitzin
Department of Psychiatry, University of Oxford, OX3 7JX

Abstract

Mental health in children and adolescents has been steadily deteriorating over the past few years [1]. The recent advent of Large Language Models (LLMs) offers much hope for cost and time efficient scaling of monitoring and intervention, yet despite specifically prevalent issues such as school bullying and eating disorders, previous studies on have not investigated performance in this domain or for open information extraction where the set of answers is not predetermined. We create a new dataset of Reddit posts from adolescents aged 12-19 annotated by expert psychiatrists for the following categories: TRAUMA, PRECARITY, CONDITION, SYMPTOMS, SUICIDALITY and TREATMENT and compare expert labels to annotations from two top performing LLMs (GPT3.5 and GPT4). In addition, we create two synthetic datasets to assess whether LLMs perform better when annotating data as they generate it. We find GPT4 to be on par with human inter-annotator agreement and performance on synthetic data to be substantially higher, however we find the model still occasionally errs on issues of negation and factuality and higher performance on synthetic data is driven by greater complexity of real data rather than inherent advantage.

1 Introduction

The recent development of powerful Large Language Models such as GPT3.5 [2] and GPT4 [3] able to perform tasks in a zero-shot manner (i.e., without having been specifically trained or fine-tuned to do so) by being simply prompted with natural language instructions shows much promise for healthcare applications and the domain of mental health. Indeed, these models display more impressive general natural language processing abilities than their predecessors and excel at tasks such as Question Answering and Named Entity Recognition [4, 5, 6, 7]. Models with the ability to process social media content for indicators of mental health issues have the potential to become invaluable cost-effective tools for applications such as public health monitoring [8] and online moderation or intervention systems [9]. In addition, synthetic data produced by LLMs can be a cost effective and privacy-preserving tool for training task specific models [10].

There have been several studies aimed at assessing the abilities of LLMs to perform a range of tasks related to mental health on datasets derived from social media. Yang et al. [11] conducted a comprehensive assessment of ChatGPT (gpt-3.5-turbo), InstructGPT3 and LlaMA7B and 13B [12]

on 11 different datasets and 5 tasks (mental health condition binary/multiclass detection, cause/factor detection, emotion detection and causal emotion entailment, i.e. determining the cause of a described emotion). They find that while the LLMs perform well (0.46-0.86 F1 depending on task), with ChatGPT substantially outperforming both LLaMA 7B and 13B, they still underperform smaller models specifically fine-tuned for each task (e.g., RoBERTa). Xu et al. [13] find similar results for Alpaca [14], FLAN-T5 [15] and LLaMA2 [16], with only fine-tuned LLMs able to perform on par with smaller, task-specific models such as RoBERTa [17, 18].

However, we find that previous studies suffer from the following shortcomings:

1. They focus on adult mental health

2. They focus on tasks with a closed (or finite) set of answers, where the model is asked to perform each task in turn

3. They do not investigate how LLMs perform on synthetic data, i.e., text they are asked to simultaneously generate and label

There is growing consensus that we are facing a child mental health crisis [1]. Before the COVID-19 pandemic there was already increasing incidence of mental health conditions in children and young people (CYP), such as depression, anxiety and eating disorders [19] as well as rising rates of self-harm and suicidal ideation [20] and cyberbullying strongly linked to adverse mental health outcomes [21]. The advent of the pandemic accelerated this already precarious situation and created additional challenges [22, 23] such as discontinuity of healthcare service provision in addition to interruption to young people's usual engagement in education and their social lives.

This age range is particularly vulnerable to onset of mental health issues, with half of conditions appearing by early adolescence and 10-20% of children and young people experiencing at least one mental health condition [24]. Females, those with low socioeconomic backgrounds, trauma, abuse or having witnessed violence [25] are at heightened risk.

On the other hand, social media now forms an important part of children and adolescents' daily lives, whose impact on mental health is debated, with potential benefits (stress reduction and support networks [26]) as well as potential risks (sleep disturbance, self esteem issues and cyberbullying [27]). Regardless of their detrimental or protective impact, social media may contribute valuable insights into CYP's mental health, with opportunities for monitoring and intervention, for example identifying those at risk of depression and mood disorders [28]. Given the mental health of CYP is a particularly pressing public health concern, we wished to investigate how LLMs perform on extracting mental health factors when faced with social media content generated by young people aged 12-19. Indeed, several issues related to mental health either exclusively apply to children and adolescents (such as school bullying and ongoing family abuse) or are particularly prevalent in this age range (such as eating disorders [29] and self-harm [30]), making both content type and factors of interest distinct from those found in adult social media posts.

In addition, previous studies focused on tasks which had either a binary or closed sets of answers (e.g., choosing between several given conditions or between several given causal factors). In contrast, we wish to examine how LLMs perform on a task of open information extraction, where they are given categories of information and asked to extract any which are found in the text (e.g., asked to detect whether there is any mental health condition indicated in the text). Furthermore, in previous studies the models were tested with each task in turn (e.g., asked to detect depression in one dataset, then detect suicidality in another dataset), whereas we gather and annotate our own dataset in order to be able to ask the LLMs to extract all categories simultaneously (e.g, extract all conditions and symptoms in a given sentence).

Finally, to our knowledge there has been no investigation on how LLM performance compares when asked to annotate text as they generate it, i.e., how their performance on synthetic data compares with their performance on real data. There is growing interest in synthetic data for healthcare [31]. Given the potential for training models and running simulations and digital twin experiments with the benefit of reduced issues of data scarcity and privacy, we believe that our work will contribute to better understanding of limitations and benefits of using synthetic data for real-world tasks.

2 Aims

In summary, we aim to:

1. Generate and annotate with high-quality expert annotations a novel dataset of social media posts which allows extraction of a wide range of mental health factors simultaneously.
2. Investigate performance of two top-performing LLMs (GPT3.5 and GPT4) on extracting mental health factors in adolescent social media posts to verify whether they can be on par with expert annotators.
3. Investigate how these LLMs perform on synthetic data, i.e., when asked to annotate text as they generate it, with the aim of assessing utility of these data in training task specific models

3 Method

3.1 Reddit dataset

We use Python's PRAW library to collect post from the Reddit website (www.reddit.com) over the last year, including posts from specific forum subthemes ('subreddits') dedicated to mental health topics: *r/anxiety, r/depression, r/mentalhealth, r/bipolarreddit, r/bipolar, r/BPD, r/schizophrenia, r/PTSD, r/autism, r/trau-matoolbox, r/socialanxiety, r/dbtselfhelp, r/offmychest* and *r/mmfb*. The distribution of subreddits in the dataset can be found in Figure 1.

As in previous works [32], we use heuristics to obtain posts from our target age range (e.g, posts containing expression such as *I am 16/just turned 16/etc.*) We gather 1000 posts written by 950 unique users. To optimise the annotation process, we select the most relevant sentences to be annotated by embedding a set of mental health keywords with Python's *sentence-transformers* library [33] calculating the cosine similarity with post sentences, choosing a threshold of 0.2 cosine similarity after trial and error. We keep the post index for each sentence to provide context. The resulting dataset contains 6500 sentences.

3.2 Ethical considerations

In conducting this research, we recognised the importance of respecting the autonomy and privacy of the Reddit users whose posts were included in our dataset. While Reddit data is publicly available and was obtained from open online forums, we acknowledge that users may not have anticipated their contributions being used for research purposes and will therefore make the data available only on demand. The verbatim example sentences given in later sections have been modified to prevent full-text searching strategies to infer the post author's immediate identity on reddit.

To protect the confidentiality of participants, we did not provide usernames or other identifying information to our annotators. Annotators were psychiatrists who were warned that the content of the posts was highly sensitive with potentially triggering topics such as self-harm and child abuse.

Reddit's data sharing and research policy allows academic researchers to access certain Reddit data for the purposes of research, subject to the platform's terms and conditions. They require researchers to obtain approval through their data access request process before using the API. The policy outlines requirements around protecting user privacy, obtaining consent, and properly attributing the data source in any published work. They reserve the right to deny data access requests or revoke access if the research is deemed to violate Reddit's policies. Researchers must also agree to Reddit's standard data use agreement when accessing the data.

Our research aims to contribute to the understanding of mental health discourse from adolescents on social media platforms. We believe the potential benefits of this work, in terms of insights that could improve mental health support and resources, outweigh the minimal risks to participants. However, we remain aware of the ethical complexities involved in using public social media data, and encourage further discussion and guidance in this emerging area of study.

3.3 Synthetic dataset

In addition to the real dataset, we generate two synthetic datasets of 500 sentences each by prompting GPT3.5 (*gpt-3.5-turbo-0125*) and GPT4 (*gpt-4-0125-preview*) to create and label Reddit-like posts of 5 sentences (temperature 0, all other parameters set to default). The instructions given were made as similar as possible to those given to annotators, and the model was expliclity told to only label factors which applied to the author of the post (e.g., not to label *My friend has depression* with CONDITION). The prompt used can be found in Appendix A.

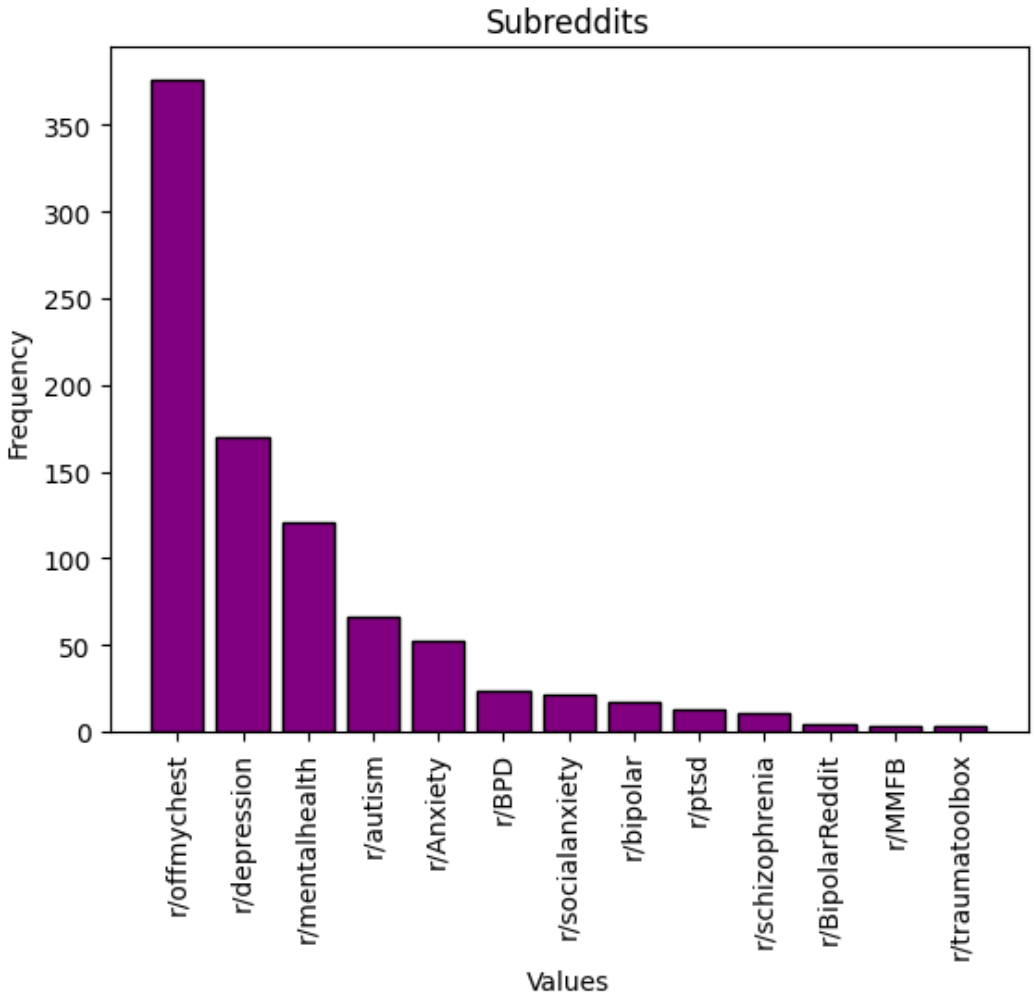


Figure 1: Distribution of subreddits

3.4 Annotation schema

Given our goal is to obtain a wide range of relevant annotations for each sentence in order to test the LLMs' ability to generalise and perform open information extraction, and the previously mentioned important factors related to trauma [34] and precarity [35], we create the following six categories in consultation with a clinical psychiatrist:

- **TRAUMA** (sexual abuse, physical abuse, emotional abuse, school bullying, death, accident, etc.)
- **PRECARITY** (socioeconomic, parental conflict, parental illness, etc.)
- **SYMPTOM** (self-harm, low self-esteem, anhedonia, panic attack, flashback, psychosis, insomnia, etc.)

- **CONDITION** (eating disorder, depression, bipolar, bpd, anxiety, ptsd, adhd, substance abuse/addiction, etc.)
- **SUICIDALITY** (no subcategories)
- **TREATMENT** (no subcategories)

Nineteen expert annotators were contacted and asked to annotate 500 sentences each for a fixed compensation of £120 ($\approx$ £60/hour). These were UK-trained psychiatrists, all of whom had obtained Membership of the Royal College of Psychiatrists by post-graduate experience and formal examinations. Thirteen annotators annotated the Reddit dataset, two annotators annotated the synthetic datasets and four annotators re-annotated samples from the Reddit and synthetic datasets for inter-annotator agreement computation (100 sentences from each dataset, 1500 sentences in total). Annotators were given the above subcategory examples but allowed to use new subcategories when appropriate (no closed set of answers). They were given the post indices to provide context (i.e., so as to be aware which sentences belonged to the same post). They were asked to annotate only school bullying as bullying, and other instances (e.g., sibling harassment) as emotional abuse. Anxiety was to be annotated as a symptom rather than condition unless specifically described as a disorder.

Experts performed the annotation by filling in the relevant columns in an Excel sheet with each sentence as a row. Importantly, given the known limitations of language models with negation [36], we wished to annotate both POSITIVE and NEGATIVE evidence in order to test LLMs' ability to handle both polarities (e.g., *I am not feeling suicidal* as negative suicidality or *We don't have any money issues* as negative socioeconomic precarity). For this purpose, annotators were asked to use the prefixes P and N (e.g., *P(adhd)* in the CONDITION column or *N(socioeconomic)* in the PRECARITY column).

3.5 Data processing and dataset statistics

In order to compare expert annotations with LLM annotations despite the wide variety of subcategories and terms used by annotators we create dictionaries mapping each term found in the dataset to a standard equivalent (e.g., *p(emotional) to p(emotional abuse), p(physical violence) to p(physical abuse), p(gun violence) and p(school shooting) to p(violence), p(rape) to p(sexual abuse), p(financial burden) and p(poor) to p(socioeconomic precarity), p(divorce) to p(family conflict), p(self hatred) to p(low self esteem)*, etc.). Parental substance abuse is considered family illness and any underspecified subcategories are marked as 'unspecified' (e.g., *p(trauma unspecified)*).

The distribution of subcategories for each category can be found in figures 2, 3, 4 and 5 in Appendix B. The most frequent subcategory in TRAUMA is emotional abuse, which occurs twice as often as physical abuse and death in the dataset. The most frequent form of PRECARITY is family conflict, then family illness (including parental substance abuse) and socioeconomic precarity. The most frequent CONDITIONS are depressive disorders, followed by substance abuse/addiction and ADHD. The most frequent SYMPTOMS are anxiety, low self-esteem, self-harm and low mood.

Interestingly, the distribution of subcategories differs quite substantially in the synthetic datasets (distributions for the GPT3.5 and GPT4 generated datasets can be found in Appendix B). Overall, the number of subcategories is reduced, indicating less diversity (however, these are smaller datasets). The top trauma subcategories are sexual abuse for GPT3.5 and school bullying for GPT4, both of which were much less prevalent in real data. The second most prevalent condition for both GPT3.5 and GPT4 is eating disorders, whereas these ranked in 8th place in real data. Finally, unlike in real data, flashbacks and panic attacks are the 3d and 4th most frequent symptoms for both GPT3.5 and GPT4-generated data, whereas self-harm ranks much lower than in real data. Given many of these subcategories were given as examples in the annotator guidelines and LLM prompt, it is likely that the LLMs used them in a more homogenous manner for generation than the distribution which would be found in real data. However, the distribution is not entirely homogenous, which suggests the LLMs did leverage some of the biases learned from their training data.

4 Results

Once both human and LLM annotations are standardised, we conduct analyses to assess performance. We provide precision, recall and F1 at the category level and accuracy at the subcategory level

collapsed across subcategories (given their high number). We compute category performance in two ways: *Positive or Negative*, where a point is awarded if the category contains an annotation in both human and LLM annotations, regardless of polarity (i.e., the annotator considered there was relevant information concerning the category TRAUMA) and *Positive Only* metrics, where negative annotations are counted as no annotations. The difference between the two metrics can be seen clearly in Table 1 (GPT3.5 results), where precision increases but recall diminishes for *Positive Only*. The increase in precision is due to the fact that GPT3.5 outputs a substantial number of negative annotations in cases where human annotators did not consider it relevant to mention the category. The reduction in recall, on the other hand, results from the fact that LLMs often confuse positive and negative annotations and will occasionally output a negative annotation for a positive one.

For real data (Tables 1 and 2), GPT3.5's performance at the category level is average, with better performance in the Positive Only metrics (0.57). GPT4 performs better, especially in Positive Only metrics (0.63) and subcategory accuracy (0.48 vs. 0.39). In general, recall is higher than precision, indicating LLMs may be overpredicting labels.

The performance for synthetic data (Tables 3 and 4) is substantially better, with no gap between the Positive or Negative and Positive Only metrics, suggesting less irrelevant negative annotations. Here again, GPT4 outperforms GPT3.5, both at the category level (0.75 vs 0.70 and 0.73 vs 0.68) and more particularly at the subcategory level, where GPT4 reaches an impressive accuracy of 0.72 (vs 0.42). The gap between recall and precision is reduced for GPT4, whereas GPT3.5 displays higher precision than recall here.

In order to assess the upper bound of human performance, we calculate inter-annotator agreement for both real and synthetic datasets using Cohen's Kappa. Values can be found in Table 5. Interestingly, while performance at the category level in real data is lower (GPT3.5) or similar (GPT4) compared to humans, GPT4 displays a substantially higher accuracy at the subcategory level (0.47 vs 0.35). For synthetic data, GPT3.5 still underperforms human agreement on all three metrics, while GPT4 is on par with humans for the Positive Only and subcategory metrics and only underperforms in the Positive and Negative metric.

Category	Positive or Negative			Positive Only			Subcategory
	Precision	Recall	F1-Score	Precision	Recall	F1-Score	Accuracy
TRAUMA	0.38	0.78	0.51	0.56	0.65	0.60	0.39
PRECARITY	0.26	0.43	0.33	0.45	0.31	0.37	0.22
CONDITION	0.33	0.85	0.48	0.54	0.72	0.62	0.55
SYMPTOMS	0.39	0.62	0.48	0.46	0.58	0.52	0.31
SUICIDALITY	0.44	0.79	0.56	0.80	0.68	0.73	/
TREATMENT	0.48	0.72	0.58	0.72	0.58	0.64	/
ALL	0.37	0.70	0.49	0.55	0.60	0.57	0.39

Table 1: GPT3.5 (real data). **Positive or Negative**: counting annotation in category regardless of polarity (category level); **Positive Only**: counting negative annotations as NaN (category level); **Subcategory**: accuracy at the subcategory level

5 Error analysis

We examine some of the sentences annotated by the LLMs in order to perform error analysis and extract the following findings (as mentioned previously some words have been paraphrased to preclude full-text search allowing user identification):

- Both GPT3.5 and GPT4 produce infelicitous negations, i.e., negative annotations which would seem irrelevant to humans, e.g., (*I have amazing people around me* =>negative parental death or *The internet is my one only coping mechanism* =>trauma unspecified)
- Despite being specifically prompted to only annotate factors related to the writer/speaker, LLMs (including GPT4) do not always comply, e.g., *She comes from what is, honestly, a horrific family situation* =>emotional abuse)

Category	Positive or Negative			Positive Only			Subcategory
	Precision	Recall	F1-Score	Precision	Recall	F1-Score	Accuracy
TRAUMA	0.44	0.89	0.59	0.57	0.84	0.68	0.57
PRECARITY	0.31	0.52	0.39	0.50	0.46	0.48	0.36
CONDITION	0.46	0.81	0.59	0.61	0.77	0.68	0.57
SYMPTOMS	0.35	0.78	0.49	0.45	0.73	0.56	0.41
SUICIDALITY	0.36	0.93	0.51	0.70	0.87	0.77	/
TREATMENT	0.39	0.87	0.54	0.64	0.81	0.71	/
ALL	0.39	0.80	0.52	0.55	0.75	0.63	0.48

Table 2: GPT4 (real data). **Positive or Negative**: counting annotation in category regardless of polarity (category level); **Positive Only**: counting negative annotations as NaN (category level); **Subcategory**: accuracy at the subcategory level

Category	Positive or Negative			Positive Only			Subcategory
	Precision	Recall	F1-Score	Precision	Recall	F1-Score	Accuracy
TRAUMA	0.90	0.49	0.64	0.90	0.49	0.64	0.38
PRECARITY	0.84	0.69	0.76	0.86	0.69	0.76	0.54
CONDITION	0.44	0.67	0.53	0.47	0.67	0.55	0.59
SYMPTOMS	0.85	0.59	0.70	0.84	0.59	0.69	0.36
SUICIDALITY	0.75	1.00	0.85	0.77	0.90	0.83	/
TREATMENT	0.68	0.84	0.75	0.76	0.57	0.65	/
ALL	0.74	0.65	0.70	0.77	0.61	0.68	0.42

Table 3: GPT3.5 (synthetic data). **Positive or Negative**: counting annotation in category regardless of polarity (category level); **Positive Only**: counting negative annotations as NaN (category level); **Subcategory**: accuracy at the subcategory level

Category	Positive or Negative			Positive Only			Subcategory
	Precision	Recall	F1-Score	Precision	Recall	F1-Score	Accuracy
TRAUMA	0.84	0.95	0.89	0.86	0.92	0.89	0.82
PRECARITY	0.85	0.84	0.85	0.91	0.82	0.86	0.80
CONDITION	0.61	0.67	0.64	0.60	0.67	0.63	0.67
SYMPTOMS	0.49	0.78	0.60	0.53	0.80	0.64	0.69
SUICIDALITY	0.81	0.94	0.87	0.78	0.82	0.80	/
TREATMENT	0.85	0.89	0.87	0.87	0.78	0.82	/
ALL	0.69	0.83	0.75	0.69	0.79	0.73	0.72

Table 4: GPT4 (synthetic data). **Positive or Negative**: counting annotation in category regardless of polarity (category level); **Positive Only**: counting negative annotations as NaN (category level); **Subcategory**: accuracy at the subcategory level

- Even GPT4 makes errors regarding negation (e.g., *I've read about people with autism getting temper tantrums/meltdowns, however, that has never really been a problem for me*=>negative autism or *i had in my head that something inside was very wrong, but i never felt completely depressed all the time so i never took bipolar seriously* =>negative bipolar disorder)
- Despite being prompted to annotate suicidality in a separate category, LLMs often annotate it in the SYMPTOM rather than SUICIDALITY category
- GPT3.5 especially often outputs irrelevant/spurious/incorrect labels (e.g., 'unemployed' as condition, 'ambition' as symptom, labelling physical conditions instead of mental conditions only, etc.)

	Positive and Negative	Positive Only	Subcategory
Annotator vs. Annotator (real data)	0.60	0.59	0.35
GPT3 vs. Annotator (real data)	0.39	0.52	0.37
GPT4 vs. Annotator (real data)	0.43	0.58	0.47
Annotator vs. Annotator (synthetic data)	0.77	0.71	0.68
GPT3 vs. Annotator (synthetic data)	0.64	0.63	0.40
GPT4 vs. Annotator (synthetic data)	0.70	0.69	0.71

Table 5: Inter-annotator agreement (Cohen's Kappa)

- Even GPT4 makes errors regarding factuality (e.g., *It was around my second year in junior high school when my father tried to take his life* =>positive death)

However, in many cases the assessment is not entirely fair, as the LLMs (particularly GPT4) often catch annotations which human annotators missed, or the difference in subcategories is subjective and open to debate (e.g., school bullying vs emotional abuse, emotional abuse vs abuse unspecified, etc.). Thus it is possible that LLMs, or most likely GPT4, in fact outperformed experts on this task.

6 Discussion

The results obtained from our comparison of LLM annotations with human annotations on both real and synthetic data allow us to make a few conclusions and recommendations.

Overall, both LLMs perform well. Inter-annotator agreement and performance indicate that GPT4 performs on par with human annotators. In fact, error analysis and manual examination of annotations suggest the LLMs potentially outperform human annotators in terms of recall (sensitivity), catching annotations which have been missed. However, while recall might be improved in LLMs versus human annotators, precision may suffer in unexpected ways, for example through errors in the use of negation and factuality, even in the case of GPT4. LLMs display a particular tendency to overpredict labels and produce negative annotations in infelicitous contexts, i.e., when humans would deem them irrelevant, creating an amount of noise. However, these negative annotations are not technically incorrect. While accuracy errors could be found in the LLM output, the experts' outputs were not entirely free of them, and previous work by [37] suggests LLMs may both be more complete AND more accurate than medical experts. There may still be a difference in the type of accuracy errors produced by LLMs, which will have to be investigated in future research.

In terms of accuracy at the subcategory level, we were surprised to find GPT4 outperformed human agreement by a large margin in real data (0.47 vs 0.35). We hypothesise this is due to the fact that human annotators display higher subjectivity in their style of annotation at the subcategory level (given the lack of predetermined subcategories) and diverge more between them. LLMs are likely to be more 'standard' and generic and thus potentially more in agreement with any given human annotator. More specifically, LLMs tend to be consistent from one annotation to the other with higher recall whereas human annotators showed less consistency. Therefore, if a sentence mentions physical, sexual and emotional abuse, annotators might only mention two out of three but when mentioning all three an LLM is more likely to be in agreement than another annotator, i.e., the LLM will catch more of the perfectly recalled annotations than the second annotator.

The better performance demonstrated on synthetic data doesn't seem due to LLMs performing better on data they are generating, but rather to the synthetic data being less complex and diverse and thus easier to annotate for both LLMs and humans, as evidenced by GPT4 reaching similar inter-annotator agreement scores to humans (with agreement both in humans and LLM/human 10% higher for synthetic data). This better performance could still warrant using synthetic data for e.g., training machine learning models (given more reliable labels) but only in cases where the potential loss in diversity is compensated by the increase in label reliability. This will likely depend on the specific application.

7 Conclusion

We presented the results of a study examining human and Large Language Models (GPT3.5 and GPT4) performance in extracting mental health factors from adolescent social media data. We performed analyses both on real and synthetic data and found GPT4 performance to be on par with human inter-annotator agreement for both datasets, with substantially better performance on the synthetic dataset. However, we find GPT4 still performing non-human errors in negation and factuality, and synthetic data to be much less diverse and differently distributed than real data. The potential for future applications in healthcare will have to be determined by weighing these factors against the substantial reductions in time and cost achieved through the use of LLMs.

Acknowledgment

I.L., D.W.J., and A.K. are partially supported by the National Institute for Health and Care Research (NIHR) AI Award grant (AI_AWARD02183) which explicitly examines the use of AI technology in mental health care provision. A.K. declare a research grant from GlaxoSmithKline (unrelated to this work). This research project is supported by the NIHR Oxford Health Biomedical Research Centre (grant NIHR203316). The views expressed are those of the authors and not necessarily those of the UK National Health Service, the NIHR or the UK Department of Health and Social Care.

References

[1] Michelle O'reilly. Social media and adolescent mental health: the good, the bad and the ugly. *Journal of Mental Health*, 29(2):200–206, 2020.

[2] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[3] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[4] Monica Agrawal, Stefan Hegselmann, Hunter Lang, Yoon Kim, and David Sontag. Large language models are few-shot clinical information extractors. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1998–2022, 2022.

[5] Yan Hu, Iqra Ameer, Xu Zuo, Xueqing Peng, Yujia Zhou, Zehan Li, Yiming Li, Jianfu Li, Xiaoqian Jiang, and Hua Xu. Zero-shot clinical entity recognition using chatgpt. *arXiv preprint arXiv:2303.16416*, 2023.

[6] Z Liu, X Yu, L Zhang, Z Wu, C Cao, H Dai, L Zhao, W Liu, D Shen, Q Li, et al. Deid-gpt: zero-shot medical text de-identification by gpt-4. arxiv. *arXiv*, 2023.

[7] Liyan Tang, Zhaoyi Sun, Betina Idnay, Jordan G Nestor, Ali Soroush, Pierre A Elias, Ziyang Xu, Ying Ding, Greg Durrett, Justin Rousseau, et al. Evaluating large language models on medical evidence summarization. medrxiv. 2023.

[8] Sarah Graham, Colin Depp, Ellen E Lee, Camille Nebeker, Xin Tu, Ho-Cheol Kim, and Dilip V Jeste. Artificial intelligence for mental health and mental illnesses: an overview. *Current psychiatry reports*, 21:1–18, 2019.

[9] Mirko Franco, Ombretta Gaggi, and Claudio E Palazzi. Analyzing the use of large language models for content moderation with chatgpt examples. In *Proceedings of the 3rd International Workshop on Open Challenges in Online Social Networks*, pages 1–8, 2023.

[10] Isabelle Lorge, Dan W Joyce, Niall Taylor, Alejo Nevado-Holgado, Andrea Cipriani, and Andrey Kormilitzin. Detecting the clinical features of difficult-to-treat depression using synthetic data from large language models. *arXiv preprint arXiv:2402.07645*, 2024.

[11] Kailai Yang, Shaoxiong Ji, Tianlin Zhang, Qianqian Xie, Ziyan Kuang, and Sophia Ananiadou. Towards interpretable mental health analysis with large language models. In *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023.

[12] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[13] Xuhai Xu, Bingsheng Yao, Yuanzhe Dong, Saadia Gabriel, Hong Yu, James Hendler, Marzyeh Ghassemi, Anind K Dey, and Dakuo Wang. Mental-llm: Leveraging large language models for mental health prediction via online text data. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 8(1):1–32, 2024.

[14] Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. Alpaca: A strong, replicable instruction-following model. *Stanford Center for Research on Foundation Models. https://crfm. stanford. edu/2023/03/13/alpaca. html*, 3(6):7, 2023.

[15] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*, 2022.

[16] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[17] Niall Taylor, Yi Zhang, Dan W Joyce, Ziming Gao, Andrey Kormilitzin, and Alejo Nevado-Holgado. Clinical prompt learning with frozen language models. *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

[18] Niall Taylor, Upamanyu Ghose, Omid Rohanian, Mohammadmahdi Nouriborji, Andrey Kormilitzin, David Clifton, and Alejo Nevado-Holgado. Efficiency at scale: Investigating the performance of diminutive language models in clinical tasks. *arXiv preprint arXiv:2402.10597*, 2024.

[19] Ramin Mojtabai, Mark Olfson, and Beth Han. National trends in the prevalence and treatment of depression in adolescents and young adults. *Pediatrics*, 138(6), 2016.

[20] Franziska Marcheselli, Ellie Brodie, Si Ning Yeoh, Nicola Pearce, Sally McManus, Katharine Sadler, Tim Vizard, Tamsin Ford, Anna Goodman, and Robert Goodman. Mental health of children and young people in england, 2017. *London: NHS*, 2018.

[21] Faye Mishna, Cheryl Regehr, Ashley Lacombe-Duncan, Joanne Daciuk, Gwendolyn Fearing, and Melissa Van Wert. Social media, cyber-aggression and student mental health on a university campus. *Journal of mental health*, 27(3):222–229, 2018.

[22] Amy Orben, Livia Tomova, and Sarah-Jayne Blakemore. The effects of social deprivation on adolescent development and mental health. *The Lancet Child & Adolescent Health*, 4(8): 634–640, 2020.

[23] Jean M Twenge and Thomas E Joiner. Us census bureau-assessed prevalence of anxiety and depressive symptoms in 2019 and during the 2020 covid-19 pandemic. *Depression and anxiety*, 37(10):954–956, 2020.

[24] Ronald C Kessler, Matthias Angermeyer, James C Anthony, RON De Graaf, Koen Demyttenaere, Isabelle Gasquet, Giovanni De Girolamo, Semyon Gluzman, OYE Gureje, Josep Maria Haro, et al. Lifetime prevalence and age-of-onset distributions of mental disorders in the world health organization's world mental health survey initiative. *World psychiatry*, 6(3):168, 2007.

[25] Kessler Rc. Lifetime prevalence and age-of-onset distributeions of dsm-iv disorders in the national comorbidity survey replication. *Arch Gen Psychiatry*, 62:593–602, 2005.

[26] Kelly A Allen, Tracii Ryan, DeLeon L Gray, Dennis M McInerney, and Lea Waters. Social media use and social connectedness in adolescents: The positives and the potential pitfalls. *The Educational and Developmental Psychologist*, 31(1):18–31, 2014.

[27] Madeleine J George and Candice L Odgers. Seven fears and the science of how mobile technologies may be influencing adolescents in the digital age. *Perspectives on psychological science*, 10(6):832–851, 2015.

[28] Minas Michikyan. Depression symptoms and negative online disclosure among young adults in college: A mixed-methods approach. *Journal of Mental Health*, 29(4):392–400, 2020.

[29] Umberto Volpe, Alfonso Tortorella, Mirko Manchia, Alessio M Monteleone, Umberto Albert, and Palmiero Monteleone. Eating disorders: What age at onset? *Psychiatry research*, 238: 225–227, 2016.

[30] Eve Griffin, Elaine McMahon, Fiona McNicholas, Paul Corcoran, Ivan J Perry, and Ella Arensman. Increasing rates of self-harm among children, adolescents and young adults: a 10-year national registry study 2007–2016. *Social psychiatry and psychiatric epidemiology*, 53: 663–671, 2018.

[31] Mauro Giuffrè and Dennis L Shung. Harnessing the power of synthetic data in healthcare: innovation, application, and privacy. *NPJ Digital Medicine*, 6(1):186, 2023.

[32] Robert Chew, Caroline Kery, Laura Baum, Thomas Bukowski, Annice Kim, Mario Navarro, et al. Predicting age groups of reddit users based on posting behavior and metadata: classification model development and validation. *JMIR Public Health and Surveillance*, 7(3):e25807, 2021.

[33] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019. URL `https://arxiv.org/abs/1908.10084`.

[34] Charles A Nelson, Zulfiqar A Bhutta, Nadine Burke Harris, Andrea Danese, and Muthanna Samara. Adversity in childhood is linked to mental and physical health throughout life. *bmj*, 371, 2020.

[35] Emla Fitzsimons, Alissa Goodman, Elaine Kelly, and James P Smith. Poverty dynamics and parental mental health: Determinants of childhood mental health in the uk. *Social Science & Medicine*, 175:43–51, 2017.

[36] Allyson Ettinger. What bert is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *Transactions of the Association for Computational Linguistics*, 8:34–48, 2020.

[37] Dave Van Veen, Cara Van Uden, Louis Blankemeier, Jean-Benoit Delbrouck, Asad Aali, Christian Bluethgen, Anuj Pareek, Malgorzata Polacin, Eduardo Pontes Reis, Anna Seehofnerová, Nidhi Rohatgi, Poonam Hosamani, William Collins, Neera Ahuja, Curtis P. Langlotz, Jason Hom, Sergios Gatidis, John Pauly, and Akshay S. Chaudhari. Adapted large language models can outperform medical experts in clinical text summarization. *Nature Medicine*, 30(4): 1134–1142, February 2024. ISSN 1546-170X. doi: 10.1038/s41591-024-02855-5. URL `http://dx.doi.org/10.1038/s41591-024-02855-5`.

A Appendix A

Write five Reddit posts from adolescents in subreddits related to mental health and annotate each sentence with the following labels:

- TRAUMA (sexual abuse, physical abuse, emotional abuse, school bullying, death, accident)

- PRECARITY (socioeconomic, parental conflict, parental illness)

- SYMPTOM (self-harm, low self-esteem, anhedonia, panic attack, flashback, psychosis, insomnia)

- CONDITION (eating disorder, depression, bipolar, anxiety, ptsd, adhd, substance abuse/addiction)

- SUICIDALITY

- TREATMENT

Do not specify the subreddit. Annotate both presence (POSITIVE) and absence (NEGATIVE) of a factor. Here are some examples:

- “I am not feeling suicidal but I can’t sleep at all [SYMPTOM:POSITIVE(insomnia), SUICIDALITY:NEGATIVE]."

- “My sister used to constantly bully me [TRAUMA:POSITIVE(emotional abuse)]."

- “I was harassed for years in secondary school [TRAUMA:POSITIVE(school bullying)] - “I went on holiday last month [NONE]."

- "My family is quite wealthy [PRECARITY:NEGATIVE(socioeconomic)]."

- "I often cut my wrists with scissors [SYMPTOM:POSITIVE(self-harm)]"

Make sure you always add POSITIVE or NEGATIVE to the factor and specify the subcategory for all factors apart from TREATMENT and SUICIDALITY.

For TREATEMENT and SUICIDALITY you only specify presence or absence, e.g. [TREATMENT:POSITIVE] or [SUICIDALITY:NEGATIVE].

Each post should be 5 sentences in length.

B Appendix B

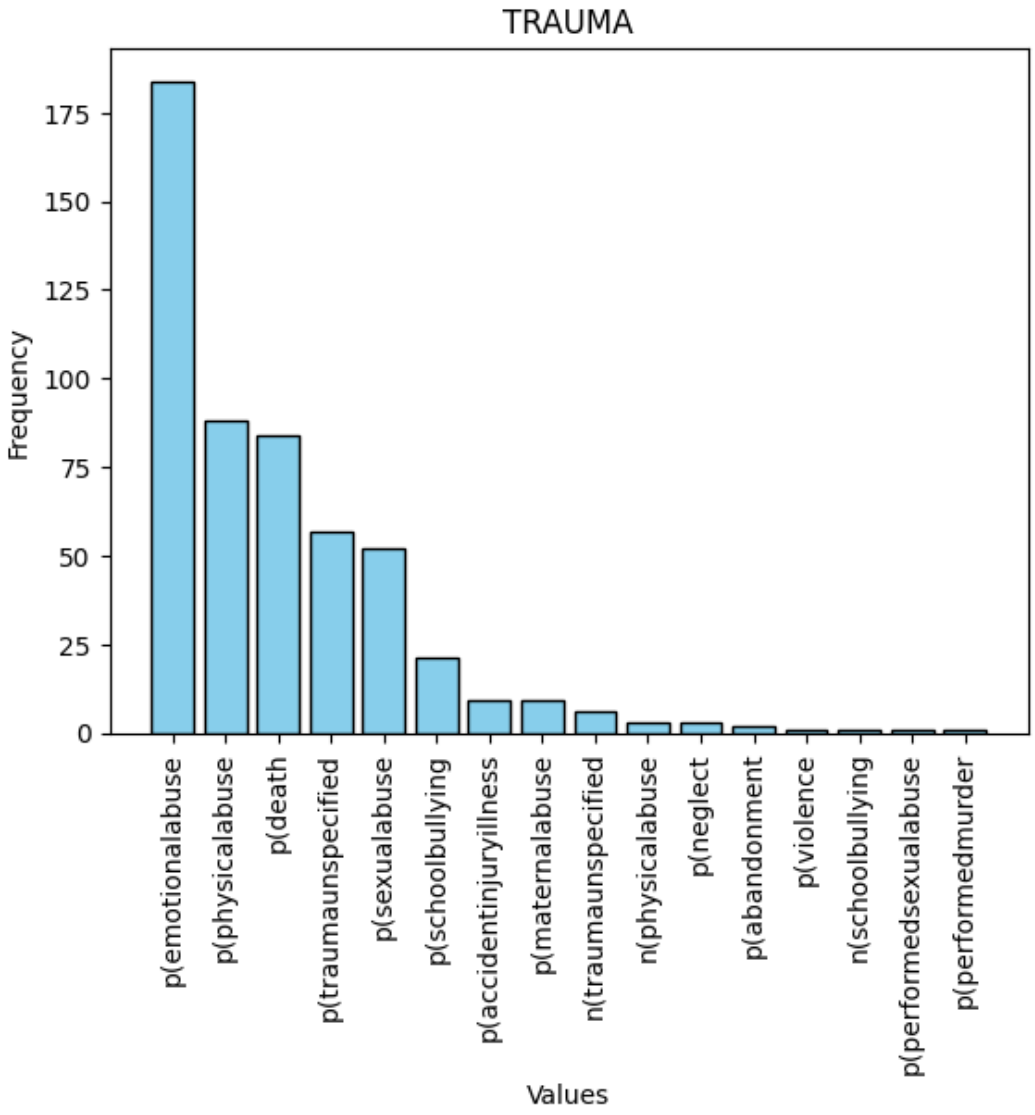


Figure 2: Distribution of trauma subcategories (real data)

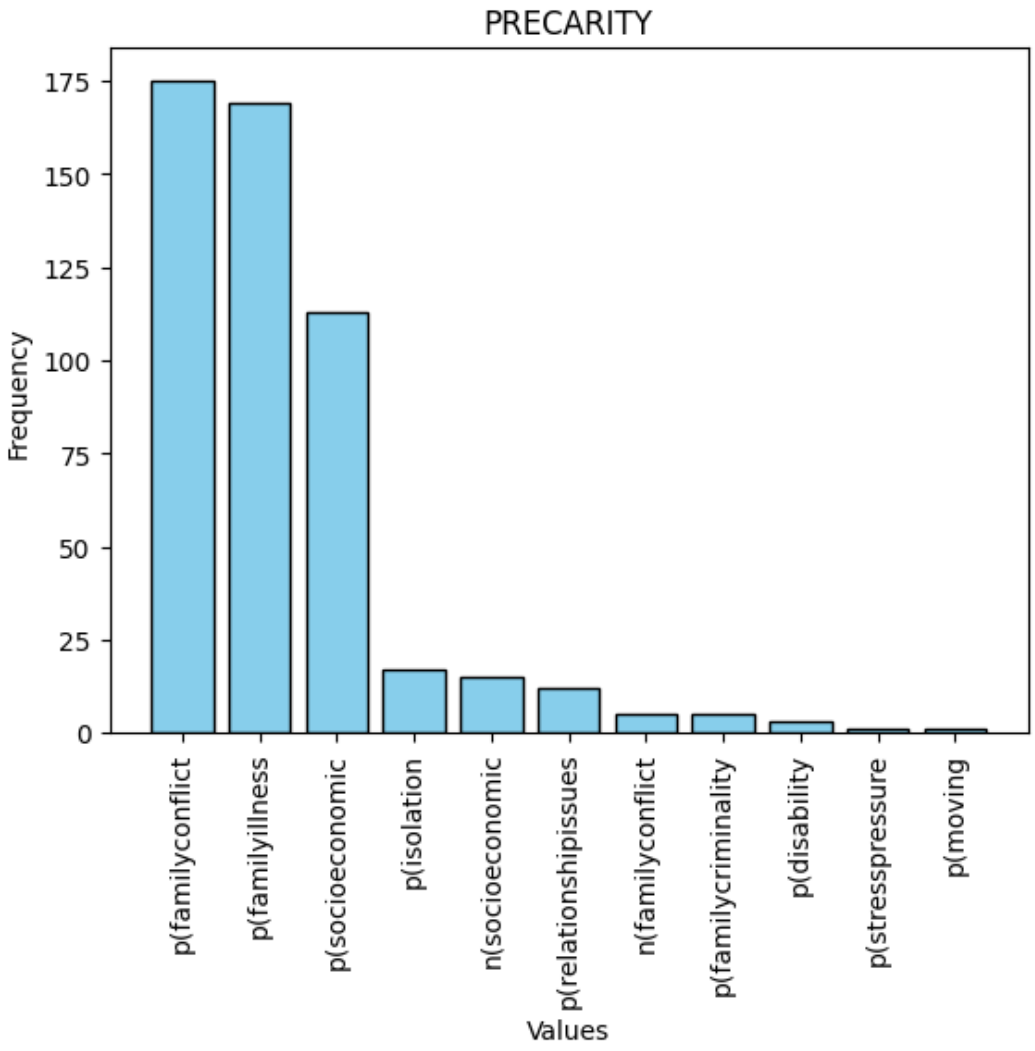


Figure 3: Distribution of precarity subcategories (real data)

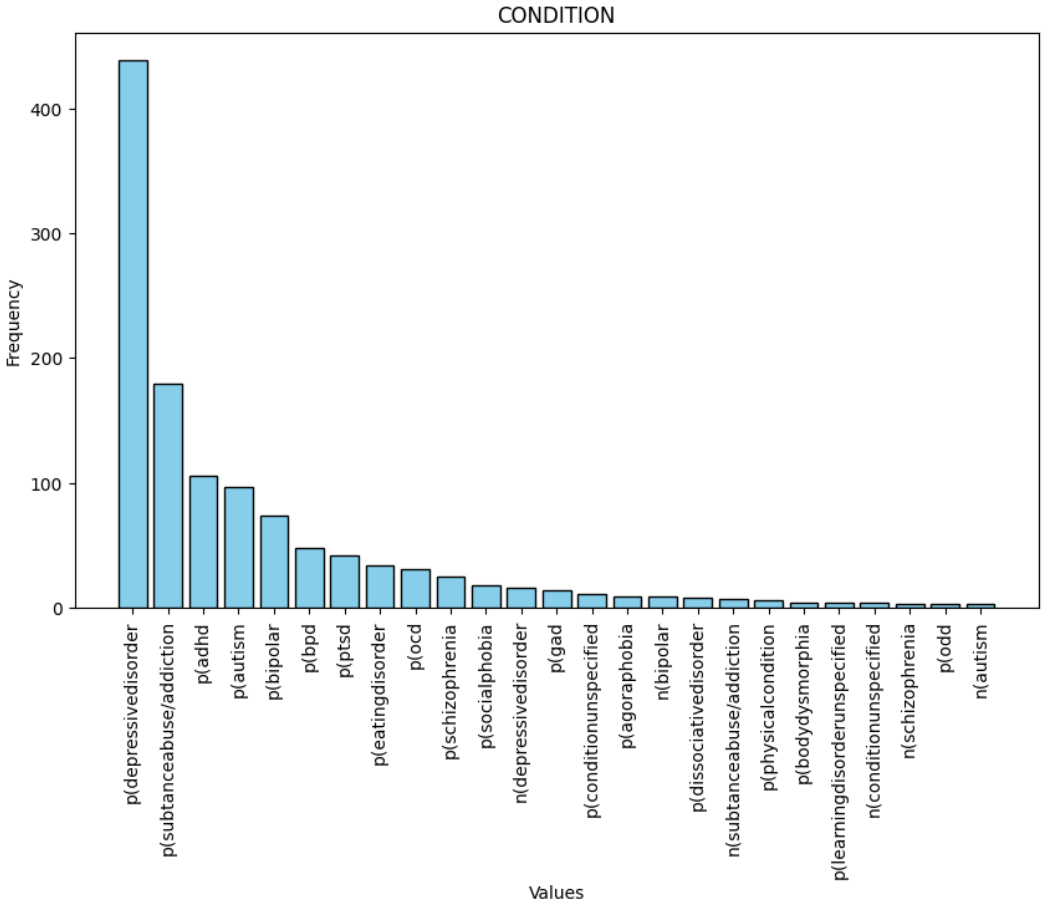


Figure 4: Distribution of condition subcategories (real data) (subcategories with n >2)

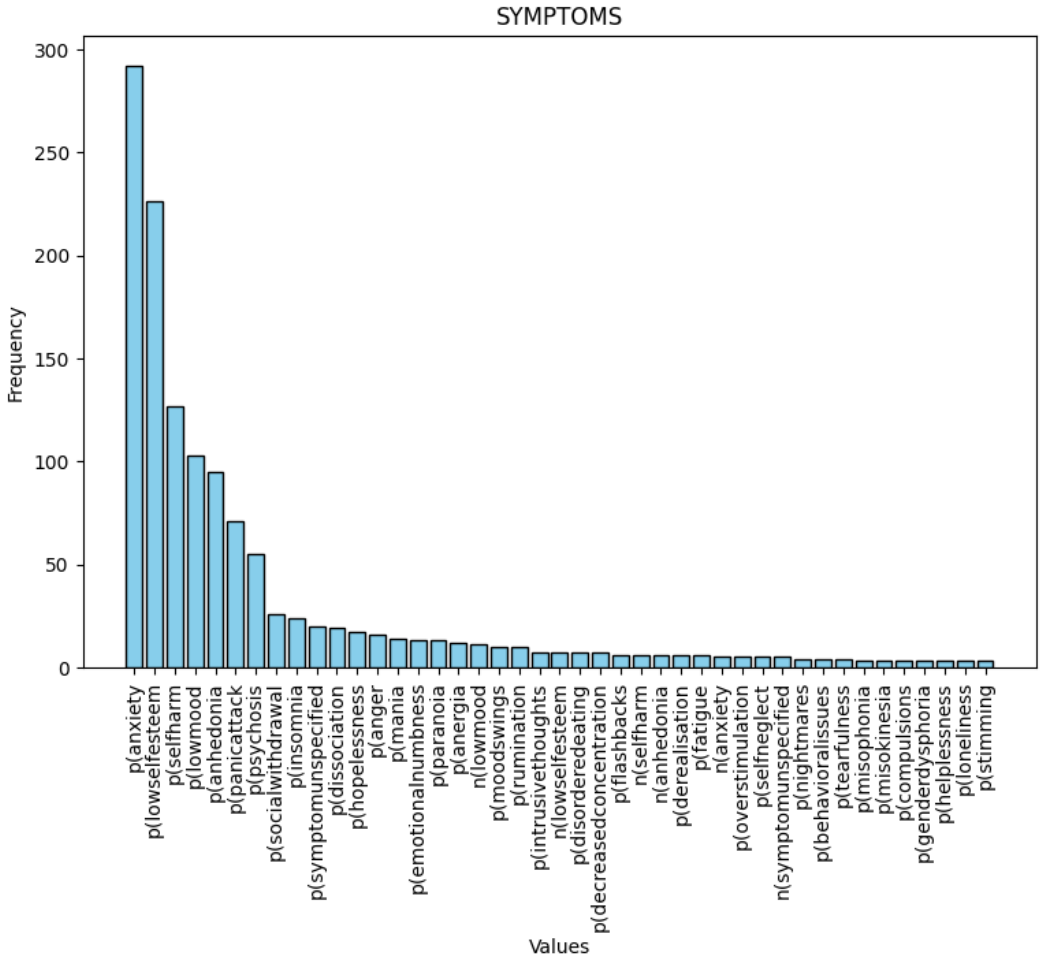
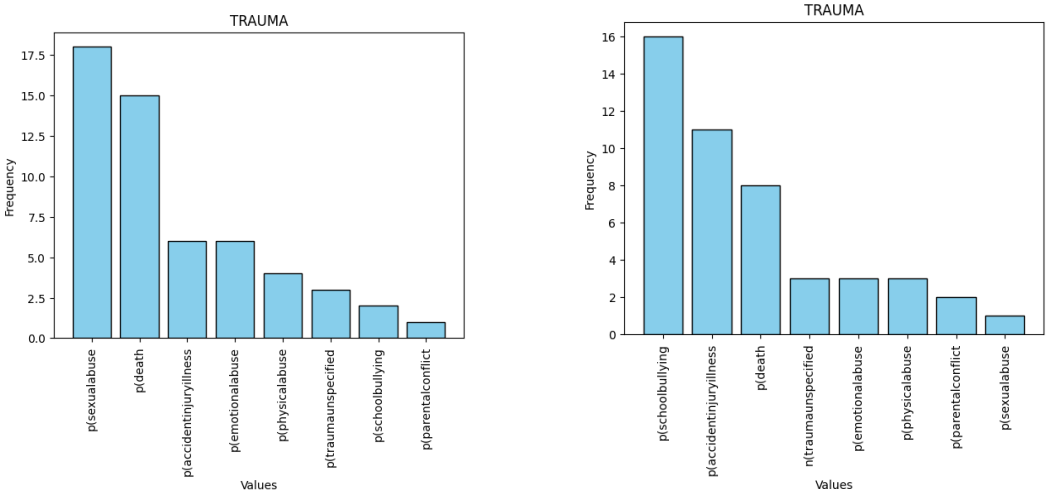
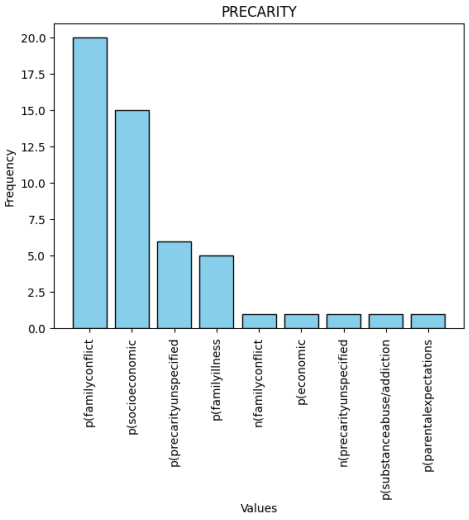


Figure 5: Distribution of symptom subcategories (real data) (subcategories with n >2)

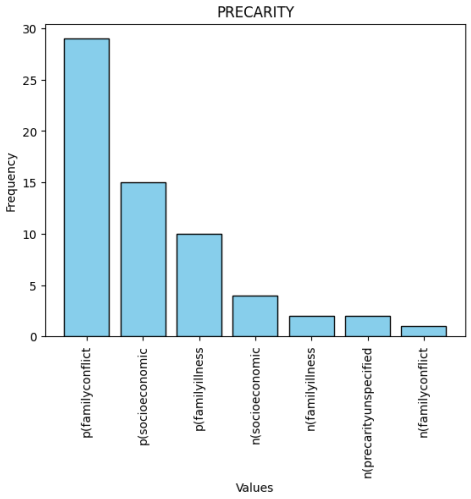


(a) GPT3.5 (b) GPT4

Figure 6: Distribution of trauma subcategories (synthetic data)

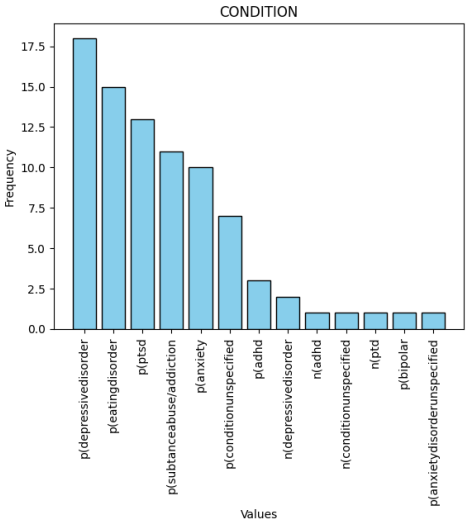


(a) GPT3.5

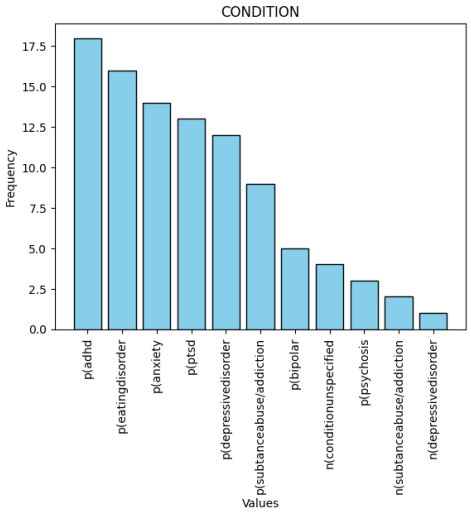


(b) GPT4

Figure 7: Distribution of precarity subcategories (synthetic data)

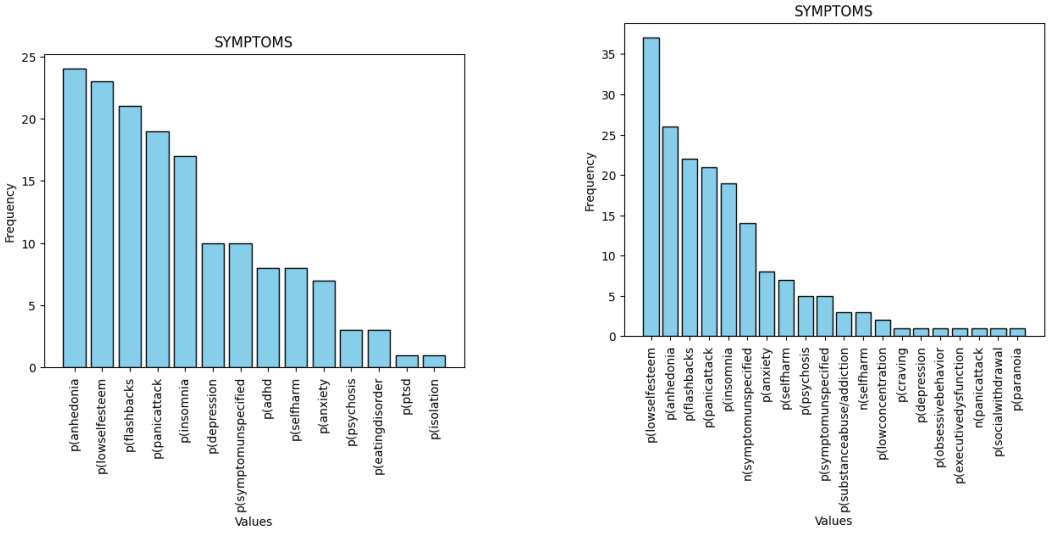


(a) GPT3.5



(b) GPT4

Figure 8: Distribution of condition subcategories (synthetic data)



(a) GPT3.5 (b) GPT4

Figure 9: Distribution of symptom subcategories (synthetic data)