Argumentative Large Language Models for Explainable and Contestable Decision-Making

Gabriel Freedman* , Adam Dejl* , Deniz Gorur* , Xiang Yin , Antonio Rago , Francesca Toni
Department of Computing, Imperial College London, UK
{gif22, adam.dejl18, d.gorur22, xy620, a.rago, ft}@imperial.ac.uk

Abstract

The diversity of knowledge encoded in large language models (LLMs) and their ability to apply this knowledge zero-shot in a range of settings makes them a promising candidate for use in decision-making. However, they are currently limited by their inability to reliably provide outputs which are explainable and contestable. In this paper, we attempt to reconcile these strengths and weaknesses by introducing a method for supplementing LLMs with argumentative reasoning. Concretely, we introduce *argumentative LLMs*, a method utilising LLMs to construct argumentation frameworks, which then serve as the basis for formal reasoning in decision-making. The interpretable nature of these argumentation frameworks and formal reasoning means that any decision made by the supplemented LLM may be naturally explained to, and contested by, humans. We demonstrate the effectiveness of argumentative LLMs experimentally in the decision-making task of claim verification. We obtain results that are competitive with, and in some cases surpass, comparable state-of-the-art techniques.

1 Introduction

Large language models (LLMs) have produced excellent results on a diverse range of reasoning tasks Brown *et al.* [2020]; Bubeck *et al.* [2023]. This capacity has made them compelling candidates for supporting automated decision systems Zhang *et al.* [2023]; Ouyang and Li [2023]; Wang *et al.* [2023]. However, their reasoning abilities currently suffer from various limitations, e.g. hallucinations and logical inconsistencies Shanahan [2024]; Berglund *et al.* [2023]; Fluri *et al.* [2023]. Deficiencies which are particularly worrying are a lack of explainability and inability to provide faithful representations of their reasoning, which raise questions regarding their trustworthiness and ability to be contested Henin and Métayer [2021]; Lyons *et al.* [2021].

In this paper we explore the following question:

> Can the reasoning abilities of LLMs improve if they are made to argue with themselves?

The question is inspired by argumentative interpretations of human reasoning Mercier and Sperber [2011, 2018] and by the fact that argumentation has been shown to excel in supporting decision making Amgoud and Prade [2009]. We take a broader than usual view of what counts as an 'improved' ability to reason. In addition to demonstrating that our *argumentative LLMs* achieve competitive scores on various reasoning benchmarks, we show how an improved ability to reason necessarily leads to more explainable and contestable decision-making Liao and Vaughan [2024].

Previous methods for improving the reasoning of LLMs do not necessitate a direct relationship between the reasoning steps and the final decision. Our argumentative approach, on the other hand, provides this as a feature of the system. This is because the system prediction is directly derived from the generated argumentation framework using a formally defined and deterministic procedure, thus providing faithful explainability. Further, argumentative LLMs also provide a guarantee of contestability, in that if a human intervenes in the reasoning process (such as by adding or removing an argument, or changing the strength of an argument), this will have a measurable effect on the output of the decision-making system. Comparable techniques lack the necessary processing stage between the LLM's output and the final decision to accommodate this flexibility.

Rather than prompting an LLM to produce 'thoughts', as in Wei *et al.* [2022] or Yao *et al.* [2023], that either enrich the context of the LLM, or provide disparate reasoning steps to compare, our approach can be seen as providing 'thoughts' for and against particular outputs, in the spirit of Miller [2023]. This makes it a natural fit for highly complex decision-making tasks, wherein an option, or set of options, must be chosen from a number of possible alternatives. In almost all real-world settings, a particular decision will have both pros and cons, which is a feature that argumentative LLMs both formalise and leverage.

In this paper, we focus on the kind of reasoning underpinning claim verification. This setting lends itself well to our framework, as claims are often under-determined, so they do not necessarily have straightforward truth values. By intrinsically considering both arguments in favour of and in conflict with the truthfulness of claims, argumentative LLMs are able to ascertain the best answer given the available evidence. For

[*] Equal contribution.

simplicity, and without loss of generality[1], we focus on a binary setting, rather than general question-answer problems as in Wei *et al.* [2022] and Yao *et al.* [2023]. In order to handle open-ended settings, it is first necessary to generate candidate answers — determining the optimal number of answers can be thought of as a hyperparameter.

In summary, we make the following contributions:

- We define argumentative LLMs, a novel method for supplementing LLMs with formal reasoning for decision-making;
- We perform an extensive evaluation of our method by comparing four variants thereof with three baselines, on three claim verification datasets, adapted from existing datasets;
- We demonstrate the explainability and contestability benefits of argumentative LLMs.

2 Related Work

A significant amount of research has been focused on improving the reasoning abilities of LLMs. This can be coarsely divided into approaches which exclusively focus on prompt optimisation Wei *et al.* [2022]; Yang *et al.* [2023], and those which endow LLMs with the ability to utilise external tools or information, or extra structural constraints Schick *et al.* [2023]; Yao *et al.* [2023]; Lewis *et al.* [2020]. Our system is more closely aligned to the latter, as it results in symbolic, deterministically evaluable graphs as its output.

Du *et al.* [2023] also use arguments to improve the reasoning ability of LLMs. However, they focus on a multi-agent setting and do not formalise the arguments produced by LLMs, or their corresponding strengths. Also at the intersection of LLMs and argumentation is work looking at the efficacy of LLMs at completing arguments Thorburn and Kruger [2022], and the persuasiveness of LLM generated arguments Hinton and Wagemans [2023]; Durmus *et al.* [2024].

Chain-of-thought approaches Wei *et al.* [2022]; Zhang and Parkes [2023] attempt to induce enhanced reasoning through a specific form of prompting. The prompt specifies (either using few-shot examples, or using a verbal description) that the problem should be broken down into discrete steps, before the final decision is outputted. However, all the reasoning takes place within the autoregressively generated output of the model. Due to the nature of the next token prediction mechanism underlying these models, this does not guarantee that the steps in the reasoning, or the final output actually follow from each other Xia *et al.* [2024]. This undermines the premise that the reasoning is faithful to the process taking place in the model or that it is directly related to the final output. Our method avoids this pitfall by building an argumentation graph which guarantees a resolution based on the constituent entities.

Other related approaches are *tree-of-thought* Yao *et al.* [2023] and *graph-of-thoughts* Besta *et al.* [2024]. Similarly to our methodology, these approaches result in graph-like structures, composed of the LLM's output, which can then be reasoned over post-hoc. In contrast to our method, the nodes of these graphs consist of decomposed components of the overall problem. Instead, our method permits a comprehensive and fully explainable reasoning process to take place concerning a single claim, which may be controversial or highly complex, in addition to composite problems like the existing methods. For example, a claim such as 'it is a good idea to drink milk when you have a cough', does not naturally lend itself to decomposition, but would benefit from argumentative reasoning, i.e., evaluating arguments for (e.g., 'it is a traditional remedy') and against (e.g., 'there have been no scientific studies confirming this').

3 Preliminaries

Claim Verification The considered task of claim verification can be divided into two primary types: *unconditioned* and *conditioned*. For *unconditioned* verification, a claim c is evaluated independently, without any contextual information. The outcome of this evaluation is binary, represented as $v(c) \in \{0, 1\}$, where $v(c) = 1$ denotes the claim is true, and $v(c) = 0$ indicates the claim is false. Meanwhile, *conditioned* verification considers a claim c given additional information or context i (with the context assumed to be truthful). The veracity of this tuple is also assessed in a binary manner, expressed as $v(c \mid i) \in \{0, 1\}$. Here, similarly, $v(c \mid i) = 1$ signifies that the claim c, given the context i, is true, while $v(c \mid i) = 0$ means it is false.

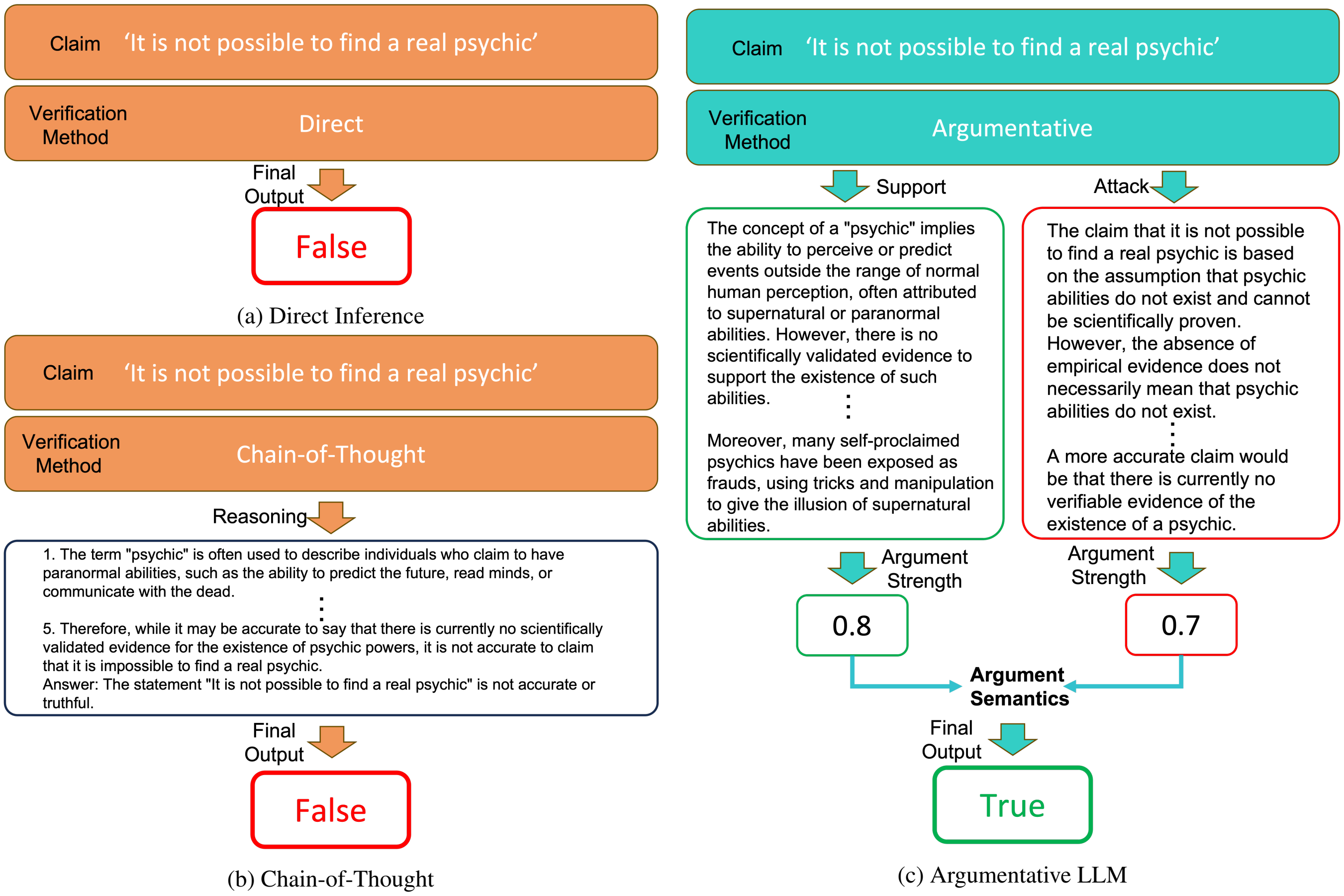
Computational Argumentation We will now cover the relevant notions from this AI discipline (see Atkinson *et al.* [2017]; Baroni *et al.* [2018a] for overviews), on which our methodology leverages. A *quantitative bipolar argumentation framework* (QBAF) Baroni *et al.* [2019] is a quadruple $\langle \mathcal{X}, \mathcal{A}, \mathcal{S}, \tau \rangle$ comprising a set of *arguments* $\mathcal{X}$, binary, directed relations of *attack* $\mathcal{A} \subseteq \mathcal{X} \times \mathcal{X}$ and *support* $\mathcal{S} \subseteq \mathcal{X} \times \mathcal{X}$, where $\mathcal{A} \cap \mathcal{S} = \emptyset$, and a total function $\tau : \mathcal{X} \to [0, 1]$, where for any $x \in \mathcal{X}$, $\tau(x)$ is the *base score* of x.[2] For any argument $x \in \mathcal{X}$, we use $\mathcal{A}(x) = \{y \in \mathcal{X} | (y, x) \in \mathcal{A}\}$ to refer to the *attackers* of x and $\mathcal{S}(x) = \{y \in \mathcal{X} | (y, x) \in \mathcal{S}\}$ to refer to the *supporters* of x. Arguments in QBAFs may be evaluated by a *gradual semantics* Baroni *et al.* [2019], i.e. a total function $\sigma : \mathcal{X} \to [0, 1]$ which, for any $x \in \mathcal{X}$, assigns a *strength* $\sigma(x)$ to x.[3]

One such gradual semantics, the *discontinuity-free quantitative argumentation debate* (DF-QuAD) algorithm Rago *et al.* [2016], is such that, for a given QBAF $\langle \mathcal{X}, \mathcal{A}, \mathcal{S}, \tau \rangle$, for any $x \in \mathcal{X}$ with $n \geq 0$ attackers with strengths $v_1, \ldots, v_n$, $m \geq 0$ supporters with strengths $v'_1, \ldots, v'_m$ and $\tau(x) = v_0$, $\sigma(x) = \mathcal{C}(v_0, \mathcal{F}(v_1, \ldots, v_n), \mathcal{F}(v'_1, \ldots, v'_m))$, where $\mathcal{C}$ is defined as follows. For $v_a = \mathcal{F}(v_1, \ldots, v_n)$ and $v_s = \mathcal{F}(v'_1, \ldots, v'_m)$: if $v_a = v_s$ then $\mathcal{C}(v_0, v_a, v_s) = v_0$; else if $v_a > v_s$ then $\mathcal{C}(v_0, v_a, v_s) = v_0 - (v_0 \cdot |v_s - v_a|)$; otherwise $\mathcal{C}(v_0, v_a, v_s) = v_0 + ((1 - v_0) \cdot |v_s - v_a|)$. Given n arguments with strengths $v_1, ..., v_n$, if $n = 0$ then $\mathcal{F}(v_1, ..., v_n) = 0$, otherwise $\mathcal{F}(v_1, ..., v_n) = 1 - \prod_{i=1}^{n}(|1 - v_i|)$.

[1] The framework is easily extended to the case where there are more than two options to decide between.

[2] Note that the codomain of the base score is defined more generally by Baroni *et al.* [2019] but we restrict to its most common form.

[3] As with the base score, we use the most commonly occurring codomain of gradual semantics.


Claim ‘It is not possible to find a real psychic’
Verification Method
Direct
Final Output
False

(a) Direct Inference

Claim ‘It is not possible to find a real psychic’
Verification Method
Chain-of-Thought
Reasoning

1. The term "psychic" is often used to describe individuals who claim to have paranormal abilities, such as the ability to predict the future, read minds, or communicate with the dead.

⋮

5. Therefore, while it may be accurate to say that there is currently no scientifically validated evidence for the existence of psychic powers, it is not accurate to claim that it is impossible to find a real psychic.
Answer: The statement "It is not possible to find a real psychic" is not accurate or truthful.

(b) Chain-of-Thought

The concept of a "psychic" implies the ability to perceive or predict events outside the range of normal human perception, often attributed to supernatural or paranormal abilities. However, there is no scientifically validated evidence to support the existence of such abilities.

⋮

Moreover, many self-proclaimed psychics have been exposed as frauds, using tricks and manipulation to give the illusion of supernatural abilities.

The claim that it is not possible to find a real psychic is based on the assumption that psychic abilities do not exist and cannot be scientifically proven. However, the absence of empirical evidence does not necessarily mean that psychic abilities do not exist.

⋮

A more accurate claim would be that there is currently no verifiable evidence of the existence of a psychic.

(c) Argumentative LLM

Figure 1: Comparison of our approach (*Argumentative LLM*) with existing alternatives. The example claim is adapted from TruthfulQA (TruthfulClaim) and the abridged outputs are generated by Mixtral.

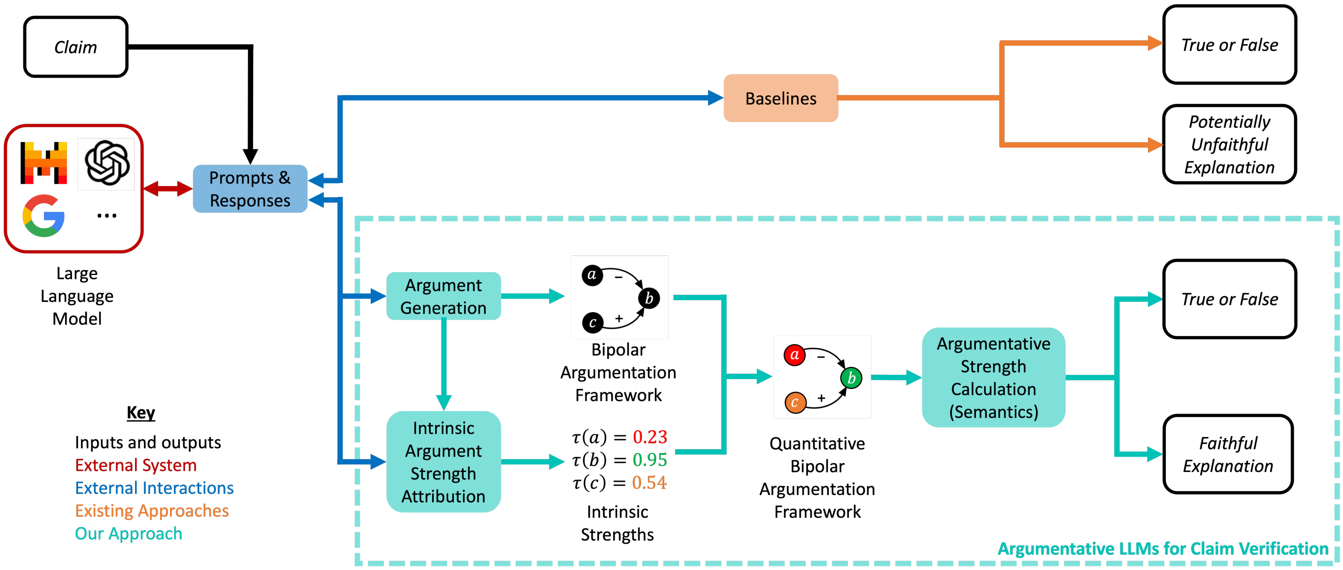


Figure 2: Pipeline for Argumentative LLMs (in comparison with baselines).

In addition to DF-QuAD, we will also use the *quadratic energy model* (QEM) semantics Potyka [2018], which is defined as follows. For a given QBAF $\langle \mathcal{X}, \mathcal{A}, \mathcal{S}, \tau \rangle$, the energy at $x \in \mathcal{X}$ is defined as $E_x = \sum_{y \in \mathcal{S}(x)} \sigma(y) - \sum_{z \in \mathcal{A}(x)} \sigma(z)$. Then, for all $v \in \mathbb{R}$, we let $h(v) = \frac{max\{v,0\}^2}{1+max\{v,0\}^2}$. Finally, the strength of an argument $x \in \mathcal{X}$ is defined as $\sigma(x) = \tau(x) + (1 - \tau(x)) \cdot h(E_x) - \tau(x) \cdot h(-E_x)$.[4]

4 Argumentative LLMs

In this section, we introduce our framework for inducing argumentative reasoning in LLMs. As indicated in the pipeline shown in Figure 2, there are three integral components of our framework: argument generation, argument strength attribution and argument semantics.

4.1 Argument Generation

Previous work has demonstrated that LLMs are able to effectively generate counter-arguments given a preexisting argument Chen *et al.* [2023]; Furman *et al.* [2023]. Leveraging this capability, we use LLMs to perform an extension of this task, where they produce arguments supporting and attacking a 'root argument'. In our cause the root argument is a (domain-agnostic) claim. We derive these claims from existing QA datasets (for an example, see Figure 1). The LLMs are fed a prompt, with the claim included, on at least two separate occasions: once prompted to generate an argument supporting the root argument, and once an argument attacking it. We refer to these as the base support and attack arguments.

Our first setting, which we refer to as having a *depth* of 1, contains only the base support and base attack arguments. When we include an additional layer of arguments (giving a depth of 2), we generate a further four arguments. These are a supporting and an attacking argument for each of the base support and base attack argument. The prompt used for generating these includes the base support or base attack argument respectively.[5]

4.2 Intrinsic Argument Strength Attribution

There have been a number of previous attempts to assess the quality, or intrinsic strength, of arguments. Note that this intrinsic strength (also called a base score) gives an argument's quality per se, i.e. before the rest of the argumentation framework is considered. These attempts have either used pairwise comparison between arguments Habernal and Gurevych [2016]; Simpson and Gurevych [2018], or human-annotated arguments Lauscher *et al.* [2020] to produce argument quality scores. However, producing such data is highly resource intensive. Instead we rely on the knowledge embedded into the LLMs, using them to attribute strengths to the arguments, zero-shot and without any task-specific finetuning. There have been some analogous uses of LLMs, such as for forecasting Halawi *et al.* [2024], where the models are used to assign numerical confidences to their outputs (within the same context window).

Incidentally, our present study can be seen as assessing if this is an 'emergent' capability of current LLMs (it is unlikely that either the pretraining or the supervised training stages contained many instances of this fairly niche task). Assigning a strength to an argument is quite subjective, and so a direct comparison between human and machine ratings may not be an ideal analysis (as it is highly likely that there would be a large variation between human scores for an individual argument). Therefore, using the scores for an objective-driven, empirical task, and ascertaining their suitability post-hoc, is perhaps a more effective method of assessing this capacity in LLMs.

In our case, the argument strengths are elicited by recursively prompting the LLMs with the arguments they have previously generated (in a separate context window). In order to capture the relative strength of the argument (relative to what it is attacking or supporting), we also include the root claim, or the base support or attack argument, in the context window.

4.3 Argument Strength Calculation

Once argumentation frameworks have been constructed, the arguments can be evaluated by means of an argument semantics. The choice of any argument semantics is dictated by the requirements of the application setting in which the argumentation frameworks are being deployed. We use gradual semantics Baroni *et al.* [2018b] that evaluate arguments quantitatively rather than extension-based semantics Dung [1995] that select sets of jointy-acceptable arguments, since only gradual semantics are applicable to arguments with continuous intrinsic strengths used in our framework. We chose the DF-QuAD semantics (as defined in Section 3) due to the dialectical properties it satisfies (see Baroni *et al.* [2019]), e.g. monotonicity, requiring that any attacker (supporter) can only decrease (increase, respectively) the strength of the argument it attacks (supports, respectively). We also experimented with the QEM semantics (defined in Section 3), given that it is shown to satisfy suitable properties. However, when tested on the validation data, there was a negligible difference between the performance of the two semantics.

4.4 Prompt Selection

We take a principled approach for prompt selection, as it has been shown that the result of slight variations in prompting on downstream task performance can be significant Santu and Feng [2023]. To reduce the impact of prompt choice on our final evaluation, we independently devise three different prompts for both our framework and the baselines (see Appendix C for details).

We evaluate all prompts, and combinations thereof, on two validation sets of 200 samples each, taken from TruthfulQA Lin *et al.* [2021] and StrategyQA Geva *et al.* [2021]. In this evaluation, we separately considered the prompts for the baselines (direct inference, estimated confidence and chain-of-thought) as well as the components of the argumentative approach (argument generation and argument strength attribution).

[4] We describe a simplification of the original algorithm for the case of trees, rather than (potentially cyclic) graphs.

[5] Note that, while our preliminary experiments are with these two depths, argumentation frameworks of any (computationally feasible) depth could be achieved with our method.

We find a large variation in performance for a particular prompt with any given dataset and model combination, both for the argumentative approach and the baselines. We choose the highest average scoring prompt over all tested models and datasets (as shown in Tables 2 and 3 in Appendix E).

5 Experimental Set-up

In this section we describe the baselines we compare against, the datasets used, the LLMs we experiment with, and the variations of our method.[6]

5.1 Baselines

We compare our method with three baselines.

Baseline 1: Direct questioning (Direct Question in short**)** We directly ask the LLMs if the given claim is true or false by prompting. The prompt used for this baseline is given in Appendix A.1. We constrained the output of the open-source models to true/false.

Baseline 2: Estimated confidence (Est. Confidence in short**)** We ask the LLMs for a confidence score on the given claim. The confidence score ranges from 0-100. The prompt used for this baseline is given in Appendix A.2. We constrain the output to values ranging from 0-100 for the open-source models in this baseline. Then to get a final decision (i.e. true/false) for the claim we check whether the outputted confidence is greater than 50. If it is greater than 50 then the claim is considered true, otherwise the claim is deemed false.

Baseline 3: Questioning with chain-of-thought (Chain-of-Thought in short**)** As our third baseline, we use (two-stage) chain-of-thought prompting Wei *et al.* [2022]. This prompt-based technique breaks down the problem into discrete steps before the final decision is outputted. Then, we pass the reasoning step back to the LLM, in a separate context window, to get the final decision. Both prompts are given in Appendix A.3.

5.2 Datasets

We focused on three datasets, which are adaptations of three existing Q/A datasets, turning Q/A pairs into claims with true/false labels. We did not use the datasets directly as the LLMs we experiment with did not perform adequately when generating arguments about the validity of the answers, rather than the questions.

Therefore, we have generated claims for the Q/A pairs. Firstly, we used LLMs for each dataset to automatically generate claims by prompting the LLM with the Q/A pair. Then, we manually checked all the claims with the Q/A pairs that we used, to see if the claim was faithful both to the question and to the answer. More details are given for each dataset below.

TruthfulClaim (adapted from TruthfulQA) TruthfulQA Lin *et al.* [2021] is a dataset curated specifically to evaluate if LLMs are able to generate truthful answers without being deceived by common misconceptions and falsehoods. The original dataset contains questions with a list of correct answers and a list of incorrect answers for each question. We transformed each answer, from the list of correct/incorrect answers with their corresponding question, to a claim generated by the process described above. We labelled the generated claims as *True* if the answer was from the correct answers list and *False* if the answer was from the incorrect answers list.

StrategyClaim (adapted from StrategyQA) StrategyQA Geva *et al.* [2021] is a dataset designed to evaluate whether LLMs can strategically reason. The original dataset is made up of binary questions and their labels as true/false. However, in this paper, we are focusing on claim verification and so we generated claims that are the affirmative answer to the question, once again using LLMs. The claims generated by the LLMs sometimes generated claims that were the negation of the question, so we manually modified those claims.

MedClaim (adapted from MedQA) MedQA Jin *et al.* [2020] is a multi-choice Q/A dataset for solving medical problems which is collected from the professional medical board exams. The MedQA dataset is slightly different from the previous two datasets as the questions are based on some contextual information. Therefore, the task we consider for the MedQA dataset becomes *conditioned claim verification*. The original dataset contains (composite) questions formed of contextual information and the final question to be answered. Each question is associated with five possible answers where only a single answer is correct. To generate the claims for this dataset, we only used the final question along with each of the possible answers, disregarding the contextual information. The claims generated in this way did not always capture the answer sufficiently well, so we manually checked and edited them where necessary. Finally, to include the contextual information during the experiments, we used the template given in Appendix D.

We randomly selected 700 claims from the TruthfulClaim and the StrategyClaim datasets (200 for the initial prompt experiments, and 500 for the main experiments), and 500 claims from the MedClaim dataset for the main experiments. All the datasets we use for our main experiments are balanced (i.e. 250 true and 250 false labels). The reason for selecting a subset of the datasets is due to the resource cost associated with experimenting with LLMs on bigger datasets.

5.3 LLMs

To run our experiment we use four models: Mistral (Mistral-7B-Instruct-v0.2) Jiang *et al.* [2023], Mixtral (Mixtral-8x7B-Instruct-v0.1) Jiang *et al.* [2024], Gemma (gemma-7b) Team *et al.* [2024], and GPT-3.5-turbo (GPT-3.5-turbo-0125) Brown *et al.* [2020]. We chose Mistral, Mixtral, and Gemma as they were the best-performing open-source[7] models of reasonable size. In order to reduce the computational costs of running the open-source models, we quantise them to 4 bits Dettmers *et al.* [2023] when running our experiments (both for the baselines and our method). As a representative of models with proprietary weights, we chose GPT-3.5-turbo as it had the best performance/cost trade-off. We did not use Llama-

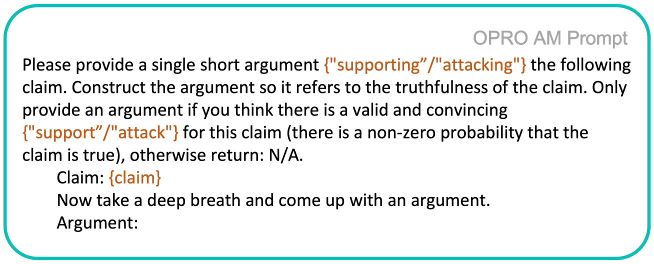
[6] All our experiments are executed with two RTX 4090 24GB GPUs on an Intel(R) Xeon(R) w5-2455X.

[7] We use a broad notion of the term "open-source", not necessarily implying the use of OSI-approved licenses, etc.

2 Touvron *et al.* [2023], as its smaller variants are typically ranked worse compared to the selected models and since the Llama-2 70B model (which is the biggest Llama-2 model) did not perform well on the validation dataset. For all the models the used parameters were temperature 0.7, max new tokens for arguments 128, max new tokens for baselines 768, top-p 0.95 and repetition penalty 1.0.

5.4 Our Method

In our experiments, we use four different variations of our argumentative method explained in Section 4. For all the variations, we use the same prompts for argument generation and argument strength attribution. The prompt for argument generation is given in Figure 3 and the one for argument strength attribution in Figure 4.



OPRO AM Prompt

Please provide a single short argument {"supporting”/"attacking"} the following claim. Construct the argument so it refers to the truthfulness of the claim. Only provide an argument if you think there is a valid and convincing {"support”/"attack"} for this claim (there is a non-zero probability that the claim is true), otherwise return: N/A.

Claim: {claim}
Now take a deep breath and come up with an argument.
Argument:

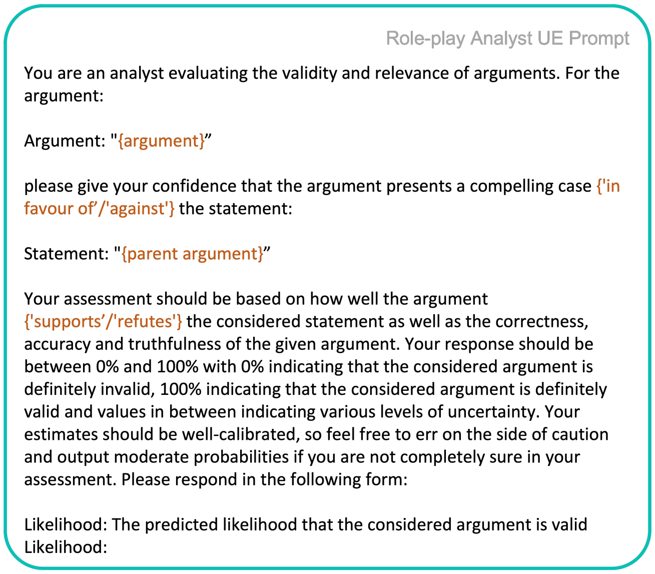
Figure 3: Prompt used for argument generation. {"supporting"/"attacking"} and {"support"/"attack"} are conditional to the required argument type (i.e. if a support argument is required, the conditionals would be "supporting" and "support", respectively). In our prompt, {claim} is replaced with the claim we want to verify.

Variation 1: 0.5 Base Argument (Depth=1) In this variation, we generate two arguments for the claim: a supporter and an attacker. The resulting argumentation framework is a tree of depth 1, composed of three arguments. We only execute the argument strength attribution component for the generated arguments. The claim is assigned a neutral base score of 0.5 to make the decision unbiased.

Variation 2: 0.5 Base Argument (Depth=2) In this variation, we generate two arguments for the claim and then generate a supporting and an attacking argument for both the supporter of the claim and the attacker of the claim. This gives us a tree of depth 2, made up of seven arguments in total. Again, we only execute the argument strength attribution for the generated arguments, assigning a 0.5 base score to the claim.

Variation 3: Estimated Base Argument (Depth=1) The argumentation framework structure in this variation is the same as in Variation 1 — a tree of depth 1 with three arguments. The only difference is that the argument strength attribution is also applied to the claim (rather than using the fixed base score of 0.5). Since the original prompt for the argument strength does not work for the claim (as it requires a parent argument), we use its adapted version, which is shown in Figure 9).

Variation 4: Estimated Base Argument (Depth=2) The tree structure in this variation is the same as in Variation 2 — a tree of depth 2 with seven arguments. The only difference is that the argument strength attribution is executed additionally for the claim (again, rather than using the fixed base score of 0.5).



Role-play Analyst UE Prompt

You are an analyst evaluating the validity and relevance of arguments. For the argument:

Argument: "{argument}”

please give your confidence that the argument presents a compelling case {'in favour of'/'against'} the statement:

Statement: "{parent argument}”

Your assessment should be based on how well the argument {'supports'/'refutes'} the considered statement as well as the correctness, accuracy and truthfulness of the given argument. Your response should be between 0% and 100% with 0% indicating that the considered argument is definitely invalid, 100% indicating that the considered argument is definitely valid and values in between indicating various levels of uncertainty. Your estimates should be well-calibrated, so feel free to err on the side of caution and output moderate probabilities if you are not completely sure in your assessment. Please respond in the following form:

Likelihood: The predicted likelihood that the considered argument is valid
Likelihood:

Figure 4: Prompt used for argument strength attribution. {"in favour of"/"against"} and {"supports"/"refutes"} are conditional to what type of argument is given (i.e. if an attack argument is given the conditionals would be "against" and "refutes", respectively). In our prompt, {argument} is replaced with the argument that needs to be evaluated and {parent argument} is replaced with the parent argument of that argument.

6 Results

We compared the performance of various methods on adapted versions of several commonly considered datasets (TruthfulClaim, StrategyClaim, and MedClaim) using several LLMs (Mistral, Mixtral, Gemma 7B, and GPT-3.5-turbo).

On the TruthfulClaim dataset, the estimated base score methods exhibited higher accuracy compared to other methods. Specifically, *Est. Base Arg (D=1)* had the highest accuracy of 0.758 and 0.81 on Mistral and Mixtral, respectively, and *Est. Base Arg (D=2)* reached the highest accuracy of 0.748 on GPT-3.5-turbo. However, on Gemma 7B, *chain-of-thought* had the highest accuracy of 0.68. For the StrategyClaim dataset, *Est. Base Arg (D=2)* achieved the highest accuracy of 0.692 on Mixtral, while *chain-of-thought* had the highest accuracy of 0.684 and 0.58 on Mistral and Gemma 7B, respectively. *Direct Question* performed the best for GPT-3.5-turbo, with an accuracy of 0.734. Regarding the MedClaim dataset, *chain-of-thought* recorded the highest accuracy of 0.612 and 0.546 on Mistral and Gemma 7B, respectively. Meanwhile, *Direct Question* had the highest accuracy of 0.67 on GPT-3.5-turbo while *Est. Base* had the highest accuracy of 0.62 on Mixtral.

Besides this, we carried out an extra experiment with GPT-4 (GPT-4-0613) to test our hypothesis that both argument generation and strength attribution were ineffective for smaller models, on the *conditioned claim verification* task (rather than

Table 1: Accuracy of three baselines and four variations of our argumentative method on claim verification tasks. The best performing method for each model-dataset combination is indicated in bold.

		Direct Question	**Est. Confidence**	**Chain-of-Thought**	**0.5 Base Arg (D=1)**	**0.5 Base Arg (D=2)**	**Est. Base Arg (D=1)**	**Est. Base Arg (D=2)**
Truthful Claim	Mistral	0.726	0.732	0.748	0.646	0.644	**0.758**	0.752
	Mixtral	0.772	0.77	0.756	0.718	0.718	**0.81**	0.806
	Gemma 7B	0.648	0.624	**0.68**	0.642	0.64	0.626	0.626
	GPT-3.5-turbo	0.698	0.728	0.744	0.604	0.606	0.728	**0.748**
Strategy Claim	Mistral	0.604	0.614	**0.684**	0.578	0.576	0.622	0.63
	Mixtral	0.68	0.67	0.64	0.62	0.618	0.684	**0.692**
	Gemma 7B	0.55	0.556	**0.58**	0.556	0.556	0.568	0.568
	GPT-3.5-turbo	**0.734**	0.696	0.716	0.558	0.558	0.696	0.708
Med Claim	Mistral	0.552	0.568	**0.612**	0.496	0.494	0.532	0.55
	Mixtral	0.598	**0.62**	0.614	0.592	0.592	0.608	0.616
	Gemma 7B	0.524	0.532	**0.546**	0.512	0.512	0.518	0.518
	GPT-3.5-turbo	**0.67**	0.572	0.666	0.564	0.56	0.574	0.566
	GPT-4	0.66	0.60	0.66	0.52	0.54	0.64	**0.68**

standard claim verification). Since MedClaim was unique in this sense we only carried out the extra experiment for this dataset, and we used only 50 samples due to financial constraints. The results indicated that *Est. Base Arg (D=2)* had the best accuracy of 0.68, followed by *Direct Question* and *chain-of-thought*, both of which achieved an accuracy of 0.66. The improved performance of GPT-4 relative to the other models tested supports this hypothesis.

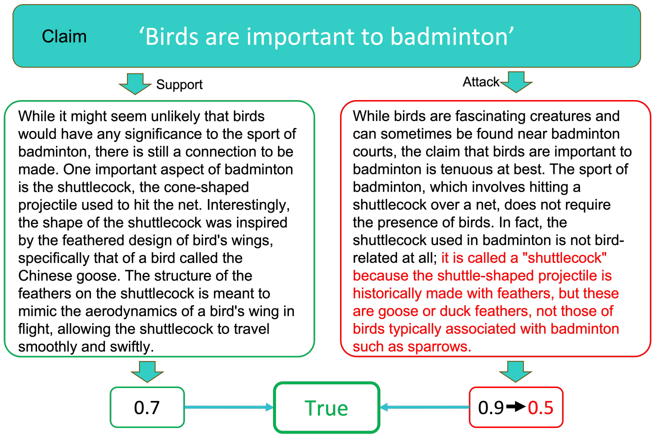
In general, the accuracy of all methods varied on different dataset across different LLMs. However, the argumentative estimated base score methods and *chain-of-thought* performed better overall compared to others. Specifically, *chain-of-thought* performed better on Mistral and Gemma 7B, while *Est. Confidence* and *Est. Base Arg (D=1)* had advantages on Mixtral, and *Direct Question* performed relatively better on GPT-3.5-turbo. Furthermore, the estimated base score methods had better accuracy overall than fixed (0.5) base score methods.

In addition to accuracy, we evaluated our argumentative methods on Brier score and AUC, compared against *Est. Confidence* as the baseline. We present the results in the Appendix (Tables 16 and 17). Overall, our argumentative methods outperformed *Est. Confidence* for all datasets and all LLMs. However, there were a few exceptions that *Est. Confidence* performed better than the argumentative methods on the TruthfulClaim and MedClaim datasets with Gemma 7B in terms of Brier score, and the StrategyClaim dataset with GPT-3.5-turbo for AUC, where argumentative methods were slightly worse than the baseline.

7 Discussion

Argumentative LLMs offer numerous benefits when compared to existing comparable techniques. As we have demonstrated with the instantiation presented in this paper, our methodology does not require any external resources or fine-tuning, to perform comparably (in terms of accuracy) at claim verification tasks with the current state-of-the-art prompting methods. Additionally, we believe that the performance of our approach could benefit from fine-tuning for the argument generation and argument strength attribution sub-components. Likewise, we expect that permitting information retrieval, both for generating the arguments and strengths, would result in further improvements.

However, perhaps the most important features of our proposed methodology cannot be adequately captured by quantitative performance metrics on benchmarks. One of these features is that the outputs generated are reasons, from which decisions can be derived, rather than decisions directly. While chain-of-thought techniques do something akin to this, the reasons are output monolithically, and cannot guarantee that they faithfully imply the final decision. Whereas the final decision output by our system is necessarily a function of the constituent reasons, due to the system architecture.



While it might seem unlikely that birds would have any significance to the sport of badminton, there is still a connection to be made. One important aspect of badminton is the shuttlecock, the cone-shaped projectile used to hit the net. Interestingly, the shape of the shuttlecock was inspired by the feathered design of bird's wings, specifically that of a bird called the Chinese goose. The structure of the feathers on the shuttlecock is meant to mimic the aerodynamics of a bird's wing in flight, allowing the shuttlecock to travel smoothly and swiftly.

While birds are fascinating creatures and can sometimes be found near badminton courts, the claim that birds are important to badminton is tenuous at best. The sport of badminton, which involves hitting a shuttlecock over a net, does not require the presence of birds. In fact, the shuttlecock used in badminton is not bird-related at all; it is called a "shuttlecock" because the shuttle-shaped projectile is historically made with feathers, but these are goose or duck feathers, not those of birds typically associated with badminton such as sparrows.

Figure 5: An illustration of a user contesting the strength of an argument attacking a claim taken from StrategyClaim (both arguments are generated by Mixtral). Before the contestation, the claim was (incorrectly) classified as false, as the attacking argument was assigned a strength of 0.9, whereas the supporting argument had a strength of 0.7. However, a human user is able to modify this score (e.g. from 0.9 to 0.5), citing the fallacious reasoning present in the attacking argument (highlighted in red). This ultimately results in the correct classification.

Furthermore, the composite nature of the reasoning results in highly explainable and contestable outputs. The explanation for why a decision has been made is transparent, and can be directly attributed to the generated arguments and their associated strengths. This explainability offers users of the system plentiful opportunity to disagree with the reasoning, either in terms of the arguments generated being relevant or true, or the strengths that have been attributed to them being representative of the extent to which they support or attack their parent argument (including the root claim). For an illustration of the latter scenario, see Figure 5.

The advantages offered by these features are not necessarily demonstrated well by the relatively simple examples found in commonly used benchmarks, such as the ones that we use for this study. The settings where they are most relevant, and important, are in highly complex, uncertain and high-stakes scenarios. These may include business, medical or legal decision-making.

Ideally, the system would be used in conjunction with a domain-expert, who could review the outputted arguments and strengths, and leverage their experience and contextual knowledge to modify them accordingly. Additionally, they can add arguments of their own. This also lends itself to collaborative use, wherein a community of people can vote on the quality of the arguments, and add their own perspectives by including additional arguments (see Appendix G for examples).

Previous work by Yin *et al.* [2023] facilitates the realisation of integrating our methodology into such a human-computer hybrid system. For very large and complex argument graphs, one is able to automatically surface the arguments which have the greatest impact on the final score. This would allow human users to manually check only the most significant arguments, and their respective scores. This makes the system amenable to human oversight, even in cases where there are potentially hundreds or thousands of relevant arguments.

Another factor which lends itself to the system's use in high-stakes scenarios is that uncertainty is inherently calculated. This is a product of the argument semantics, which uses the constituent argument strengths to output a final score for the root argument. This score can be easily interpreted as uncertainty about the final decision. This is very useful in situations where a particular decision can have a highly detrimental outcome, such as in a medical setting. As shown by the Brier score and AUC results in Tables 16 and 17, generating additional arguments improves the quality of probability estimates, compared to models directly reporting their confidence (Est. Confidence).

Lastly, the results of the prompt experiments (shown in Appendix E), emphasise the highly conditional performance of LLMs on the combination of the prompt and dataset being used. Moreover, this relationship is inconsistent between different models. This suggests that a model performing very well at a task in a particular setting does not guarantee that this performance will transfer to a different setting. Our proposed system combats this issue for the reasons noted above, namely that the outputs are entirely explainable and contestable. This provides human users with sufficient agency to guard against and remedy any unexpected dips in performance due to a change in the input data distribution.

8 Conclusions & Future Work

In this paper we introduce a methodology for harnessing the general reasoning capacity of LLMs - without requiring any fine-tuning or external resources - making them explainable, contestable and improving their reasoning in some circumstances. Furthermore, our system innately permits human-computer collaboration, and provides accurate uncertainty estimates as an output.

The instantiation of the system in this paper is very basic. This is suitable for the simple claims that make up the existing benchmarks we use. We leave to future work the use of more general argumentative explanations in the spirit of Kotonya and Toni [2024].

Similarly, we conduct all experiments without any task-specific training and by using the most basic method of argument strength attribution. The reason for this decision was twofold - in order to assess the ability of 'out-the-box' LLMs to perform argumentative reasoning, and to demonstrate the viability of our methodology in its simplest form. While we have demonstrated that this is a reasonably effective approach, we envision that both fine-tuning, and employing more tailored methods will result in improved results.

There are numerous methodologies for argument strength attribution which warrant further analysis. These include techniques which adapt our chosen method of directly prompting an LLM by, for example, sampling multiple outputs of the same LLM or taking the weighted value of the relevant logits in the final layer of the model.

Furthermore, an adapted version of the 'semantic uncertainty' Kuhn *et al.* [2023] methodology may be devised, wherein one directly clusters semantically similar sampled arguments, rather than having to prompt models for numerical scores. We also experimented with verbal confidence scores to assign argument strengths. While we did not observe promising results, this approach may respond well to supervised finetuning.

Another promising direction for future work is the ensembling of many different LLMs, both for argument generation and strength attribution. This is a way to harness the heterogeneous knowledge encoded in disparate models. In this vein, using information retrieval or retrieval augmented generation Lewis *et al.* [2020], is a way to increase the breadth and reliability of the arguments generated.

9 Limitations

Any study that is attempting to make general claims about LLMs should strive to use as many, and as diverse a range of them as possible. We have tried to do so, but due to the overwhelming number of both open-sourced and closed-sourced models available, we have only been able to test a fraction of the total available. However, by using the models that rank highest across various benchmarks, we have attempted to demonstrate that this methodology can be effectively employed with the state of the art.

10 Ethics Statement

There are potential risks of LLMs such as social bias and generation of misinformation. In this work, we intentionally devise our methodology to be used with human oversight. This means users have recourse in the case that any biased output is produced by an LLM being utilised as a part of our system. However, in cases where the argumentation framework that is produced is too large for humans to review every argument, there is some risk that biased reasoning could impact the final decision.

Making the reasons for a LLM-driven decision explicit increases explainability, and thus safety. However, the ability to contest decision may be co-opted by bad actors, who intentionally subvert the reasoning process. This is why the our proposed methodology is intended to be used with trusted human oversight.

References

Leila Amgoud and Henri Prade. Using arguments for making and explaining decisions. *Artif. Intell.*, 173(3-4):413–436, 2009.

Katie Atkinson, Pietro Baroni, Massimiliano Giacomin, Anthony Hunter, Henry Prakken, Chris Reed, Guillermo Ricardo Simari, Matthias Thimm, and Serena Villata. Towards artificial argumentation. *AI Mag.*, 38(3):25–36, 2017.

Pietro Baroni, Dov Gabbay, Massimiliano Giacomin, and Leendert van der Torre, editors. *Handbook of Formal Argumentation*. College Publications, 2018.

Pietro Baroni, Antonio Rago, and Francesca Toni. How many properties do we need for gradual argumentation? In Sheila A. McIlraith and Kilian Q. Weinberger, editors, *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 1736–1743. AAAI Press, 2018.

Pietro Baroni, Antonio Rago, and Francesca Toni. From fine-grained properties to broad principles for gradual argumentation: A principled spectrum. *Int. J. Approx. Reason.*, 105:252–286, 2019.

Lukas Berglund, Meg Tong, Max Kaufmann, Mikita Balesni, Asa Cooper Stickland, Tomasz Korbak, and Owain Evans. The reversal curse: Llms trained on "a is b" fail to learn "b is a". *CoRR*, abs/2309.12288, 2023.

Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Michal Podstawski, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Hubert Niewiadomski, Piotr Nyczyk, and Torsten Hoefler. Graph of thoughts: Solving elaborate problems with large language models. In Michael J. Wooldridge, Jennifer G. Dy, and Sriraam Natarajan, editors, *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pages 17682–17690. AAAI Press, 2024.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, T. J. Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeff Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. *Neural Information Processing Systems*, abs/2005.14165, 2020.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, John A. Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuan-Fang Li, Scott M. Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. Sparks of artificial general intelligence: Early experiments with gpt-4. *ArXiv*, abs/2303.12712, 2023.

Guizhen Chen, Liying Cheng, Anh Tuan Luu, and Lidong Bing. Exploring the potential of large language models in computational argumentation. *ArXiv*, abs/2311.09022, 2023.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. QLoRA: Efficient finetuning of quantized LLMs. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine, editors, *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate. *CoRR*, abs/2305.14325, 2023.

Phan Minh Dung. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artif. Intell.*, 77(2):321–358, 1995.

Esin Durmus, Liane Lovitt, Alex Tamkin, Stuart Ritchie, Jack Clark, and Deep Ganguli. Measuring the persuasiveness of language models, 2024.

Lukas Fluri, Daniel Paleka, and Florian Tramèr. Evaluating superhuman models with consistency checks. *CoRR*, abs/2306.09983, 2023.

Damián Ariel Furman, Pablo Torres, José A. Rodríguez, Diego Letzen, Maria Vanina Martinez, and Laura Alonso Alemany.

High-quality argumentative information in low resources approaches improve counter-narrative generation. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 2942–2956. Association for Computational Linguistics, 2023.

Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. Did aristotle use a laptop? A question answering benchmark with implicit reasoning strategies. *CoRR*, abs/2101.02235, 2021.

Ivan Habernal and Iryna Gurevych. Which argument is more convincing? analyzing and predicting convincingness of web arguments using bidirectional LSTM. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*. The Association for Computer Linguistics, 2016.

Danny Halawi, Fred Zhang, Chen Yueh-Han, and Jacob Steinhardt. Approaching human-level forecasting with language models, 2024.

Clément Henin and Daniel Le Métayer. Beyond explainability: justifiability and contestability of algorithmic decision systems. *AI & SOCIETY*, 37:1397 – 1410, 2021.

Martin Hinton and Jean H. M. Wagemans. How persuasive is AI-generated argumentation? an analysis of the quality of an argumentative text produced by the GPT-3 AI text generator. *Argument Comput.*, 14(1):59–74, 2023.

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, 2023.

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, 2024.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. What disease does this patient have? A large-scale open domain question answering dataset from medical exams. *CoRR*, abs/2009.13081, 2020.

Neema Kotonya and Francesca Toni. Towards a framework for evaluating explanations in automated fact verification. In *LREC-COLING 2024*, 2024. To appear.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023.

Anne Lauscher, Lily Ng, Courtney Napoles, and Joel Tetreault. Rhetoric, logic, and dialectic: Advancing theory-based argument quality assessment in natural language processing. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4563–4574, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics.

Patrick S. H. Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

Q Vera Liao and Jennifer Wortman Vaughan. AI transparency in the age of LLMs: A human-centered research roadmap. *Harvard Data Science Review*, 2024.

Stephanie Lin, Jacob Hilton, and Owain Evans. Truthfulqa: Measuring how models mimic human falsehoods. *CoRR*, abs/2109.07958, 2021.

Henrietta Lyons, Eduardo Velloso, and Tim Miller. Conceptualising contestability. *Proceedings of the ACM on Human-Computer Interaction*, 5:1 – 25, 2021.

Hugo Mercier and Dan Sperber. Why do humans reason? arguments for an argumentative theory. *Behavioral and Brain Sciences*, 34(2):57–74, 2011.

Hugo Mercier and Dan Sperber. *The Enigma of Reason*. Penguin, 2018.

Tim Miller. Explainable AI is dead, long live explainable ai!: Hypothesis-driven decision support using evaluative AI. In *FAccT 2023*, 2023.

Siqi Ouyang and Lei Li. Autoplan: Automatic planning of interactive decision-making tasks with large language models. In *Conference on Empirical Methods in Natural Language Processing*, 2023.

Nico Potyka. Continuous dynamical systems for weighted bipolar argumentation. In Michael Thielscher, Francesca Toni, and Frank Wolter, editors, *Principles of Knowledge Representation and Reasoning: Proceedings of the Sixteenth International Conference, KR 2018, Tempe, Arizona, 30 October - 2 November 2018*, pages 148–157. AAAI Press, 2018.

Antonio Rago, Francesca Toni, Marco Aurisicchio, and Pietro Baroni. Discontinuity-free decision support with quantitative argumentation debates. In Chitta Baral, James P. Delgrande, and Frank Wolter, editors, *Principles of Knowledge Representation and Reasoning: Proceedings of the Fifteenth International Conference, KR 2016, Cape Town, South Africa, April 25-29, 2016*, pages 63–73. AAAI Press, 2016.

Shubhra Kanti Karmaker Santu and Dongji Feng. Teler: A general taxonomy of LLM prompts for benchmarking complex tasks. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 14197–14203. Association for Computational Linguistics, 2023.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language models can teach themselves to use tools. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine, editors, *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023.

Murray Shanahan. Talking about large language models. *Commun. ACM*, 67(2):68–79, 2024.

Edwin Simpson and Iryna Gurevych. Finding convincing arguments using scalable bayesian preference learning. *Trans. Assoc. Comput. Linguistics*, 6:357–371, 2018.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, Léonard Hussenot, Pier Giuseppe Sessa, Aakanksha Chowdhery, Adam Roberts, Aditya Barua, Alex Botev, Alex Castro-Ros, Ambrose Slone, Amélie Héliou, Andrea Tacchetti, Anna Bulanova, Antonia Paterson, Beth Tsai, Bobak Shahriari, Charline Le Lan, Christopher A. Choquette-Choo, Clément Crepy, Daniel Cer, Daphne Ippolito, David Reid, Elena Buchatskaya, Eric Ni, Eric Noland, Geng Yan, George Tucker, George-Christian Muraru, Grigory Rozhdestvenskiy, Henryk Michalewski, Ian Tenney, Ivan Grishchenko, Jacob Austin, James Keeling, Jane Labanowski, Jean-Baptiste Lespiau, Jeff Stanway, Jenny Brennan, Jeremy Chen, Johan Ferret, Justin Chiu, Justin Mao-Jones, Katherine Lee, Kathy Yu, Katie Millican, Lars Lowe Sjoesund, Lisa Lee, Lucas Dixon, Machel Reid, Maciej Mikuła, Mateo Wirth, Michael Sharman, Nikolai Chinaev, Nithum Thain, Olivier Bachem, Oscar Chang, Oscar Wahltinez, Paige Bailey, Paul Michel, Petko Yotov, Rahma Chaabouni, Ramona Comanescu, Reena Jana, Rohan Anil, Ross McIlroy, Ruibo Liu, Ryan Mullins, Samuel L Smith, Sebastian Borgeaud, Sertan Girgin, Sholto Douglas, Shree Pandya, Siamak Shakeri, Soham De, Ted Klimenko, Tom Hennigan, Vlad Feinberg, Wojciech Stokowiec, Yu hui Chen, Zafarali Ahmed, Zhitao Gong, Tris Warkentin, Ludovic Peran, Minh Giang, Clément Farabet, Oriol Vinyals, Jeff Dean, Koray Kavukcuoglu, Demis Hassabis, Zoubin Ghahramani, Douglas Eck, Joelle Barral, Fernando Pereira, Eli Collins, Armand Joulin, Noah Fiedel, Evan Senter, Alek Andreev, and Kathleen Kenealy. Gemma: Open models based on gemini research and technology, 2024.

Luke Thorburn and Ariel Kruger. Optimizing language models for argumentative reasoning. In *Proceedings of the 1st Workshop on Argumentation & Machine Learning co-located with 9th International Conference on Computational Models of Argument (COMMA 2022), Cardiff, Wales, September 13th, 2022*, volume 3208 of *CEUR Workshop Proceedings*, pages 27–44. CEUR-WS.org, 2022.

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. *CoRR*, abs/2307.09288, 2023.

Zefan Wang, Zichuan Liu, Yingying Zhang, Aoxiao Zhong, Lunting Fan, Lingfei Wu, and Qingsong Wen. Rcagent: Cloud root cause analysis by autonomous agents with tool-augmented large language models. *ArXiv*, abs/2310.16340, 2023.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *NeurIPS 2022*, 2022.

Shijie Xia, Xuefeng Li, Yixin Liu, Tongshuang Wu, and Pengfei Liu. Evaluating mathematical reasoning beyond accuracy, 2024.

Chengrun Yang, Xuezhi Wang, Yifeng Lu, Hanxiao Liu, Quoc V. Le, Denny Zhou, and Xinyun Chen. Large language models as optimizers. *ArXiv*, abs/2309.03409, 2023.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. In *NeurIPS 2023*, 2023.

Xiang Yin, Nico Potyka, and Francesca Toni. Argument attribution explanations in quantitative bipolar argumentation frameworks (technical report). In *European Conference on Artificial Intelligence*, 2023.

Hugh Zhang and David C. Parkes. Chain-of-thought reasoning is a policy improvement operator. *CoRR*, abs/2309.08589, 2023.

Haodi Zhang, Jiahong Li, Yichi Wang, and Yuanfeng Songi. Integrating automated knowledge extraction with large language models for explainable medical decision-making. *2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 1710–1717, 2023.

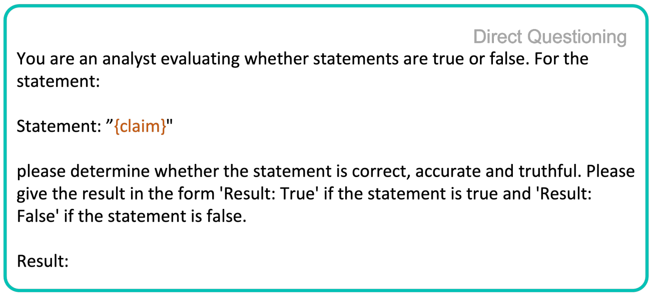
Appendix

A Baseline Prompts

We show the prompts used for the baselines.

A.1 Direct questioning

The direct questioning prompt is shown in Figure 6.



Direct Questioning

You are an analyst evaluating whether statements are true or false. For the statement:

Statement: "{claim}"

please determine whether the statement is correct, accurate and truthful. Please give the result in the form 'Result: True' if the statement is true and 'Result: False' if the statement is false.

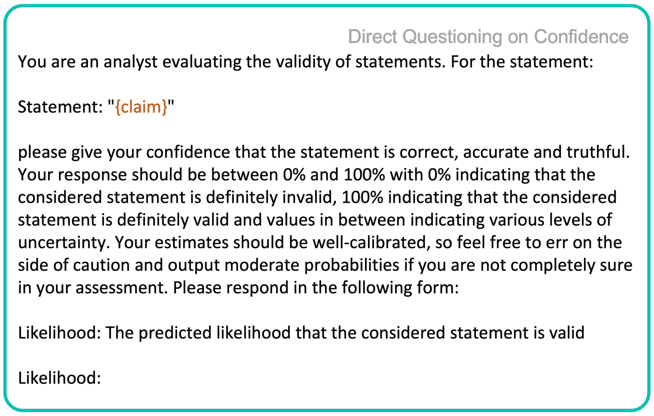
Result:

Figure 6: Prompt used for direct questioning baseline.

In our prompt, {claim} is replaced with the claim we want to verify.

A.2 Direct questioning on confidence

The direct questioning on confidence prompt is shown in Figure 7.



Direct Questioning on Confidence

You are an analyst evaluating the validity of statements. For the statement:

Statement: "{claim}"

please give your confidence that the statement is correct, accurate and truthful. Your response should be between 0% and 100% with 0% indicating that the considered statement is definitely invalid, 100% indicating that the considered statement is definitely valid and values in between indicating various levels of uncertainty. Your estimates should be well-calibrated, so feel free to err on the side of caution and output moderate probabilities if you are not completely sure in your assessment. Please respond in the following form:

Likelihood: The predicted likelihood that the considered statement is valid

Likelihood:

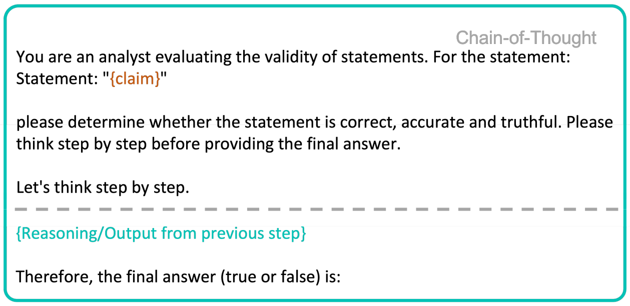
Figure 7: Prompt used for direct questioning on confidence baseline.

In our prompt, {claim} is replaced with the claim we want to verify.

A.3 Questioning with Chain-of-thought

For chain-of-thought the first prompt used to obtain the discrete steps and the prompt to get the final decision are given in Figure 8.

The prompt above the line is to obtain the reasoning steps and {claim} is replaced with the claim we want to verify. The prompt below the line is to get the final decision and {Reasoning/Output from previous step} is replaced with the reasoning obtained from the prompt above the line.



Chain-of-Thought

You are an analyst evaluating the validity of statements. For the statement:
Statement: "{claim}"

please determine whether the statement is correct, accurate and truthful. Please think step by step before providing the final answer.

Let's think step by step.

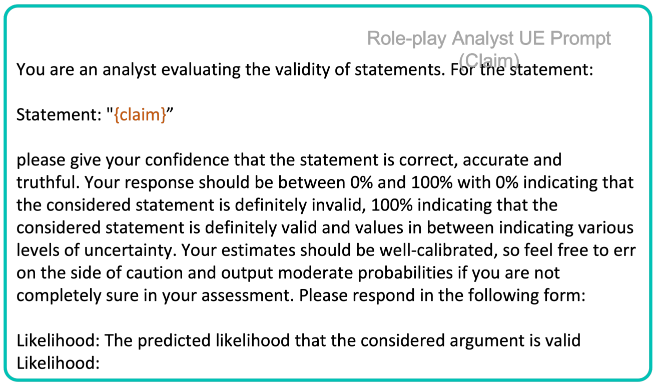
{Reasoning/Output from previous step}

Therefore, the final answer (true or false) is:

Figure 8: Prompts used for chain-of-thought baseline.

B Prompt for Argument Strength Attribution for Claim

The prompt for argument strength attribution does not work for the claim as it requires a parent argument to be present. So, we altered the prompt for only claim argument strength attribution (the prompt could be seen in Figure 9).



Role-play Analyst UE Prompt (Claim)

You are an analyst evaluating the validity of statements. For the statement:

Statement: "{claim}"

please give your confidence that the statement is correct, accurate and truthful. Your response should be between 0% and 100% with 0% indicating that the considered statement is definitely invalid, 100% indicating that the considered statement is definitely valid and values in between indicating various levels of uncertainty. Your estimates should be well-calibrated, so feel free to err on the side of caution and output moderate probabilities if you are not completely sure in your assessment. Please respond in the following form:

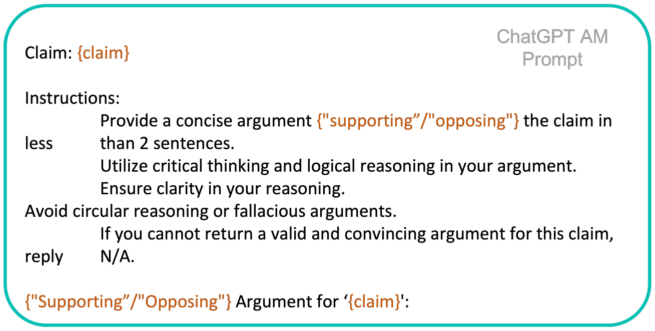
Likelihood: The predicted likelihood that the considered argument is valid
Likelihood:

Figure 9: Prompt used for argument strength attribution for the claim. In our prompt, {claim} is replaced with the claim we want to verify.

C Considered Prompts

C.1 ChatGPT prompts

ChatGPT prompts were generated mostly using ChatGPT. First, the prompt is initialised by giving ChatGPT the instructions, then the prompt is optimised by giving ChatGPT some outputs and asking it to improve the prompts. The ChatGPT Argument Generator prompt can be found in Figure 10 and the ChatGPT Argument Strength Attribution prompt can be found in Figure 11.



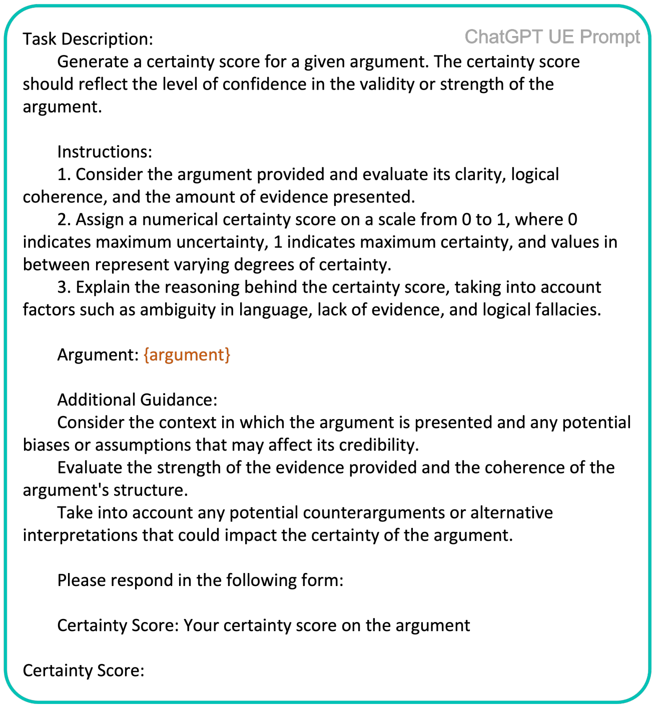
ChatGPT AM Prompt

Claim: {claim}

Instructions:
Provide a concise argument {"supporting”/"opposing"} the claim in
less than 2 sentences.
Utilize critical thinking and logical reasoning in your argument.
Ensure clarity in your reasoning.
Avoid circular reasoning or fallacious arguments.
If you cannot return a valid and convincing argument for this claim,
reply N/A.

{"Supporting”/"Opposing"} Argument for ‘{claim}':

Figure 10: ChatGPT Argument Generator prompt



ChatGPT UE Prompt

Task Description:
Generate a certainty score for a given argument. The certainty score should reflect the level of confidence in the validity or strength of the argument.

Instructions:
1. Consider the argument provided and evaluate its clarity, logical coherence, and the amount of evidence presented.
2. Assign a numerical certainty score on a scale from 0 to 1, where 0 indicates maximum uncertainty, 1 indicates maximum certainty, and values in between represent varying degrees of certainty.
3. Explain the reasoning behind the certainty score, taking into account factors such as ambiguity in language, lack of evidence, and logical fallacies.

Argument: {argument}

Additional Guidance:
Consider the context in which the argument is presented and any potential biases or assumptions that may affect its credibility.
Evaluate the strength of the evidence provided and the coherence of the argument's structure.
Take into account any potential counterarguments or alternative interpretations that could impact the certainty of the argument.

Please respond in the following form:

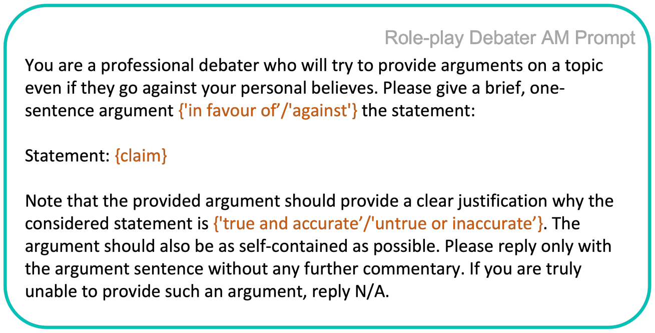
Certainty Score: Your certainty score on the argument

Certainty Score:

Figure 11: ChatGPT Argument Strength Attribution prompt

C.2 Role-player prompts

Role-player prompts followed a prompting strategy where the LLMs were expected to act like debater for the Argument Generator component and analyst for the Argument Strength Attribution component. The Argument Generator prompt, Debater, can be found in Figure 12 and the Argument Strength Attribution, Analyst, can be found in Figure 13.



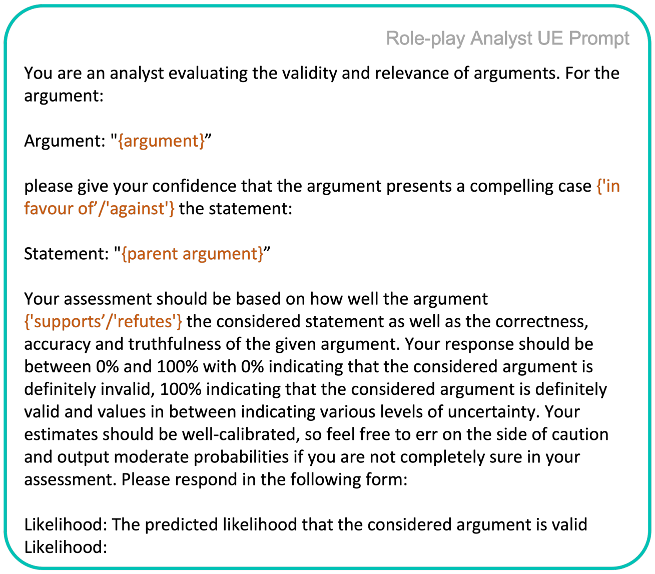
Role-play Debater AM Prompt

You are a professional debater who will try to provide arguments on a topic even if they go against your personal believes. Please give a brief, one-sentence argument {'in favour of’/'against'} the statement:

Statement: {claim}

Note that the provided argument should provide a clear justification why the considered statement is {'true and accurate’/'untrue or inaccurate’}. The argument should also be as self-contained as possible. Please reply only with the argument sentence without any further commentary. If you are truly unable to provide such an argument, reply N/A.

Figure 12: Role-Player: Debater Argument Miner prompt



Role-play Analyst UE Prompt

You are an analyst evaluating the validity and relevance of arguments. For the argument:

Argument: "{argument}”

please give your confidence that the argument presents a compelling case {'in favour of’/'against'} the statement:

Statement: "{parent argument}”

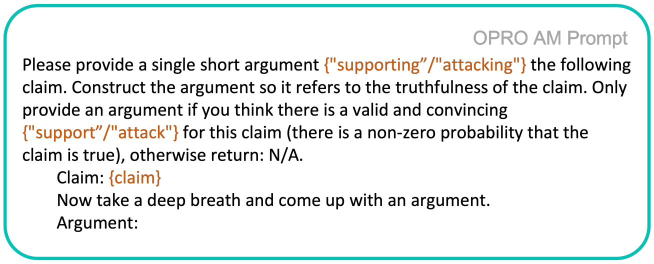
Your assessment should be based on how well the argument {'supports’/'refutes'} the considered statement as well as the correctness, accuracy and truthfulness of the given argument. Your response should be between 0% and 100% with 0% indicating that the considered argument is definitely invalid, 100% indicating that the considered argument is definitely valid and values in between indicating various levels of uncertainty. Your estimates should be well-calibrated, so feel free to err on the side of caution and output moderate probabilities if you are not completely sure in your assessment. Please respond in the following form:

Likelihood: The predicted likelihood that the considered argument is valid
Likelihood:

Figure 13: Role-Player: Analyst Uncertainty Estimator prompt

C.3 OPRO prompts

OPRO prompts follow the Optimization by PROmpting (OPRO) strategy Yang *et al.* [2023]. The OPRO Argument Generator prompt can be found in Figure 14 and the OPRO Argument Strength Attribution prompt can be found in Figure 15.



OPRO AM Prompt

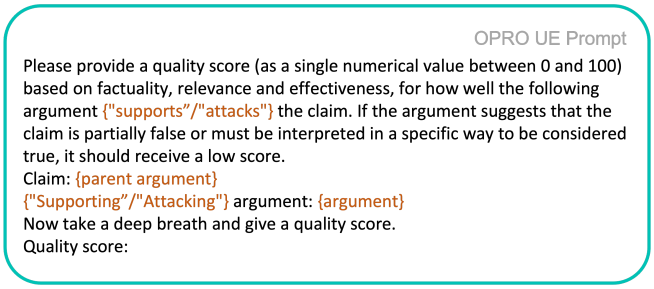
Please provide a single short argument {"supporting"/"attacking"} the following claim. Construct the argument so it refers to the truthfulness of the claim. Only provide an argument if you think there is a valid and convincing {"support"/"attack"} for this claim (there is a non-zero probability that the claim is true), otherwise return: N/A.

Claim: {claim}

Now take a deep breath and come up with an argument.

Argument:

Figure 14: OPRO Argument Miner prompt



OPRO UE Prompt

Please provide a quality score (as a single numerical value between 0 and 100) based on factuality, relevance and effectiveness, for how well the following argument {"supports"/"attacks"} the claim. If the argument suggests that the claim is partially false or must be interpreted in a specific way to be considered true, it should receive a low score.

Claim: {parent argument}

{"Supporting"/"Attacking"} argument: {argument}

Now take a deep breath and give a quality score.

Quality score:

Figure 15: OPRO Uncertainty Estimator prompt

D MedClaim Template

For the MedClaim dataset, to include the contextual information during the experiments, we use the following template for the claims, where {information} is the contextual information and {claim} is the claim:

> Consider the following background information: {information} Given the background information the following is correct: {claim}

E Prompt Experiment Results

In this section we give the results of the prompt experiments conducted on the two validation datasets, both consisting of 200 samples.

Table 2 shows the average results for the prompts used for baselines.

Table 3 shows the average results for the prompts used for variations of our method.

Table 4 shows the results for the prompts used for baselines using Mixtral on the TruthfulClaim dataset.

Table 5 shows the results for the prompts used for variations with depth=1 of our method using Mixtral on the TruthfulClaim dataset.

Table 6 shows the results for the prompts used for variations with depth=2 of our method using Mixtral on the TruthfulClaim dataset.

Table 7 shows the results for the prompts used for baselines using Mixtral on the StrategyClaim dataset.

Table 8 shows the results for the prompts used for variations with depth=1 of our method using Mixtral on the StrategyClaim dataset.

Table 9 shows the results for the prompts used for variations with depth=2 of our method using Mixtral on the StrategyClaim dataset.

Table 10 shows the results for the prompts used for baselines using Mistral on the TruthfulClaim dataset.

Table 11 shows the results for the prompts used for variations with depth=1 of our method using Mistral on the TruthfulClaim dataset.

Table 12 shows the results for the prompts used for variations with depth=2 of our method using Mistral on the TruthfulClaim dataset.

Table 13 shows the results for the prompts used for baselines using Mistral on the StrategyClaim dataset.

Table 14 shows the results for the prompts used for variations with depth=1 of our method using Mistral on the StrategyClaim dataset.

Table 15 shows the results for the prompts used for variations with depth=2 of our method using Mistral on the StrategyClaim dataset.

F Final Results

This section gives the final results of all experiments run on the three held-out test datasets, consisting of 500 samples each.

Table 16 gives the main experiment Brier scores of Direct Questioning on Confidence baseline and variations of our method on all three datasets and all four models.

Table 17 gives the main experiment AUC scores of Direct Questioning on Confidence baseline and variations of our method on all three datasets and all four models.

Baseline Prompt	Direct Question	Chain-of-Thought
ChatGPT	0.663	0.671
analyst	**0.669**	**0.681**
OPRO	0.613	0.633

Table 2: Baseline prompt experiment results — average over both models and datasets

AM	UE	0.5 Base + Arg (D=1)	Estimated Base + Arg (D=1)	Estimated Base Only (baseline)
ChatGPT	ChatGPT	0.573	0.604	0.596
ChatGPT	OPRO	0.63	0.649	0.645
ChatGPT	analyst	0.615	0.663	0.65
OPRO	ChatGPT	0.586	0.592	0.596
OPRO	OPRO	0.584	0.64	0.645
OPRO	analyst	0.601	**0.679**	**0.65**
debater	ChatGPT	0.549	0.591	0.596
debater	OPRO	0.624	0.644	0.645
debater	analyst	0.573	0.64	0.65

Table 3: Argumentation prompt experiment results — average over models, depths and datasets

Baseline Prompt	Direct Question	Chain-of-Thought
ChatGPT	**0.815**	**0.76**
analyst	0.81	0.755
OPRO	0.67	0.685

Table 4: Mixtral baseline prompt experiment results, 4bit - TruthfulClaim on 200 datapoints

AM	UE	0.5 Base + Arg (D=1)	Estimated Base + Arg (D=1)	Estimated Base Only (baseline)
ChatGPT	ChatGPT	0.56	0.69	0.695
ChatGPT	OPRO	0.71	0.73	0.725
ChatGPT	analyst	0.685	0.765	**0.745**
OPRO	ChatGPT	0.665	0.7	0.695
OPRO	OPRO	0.685	0.73	0.725
OPRO	analyst	0.69	**0.795**	**0.745**
debater	ChatGPT	0.68	0.695	0.695
debater	OPRO	**0.77**	0.72	0.725
debater	analyst	0.68	0.76	**0.745**

Table 5: Mixtral prompt experiment results, depth 1 - TQA on 200 datapoints

AM	UE	0.5 Base + Arg (D=2)	Estimated Base + Arg (D=2)	Estimated Base Only (baseline)
ChatGPT	ChatGPT	0.56	0.69	0.695
ChatGPT	OPRO	0.71	0.73	0.725
ChatGPT	analyst	0.685	0.76	**0.745**
OPRO	ChatGPT	0.665	0.71	0.695
OPRO	OPRO	0.685	0.73	0.725
OPRO	analyst	0.69	**0.785**	**0.745**
debater	ChatGPT	0.68	0.695	0.695
debater	OPRO	**0.77**	0.715	0.725
debater	analyst	0.675	0.755	**0.745**

Table 6: Mixtral prompt experiment results, depth 2 - TQA on 200 datapoints

Baseline Prompt	Direct Question	Chain-of-Thought
ChatGPT	0.655	0.6
analyst	**0.66**	**0.64**
OPRO	0.55	0.58

Table 7: Mixtral baseline prompt experiment results, 4bit - SQA on 200 datapoints

AM	UE	0.5 Base + Arg (D=1)	Estimated Base + Arg (D=1)	Estimated Base Only (baseline)
ChatGPT	ChatGPT	0.57	0.55	0.535
ChatGPT	OPRO	0.565	0.64	0.625
ChatGPT	analyst	0.59	0.62	**0.635**
OPRO	ChatGPT	**0.6**	0.545	0.535
OPRO	OPRO	0.58	0.63	0.625
OPRO	analyst	0.57	**0.655**	**0.635**
debater	ChatGPT	0.5	0.525	0.535
debater	OPRO	0.58	0.635	0.625
debater	analyst	0.55	0.61	**0.635**

Table 8: Mixtral prompt experiment results, depth 1 - SQA on 200 datapoints

AM	UE	0.5 Base + Arg (D=2)	Estimated Base + Arg (D=2)	Estimated Base Only (baseline)
ChatGPT	ChatGPT	0.57	0.55	0.535
ChatGPT	OPRO	0.565	0.635	0.625
ChatGPT	analyst	0.59	0.625	**0.635**
OPRO	ChatGPT	**0.6**	0.55	0.535
OPRO	OPRO	0.58	0.63	0.625
OPRO	analyst	0.57	**0.655**	**0.635**
debater	ChatGPT	0.5	0.535	0.535
debater	OPRO	0.58	0.635	0.625
debater	analyst	0.55	0.625	**0.635**

Table 9: Mixtral prompt experiment results, depth 2 - SQA on 200 datapoints

Baseline Prompt	Direct Question	Chain-of-Thought
ChatGPT	0.625	0.685
analyst	0.665	**0.71**
OPRO	**0.68**	0.65

Table 10: Mistral baseline prompt experiment results, 4bit - TQA on 200 datapoints

AM	UE	0.5 Base + Arg (D=1)	Estimated Base + Arg (D=1)	Estimated Base Only
ChatGPT	ChatGPT	0.615	0.67	0.65
ChatGPT	OPRO	**0.67**	**0.73**	**0.725**
ChatGPT	analyst	0.635	0.715	0.705
OPRO	ChatGPT	0.61	0.665	0.65
OPRO	OPRO	0.605	0.715	**0.725**
OPRO	analyst	0.62	**0.73**	0.705
debater	ChatGPT	0.49	0.645	0.65
debater	OPRO	0.615	0.725	**0.725**
debater	analyst	0.545	0.665	0.705

Table 11: Mistral prompt experiment results, 4bit, depth 1 - TQA on 200 datapoints

AM	UE	0.5 Base + Arg (D=2)	Estimated Base + Arg (D=2)	Estimated Base Only
ChatGPT	ChatGPT	0.615	0.66	0.65
ChatGPT	OPRO	**0.67**	**0.73**	**0.725**
ChatGPT	analyst	0.635	0.71	0.705
OPRO	ChatGPT	0.61	0.655	0.65
OPRO	OPRO	0.605	0.715	**0.725**
OPRO	analyst	0.62	0.725	0.705
debater	ChatGPT	0.49	0.655	0.65
debater	OPRO	0.62	0.725	**0.725**
debater	analyst	0.55	0.695	0.705

Table 12: Mistral prompt experiment results, 4bit, depth 2 - TQA on 200 datapoints

Baseline Prompt	Direct Question	Chain-of-Thought
ChatGPT	**0.56**	**0.64**
analyst	0.54	0.62
OPRO	0.55	0.615

Table 13: Mistral baseline prompt experiment results, 4bit - SQA on 200 datapoints

AM	UE	0.5 Base + Arg (D=1)	Estimated Base + Arg (D=1)	Estimated Base Only
ChatGPT	ChatGPT	0.55	0.515	0.505
ChatGPT	OPRO	**0.575**	0.5	0.505
ChatGPT	analyst	0.55	**0.56**	**0.515**
OPRO	ChatGPT	0.47	0.445	0.505
OPRO	OPRO	0.465	0.485	0.505
OPRO	analyst	0.525	0.54	**0.515**
debater	ChatGPT	0.525	0.5	0.505
debater	OPRO	0.53	0.495	0.505
debater	analyst	0.515	0.515	**0.515**

Table 14: Mistral prompt experiment results, 4bit, depth 1 - SQA on 200 datapoints

AM	UE	0.5 Base + Arg (D=2)	Estimated Base + Arg (D=2)	Estimated Base Only
ChatGPT	ChatGPT	0.55	0.505	0.505
ChatGPT	OPRO	**0.575**	0.5	0.505
ChatGPT	analyst	0.55	**0.55**	**0.515**
OPRO	ChatGPT	0.47	0.465	0.505
OPRO	OPRO	0.465	0.485	0.505
OPRO	analyst	0.525	0.545	**0.515**
debater	ChatGPT	0.525	0.48	0.505
debater	OPRO	0.53	0.5	0.505
debater	analyst	0.515	0.495	**0.515**

Table 15: Mistral prompt experiment results, 4bit, depth 2 - SQA on 200 datapoints

Table 16: Brier scores of a baseline and four variations of our argumentative method on claim verification tasks. The best performing method for each model-dataset combination is indicated in bold.

		Est. Confidence	**0.5 Base Arg (D=1)**	**0.5 Base Arg (D=2)**	**Est. Base Arg (D=1)**	**Est. Base Arg (D=2)**
Truthful Claim	Mistral	0.205	0.21	0.215	**0.195**	0.198
	Mixtral	0.169	0.187	0.195	**0.153**	0.155
	Gemma 7B	**0.238**	0.272	0.294	0.273	0.286
	GPT-3.5-turbo	0.205	0.219	0.222	0.191	**0.183**
Strategy Claim	Mistral	0.335	0.266	**0.258**	0.321	0.321
	Mixtral	0.258	0.23	**0.229**	0.26	0.259
	Gemma 7B	0.304	**0.27**	0.286	0.321	0.332
	GPT-3.5-turbo	0.245	0.256	**0.243**	0.252	**0.243**
Med Claim	Mistral	0.353	0.362	**0.305**	0.378	0.331
	Mixtral	0.268	0.282	0.257	0.273	**0.256**
	Gemma 7B	**0.302**	0.373	0.42	0.41	0.443
	GPT-3.5-turbo	0.314	**0.245**	0.248	0.305	0.315

Table 17: AUC of a baseline and four variations of our argumentative method on claim verification tasks. The best performing method for each model-dataset combination is indicated in bold.

		Est. Confidence	**0.5 Base Arg (D=1)**	**0.5 Base Arg (D=2)**	**Est. Base Arg (D=1)**	**Est. Base Arg (D=2)**
Truthful Claim	Mistral	0.792	0.748	0.741	**0.809**	0.806
	Mixtral	0.831	0.834	0.831	**0.852**	0.85
	Gemma 7B	**0.691**	0.637	0.625	**0.691**	0.675
	GPT-3.5-turbo	0.795	0.75	0.735	0.807	0.825
Strategy Claim	Mistral	0.645	0.643	0.641	**0.656**	0.653
	Mixtral	0.727	0.759	0.749	**0.759**	0.753
	Gemma 7B	0.584	0.551	0.551	**0.593**	0.592
	GPT-3.5-turbo	**0.747**	0.654	0.655	0.741	0.741
Med Claim	Mistral	0.575	0.475	0.514	0.563	**0.584**
	Mixtral	0.659	0.615	0.608	**0.671**	0.67
	Gemma 7B	0.532	0.528	0.523	**0.534**	**0.534**
	GPT-3.5-turbo	0.638	0.601	0.585	0.644	**0.645**

G Contestation Examples

In this section we show different ways of contesting our model. Figure 16 and Figure 17 are illustrations of different methods by which the output of our system can be contested, and modified, by human users.

H Licenses

Following are the licenses for all the datasets we adapt, and models we experiment with. The purposes we use the models and data are all covered by their respective licenses. Datasets: TruthfulQA - Apache 2.0, StrategyQA - MIT, MedQA - N/A. Models: Mistral - Apache 2.0, Mixtral - Apache 2.0, Gemma - Apache 2.0, GPT-3.5-turbo/4 - Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International.

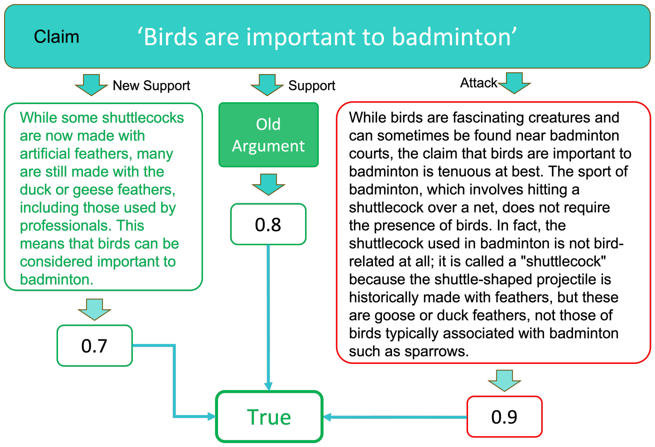

Claim
'Birds are important to badminton'
New Support
Support
Attack
While some shuttlecocks are now made with artificial feathers, many are still made with the duck or geese feathers, including those used by professionals. This means that birds can be considered important to badminton.
Old Argument
0.8
While birds are fascinating creatures and can sometimes be found near badminton courts, the claim that birds are important to badminton is tenuous at best. The sport of badminton, which involves hitting a shuttlecock over a net, does not require the presence of birds. In fact, the shuttlecock used in badminton is not bird-related at all; it is called a "shuttlecock" because the shuttle-shaped projectile is historically made with feathers, but these are goose or duck feathers, not those of birds typically associated with badminton such as sparrows.
0.7
True
0.9

Figure 16: An illustration of a user adding an additional supporting argument. Please note that the 'addition sign' is purely illustrative, and not indicative of the actual process that takes place in the argument semantics. However, the effect of changing the classification from false to true is a realistic demonstration of what would happen in this case.

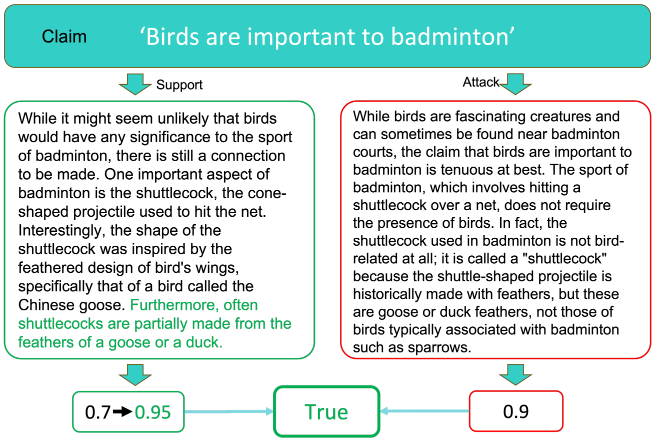

Claim
'Birds are important to badminton'
Support
Attack
While it might seem unlikely that birds would have any significance to the sport of badminton, there is still a connection to be made. One important aspect of badminton is the shuttlecock, the cone-shaped projectile used to hit the net. Interestingly, the shape of the shuttlecock was inspired by the feathered design of bird's wings, specifically that of a bird called the Chinese goose. Furthermore, often shuttlecocks are partially made from the feathers of a goose or a duck.
While birds are fascinating creatures and can sometimes be found near badminton courts, the claim that birds are important to badminton is tenuous at best. The sport of badminton, which involves hitting a shuttlecock over a net, does not require the presence of birds. In fact, the shuttlecock used in badminton is not bird-related at all; it is called a "shuttlecock" because the shuttle-shaped projectile is historically made with feathers, but these are goose or duck feathers, not those of birds typically associated with badminton such as sparrows.
0.7→0.95
True
0.9

Figure 17: An illustration of a user modifying a supporting information with extra information. Due to the improvement in the argument, the argument strength is increased, leading to a change in classification from false to true.