Argumentative Large Language Models for Explainable and Contestable Claim Verification

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 profusion of knowledge encoded in large language models (LLMs) and their ability to apply this knowledge zero-shot in a range of settings makes them promising candidates for use in decision-making. However, they are currently limited by their inability to provide outputs which can be faithfully explained and effectively contested to correct mistakes. In this paper, we attempt to reconcile these strengths and weaknesses by introducing *argumentative LLMs (ArgLLMs)*, a method for augmenting LLMs with argumentative reasoning. Concretely, ArgLLMs construct argumentation frameworks, which then serve as the basis for formal reasoning in support of decision-making. The interpretable nature of these argumentation frameworks and formal reasoning means that any decision made by ArgLLMs may be explained and contested. We evaluate ArgLLMs' performance experimentally in comparison with state-of-the-art techniques, in the context of the decision-making task of claim verification. We also define novel properties to characterise contestability and assess ArgLLMs formally in terms of these properties.

Code and Datasets — https://github.com/CLArg-group/argumentative-llms

Extended version — https://arxiv.org/abs/2405.02079

1 Introduction

The profusion of knowledge encoded in large language models (LLMs) and their ability to apply this knowledge zero-shot in a range of settings (e.g. as in Brown et al. (2020); Bubeck et al. (2023); Achiam et al. (2023)) makes them promising candidates for supporting automated decision making (Zhang et al. 2023; Ouyang and Li 2023; Wang et al. 2023). However, they are currently limited by their inability to explain their outputs faithfully, i.e. in a way that reflects their "reasoning" and knowledge. LLMs also lack contestability, meaning that there is no mechanism for external agents to reliably dispute and correct the reasoning steps taken by the model: although users may attempt contestation through prompting, the highly stochastic nature of LLMs provides no guarantee of this achieving the intended outcome. These abilities (to be explainable and contestable)

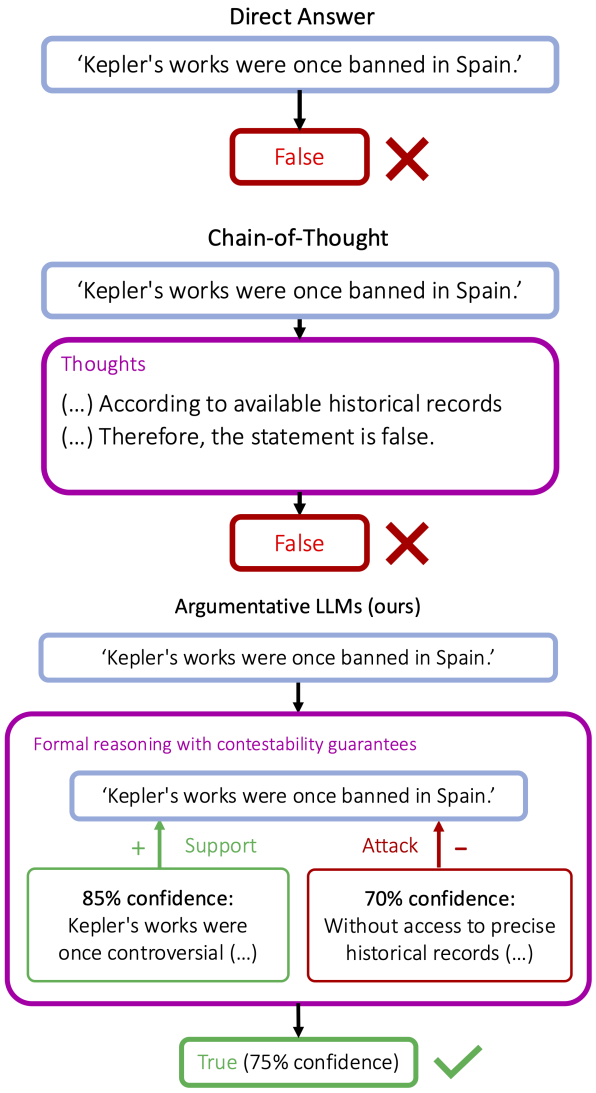

Figure 1: Comparison of our approach (*ArgLLM*, here in combination with Mixtral) with existing alternatives. The example claim is adapted from TruthfulQA.

*These authors contributed equally.

are important for AI systems in general (Henin and Métayer 2022; Lyons, Velloso, and Miller 2021; Leofante et al. 2024) and for LLMs in particular, given their limitations, e.g. hallucinations and logical inconsistencies (Shanahan 2024; Berglund et al. 2023; Fluri, Paleka, and Tramèr 2023). In this paper we explore the following question, grounded in the context of claim verification as a form of decision-making:

> Can LLMs be augmented with the ability to construct formal arguments in order to give explainable and contestable outputs?

The question is inspired by argumentative interpretations of human reasoning (Mercier and Sperber 2011, 2018) and by the fact that formal argumentation (as understood in Atkinson et al. (2017)) has been shown to excel in supporting decision making (Amgoud and Prade 2009) and explainability (Cyras et al. 2021). It is also advocated as a suitable approach towards contestability (Leofante et al. 2024). Further, Hammond and Leake (2023) argue that the inherent limitations of LLMs can be remedied by the use of symbolic AI techniques, which include formal argumentation. In a nutshell, this (also referred to as 'computational argumentation'), represents information in terms of 'arguments' and dialectical relations (of 'attack' and, possibly, 'support') between them; it is equipped with 'semantics' (and algorithms) to reach consensus on conclusions to be formally drawn.

Concretely, we propose *argumentative LLMs (ArgLLMs)*. While existing approaches (Wei et al. 2022; Yao et al. 2023) prompt LLMs to produce 'thoughts' that either enrich the context of the LLMs or provide disparate reasoning steps, ArgLLMs provide 'arguments' for and against particular outputs, in the spirit of Miller (2023), as illustrated for claim verification in Figure 1. This feature of ArgLLMs makes them a natural fit for complex decision-making tasks (such as claim verification), wherein an option must be chosen from a number of possible alternatives. Indeed, in many real-world settings, a particular decision will have both pros and cons, which ArgLLMs formalise and leverage.

Overall, we aim to achieve a broader 'improvement' in the LLMs' reasoning, in the spirit of Liao and Wortman Vaughan (2024): in addition to achieving reasonable performance, we target explainability and contestability. While existing methods for improving the reasoning of LLMs (such as Wei et al. (2022); Lewis et al. (2020)) do not necessitate a direct relationship between the reasoning steps and the final decision (Turpin et al. 2023), our argumentative approach provides this as a feature of the system. Indeed, we use the outputs of LLMs to perform deterministic inference under *gradual semantics* with *quantitative bipolar argumentation frameworks* (QBAFs) (Baroni, Rago, and Toni 2019). QBAFs comprise of arguments, each equipped with an *intrinsic strength*, and attack and support relations. They are inherently interpretable and thus, as stated in Rudin (2019), "provide their own explanations, which are faithful to what the model actually computes". This is analogous to how the rules forming a decision-tree model necessarily constitute a faithful explanation of the model.

Further, ArgLLMs also provide a guarantee of contestability, in that an intervention in the reasoning process by modifying the QBAF (such as by adding support for an argument or by changing its intrinsic strength) will have a measurable effect on the output.

Throughout the paper, we focus on claim verification as the decision-making task of interest. This setting lends itself well to ArgLLMs as claims are often under-determined, so they may not have straightforward truth values. By intrinsically considering both arguments in favour of and in conflict with the truthfulness of claims, ArgLLMs are able to ascertain the best answer given the available evidence. For simplicity, and without loss of generality,[1] we focus on the binary setting, rather than general question-answering as in Wei et al. (2022) and Yao et al. (2023).

In summary, we make the following contributions:

- We define ArgLLMs, a novel interpretable method augmenting LLMs with formal, argumentative reasoning.
- We perform an evaluation of ArgLLMs' claim verification abilities by comparing four variants thereof with three baselines (two based on direct prompting, plus the chain-of-thought approach (Wei et al. 2022)), on three novel claim verification datasets, adapted from existing datasets (TruthfulQA (Lin, Hilton, and Evans 2021), StrategyQA (Geva et al. 2021) and MedQA (Jin et al. 2020)). The evaluation shows that ArgLLMs deliver performance comparable to the baselines with the added benefit of being faithfully explainable.
- We formally characterise contestability properties for ArgLLMs, showing formally in which sense ArgLLMs may be deemed contestable.

Figure 2 gives an overview of our approach, which can be deployed in combination with any LLM.

2 Related Work

Existing research into LLM optimisation has primarily focused on improving performance on reasoning benchmarks, with improvements in explainability treated as a desirable byproduct. This existing research can be coarsely divided into approaches which exclusively focus on prompt-engineering (Wei et al. 2022; Yang et al. 2024), 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 method is more closely aligned to the latter, resulting in symbolic, deterministically evaluable arguments as its output.

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 (using either few-shot examples or 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

[1] ArgLLMs can be easily extended to the case where there are multiple possible options by generating a set of candidate answers and applying our method to each of these.

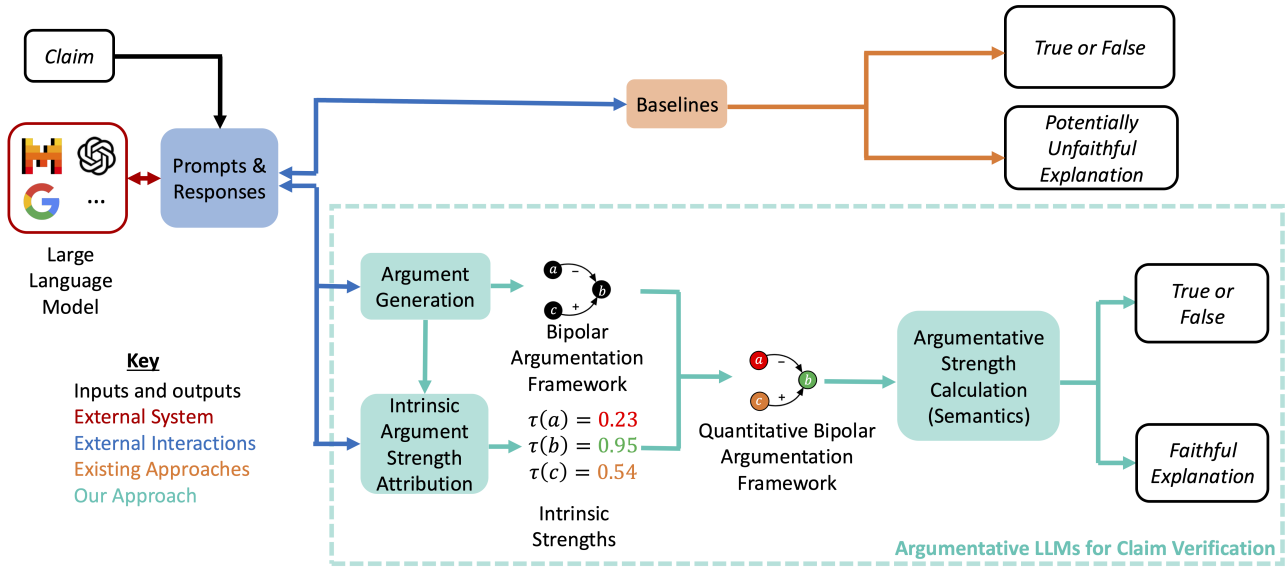

Figure 2: Pipeline for ArgLLMs (in comparison with baselines, see §4 and §5 for the details).

follow from each other (Turpin et al. 2023; 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.

Tafjord, Mishra, and Clark (2022) train an extra model to measure entailment between reasoning steps and the final output. While this provides some level of assurance that the output will be causally related to the explanation, it relies on an external opaque model, which cannot provide any guarantees. Instead, our method builds an argumentation framework which guarantees a resolution solely based on the constituent entities, offering the added advantages of innately dealing with conflict and being amenable to contestation.

Other related approaches are *tree-of-thought* (Yao et al. 2023) and *graph-of-thoughts* (Besta et al. 2024). Similarly to our method, they 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, ArgLLMs permit a comprehensive and explainable reasoning process to take place concerning a single claim, which may be controversial or highly complex. 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'). In addition, to the best of our knowledge, no existing approach attempts to integrate any formal guarantees of contestability, which ArgLLMs naturally provide.

LLMs have also been used to extract logical formulas from text (Ishay, Yang, and Lee 2023), by combining LLMs' prowess in dealing with natural language with the complex reasoning capabilities of reasoning by answer set programming. ArgLLMs differ from this approach in that we extract symbolic representations directly from LLMs (rather than from externally provided text) and deploy a gradual approach to reasoning (rather than a model-theoretic one), with the potential to handle uncertainty.

Also relevant to our work are approaches focused specifically on automated claim verification using LLMs, such as FOLK (Wang and Shu 2023) or HiSS (Zhang and Gao 2023). However, in contrast with ArgLLMs, these works rely on access to external sources or knowledge bases.

3 Preliminaries

Claim Verification We consider two types of claim verification: *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 $v(c) = 0$ denotes 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 $c \mid i$ is also assessed in a binary manner, expressed as $v(c \mid i) \in \{0, 1\}$. Here, similarly, $v(c \mid i) = 1$ denotes the claim c, given the context i, is true, while $v(c \mid i) = 0$ denotes it is false. For ease of reference, we use x to denote either c or $c \mid i$.

Computational Argumentation A QBAF (Baroni, Rago, and Toni 2019) is a quadruple $\langle \mathcal{A}, \mathcal{R}^-, \mathcal{R}^+, \tau \rangle$ comprising: a set of *arguments* $\mathcal{A}$; binary, directed relations of *attack* $\mathcal{R}^- \subseteq \mathcal{A} \times \mathcal{A}$ and *support* $\mathcal{R}^+ \subseteq \mathcal{A} \times \mathcal{A}$, where $\mathcal{R}^- \cap \mathcal{R}^+ = \emptyset$; and a total function $\tau : \mathcal{A} \to [0, 1]$, where for any $\alpha \in \mathcal{A}$,

$\tau(\alpha)$ is the *base score* of α.[2] For any argument $\alpha \in \mathcal{A}$, we use $\mathcal{R}^-(\alpha) = \{\beta \in \mathcal{A} | (\beta, \alpha) \in \mathcal{R}^-\}$ to refer to the *attackers* of α and $\mathcal{R}^+(\alpha) = \{\beta \in \mathcal{A} | (\beta, \alpha) \in \mathcal{R}^+\}$ to refer to the *supporters* of α. Arguments in QBAFs may be evaluated by a *gradual semantics* (Baroni, Rago, and Toni 2019), i.e. a total function $\sigma : \mathcal{A} \to [0, 1]$ which, for any $\alpha \in \mathcal{A}$, assigns a *strength* $\sigma(\alpha)$ to α.[3] While $\tau(\alpha)$ can be seen as the *intrinsic strength* for α, the strength $\sigma(\alpha)$ can be seen as 'dialectical', following the debate captured by $\mathcal{R}^-$ and $\mathcal{R}^+$. If we ignore τ in a QBAF, we obtain a *bipolar argumentation framework (BAF)* (Cayrol and Lagasquie-Schiex 2005).

One gradual semantics, the *discontinuity-free quantitative argumentation debate* (DF-QuAD) algorithm (Rago et al. 2016), is such that, for a given QBAF $\langle \mathcal{A}, \mathcal{R}^-, \mathcal{R}^+, \tau \rangle$, for any $\alpha \in \mathcal{A}$ with $n \geq 0$ attackers with strengths $v_1, \ldots, v_n$, $m \geq 0$ supporters with strengths $v'_1, \ldots, v'_m$ and $\tau(\alpha) = v_0$, $\sigma(\alpha) = \mathcal{C}(v_0, \mathcal{F}(v_1, \ldots, v_n), \mathcal{F}(v'_1, \ldots, v'_m))$, where $\mathcal{C}$ and $\mathcal{F}$ are 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 $v_1, \ldots, v_n$, if $n = 0$ then $\mathcal{F}(v_1, \ldots, v_n) = 0$, otherwise $\mathcal{F}(v_1, \ldots, v_n) = 1 - \prod_{i=1}^{n}(|1 - v_i|)$. Note that DF-QuAD (like all other gradual semantics) is deterministic.

In this paper, we assume all (Q)BAFs are restricted, similarly to (Rago, Li, and Toni 2023), as follows. For $\mathcal{Q} = \langle \mathcal{A}, \mathcal{R}^-, \mathcal{R}^+, \tau \rangle$ a QBAF and $\mathcal{B} = \langle \mathcal{A}, \mathcal{R}^-, \mathcal{R}^+ \rangle$ a BAF, for any $\alpha, \beta \in \mathcal{A}$, let a *path* from α to β be $p = \langle (\alpha_0, \alpha_1), \ldots, (\alpha_{n-1}, \alpha_n) \rangle$ for some $n > 0$ (referred to as the *length* of p, denoted $|p|$) where $\alpha_0 = \alpha$, $\alpha_n = \beta$ and, for any $1 \leq i \leq n$, $(\alpha_{i-1}, \alpha_i) \in \mathcal{R}^- \cup \mathcal{R}^+$. Let $\mathsf{paths}(\alpha, \beta)$ and $|\mathsf{paths}(\alpha, \beta)|$ indicate the set of all paths from α to β, and the number of paths in $\mathsf{paths}(\alpha, \beta)$, respectively. Then, for $\alpha^* \in \mathcal{A}$, $\mathcal{Q}/\mathcal{B}$ is a *QBAF/BAF for* α^* iff

i) $\forall \alpha \in \mathcal{A}$, $\mathsf{paths}(\alpha^*, \alpha) = \emptyset$ and $\mathsf{paths}(\alpha, \alpha) = \emptyset$;

ii) $\forall \alpha \in \mathcal{A} \setminus \{\alpha^*\}$, $|\mathsf{paths}(\alpha, \alpha^*)| = 1$.

Intuitively, (Q)BAFs for α^* can be seen as trees with root α^*. We call *leaves* any unattacked and unsupported arguments in (Q)BAFs (i.e. α is a leaf if $\mathcal{R}^-(\alpha) \cup \mathcal{R}^+(\alpha) = \emptyset$).

We also use *pro* and *con* arguments in QBAFs as in Rago, Li, and Toni (2023). Let $\mathcal{Q}$ be a QBAF for α^*. Then, the *pro arguments* and *con arguments* for $\mathcal{Q}$ are, respectively:

- $\mathsf{pro}(\mathcal{Q}) = \{\alpha \in \mathcal{A} | \exists p \in \mathsf{paths}(\alpha, \alpha^*), \text{with} |p \cap \mathcal{R}^-| \text{even}\}$;
- $\mathsf{con}(\mathcal{Q}) = \{\alpha \in \mathcal{A} | \exists p \in \mathsf{paths}(\alpha, \alpha^*), \text{with} |p \cap \mathcal{R}^-| \text{odd}\}$.

In the remainder, unless specified otherwise, we let $\mathcal{Q}$ denote the QBAF $\langle \mathcal{A}, \mathcal{R}^-, \mathcal{R}^+, \tau \rangle$. For any gradual semantics σ, we let $\sigma_{\mathcal{Q}}(\alpha)$ denote the strength of $\alpha \in \mathcal{A}$ in $\mathcal{Q}$.

4 Argumentative LLMs

Our method is based on extracting interpretable QBAFs from LLMs and formally reasoning with them with gradual semantics. As shown in Figure 2, there are three integral components in ArgLLMs: argument generation, intrinsic argument strength attribution and argumentative strength calculation. We outline them here, for any input claim x (see §3), and give implementation details for claim verification in §5.

Argument generation Argument generation returns a BAF $\mathcal{B}$ for x (see §3):

$$\Gamma(x, G, \theta) \to \mathcal{B}.$$

Here, Γ is a *generating function*, G denotes the underpinning *generative model*, and θ represents the *parameters* associated with the argument generation, e.g. the split between attackers and supporters for each argument and the desired *width* and *depth* of the BAF, determined respectively by the number and length of paths from x to leaves in $\mathcal{B}$. For illustration, in Figure 1, we set θ to generate a BAF with width 2 and depth 1. Note that, in general, θ may influence G to determine depth and width dynamically, e.g. based on the semantic originality of the incrementally generated arguments.

Previous work has shown 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 (in conjunction with prompts) as the underpinning generative model G, to produce both attacking and supporting arguments to populate a BAF.

Intrinsic argument strength attribution Intrinsic argument strength attribution adds base scores to the BAF, to obtain a QBAF $\mathcal{Q}$ for x:

$$\mathcal{E}(\mathcal{B}, E) \to \mathcal{Q}.$$

Here, $\mathcal{E}$ is an *evaluative function* and E is an *evaluative model*. For illustration, in Figure 1, we obtain values of 0.85 for the supporter and 0.70 for the attacker (reported as confidence values in the visualisation shown).

There have been a number of previous attempts to assess the quality of arguments, which can be seen as a proxy for their intrinsic strength. 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). However, producing such data is resource intensive. Instead, in all instantiations of ArgLLMs in this paper, we rely on the knowledge embedded into the LLMs, using them as the evaluative model, E, zero-shot and without any task-specific fine-tuning. 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 to assess argument quality, as we deem unlikely that either the pre-training or the supervised training stages contained many instances of this fairly niche task. Assigning an intrinsic 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). Thus, using the scores for an objective-driven, empirical task (claim verification), and ascertaining their suitability post-hoc, is perhaps a more effective method of assessing this capacity in LLMs.

[2] The codomain of τ is defined more generally by Baroni, Rago, and Toni (2019), but we use here its most common form.

[3] As with τ, we use the most common codomain for σ.

Argumentative strength calculation Argumentative strength calculation amounts to applying a gradual semantics to resolve the conflicts within the QBAF and obtain an assessment of x:

$$\Sigma(x, \mathcal{Q}, \sigma) \rightarrow \sigma_{\mathcal{Q}}(x).$$

Here, Σ represents the *strength calculation function*, and σ (DF-QuAD in the main experiments in §5) is a chosen gradual semantics. The output $\sigma_{\mathcal{Q}}(x)$ represents the evaluated strength (or degree of acceptability) of the input x according to $\sigma_{\mathcal{Q}}$, taking into account the structural (dialectical) and quantitative (base score) aspects of the QBAF. For illustration, in Figure 1, we obtain a value of 0.75 for the claim (reported as confidence value in the figure).

The strength of x can be seen as the final output of ArgLLMs but it can also be used to determine further outputs, e.g., in the case of claim verification, to determine whether the input claim x is true or false. In Figure 1, given that the computed value for the claim is above 0.5, the ArgLLM returns the label True (we use 0.5 as threshold in all our experiments in §5). While we did not empirically validate or tune the value of this threshold, we chose it for suitability for the task of claim verification and ease of empirical validation.

In addition to the computed value/decision, ArgLLMs also return the computed QBAF as the reasoning trail for the outputs and thus can be deemed to be interpretable. The output can be directly attributed to the generated arguments and their associated strengths, and the QBAF can serve as an explanation for why a decision has been made. In addition, should the QBAF be large, explanations in various forms can be extracted from the QBAFs, e.g. argument attributions (Yin, Potyka, and Toni 2023; Kampik et al. 2024), relation attributions (Yin, Potyka, and Toni 2024b) and counterfactuals (Yin, Potyka, and Toni 2024a). These explanation would allow human users to manually check only the most significant components of the QBAFs, and their respective scores. This makes ArgLLMs amenable to human oversight, even in cases where there may be very many arguments.

5 Performance Evaluation

In this section, we describe the experiments evaluating the reasoning performance of ArgLLMs on our selected task of claim verification.[4] We derive claims from existing QA datasets (for an example, see Figure 1). Also, we experiment with a range of parameters, θ, in the argument generation component and with several choices for E. We report results for DF-QuAD as σ in the Σ component, but include results for another choice of gradual semantics (quadratic energy (Potyka 2018)) in the Supplementary Material in the extended version (referred to in the remainder in short as SM).

5.1 Experimental Set-up

Datasets We focus on three claim verification datasets adapted from existing Q/A datasets: TruthfulClaim (adapted from TruthfulQA (Lin, Hilton, and Evans 2021)), StrategyClaim (adapted from StrategyQA (Geva et al. 2021)) and MedClaim (adapted from MedQA (Jin et al. 2020)). For the purposes of our experiments, we transformed the question-answer pairs in each of the original datasets into claims with true/false labels. We did not use the original data directly as the LLMs we experimented with did not perform adequately when evaluating the validity of question-answer pairs rather than self-contained claims. To generate the claims for each Q/A pair, we used LLMs, manually checking and amending the results to ensure their correctness. The three datasets have different flavours, reflecting the original datasets': TruthfulQA was curated specifically to evaluate if LLMs are able to identify truthful answers without being deceived by common misconceptions and falsehoods; StrategyQA was designed to evaluate whether LLMs can strategically reason; and MedQA evaluates models on claims associated with medical problems from the professional medical board exams. TruthfulClaim and StrategyClaim embed unconditioned claim verification, whereas MedClaim embeds conditioned claim verification. For our experiments, we select 700 claims from TruthfulClaim and StrategyClaim (200 for the prompt selection experiments, as discussed later, 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 restriction to subsets of the datasets is due to the resource cost with LLMs.

LLMs We use seven main models: Mistral (Mistral-7B-Instruct-v0.2) (Jiang et al. 2023), Mixtral (Mixtral-8x7B-Instruct-v0.1) (Jiang et al. 2024), Gemma (gemma-7b-it) (Mesnard et al. 2024), Gemma 2 (gemma-2-9b-it) (Riviere et al. 2024), Llama 3 (Meta-Llama-3-8B-Instruct) (Dubey et al. 2024), GPT-3.5-turbo (GPT-3.5-turbo-0125) (Brown et al. 2020) and GPT-4o mini (gpt-4o-mini) (OpenAI 2024). We chose Mistral, Mixtral, Gemma and Llama 3 as they were among the most known and best-performing open-source[5] models of reasonable size at the time of our evaluation. In order to reduce the computational costs of running the open-source models, we quantise them to 4 bits (Dettmers et al. 2023) for both the baselines and our method. As representatives of models with proprietary weights, we chose GPT-3.5-turbo and GPT-4o mini, since they had the best performance/cost trade-off. For all the models, we use parameters: temperature 0.7, max new tokens for arguments 128, max new tokens for baselines 768, top-p 0.95 and repetition penalty 1.0.

Baselines We compare our method with three baselines: direct questioning ("Direct Question" in short), estimated confidence ("Est. Confidence" in short) and questioning with chain-of-thought (Wei et al. 2022) ("Chain-of-Thought" in short). The direct questioning baseline consists of directly asking the LLMs if the given claim is true or false by prompting, where we constrain the output of the open-source LLMs to true or false. For the estimated confidence,

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

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

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 Γ. {"supporting"/"attacking"} and {"support"/"attack"} are determined by θ.

we ask the LLMs for a numerical confidence score for the given claim, ranging from 0 to 100. Similarly to direct questioning, we constrain the output of the open-source LLMs to the associated values. We consider the claim to be predicted as truthful if the outputted confidence is greater than 50 and false otherwise. Finally, Chain-of-Thought breaks down the problem into multiple steps before the final decision is outputted. Then, we pass the reasoning back to the LLM, in a separate context window, to get the final decision.

Prompt Selection We take a principled approach for prompt selection, as 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 devised three different prompts for both ArgLLMs and the baselines.

We evaluate all prompts, and the combinations thereof, in a pilot experiment on two validation sets of 200 samples each, taken from TruthfulClaim and StrategyClaim, using Mistral and Mixtral models (selected as two substantially different open-source models). In this evaluation, we separately considered the prompts for the baselines and ArgLLM components. We find a large variation in performance for any prompt with any given dataset and model combination, both for ArgLLMs and the baselines. We choose the highest average scoring prompts over all tested models and datasets. The results of these experiments are in the SM.

ArgLLM Variations In our experiments, we use four different variations of ArgLLMs. For all the variations, we use the same prompts for argument generation (Γ), given in Figure 3, and intrinsic argument strength attribution ($\mathcal{E}$), given in Figure 4. In order to assess accuracy for claim verification we use a similar threshold as for the Est. Confidence baseline — if the the input claim's final strength is greater than 0.5 it is classified as true, and otherwise as false.

The four variations we consider arise from the hyperparameters θ and the choice of intrinsic strength for the input claim by $\mathcal{E}$. For θ, we consider two options — Depth=1 and Depth=2: BAFs with Depth=1 are composed of the claim along with two generated arguments (a supporter and an attacker); in BAFs with Depth=2, we recursively generate a supporter and an attacker for each of the arguments in

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 $\mathcal{E}$. {"in favour of"/"against"} and {"supports"/"refutes"} depend on the type of {argument}.

Depth=1, giving seven arguments in total. The input claim can either be given a neutral score of 0.5 (0.5 Base Arg in short), which means that its final strength will be solely determined by the remaining arguments in the QBAF, or can be assigned a confidence score as for the other arguments (Est. Base Arg in short). By considering the combinations of the above choices, we get the $2 \times 2 = 4$ variations.

5.2 Evaluation Results

Generally, accuracy (see Table 1) varied on different datasets across different LLMs. However, Chain-of-Thought, Direct Questioning, and ArgLLMs with estimated base score for the topic argument performed best.

Table 1 also includes results on an extra MedClaim experiment with GPT-4o (gpt-4o-2024-08-06), testing the performance of ArgLLMs when used with a larger model. The results indicate that Direct Question and Chain-of-Thought had the best accuracy, followed by Est. Base Arg (D=1). This result might indicate that even powerful models have relatively limited zero-shot capabilities in argument generation and estimation, highlighting the potential for task-specific finetuning.

Overall, as the table shows, ArgLLMs perform comparably with Chain-of-Thought. However, perhaps the most important features of ArgLLMs cannot be adequately captured by quantitative metrics such as accuracy. One such feature is that the outputs generated are faithful explanations in

		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	0.73	0.73	0.75	0.65	0.67	**0.76**	0.75
	Mixtral	0.77	0.77	0.76	0.72	0.72	**0.81**	0.78
	Gemma 7B	0.65	0.62	**0.68**	0.64	0.62	0.63	0.62
	Gemma 2 9B	**0.78**	0.74	**0.78**	0.68	0.68	0.73	0.73
	Llama 3 8B	0.66	0.69	**0.70**	0.63	0.61	0.68	0.69
	GPT-3.5-turbo	0.70	0.73	**0.74**	0.60	0.64	0.73	0.72
	GPT-4o mini	0.78	0.79	0.79	0.74	0.78	**0.81**	**0.81**
Strategy Claim	Mistral	0.60	0.61	**0.68**	0.58	0.61	0.62	0.60
	Mixtral	0.68	0.67	0.64	0.62	0.66	0.68	**0.70**
	Gemma 7B	0.55	0.56	0.58	0.56	**0.59**	0.57	0.57
	Gemma 2 9B	0.71	0.71	**0.72**	0.65	0.67	0.71	0.70
	Llama 3 8B	0.61	0.59	**0.65**	0.54	0.53	0.61	0.58
	GPT-3.5-turbo	**0.73**	0.70	0.72	0.56	0.57	0.70	0.67
	GPT-4o mini	**0.77**	0.74	0.75	0.66	0.67	0.74	0.75
Med Claim	Mistral	0.55	0.57	**0.61**	0.50	0.51	0.53	0.52
	Mixtral	0.60	0.62	0.61	0.59	0.56	0.61	**0.63**
	Gemma 7B	0.52	0.53	**0.55**	0.51	0.52	0.52	0.52
	Gemma 2 9B	0.61	0.58	**0.62**	0.57	0.59	0.60	0.59
	Llama 3 8B	0.54	0.56	**0.58**	0.51	0.53	0.53	0.52
	GPT-3.5-turbo	**0.67**	0.57	**0.67**	0.56	0.55	0.57	0.57
	GPT-4o mini	**0.74**	0.71	0.72	0.62	0.65	0.71	0.71
	GPT-4o	**0.85**	0.82	**0.85**	0.74	0.76	0.83	0.80

Table 1: Accuracy of three baselines and four variations of ArgLLMs (all using σ=DF-QuAD) on claim verification tasks. The best result for each model-dataset combination is indicated in bold. Values within 0.03 of the best results are underlined.

terms of arguments, from which decisions can be deterministically derived, rather than decisions alone. While Chain-of-Thought can also be seen as providing reasons, these may not be faithful to the true, stochastic reasoning process of the model (as noted by Turpin et al. (2023)). In contrast, the final decision outputted by ArgLLMs is faithfully determined by the arguments and the argumentation semantics.

6 Contestability

Another unique feature of ArgLLMs is contestability. The (faithful) explainability of ArgLLMs offers plentiful opportunity to disagree with the reasoning, either in terms of the arguments generated being relevant or true, or the intrinsic strengths that have been attributed to them being representative of the extent to which they support or attack their 'parent' argument. For illustration, consider Figure 5. Here, a user presented with the output of the ArgLLM may disagree with some parts of the argument attacking the claim. This may lead the user to decrease the argument intrinsic strength and contest the False label originally determined by the ArgLLM. A further illustration can be found in the SM.

Here, we frame formally the contestability functionality for ArgLLMs, by proposing two notions of contestability in the QBAF setting. This formal analysis applies to QBAFs of any depth and width. Since QBAFs are composed of arguments/relations and base scores (intrinsic argument strength), ArgLLMs can be contested on two fronts.

We start with a property governing contestability on base scores:

Property 1. *A gradual semantics σ satisfies* base score contestability *iff for any QBAF $\mathcal{Q}$ for α^*, for any $\mathcal{Q}'$ with $\mathcal{A}' = \mathcal{A}$, $\mathcal{R}^{-\prime} = \mathcal{R}^-$, $\mathcal{R}^{+\prime} = \mathcal{R}^+$, and, for $\beta \in \mathcal{A}$, $\tau(\beta) < \tau'(\beta)$ while $\tau(\gamma) = \tau'(\gamma)$ for all $\gamma \in \mathcal{A} \setminus \{\beta\}$:*

- *if $\beta \in \mathsf{pro}(\mathcal{Q})$, then $\sigma_{\mathcal{Q}}(\alpha^*) \leq \sigma_{\mathcal{Q}'}(\alpha^*)$;*
- *if $\beta \in \mathsf{con}(\mathcal{Q})$, then $\sigma_{\mathcal{Q}}(\alpha^*) \geq \sigma_{\mathcal{Q}'}(\alpha^*)$.*

This property formalises that increasing the base score of pro/con arguments in a QBAF for an argument will not decrease/increase (respectively) the argument's final strength[6].

Example 1. *Let $\mathcal{Q}$ be as in Figure 6, and σ be DF-QuAD. Since there is only one (odd number) attack from β to α, $\beta \in \mathsf{pro}(\mathcal{Q})$. Then, increasing $\tau(\beta)$ will not decrease $\sigma_{\mathcal{Q}}(\alpha)$.*

The next property states that the output of ArgLLMs can be contested by adding or removing arguments and relations.

Property 2. *A gradual semantics σ satisfies* argument relation contestability *iff for any QBAF $\mathcal{Q}$ for α^*, for any $\mathcal{Q}'$ with $\mathcal{A}' = \mathcal{A} \cup \{\beta\}$, $\mathcal{R}^{-\prime} \cup \mathcal{R}^{+\prime} = \mathcal{R}^- \cup \mathcal{R}^+ \cup \{(\beta, \alpha)\}$ for some $\alpha \in \mathcal{A}$, and $\tau'(\gamma) = \tau(\gamma)$ for all $\gamma \in \mathcal{A}$:*

- *if $\beta \in \mathsf{pro}(\mathcal{Q}')$, then $\sigma_{\mathcal{Q}}(\alpha^*) \leq \sigma_{\mathcal{Q}'}(\alpha^*)$;*
- *if $\beta \in \mathsf{con}(\mathcal{Q}')$, then $\sigma_{\mathcal{Q}}(\alpha^*) \geq \sigma_{\mathcal{Q}'}(\alpha^*)$.*

Example 2. *Let $\mathcal{Q}$ be as in Figure 6, and σ be DF-QuAD. If we contest $\mathcal{Q}$ by adding an attacker δ to γ (to obtain $\mathcal{Q}'$ in Figure 7), then $\delta \in \mathsf{pro}(\mathcal{Q}')$ because there are two (even number) attacks from δ to α. Then, $\sigma_{\mathcal{Q}}(\alpha) \leq \sigma_{\mathcal{Q}'}(\alpha)$.*

We now give a theoretical result demonstrating DF-QuAD's suitability for providing contestability in ArgLLMs (the proof can be found in the SM).

Proposition 1. *DF-QuAD satisfies base score contestability and argument relation contestability.*

[6]Pro/con arguments no longer affect $\sigma_{\mathcal{Q}}(\alpha^*)$ when it is 1 or 0.

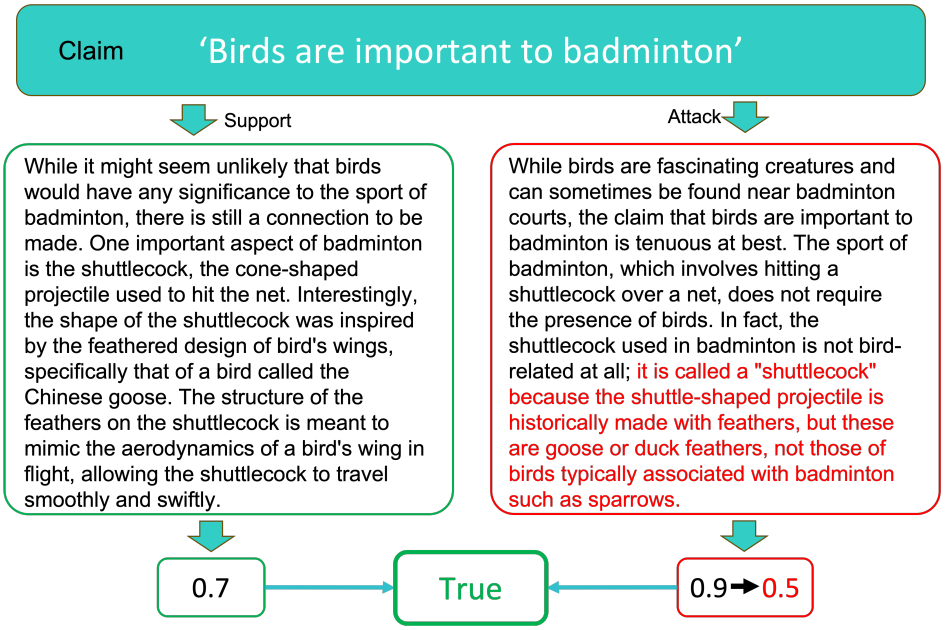

Figure 5: An example of contestation, in the ArgLLM with Mixtral, for a claim taken from StrategyClaim. Before contestation, the claim was (incorrectly) classified as False, but after contesting the intrinsic strength of the attacking argument from 0.9 to 0.5 (citing the fallacious reasoning highlighted in red), the correct True classification results.

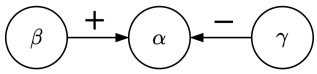

Figure 6: Base score contestability (Example 1).

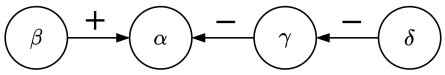

Figure 7: Argument relation contestability (Example 2).

Note that stronger versions of Properties 1 and 2 can be defined, with strict inequalities, satisfied by the quadratic energy model gradual semantics (Potyka 2018) (see the SM).

7 Conclusions & Future Work

We have proposed ArgLLMs, harnessing the knowledge encoded in any LLMs to obtain interpretable outputs that can be faithfully explained and provably contested - without requiring any fine-tuning or external resources. We have shown that ArgLLMs's added value does not lead to compromising prediction accuracy in the task of claim verification in comparison with state-of-the-art prompting methods.

Our work opens several avenues for future study. Sampling multiple outputs of the same LLM or taking the weighted value of the relevant logits in the final layer could provide more sophisticated means for the intrinsic argument strength attribution component. Furthermore, an adapted version of the 'semantic uncertainty' (Kuhn, Gal, and Farquhar 2023) method may be devised, wherein one directly clusters semantically similar sampled arguments, rather than having to prompt models for numerical scores.

Another potential direction for the first two components, argument generation and intrinsic strength attribution, is the ensembling of many different LLMs, to harness the heterogeneous knowledge encoded therein. In a similar vein, using information retrieval or retrieval augmented generation (Lewis et al. 2020) may improve the generated QBAFs. We also plan to conduct more experiments on how tuning hyperparameters, e.g. temperature, would impact performance.

It would also be interesting to study the generated QBAFs in terms of the properties for general argumentative explanations in the spirit of Kotonya and Toni (2024), and to consider other formal properties of contestability and results for other gradual semantics for QBAFs. We also plan to undertake human evaluations of ArgLLMs. Finally, the transparency and contestability properties of ArgLLMs make them an ideal candidate for highly complex, uncertain and high-stakes scenarios, including business, medical or legal decision-making where ArgLLMs could be used in conjunction with domain-experts providing arguments of their own.

Acknowledgments

This research was partially supported by ERC under the EU's Horizon 2020 research and innovation programme (grant agreement No. 101020934, ADIX), by J.P. Morgan and the Royal Academy of Engineering under the Research Chairs and Senior Research Fellowships scheme, by Imperial College through an Imperial College Research Fellowship and by UKRI through the CDT in Safe and Trusted Artificial Intelligence (Grant No. EP/S023356/1).

References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. GPT-4 Technical Report. *CoRR*, abs/2303.08774.

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

Atkinson, K.; Baroni, P.; Giacomin, M.; Hunter, A.; Prakken, H.; Reed, C.; Simari, G. R.; Thimm, M.; and Villata, S. 2017. Towards Artificial Argumentation. *AI Mag.*, 38(3): 25–36.

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

Berglund, L.; Tong, M.; Kaufmann, M.; Balesni, M.; Stickland, A. C.; Korbak, T.; and Evans, O. 2023. The Reversal Curse: LLMs trained on "A is B" fail to learn "B is A". *CoRR*, abs/2309.12288.

Besta, M.; Blach, N.; Kubicek, A.; Gerstenberger, R.; Podstawski, M.; Gianinazzi, L.; Gajda, J.; Lehmann, T.; Niewiadomski, H.; Nyczyk, P.; and Hoefler, T. 2024. Graph of Thoughts: Solving Elaborate Problems with Large Language Models. In *AAAI*, 17682–17690.

Brown, T. B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language Models are Few-Shot Learners. In *NeurIPS*.

Bubeck, S.; Chandrasekaran, V.; Eldan, R.; Gehrke, J.; Horvitz, E.; Kamar, E.; Lee, P.; Lee, Y. T.; Li, Y.; Lundberg, S. M.; et al. 2023. Sparks of Artificial General Intelligence: Early experiments with GPT-4. *CoRR*, abs/2303.12712.

Cayrol, C.; and Lagasquie-Schiex, M. 2005. On the Acceptability of Arguments in Bipolar Argumentation Frameworks. In *ECSQARU*, 378–389.

Chen, G.; Cheng, L.; Tuan, L. A.; and Bing, L. 2023. Exploring the Potential of Large Language Models in Computational Argumentation. *CoRR*, abs/2311.09022.

Cyras, K.; Rago, A.; Albini, E.; Baroni, P.; and Toni, F. 2021. Argumentative XAI: A Survey. In *IJCAI*, 4392–4399.

Dettmers, T.; Pagnoni, A.; Holtzman, A.; and Zettlemoyer, L. 2023. QLoRA: Efficient Finetuning of Quantized LLMs. In *NeurIPS*.

Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Yang, A.; Fan, A.; et al. 2024. The Llama 3 Herd of Models. *CoRR*, abs/2407.21783.

Fluri, L.; Paleka, D.; and Tramèr, F. 2023. Evaluating Superhuman Models with Consistency Checks. *CoRR*, abs/2306.09983.

Furman, D. A.; Torres, P.; Rodríguez, J. A.; Letzen, D.; Martinez, M. V.; and Alemany, L. A. 2023. High-quality argumentative information in low resources approaches improve counter-narrative generation. In *EMNLP*, 2942–2956.

Geva, M.; Khashabi, D.; Segal, E.; Khot, T.; Roth, D.; and Berant, J. 2021. Did Aristotle Use a Laptop? A Question Answering Benchmark with Implicit Reasoning Strategies. *CoRR*, abs/2101.02235.

Habernal, I.; and Gurevych, I. 2016. Which argument is more convincing? Analyzing and predicting convincingness of Web arguments using bidirectional LSTM. In *ACL*.

Halawi, D.; Zhang, F.; Yueh-Han, C.; and Steinhardt, J. 2024. Approaching Human-Level Forecasting with Language Models. *CoRR*, abs/2402.18563.

Hammond, K. J.; and Leake, D. B. 2023. Large Language Models Need Symbolic AI. In *NeSy*, 204–209.

Henin, C.; and Métayer, D. L. 2022. Beyond explainability: justifiability and contestability of algorithmic decision systems. *AI Soc.*, 37(4): 1397–1410.

Ishay, A.; Yang, Z.; and Lee, J. 2023. Leveraging Large Language Models to Generate Answer Set Programs. In *KR*, 374–383.

Jiang, A. Q.; Sablayrolles, A.; Mensch, A.; Bamford, C.; Chaplot, D. S.; de Las Casas, D.; Bressand, F.; Lengyel, G.; Lample, G.; Saulnier, L.; et al. 2023. Mistral 7B. *CoRR*, abs/2310.06825.

Jiang, A. Q.; Sablayrolles, A.; Roux, A.; Mensch, A.; Savary, B.; Bamford, C.; Chaplot, D. S.; de Las Casas, D.; Hanna, E. B.; Bressand, F.; et al. 2024. Mixtral of Experts. *CoRR*, abs/2401.04088.

Jin, D.; Pan, E.; Oufattole, N.; Weng, W.; Fang, H.; and Szolovits, P. 2020. What Disease does this Patient Have? A Large-scale Open Domain Question Answering Dataset from Medical Exams. *CoRR*, abs/2009.13081.

Kampik, T.; Potyka, N.; Yin, X.; Cyras, K.; and Toni, F. 2024. Contribution functions for quantitative bipolar argumentation graphs: A principle-based analysis. *Int. J. Approx. Reason.*, 173: 109255.

Kotonya, N.; and Toni, F. 2024. Towards a Framework for Evaluating Explanations in Automated Fact Verification. In *LREC/COLING*, 16364–16377.

Kuhn, L.; Gal, Y.; and Farquhar, S. 2023. Semantic Uncertainty: Linguistic Invariances for Uncertainty Estimation in Natural Language Generation. In *ICLR*.

Lauscher, A.; Ng, L.; Napoles, C.; and Tetreault, J. R. 2020. Rhetoric, Logic, and Dialectic: Advancing Theory-based Argument Quality Assessment in Natural Language Processing. In *COLING*, 4563–4574.

Leofante, F.; Ayoobi, H.; Dejl, A.; Freedman, G.; Gorur, D.; Jiang, J.; Paulino-Passos, G.; Rago, A.; Rapberger, A.; Russo, F.; Yin, X.; Zhang, D.; and Toni, F. 2024. Contestable AI Needs Computational Argumentation. In *Proceedings of the 21st International Conference on Principles of Knowledge Representation and Reasoning*, 888–896.

Lewis, P. S. H.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Küttler, H.; Lewis, M.; Yih, W.; Rocktäschel, T.; Riedel, S.; and Kiela, D. 2020. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *NeurIPS*.

Liao, Q. V.; and Wortman Vaughan, J. 2024. AI Transparency in the Age of LLMs: A Human-Centered Research Roadmap. *Harvard Data Science Review*, (Special Issue 5).

Lin, S.; Hilton, J.; and Evans, O. 2021. TruthfulQA: Measuring How Models Mimic Human Falsehoods. *CoRR*, abs/2109.07958.

Lyons, H.; Velloso, E.; and Miller, T. 2021. Conceptualising Contestability: Perspectives on Contesting Algorithmic Decisions. *Proc. ACM Hum. Comput. Interact.*, 5(CSCW1): 106:1–106:25.

Mercier, H.; and Sperber, D. 2011. Why do humans reason? Arguments for an argumentative theory. *Behavioral and Brain Sciences*, 34(2): 57–74.

Mercier, H.; and Sperber, D. 2018. *The Enigma of Reason*. Penguin.

Mesnard, T.; Hardin, C.; Dadashi, R.; Bhupatiraju, S.; Pathak, S.; Sifre, L.; Rivière, M.; Kale, M. S.; Love, J.; Tafti, P.; et al. 2024. Gemma: Open Models Based on Gemini Research and Technology. *CoRR*, abs/2403.08295.

Miller, T. 2023. Explainable AI is Dead, Long Live Explainable AI!: Hypothesis-driven Decision Support using Evaluative AI. In *FAccT*, 333–342.

OpenAI. 2024. GPT-4o mini: advancing cost-efficient intelligence. https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/.

Ouyang, S.; and Li, L. 2023. AutoPlan: Automatic Planning of Interactive Decision-Making Tasks With Large Language Models. In *EMNLP*, 3114–3128.

Potyka, N. 2018. Continuous Dynamical Systems for Weighted Bipolar Argumentation. In *KR*, 148–157.

Rago, A.; Li, H.; and Toni, F. 2023. Interactive Explanations by Conflict Resolution via Argumentative Exchanges. In *KR*, 582–592.

Rago, A.; Toni, F.; Aurisicchio, M.; and Baroni, P. 2016. Discontinuity-Free Decision Support with Quantitative Argumentation Debates. In *KR*, 63–73.

Riviere, M.; Pathak, S.; Sessa, P. G.; Hardin, C.; Bhupatiraju, S.; Hussenot, L.; Mesnard, T.; Shahriari, B.; Ramé, A.; et al. 2024. Gemma 2: Improving Open Language Models at a Practical Size. *CoRR*, abs/2408.00118.

Rudin, C. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat. Mach. Intell.*, 1(5): 206–215.

Santu, S. K. K.; and Feng, D. 2023. TELeR: A General Taxonomy of LLM Prompts for Benchmarking Complex Tasks. In *EMNLP*, 14197–14203.

Schick, T.; Dwivedi-Yu, J.; Dessì, R.; Raileanu, R.; Lomeli, M.; Hambro, E.; Zettlemoyer, L.; Cancedda, N.; and Scialom, T. 2023. Toolformer: Language Models Can Teach Themselves to Use Tools. In *NeurIPS*.

Shanahan, M. 2024. Talking about Large Language Models. *Commun. ACM*, 67(2): 68–79.

Simpson, E.; and Gurevych, I. 2018. Finding Convincing Arguments Using Scalable Bayesian Preference Learning. *Trans. Assoc. Comput. Linguistics*, 6: 357–371.

Tafjord, O.; Mishra, B. D.; and Clark, P. 2022. Entailer: Answering Questions with Faithful and Truthful Chains of Reasoning. In *EMNLP*, 2078–2093.

Turpin, M.; Michael, J.; Perez, E.; and Bowman, S. R. 2023. Language Models Don't Always Say What They Think: Unfaithful Explanations in Chain-of-Thought Prompting. In *NeurIPS*.

Wang, H.; and Shu, K. 2023. Explainable Claim Verification via Knowledge-Grounded Reasoning with Large Language Models. In *EMNLP*, 6288–6304.

Wang, Z.; Liu, Z.; Zhang, Y.; Zhong, A.; Fan, L.; Wu, L.; and Wen, Q. 2023. RCAgent: Cloud Root Cause Analysis by Autonomous Agents with Tool-Augmented Large Language Models. *CoRR*, abs/2310.16340.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Ichter, B.; Xia, F.; Chi, E. H.; Le, Q. V.; and Zhou, D. 2022. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In *NeurIPS*.

Xia, S.; Li, X.; Liu, Y.; Wu, T.; and Liu, P. 2024. Evaluating Mathematical Reasoning Beyond Accuracy. *CoRR*, abs/2404.05692.

Yang, C.; Wang, X.; Lu, Y.; Liu, H.; Le, Q. V.; Zhou, D.; and Chen, X. 2024. Large Language Models as Optimizers. In *ICLR*.

Yao, S.; Yu, D.; Zhao, J.; Shafran, I.; Griffiths, T.; Cao, Y.; and Narasimhan, K. 2023. Tree of Thoughts: Deliberate Problem Solving with Large Language Models. In *NeurIPS*.

Yin, X.; Potyka, N.; and Toni, F. 2023. Argument Attribution Explanations in Quantitative Bipolar Argumentation Frameworks. In *ECAI*, 2898–2905.

Yin, X.; Potyka, N.; and Toni, F. 2024a. CE-QArg: Counterfactual Explanations for Quantitative Bipolar Argumentation Frameworks. In *KR*, 697–707.

Yin, X.; Potyka, N.; and Toni, F. 2024b. Explaining Arguments' Strength: Unveiling the Role of Attacks and Supports. In *IJCAI*, 3622–3630.

Zhang, H.; Li, J.; Wang, Y.; and Song, Y. 2023. Integrating Automated Knowledge Extraction with Large Language Models for Explainable Medical Decision-Making. In *BIBM*, 1710–1717.

Zhang, H.; and Parkes, D. C. 2023. Chain-of-Thought Reasoning is a Policy Improvement Operator. *CoRR*, abs/2309.08589.

Zhang, X.; and Gao, W. 2023. Towards LLM-based Fact Verification on News Claims with a Hierarchical Step-by-Step Prompting Method. In *IJCNLP*, 996–1011.