Assessing the Reasoning Abilities of ChatGPT in the Context of Claim Verification

Anonymous ACL submission

Abstract

The reasoning capabilities of LLMs are currently hotly debated. We examine the issue from the perspective of claim/rumour verification. We propose the first logical reasoning framework designed to break down any claim or rumor paired with evidence into the atomic reasoning steps necessary for verification. Based on our framework, we curate two annotated collections of such claim/evidence pairs: a synthetic dataset from Wikipedia and a real-world set stemming from rumours circulating on Twitter. We use them to evaluate the reasoning capabilities of GPT-3.5-Turbo and GPT-4 (hereinafter referred to as ChatGPT) within the context of our framework, providing a thorough analysis. Our results show that ChatGPT struggles in abductive reasoning, although this can be somewhat mitigated by using manual Chain of Thought (CoT) as opposed to Zero Shot (ZS) and ZS CoT approaches. Our study contributes to the growing body of research suggesting that ChatGPT's reasoning processes are unlikely to mirror human-like reasoning, and that LLMs need to be more rigorously evaluated in order to distinguish between hype and actual capabilities, especially in high stake real-world tasks such as claim verification.

1 Introduction

Large Language Models (LLMs) can perform well on non-trivial tasks and solve difficult problems. These capabilities range from solving MBA exams (Terwiesch, 2023) to passing professional medical tests (Kung et al., 2023; Nori et al., 2023) to quantitative reasoning (Lewkowycz et al., 2022). Nevertheless, there is much ongoing debate surrounding their evaluation and reasoning capabilities. For instance, initial claims of Theory of Mind (ToM) capabilities (Bubeck et al., 2023; Kosinski, 2023) of LLMs have turned out to be hastily drawn conclusions (Ullman, 2023; Sileo and Lernould, 2023) and later studies have found that even though LLMs manifest some form of ToM capabilities these are not robust (Shapira et al., 2023). Later work showed that it is possible to improve LLMs' ToM capabilities by using Chain of Thought (CoT) (Moghaddam and Honey, 2023; Zhou et al., 2023) and prompt planning (Sclar et al., 2023) techniques. LLMs have also been successfully used in communication games such as werewolf (Xu et al., 2023), diplomacy (Bakhtin et al., 2022), prisoner's dilemma (Akata et al., 2023), and negotiation games (Gandhi et al., 2023) that require social, game theoretic (Mao et al., 2024), and strategic reasoning capabilities. Additionally, there were also claims of emergent reasoning abilities (Wei et al., 2022), recently shown to be due to metric non-linearity (Schaeffer et al., 2023), with reasoning capabilities attributable to in-context learning (Lu et al., 2023b). Furthermore, Shapira et al. (2023) found that LLMs show Clever Hans (Kavumba et al., 2019) behaviour and rely on shortcuts, heuristics, and spurious correlations. These mixed and initial over-promised results show that work on the reasoning abilities of LLMs is still inconclusive (Huang and Chang, 2023) and we are still far from understanding their reasoning capabilities.

With the rise in popularity of LLMs comes both the potential for increased productivity and the scope for bad actors to proliferate misinformation (Guo et al., 2023). It is thus essential to understand the reasoning capabilities and limitations of LLMs and how they can be useful for tasks mitigating the spread of misinformation. Here we focus on information verification, which requires both accurate classification and strong rationale generation to be effective (Schlichtkrull et al., 2023). We extend the current discussion around reasoning abilities of LLMs, aiming to elucidate their capabilities in evaluating claims and rumours that circulate online more precisely. To this end, we create two small datasets, one of *claims* from Wikipedia and another of *rumours* from the PHEME dataset (Zu-

biaga et al., 2016), containing different types of claims in terms of the reasoning skills required to resolve them. It is important to draw a distinction between claims and rumours, since the former are usually simple checkable facts with a closed-world assumption, whilst the latter are usually complex, containing multiple parts - here rumours comprise social media posts of questionable veracity.

We propose a logical reasoning framework in Section 2.1, capable of breaking any claim or rumour down into atomic reasoning steps, and use it to manually annotate both of our datasets. ChatGPT is used to provide veracity assessments for both. We find that whilst ChatGPT is highly accurate at verifying our Wikipedia-based dataset of claims, it struggles with the real-world rumours of the PHEME-based dataset. Furthermore, when mathematical aspects are taken out of the picture, the model performs poorly when verification requires abductive (as opposed to deductive) reasoning. We make the following contributions:

- We propose the first logical reasoning framework capable of breaking any claim or rumour down into atomic reasoning steps.
- We create two small datasets, manually annotated based on our framework, to evaluate the reasoning abilities of LLMs.
- We use our datasets and framework to conduct an in-depth investigation of the reasoning abilities and limitations of multiple versions of ChatGPT under Zero Shot and Chain of Thought paradigms.
- We demonstrate that ChatGPT is better at deductive than abductive reasoning, verifying simple claims rather than rumours, and provide further evidence to suggest that it does not possess human-like reasoning abilities.
- We show that Zero Shot Chain of Thought produces better quality explanations compared to other prompt paradigms.

2 Methodology

2.1 Logical Reasoning Framework

The term reasoning is often used interchangeably to denote critical thinking, decision making and logical reasoning. Hence, we provide a relaxed definition of reasoning followingWason and Johnson-Laird (1972) and Galotti (1989), as the process based on critical thinking that results in either some form of decision making or logical conclusion. In other words, reasoning is a process employing inference between claims/arguments and premises/evidence to come to a conclusion/decision. Inference can be broken down into three inter-related components which we refer to as the reasoning *path*, reasoning *modes*, and reasoning *processes*.

The reasoning *path* is a series of hops linking the claim and evidence pair to a conclusion (here veracity label), as can be seen in Figure 1. Paths can be single or multi-hop, in which case the resolution of the original pair results in a series of sub-problems.

Although it is possible for a claim/evidence pair to have multiple valid reasoning paths, the vast majority of entries in our dataset have just one. We consider a hop to take place whenever we use reasoning to obtain new information, rather like the steps involved in solving a simultaneous equation. We do not consider the understanding of word definitions nor the introduction of common-sense knowledge to constitute a hop (although using these to derive new information counts as a hop).

Each hop in a reasoning path has an associated *process* and *mode*, which can be viewed as mutually exclusive sets of labels. First, we consider process, which can take many different forms depending on the claim structure. For instance, it can be causal reasoning, identifying the causal link between an action or event and its effect, or mathematical reasoning in which calculations are performed. Temporal reasoning involves understanding the ordering of events and performing the necessary mathematical reasoning about dates and times

The reasoning mode of a hop can be either deductive, abductive, inductive, or analogical. Below we provide informal definitions of these modes. More rigorous definitions can be found in Appendix A.

Deductive: Deductive reasoning is a logical reasoning mode by which a conclusion is drawn from a set of premises. The conclusion is valid if it reasonably follows the associated premises. Thus, if the premise holds then the conclusion also holds. For example:

- **Claim:** Schools closed, Dammartin-en-Goele residents told to stay indoors, town 'like warzone'
- **Evidence:** Schools went into lockdown and the town appealed to residents to stay inside resident's houses.
- **Conclusion:** The evidence explicitly references the school closing down and also resi-

dents being told to shelter at home. Therefore, we deductively infer that the rumour is true as the conclusion based on the claim logically follows the evidence.

Abductive: Abductive reasoning is the logical reasoning mode by which the most plausible conclusion is drawn from a set of hypotheses, based on partial observations/evidence. In simple terms, abductive reasoning is a best guess given some evidence that doesn't account for every possibility. This means, abductive reasoning can also lead to false conclusions.

- **Claim:** Sydney airspace wasn't closed. A second terror suspect wasn't arrested. Myths around sydneysiege debunked.
- **Evidence:** The suspect, who first took the cafe during Monday morning rush hour, was identified by police as local Muslim cleric Man Haron Monis, who had taken to calling 'Sheikh Haron'. Officials said local Muslim cleric Man Haron Monis, who had taken to calling himself 'Sheikh Haron' died after being shot during the raid. An armed intruder, identified by police as Man Haron Monis, held an 'undisclosed number' of hostages inside the cafe in central Sydney for 16 hours before police stormed the cafe in central Sydney at about 2:10am AEDT on Tuesday.
- **Conclusion:** Here, the claim mentions a second terror suspect. However, the evidence doesn't confirm anything regarding another suspect. Therefore, given the partial set of evidence we have, we can guess, using abduction, that the rumour[1],is false based on the hypothesis that the local muslim cleric was working by himself as he was the only suspect in the cafe.

Inductive: In induction inference is drawn from a complete set of observations (for a specific domain) and based on them a generalization is derived - a rule that can be used beyond the initial set of observations. This is in contrast to a abductive reasoning, whose conclusion is merely a best guess based on incomplete information. We discuss this more in Appendix A. As per Flach and Kakas (2000), for inductive reasoning, the evidence can be true but only provide partial support for the conclusion, as the conclusion can go beyond the information found in the evidence. Thus, its conclusion can be generalized over a broader domain that is not cover by initial evidence and therefore, it can also be false. An example is provided in A.

- **Claim:** Injecting or consuming bleach or disinfectant kills the virus (Covid-19).
- **Evidence 1:** Applying alcohol or chlorine to the skin can cause harm, especially if it enters the eyes or mouth.
- **Evidence 2:** These chemicals can disinfect surfaces, but people should not use them on their bodies.
- **Evidence 3:** Also, these products cannot kill viruses inside the body.
- **Conclusion:** From the evidence we can inductively draw a general conclusion that that claim is false as evident bleach and disinfectant cause harm to human bodies.Therefore, the general rule here is that, these can not kill any kind viruses in side the human body.

Analogical reasoning: Analogical reasoning is the logical reasoning mode by which a comparison is made between objects or entities with respect to similarity. Analogical reasoning is unlikely to occur in the domain of rumour verification. Nevertheless, we have included it here due to the sake of completion. We provide synthetic examples in Appendix A.

We provide an example of rumour resolution using our framework in Figure 1. It involves a multi-hop reasoning path consisting of 2 hops. Both hops use a causal reasoning process. In the first hop, through abductive reasoning, we establish the relation between a pilot being locked out of the cockpit and the pounding on the door. In the second hop, we link the context of the evidence to the rumour via deductive reasoning.

2.2 Dataset

As discussed previously, we have constructed two small datasets to test the abilities of LLMs to verify claims. Despite their size, they contain sufficiently diverse examples to challenge the models in multiple ways. The Wikipedia-based dataset focuses on fact checking simple claims from material which most likely formed part of the models' training data (Balloccu et al., 2024), whereas the PHEME-based dataset consists of substantially more complex claims from social media. Additionally, most entries in both datasets have a single reasoning path as most evidence provides clear veracity for its associated claim.

[1]the claim refutes the rumour, so we abduce the claim is true whereas the rumour is false

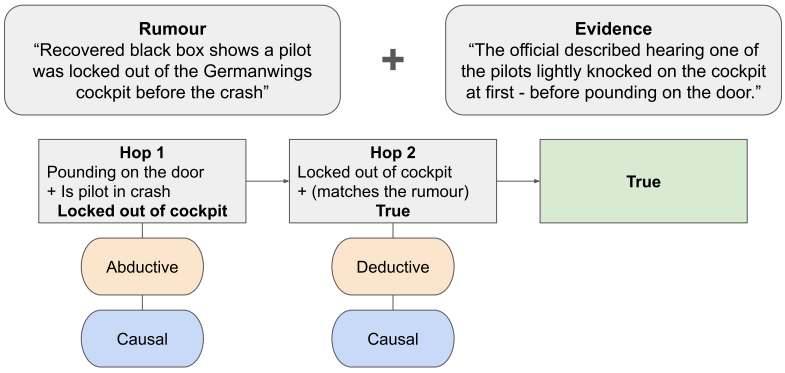

Figure 1: Resolution of a multi-hop compound reasoning type rumour using our proposed framework. We apply two sets of mutually exclusive labels: {*Abductive*, *Deductive*, *Inductive*} and {*Causal*, *Mathematical*}.

2.2.1 Wikipedia-based Dataset

Inspired by FEVER (Thorne et al., 2018), we randomly selected 20 articles from the top 5000 most popular ones on Wikipedia. For each selected article we create four separate claims, two of which involve numbers or dates.

To ensure the facts pertain to reasonably important information from each article, we used only information from the first paragraph, unless there was a compelling reason to do otherwise. Facts were rephrased to avoid direct quotations from Wikipedia and thus prevent models from providing responses based on pure memorization (Carlini et al., 2023; Wu et al., 2023b; Balloccu et al., 2024). However, when it came to incorporating evidence, we mostly retained the information in its original format from Wikipedia with minimal modification.

Each claim was manually assigned a veracity label *False True*, or *Unverified*. Claims were further annotated with their 'reasoning paths' via our framework in Section 2.1 by two annotators, reaching an inter-annotator agreement of 0.9. Due to imbalanced distribution of reasoning modes, we used Bennetts S score instead of Cohen's Kappa. Disputes were resolved by discussion with an independent expert. Statistics for this dataset can be found in Table 1.

Claim Type	False	True	Unverified
Deductive	17	37	-
Abductive	1	2	-
Total	18	39	20

Table 1: Statistics of the Wikipedia-based dataset, comprised of claims. Unverified claims have no reasoning paths.

2.2.2 PHEME-based Claim Collection

Claims from the PHEME dataset with sufficient evidence were manually chosen to provide a diverse set of test cases for LLMs. The resulting dataset is notably less challenging than PHEMEPlus (Dougrez-Lewis et al., 2022) and similar datasets that employ automated evidence retrieval methods, due to our requirement that every rumour comes with sufficient evidence allowing accurate veracity classification (except the '*Unverified*' class).

Evidence was collected in the form of direct quotes from sources, and annotation of 'reasoning paths' was carried out as in the previous section with pre-resolution and our inter-annotator agreement for PHEME was 0.86. Statistics for this dataset can be found in Table 2.

Rumour Type	False	True	Unverified
Deductive	17	25	-
Abductive	5	3	-
Total	22	28	34

Table 2: Statistics of the PHEME-based dataset, comprised of rumours. Unverified rumours have no reasoning paths.

2.3 Experimental Setup

2.3.1 Task Definition

We provide ChatGPT with a claim/rumour and relevant evidence pair, and ask it to identify the rumour as one of {*False*, *True*, *Unverified*} as well as provide justification.

2.3.2 Task Details

Prompting structure: We conduct experiments under three different prompting paradigms, the baseline being Zero-Shot prompting. Prompts are constructed as follows:

```
Q: As an expert journalist, classify the following rumour as true or false using only the provided evidence.
RUMOUR: Recovered black box shows a pilot was locked out of the Germanwings cockpit before the crash.
EVIDENCE: Germanwings plane crash: Pilot locked out of cockpit before aircraft hit French Alps, says investigator, German state prosecutor The official described hearing one of the pilots the cockpit lightly knocked on the door at first - before pounding on the door.
A:
```

Our second method is Zero-Shot CoT. We follow Kojima et al. (2023) and append '*Let's think step by step.*' to the above prompt.

Our final prompting method is manual CoT, where we follow Wei et al. (2023) and construct seven-shot examples as our rationales. The full set of examples is provided in Appendix D.

Closed World Assumption: We instruct ChatGPT to use "only the evidence available". This instruction appears to have been well followed, with very few cases that could suggest outside information creeping in to explanations. The model still draws upon common-sense knowledge despite this instruction, although we consider this both reasonable and necessary.

Role-playing: Significant efforts have been made to regulate bias and ensure the safety of proprietary LLMs (Perez et al., 2022; Ganguli et al., 2022). Therefore, additional instructions may be required for them to generate useful responses. In our case, we instruct ChatGPT to evaluate the claims/rumours as "an expert journalist", as without this instruction, the model usually refuses to answer prompts on the grounds of ethics or due to other constraints supposedly imposed by OpenAI, but adding this context enabled us to sidestep these constraints.

Explanations are an important aspect of rumour verification. Confidence and acceptability of the source of verification is fundamental to classification as we can't validate claims without plausible trustable explanations. Therefore, it is essential to evaluate explanations generated by LLMs for veracity predictions as LLMs are know to hallucinate (Bouyamourn, 2023; Rawte et al., 2023). Such explanations have also been shown to be self-contradictory (Mündler et al., 2023). Furthermore, it has been found that CoT methods can exacerbate hallucination (Gao et al., 2023a) and Shapira et al. (2023) found that LLMs show confidence in their predictions even when wrong.

3 Results and Discussion

3.1 Results Overview

ChatGPT performs better on the Wikipeda-based claims than the PHEME-based rumours, probably due the model having more exposure to data surrounding these claims and their more straightforward nature. In Table 5, both models hit the performance ceiling for causal reasoning on these relatively simple claims, not encountering the Reversal Curse (Berglund et al., 2023). This suggests they are well acquainted with these topics and thus we focus our analyses on the PHEME-based rumours. The LLMs tend to particularly struggle when rumour verification requires abductive reasoning, although this may be alleviated to some extent by using Manual CoT. Zero Shot reasoning outperforms Zero Shot CoT but not Manual CoT.

3.2 Discussion

Reasoning Performance When mathematical claims/rumours are excluded as in Table 5 (we already know from López Espejel et al. (2023) that ChatGPT underperforms on them), it emerges that the LLMs perform substantially better on deductive than abductive reasoning. Abductive performance is best with Manual CoT. We hypothesise that since deductive reasoning is by far the most common mode LLMs are likely to encounter in their training data, adding these examples to the prompt may have prompted the model to pursue such lines of reasoning. However, abductive performance of both models on the PHEME-based dataset remains unchanged when the abductive examples are removed, suggesting that the presence of deductive examples alone is sufficient to improve abductive reasoning, or perhaps that there were not enough abductive examples to impact performance.

One potential difficulty with abductive reasoning is that the conclusion is by definition uncertain, even though our dataset is designed such that every

	F	T	U
F	12	2	6
T	0	36	4
U	3	2	15

Zero Shot

	F	T	U
F	17	0	3
T	2	31	7
U	10	0	10

ZS CoT (Kojima et al., 2023)

	F	T	U
F	14	0	6
T	1	31	8
U	1	0	19

Manual CoT (Wei et al., 2023)

Table 3: 3-class classification results of GPT-3.5-Turbo on the Wikipedia-based dataset under the Zero Shot (ZS) and Chain of Thought (CoT) paradigms. F = *False*, T = *True*, U = *Unverified*. Bold letters indicate the ground-truth.

	F	T	U
F	7	9	9
T	2	27	1
U	14	6	14

Zero Shot

	F	T	U
F	10	4	11
T	1	17	12
U	6	6	22

ZS CoT (Kojima et al., 2023)

	F	T	U
F	7	7	11
T	2	23	5
U	4	8	22

Manual CoT (Wei et al., 2023)

Table 4: 3-class classification results of GPT-3.5-Turbo on the PHEME-based dataset under the Zero Shot (ZS) and Chain of Thought (CoT) paradigms. F = *False*, T = *True*, U = *Unverified*. Bold letters indicate the ground-truth.

	Deductive				Abductive				Deductive				Abductive			
	gpt-3.5-turbo		gpt-4		gpt-3.5-turbo		gpt-4		gpt-3.5-turbo		gpt-4		gpt-3.5-turbo		gpt-4	
	✓	x	✓	x	✓	x	✓	x	✓	x	✓	x	✓	x	✓	x
ZS	22	2	24	0	3	0	3	0	24	3	25	2	3	4	3	4
ZS CoT	23	1	24	0	3	0	3	0	19	8	24	3	4	3	2	5
Manual CoT	22	2	24	0	3	0	3	0	21	6	22	5	5	2	6	1
	Wikipedia-based dataset								PHEME-based dataset							

Table 5: 2-class {*False*, *True*} classification results of GPT-3.5-Turbo and GPT-4 on Wiki and PHEME under zero-shot (ZS) and chain-of-thought (CoT) paradigms, stratified by reasoning mode, and excluding mathematical rumours (causal only). Abductive claims/rumours include any with an abductive step in their reasoning path.

rumour/claim can be proven "beyond reasonable doubt" with the evidence available. We conduct our reasoning experiments in a 2-class setting to prevent the model from giving the *Unverified* label in such cases - when introducing this possibility, most but not all abductive rumours are indeed deemed *Unverified* and Manual CoT no longer yields improvements over the other paradigms. Possible reason is bias towards deductive reasoning.

Prompt Structure Contrary to our expectations and empirical evidence (Kojima et al., 2023; Zhou et al., 2023), ChatGPT performed better under the Zero Shot paradigm than Zero Shot (ZS) CoT for PHEME. However, performance is restored, when 7-shot manual CoT is used instead. The performance improvement with manual CoT also aligns with literature (Wei et al., 2023; Chen et al., 2023; Gao et al., 2023b). One likely reason of failure of ZS CoT might be due to the nature of rumours and ZS CoT creating ambiguous reasoning paths for ChatGPT. This might be explained through Prystawski et al. (2023)'s findings, who show that ZS CoT works well when observations tend to occur in partially overlapping neighborhoods of related concepts. For instance, ZS CoT has the most unverified predictions out of all the other paradigms (GPT-4 results are provided in Appendix E). This might be due to rumourous samples not having local structures within training data in order for the model to find a valid path. As for manual CoT performance, the provided examples are likely bridging reasoning paths to reduce confusion within the models' own paths thus resulting in better performance.

Explanations ChatGPT usually generates a convincing explanation for its classification. However, the quality of explanation varies based on the prompt paradigm. For PHEME, ZS CoT generated the most verbose explanation (A1) whereas manual CoT explanations are much more brief. GPT-4 sometimes did not produce explanations even with

ZS CoT and occasionally it performs classification without any provided explanation. ChatGPT was able to breakdown rumours and provide partial verification with ZS CoT, which was not observed for the other two prompting methods. For example, given the rumour/evidence pair in A2, the explanation by GPT-4 under ZS CoT was: "*The rumour is partially true. The evidence confirms that there was a shooting at the Canada War Memorial and a person (a Canadian soldier) was shot. However, the evidence only mentions one gunman, not 'numerous gunmen'*". We also noticed self-contradictory explanations A3. For the wiki dataset, most generated explanations were of similar quality. We provide further examples of provided explanations in Appendix C.

Inconsistencies Apart from GPT-4 hitting the performance ceiling on the Wikipedia-based dataset, we otherwise observe a curious pattern of errors. Suppose there exist many claims for some model to evaluate, paired with suitable pieces of evidence, and of varying difficulty to correctly classify. We would expect a model with the ability to reason *per se* (or indeed a human) to achieve near perfect results for sufficiently simple claims, until a threshold difficulty is reached where accuracy decays towards random chance.

Although the LLMs are reasonably accurate in Tables 3 and 4, and are often assumed to be capable of human-like reasoning *per se* (Dasgupta et al., 2023), the pattern of errors we observe differs from that specified in the previous paragraph. Loosely speaking, the models frequently make errors on rumours we consider to be simple, whilst still maintaining reasonable accuracy in more complex cases. It is more likely that ChatGPT is merely displaying some reasoning-like ability via statistical means, akin to pulling an answer out of a hat containing previously observed relevant answers (statistically blurred together).

This hypothesis is further evidenced by the greatly increased failure rate when certain words are included in prompts, specifically those which can appear in a great number of contexts, and thus could be relatively impervious to learning via ChatGPT's (hypothesised) usual statistical means. These words include soft references to amounts such as "several", and positional words like "above". The same logic applies to numbers, although we would not expect this contextual issue to be the key driver of poor arithmetic reasoning.

When tested with nonsensical claims which are impossible to verify, ChatGPT usually does not recognise them as such and attempts to resolve them. For example, given the nonsense claim "The Crown television only eats past future centuries." and evidence "The sixth season, which will close the series, will cover the Queen's reign of the previous century.", GPT-4 predicts *False*, explaining "The evidence provided contradicts the rumour. The Crown television series does not only cover future centuries, as it is shown to cover the Queen's reign of the previous century. Therefore, the rumour is false."

4 Related work

4.1 LLMs for Reasoning

Most recent literature appears to be centered around evaluation of different types of reasoning capabilities of LLMs. For instance, on the task of abductive reasoning, Zhao et al. (2023b) found that GPT-3 lacked the ability of abductive reasoning, whereas Shi et al. (2023) showed that GPT-3.5 could be used as an abductive reasoner to generate possible cause events in order to guide an event sequence model. As for analogical reasoning, both Webb et al. (2023) and Hu et al. (2023) found that GPT-3 was capable of performing analogical reasoning. However, Hu et al. (2023) also found that 100 times smaller PLM models could also achieve GPT-3 level performance. Yu et al. (2023a) also showed that ChatGPT (GPT3.5-turbo) was capable of analogical/comparative reasoning. They also reserved concerns over data contamination issues. Saparov et al. (2023) showed that LLMs were capable of deductive reasoning as they were able to generalize to compositional proofs. However they also showed that, LLMs have difficulty generalizing to longer proofs such as, proof by cases and proof by contradiction. Akyürek et al. (2024) showed the deductive reasoning capabilities of LLMs through performing deductive closure of a given training set.

The evaluation of syllogism, Ye et al. (2023) showed that LLMs failed at inductive reasoning even with CoT and had consistent, systematic failure patterns unique to each LLM family. For causal reasoning, Gao et al. (2023a) performed a comprehensive evaluation and found that ChatGPT was not a good causal reasoner, although it was a good causal explainer. They also showed that ChatGPT had serious causal hallucination issues and was also

sensitive toward causal words in the prompts. Lu et al. (2023a) showed that LLMs could perform complex compositional reasoning through a LLM based planner and different tools that ranged from LLMs, off-the-shelf vision models, web search engines to Python functions and heuristic-based modules. Wu et al. (2023b) showed that LLMs were able to show nontrivial performance on different counterfactual reasoning tasks. However, they also found that LLMs often relied on memorization, shortcuts and non-transferable procedures for solving these tasks. Li et al. (2023a) also showed that GPT-3 was somewhat sensitive towards counterfactual cues but also susceptible to lexical associative cues. Multiple recent studies have also evaluated LLMs on multihop reasoning capabilities.Stechly et al. (2023) showed that GPT-4 lacked multi-hop reasoning abilities when it came to complex tasks like graph coloring problems. Lu et al. (2022) found that LLMs could perform multi-hop multi-modal reasoning on multiple choice scientific question answering and CoT improved this performance. Using mechanistic interprebility methods, Hou et al. (2023) showed that LLaMA(7B) was capable of step by step reasoning. Apart from above discussion, others have also explored more complex reasoning processes. A more in depth discussion of these are provided in Appendix B.

4.2 Chain of Thought (CoT) prompting for reasoning

The reasoning ability of LLMs has been enhanced by the Chain of Thought (CoT) approach (Wei et al., 2023), in which models are prompted in a few shot manner with rationales for each example instead of the standard in-context learning prompt. Kojima et al. (2023) showed that LLMs could reason even in a ZS manner just by adding, "Let's think step by step" to the prompt. Wang et al. (2023b) found that sampling multiple CoTs through majority voting also improved reasoning performance, while Xue et al. (2023) also proposed a voting method for exampler choice. Their voting method is dynamic, setting a consistency threshold to generate examplers in an iterative manner to choose the best set of exampler. Conversely, Xi et al. (2023) proposed to iteratively update/refine an example instead of generating multiple CoTs. Instead of voting or refinement, Li et al. (2023c) proposed to generate multiple prompts for creating multiple reasoning paths in order to perform a reasoning-step aware voting to choose the best CoT reasoning path. CoT is still a very active research topic and there have been a plethora of more advanced methods proposed such as Tree of Thought (Yao et al., 2023), Graph of Thought (Besta et al., 2024), and linguistic planning (Tian et al., 2023; Wang et al., 2023a). In order to get a more complete understanding, readers are suggested to check these survey's out (Qiao et al., 2023; Chu et al., 2023; Yu et al., 2023b).

4.3 LLMs for Fact Checking and Rumour Veracity Classification

Early work focused on simply asking a LLM to verify a fact in question (Lee et al., 2020). More recently, approaches by Li et al. (2023b) and Cheung and Lam (2023) augment an LLM's knowledge with up-to-date information retrieved via web search, and presumably could be further improved using the multi-CoT techniques used in (Cao, 2023) and (Li et al., 2023c). Particularly relevant is the claim-splitting approach of (Li et al., 2023b), whereby a claim with multiple parts is split up by the LLM for individual verification, although this does not solve the problem of disregarding incidental trivial claims which might lead a false multi-claim to be classified as "mostly true". Wang and Shu (2023) showed that formulating claims into first order logic helped with rumour reasoning in a Retrieval-Augmented Generation setting. There have also been studies around misinformation detection. (Alhindi et al., 2023) showed the use of LLMs in generating fallacy based misinformation as form of data augmentation. Lin et al. (2023) used LLMs to perform multimodal abductive reasoning by detecting harmful memes that spread misinformation and hate on the internet.

5 Conclusion

We have created a novel extendable logical reasoning framework which is capable of deconstructing any claim-evidence pair into the atomic reasoning steps required for verifying the claims. We have used the framework to create two small annotated datasets. Evaluating ChatGPT on the claims withing the datasets we have shown it performs particularly poorly on reasoning paths containing abductive hops, and contribute further evidence suggesting that these models are unlikely to (yet) possess human-like reasoning abilities.

Ethics Statement

The PHEME dataset is a pre-existing dataset of rumours, for which ethical approval was obtained by the original research team. Our Wikipedia-based dataset is constructed analogously to FEVER using publicly available information from Wikipedia. All of the evidence we use is freely and readily available online via Google Search.

Limitations

When ChatGPT generates an explanation, there is no guarantee that it is true to the final label assigned by the model. We mitigate this issue by obtaining both the label and explanation in the same prompt, although it should still be treated as merely "a plausible post-hoc explanation generated by the model" rather than the specific reason behind its decision.

The LLMs used in this paper are closed source, and there is no way of looking behind the veil of OpenAI. Model outputs have been known to vary between runs even when the temperature is set to 0. Fortunately, throughout our experiments, this was not an issue and the results were almost entirely stable.

References

Elif Akata, Lion Schulz, Julian Coda-Forno, Seong Joon Oh, Matthias Bethge, and Eric Schulz. 2023. Playing repeated games with large language models. *ArXiv*, abs/2305.16867.

Afra Feyza Akyürek, Ekin Akyürek, Leshem Choshen, Derry Wijaya, and Jacob Andreas. 2024. Deductive closure training of language models for coherence, accuracy, and updatability.

Tariq Alhindi, Smaranda Muresan, and Preslav Nakov. 2023. Large language models are few-shot training example generators: A case study in fallacy recognition.

Konstantine Arkoudas. 2023. Gpt-4 can't reason.

Anton Bakhtin, Noam Brown, Emily Dinan, Gabriele Farina, Colin Flaherty, Daniel Fried, Andrew Goff, Jonathan Gray, Hengyuan Hu, Athul Paul Jacob, Mojtaba Komeili, Karthik Konath, Minae Kwon, Adam Lerer, Mike Lewis, Alexander H. Miller, Sasha Mitts, Adithya Renduchintala, Stephen Roller, Dirk Rowe, Weiyan Shi, Joe Spisak, Alexander Wei, David Wu, Hugh Zhang, and Markus Zijlstra. 2022. Human-level play in the game of diplomacy by combining language models with strategic reasoning. *Science*, 378(6624):1067–1074.

Simone Balloccu, Patrícia Schmidtová, Mateusz Lango, and Ondřej Dušek. 2024. Leak, cheat, repeat: Data contamination and evaluation malpractices in closed-source llms.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, and Pascale Fung. 2023. A multitask, multilingual, multimodal evaluation of ChatGPT on reasoning, hallucination, and interactivity. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 675–718, Nusa Dua, Bali. Association for Computational Linguistics.

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

Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Michal Podstawski, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Hubert Niewiadomski, Piotr Nyczyk, and Torsten Hoefler. 2024. Graph of thoughts: Solving elaborate problems with large language models.

Ali Borji. 2023. A categorical archive of chatgpt failures.

Adam Bouyamourn. 2023. Why LLMs hallucinate, and how to get (evidential) closure: Perceptual, intensional, and extensional learning for faithful natural language generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3181–3193, Singapore. Association for Computational Linguistics.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4.

Lang Cao. 2023. Enhancing reasoning capabilities of large language models: A graph-based verification approach.

Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, and Chiyuan Zhang. 2023. Quantifying memorization across neural language models.

Wenhu Chen, Xueguang Ma, Xinyi Wang, and William W. Cohen. 2023. Program of thoughts prompting: Disentangling computation from reasoning for numerical reasoning tasks.

Tsun-Hin Cheung and Kin-Man Lam. 2023. Factllama: Optimizing instruction-following language models with external knowledge for automated fact-checking.

Zheng Chu, Jingchang Chen, Qianglong Chen, Weijiang Yu, Tao He, Haotian Wang, Weihua Peng, Ming Liu, Bing Qin, and Ting Liu. 2023. A survey of chain of thought reasoning: Advances, frontiers and future.

Ishita Dasgupta, Andrew K. Lampinen, Stephanie C. Y. Chan, Hannah R. Sheahan, Antonia Creswell, Dharshan Kumaran, James L. McClelland, and Felix Hill. 2023. Language models show human-like content effects on reasoning tasks.

John Dougrez-Lewis, Elena Kochkina, Miguel Arana-Catania, Maria Liakata, and Yulan He. 2022. PHEMEPlus: Enriching social media rumour verification with external evidence. In *Proceedings of the Fifth Fact Extraction and VERification Workshop (FEVER)*, pages 49–58, Dublin, Ireland. Association for Computational Linguistics.

Peter A. Flach and Antonis C. Kakas. 2000. *Abductive and Inductive Reasoning: Background and Issues*, page 1–27. Springer Netherlands.

Hao Fu, Yao; Peng and Tushar Khot. 2022. How does gpt obtain its ability? tracing emergent abilities of language models to their sources. *Yao Fu's Notion*.

Kathleen M. Galotti. 1989. Approaches to studying formal and everyday reasoning. *Psychological Bulletin*, 105:331–351.

Kanishk Gandhi, Dorsa Sadigh, and Noah D. Goodman. 2023. Strategic reasoning with language models. *ArXiv*, abs/2305.19165.

Deep Ganguli, Liane Lovitt, Jackson Kernion, Amanda Askell, Yuntao Bai, Saurav Kadavath, Ben Mann, Ethan Perez, Nicholas Schiefer, Kamal Ndousse, Andy Jones, Sam Bowman, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Nelson Elhage, Sheer El-Showk, Stanislav Fort, Zac Hatfield-Dodds, Tom Henighan, Danny Hernandez, Tristan Hume, Josh Jacobson, Scott Johnston, Shauna Kravec, Catherine Olsson, Sam Ringer, Eli Tran-Johnson, Dario Amodei, Tom Brown, Nicholas Joseph, Sam McCandlish, Chris Olah, Jared Kaplan, and Jack Clark. 2022. Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned.

Jinglong Gao, Xiao Ding, Bing Qin, and Ting Liu. 2023a. Is ChatGPT a good causal reasoner? a comprehensive evaluation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 11111–11126, Singapore. Association for Computational Linguistics.

Luyu Gao, Aman Madaan, Shuyan Zhou, Uri Alon, Pengfei Liu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023b. PAL: Program-aided language models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 10764–10799. PMLR.

Biyang Guo, Xin Zhang, Ziyuan Wang, Minqi Jiang, Jinran Nie, Yuxuan Ding, Jianwei Yue, and Yupeng Wu. 2023. How close is chatgpt to human experts? comparison corpus, evaluation, and detection.

Wes Gurnee and Max Tegmark. 2023. Language models represent space and time.

Danijar Hafner. 2022. Benchmarking the spectrum of agent capabilities.

Shibo Hao, Yi Gu, Haodi Ma, Joshua Hong, Zhen Wang, Daisy Wang, and Zhiting Hu. 2023. Reasoning with language model is planning with world model. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 8154–8173, Singapore. Association for Computational Linguistics.

Yifan Hou, Jiaoda Li, Yu Fei, Alessandro Stolfo, Wangchunshu Zhou, Guangtao Zeng, Antoine Bosselut, and Mrinmaya Sachan. 2023. Towards a mechanistic interpretation of multi-step reasoning capabilities of language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4902–4919, Singapore. Association for Computational Linguistics.

Xiaoyang Hu, Shane Storks, Richard Lewis, and Joyce Chai. 2023. In-context analogical reasoning with pre-trained language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1953–1969, Toronto, Canada. Association for Computational Linguistics.

Jie Huang and Kevin Chen-Chuan Chang. 2023. Towards reasoning in large language models: A survey.

Xiaoxi Kang, Lizhen Qu, Lay-Ki Soon, Adnan Trakic, Terry Zhuo, Patrick Emerton, and Genevieve Grant. 2023. Can ChatGPT perform reasoning using the IRAC method in analyzing legal scenarios like a lawyer? In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13900–13923, Singapore. Association for Computational Linguistics.

Pride Kavumba, Naoya Inoue, Benjamin Heinzerling, Keshav Singh, Paul Reisert, and Kentaro Inui. 2019. When choosing plausible alternatives, clever hans can be clever. In *Proceedings of the First Workshop on Commonsense Inference in Natural Language Processing*, pages 33–42, Hong Kong, China. Association for Computational Linguistics.

Akira Kawabata and Saku Sugawara. 2023. Evaluating the rationale understanding of critical reasoning in logical reading comprehension. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 116–143, Singapore. Association for Computational Linguistics.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2023. Large language models are zero-shot reasoners.

Michal Konkol, Tomáš Brychcín, Michal Nykl, and Tomáš Hercig. 2017. Geographical evaluation of word embeddings. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 224–232, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Michal Kosinski. 2023. Theory of mind might have spontaneously emerged in large language models.

Tiffany H. Kung, Morgan Cheatham, Arielle Medenilla, Czarina Sillos, Lorie De Leon, Camille Elepaño, Maria Madriaga, Rimel Aggabao, Giezel Diaz-Candido, James Maningo, and Victor Tseng. 2023. Performance of chatgpt on usmle: Potential for ai-assisted medical education using large language models. *PLOS Digital Health*, 2(2):e0000198.

Nayeon Lee, Belinda Z. Li, Sinong Wang, Wen-tau Yih, Hao Ma, and Madian Khabsa. 2020. Language models as fact checkers?

Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, Yuhuai Wu, Behnam Neyshabur, Guy Gur-Ari, and Vedant Misra. 2022. Solving quantitative reasoning problems with language models.

Jiaxuan Li, Lang Yu, and Allyson Ettinger. 2023a. Counterfactual reasoning: Testing language models' understanding of hypothetical scenarios. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 804–815, Toronto, Canada. Association for Computational Linguistics.

Miaoran Li, Baolin Peng, and Zhu Zhang. 2023b. Self-checker: Plug-and-play modules for fact-checking with large language models.

Yifei Li, Zeqi Lin, Shizhuo Zhang, Qiang Fu, Bei Chen, Jian-Guang Lou, and Weizhu Chen. 2023c. Making language models better reasoners with step-aware verifier. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5315–5333, Toronto, Canada. Association for Computational Linguistics.

Hongzhan Lin, Ziyang Luo, Jing Ma, and Long Chen. 2023. Beneath the surface: Unveiling harmful memes with multimodal reasoning distilled from large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9114–9128, Singapore. Association for Computational Linguistics.

Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering.

Pan Lu, Baolin Peng, Hao Cheng, Michel Galley, Kai-Wei Chang, Ying Nian Wu, Song-Chun Zhu, and Jianfeng Gao. 2023a. Chameleon: Plug-and-play compositional reasoning with large language models. *arXiv preprint arXiv:2304.09842*.

Sheng Lu, Irina Bigoulaeva, Rachneet Sachdeva, Harish Tayyar Madabushi, and Iryna Gurevych. 2023b. Are emergent abilities in large language models just in-context learning?

Jessica López Espejel, El Hassane Ettifouri, Mahaman Sanoussi Yahaya Alassan, El Mehdi Chouham, and Walid Dahhane. 2023. Gpt-3.5, gpt-4, or bard? evaluating llms reasoning ability in zero-shot setting and performance boosting through prompts. *Natural Language Processing Journal*, 5:100032.

Shaoguang Mao, Yuzhe Cai, Yan Xia, Wenshan Wu, Xun Wang, Fengyi Wang, Tao Ge, and Furu Wei. 2024. Alympics: Llm agents meet game theory – exploring strategic decision-making with ai agents.

Shima Rahimi Moghaddam and Christopher J. Honey. 2023. Boosting theory-of-mind performance in large language models via prompting.

Terufumi Morishita, Gaku Morio, Atsuki Yamaguchi, and Yasuhiro Sogawa. 2023. Learning deductive reasoning from synthetic corpus based on formal logic. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 25254–25274. PMLR.

Niels Mündler, Jingxuan He, Slobodan Jenko, and Martin Vechev. 2023. Self-contradictory hallucinations of large language models: Evaluation, detection and mitigation.

Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. 2023. Capabilities of gpt-4 on medical challenge problems.

Liangming Pan, Alon Albalak, Xinyi Wang, and William Wang. 2023. Logic-LM: Empowering large language models with symbolic solvers for faithful logical reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3806–3824, Singapore. Association for Computational Linguistics.

Gabriele Paul. 1993. Approaches to abductive reasoning: an overview. *Artificial Intelligence Review*, 7(2):109–152.

Ethan Perez, Saffron Huang, Francis Song, Trevor Cai, Roman Ring, John Aslanides, Amelia Glaese, Nat McAleese, and Geoffrey Irving. 2022. Red teaming language models with language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3419–3448, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Anya Plutynski. 2011. Four problems of abduction: A brief history. *HOPOS: The Journal of the International Society for the History of Philosophy of Science*, 1(2):227–248.

Ben Prystawski, Michael Y. Li, and Noah D. Goodman. 2023. Why think step by step? reasoning emerges from the locality of experience.

Shuofei Qiao, Yixin Ou, Ningyu Zhang, Xiang Chen, Yunzhi Yao, Shumin Deng, Chuanqi Tan, Fei Huang, and Huajun Chen. 2023. Reasoning with language model prompting: A survey. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5368–5393, Toronto, Canada. Association for Computational Linguistics.

Chengwei Qin, Aston Zhang, Zhuosheng Zhang, Jiaao Chen, Michihiro Yasunaga, and Diyi Yang. 2023. Is ChatGPT a general-purpose natural language processing task solver? In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1339–1384, Singapore. Association for Computational Linguistics.

Vipula Rawte, Swagata Chakraborty, Agnibh Pathak, Anubhav Sarkar, S.M Towhidul Islam Tonmoy, Aman Chadha, Amit Sheth, and Amitava Das. 2023. The troubling emergence of hallucination in large language models - an extensive definition, quantification, and prescriptive remediations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2541–2573, Singapore. Association for Computational Linguistics.

Abulhair Saparov, Richard Yuanzhe Pang, Vishakh Padmakumar, Nitish Joshi, Seyed Mehran Kazemi, Najoung Kim, and He He. 2023. Testing the general deductive reasoning capacity of large language models using ood examples.

Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo. 2023. Are emergent abilities of large language models a mirage?

Michael Schlichtkrull, Zhijiang Guo, and Andreas Vlachos. 2023. Averitec: A dataset for real-world claim verification with evidence from the web.

Melanie Sclar, Sachin Kumar, Peter West, Alane Suhr, Yejin Choi, and Yulia Tsvetkov. 2023. Minding language models' (lack of) theory of mind: A plug-and-play multi-character belief tracker.

Natalie Shapira, Mosh Levy, Seyed Hossein Alavi, Xuhui Zhou, Yejin Choi, Yoav Goldberg, Maarten Sap, and Vered Shwartz. 2023. Clever hans or neural theory of mind? stress testing social reasoning in large language models. *ArXiv*, abs/2305.14763.

Xiaoming Shi, Siqiao Xue, Kangrui Wang, Fan Zhou, James Y. Zhang, Jun Zhou, Chenhao Tan, and Hongyuan Mei. 2023. Language models can improve event prediction by few-shot abductive reasoning.

Chenglei Si, Weijia Shi, Chen Zhao, Luke Zettlemoyer, and Jordan Boyd-Graber. 2023. Getting MoRE out of mixture of language model reasoning experts. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8234–8249, Singapore. Association for Computational Linguistics.

Damien Sileo and Antoine Lernould. 2023. MindGames: Targeting theory of mind in large language models with dynamic epistemic modal logic. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4570–4577, Singapore. Association for Computational Linguistics.

Kaya Stechly, Matthew Marquez, and Subbarao Kambhampati. 2023. Gpt-4 doesn't know it's wrong: An analysis of iterative prompting for reasoning problems.

Christian Terwiesch. 2023. Would chat gpt get a wharton mba? a prediction based on its performance in the operations management course. Technical report, Mack Institute for Innovation Management at the Wharton School, University of Pennsylvania.

Rosamond Thalken, Edward Stiglitz, David Mimno, and Matthew Wilkens. 2023. Modeling legal reasoning: LM annotation at the edge of human agreement. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9252–9265, Singapore. Association for Computational Linguistics.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819, New Orleans, Louisiana. Association for Computational Linguistics.

Qingyuan Tian, Hanlun Zhu, Lei Wang, Yang Li, and Yunshi Lan. 2023. R^3 prompting: Review, rephrase and resolve for chain-of-thought reasoning in large language models under noisy context. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1670–1685, Singapore. Association for Computational Linguistics.

Tomer Ullman. 2023. Large language models fail on trivial alterations to theory-of-mind tasks.

Haoran Wang and Kai Shu. 2023. Explainable claim verification via knowledge-grounded reasoning with large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6288–6304, Singapore. Association for Computational Linguistics.

Lei Wang, Wanyu Xu, Yihuai Lan, Zhiqiang Hu, Yunshi Lan, Roy Ka-Wei Lee, and Ee-Peng Lim. 2023a. Plan-and-solve prompting: Improving zero-shot chain-of-thought reasoning by large language

models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2609–2634, Toronto, Canada. Association for Computational Linguistics.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023b. Self-consistency improves chain of thought reasoning in language models.

Peter Cathcart Wason and Philip Nicholas Johnson-Laird. 1972. *Psychology of Reasoning: Structure and Content*. Harvard University Press, Cambridge, MA, USA.

Taylor Webb, Keith J. Holyoak, and Hongjing Lu. 2023. Emergent analogical reasoning in large language models. *Nature Human Behaviour*, 7(9):1526–1541.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022. Emergent abilities of large language models.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models.

Yue Wu, Shrimai Prabhumoye, So Yeon Min, Yonatan Bisk, Ruslan Salakhutdinov, Amos Azaria, Tom Mitchell, and Yuanzhi Li. 2023a. Spring: Studying the paper and reasoning to play games.

Zhaofeng Wu, Linlu Qiu, Alexis Ross, Ekin Akyürek, Boyuan Chen, Bailin Wang, Najoung Kim, Jacob Andreas, and Yoon Kim. 2023b. Reasoning or reciting? exploring the capabilities and limitations of language models through counterfactual tasks.

Zhiheng Xi, Senjie Jin, Yuhao Zhou, Rui Zheng, Songyang Gao, Jia Liu, Tao Gui, Qi Zhang, and Xuanjing Huang. 2023. Self-Polish: Enhance reasoning in large language models via problem refinement. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 11383–11406, Singapore. Association for Computational Linguistics.

Yuzhuang Xu, Shuo Wang, Peng Li, Fuwen Luo, Xiaolong Wang, Weidong Liu, and Yang Liu. 2023. Exploring large language models for communication games: An empirical study on werewolf. *ArXiv*, abs/2309.04658.

Mingfeng Xue, Dayiheng Liu, Wenqiang Lei, Xingzhang Ren, Baosong Yang, Jun Xie, Yidan Zhang, Dezhong Peng, and Jiancheng Lv. 2023. Dynamic voting for efficient reasoning in large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3085–3104, Singapore. Association for Computational Linguistics.

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

Mengyu Ye, Tatsuki Kuribayashi, Jun Suzuki, Goro Kobayashi, and Hiroaki Funayama. 2023. Assessing step-by-step reasoning against lexical negation: A case study on syllogism. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14753–14773, Singapore. Association for Computational Linguistics.

Mengxia Yu, Zhihan Zhang, Wenhao Yu, and Meng Jiang. 2023a. Pre-training language models for comparative reasoning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12421–12433, Singapore. Association for Computational Linguistics.

Zihan Yu, Liang He, Zhen Wu, Xinyu Dai, and Jiajun Chen. 2023b. Towards better chain-of-thought prompting strategies: A survey.

Hongyu Zhao, Kangrui Wang, Mo Yu, and Hongyuan Mei. 2023a. Explicit planning helps language models in logical reasoning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11155–11173, Singapore. Association for Computational Linguistics.

Wenting Zhao, Justin Chiu, Claire Cardie, and Alexander Rush. 2023b. Abductive commonsense reasoning exploiting mutually exclusive explanations. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14883–14896, Toronto, Canada. Association for Computational Linguistics.

Pei Zhou, Aman Madaan, Srividya Pranavi Potharaju, Aditya Gupta, Kevin R. McKee, Ari Holtzman, Jay Pujara, Xiang Ren, Swaroop Mishra, Aida Nematzadeh, Shyam Upadhyay, and Manaal Faruqui. 2023. How far are large language models from agents with theory-of-mind?

Arkaitz Zubiaga, Geraldine Wong Sak Hoi, Maria Liakata, and Rob Procter. 2016. Pheme dataset of rumours and non-rumours.

A Logical Reasoning Framework

Deductive Reasoning: Deductive reasoning or top-down logic is a logical reasoning process where we use inference rules such as modus ponens to deduce the veracity of a conclusion based on multiple hypothesis. For example, given the rumour/evidence pair, we can construct them as

$P \implies Q$: *Schools closed , Dammartin-en-Goele residents told to stay indoors , town 'like warzone'.* **[Rumour]**

P: *Schools went into lockdown and the town appealed to residents to stay inside residents's houses.* **[Evidence]**

Q: The schools have been closed and citizens have been told to stay home. Thus, the town is like in a warzone situation. **[Conclusion]**

One of the core element of deductive inference is that if the premises are true then the conclusion is true by design. Additionally, deductive reasoning is done based off of deduction rules and as per formal logic there are infinite deduction rules (Morishita et al., 2023). The most used ones deductive rules are modus poens, syllogism and elimination. The readers are referred to these studies (Morishita et al., 2023; Saparov et al., 2023) for a more in depth handling of deduction rules.

Inductive Reasoning: Inductive reasoning is the reasoning process where we use observations and outcomes to infer a generalizable rule. Hence, the logical structure can be represented as,

- $\forall x, observations(x) \implies conclusion$ or
- $\exists x, observations(x) \implies conclusion$

or in many different forms. One core component of inductive reasoning is that it's conclusion can be false. As per Flach and Kakas (2000), if premises for any stated argument only provides partial support for the conclusion then that is an inductive argument given the premises are true. An example would be,

Observation1 : Eagles have wings. Eagles are birds and eagles can fly.

Observation2 : Ducks have wings. Ducks are birds and ducks can fly. and

Observation 3a : Pigeons have wings. Pigeons are birds and pigeons can fly.

or

Observation 3b : Bats have wings. Bats are mammals and bats can fly.

Conclusion a : All birds have wings and all birds can fly.

or

Conclusion b : Those who have wings can fly.

Here, we can see from the examples that, conclusion is correct within our premise. However, beyond our premise, we know that there are flightless birds like Emu, Ostrich, and Penguins and there is wingless bird like Kiwi. Similar can be said about conclusion b. Conclusion b is true until we start including flightless birds.

Abductive Reasoning: There is much debate regarding definition of abductive reasoning (Plutynski, 2011). Therefore, we provide the definition from Paul (1993). Paul (1993) provides three different approches of defining abductive reasoning. These are,

- set-cover-based approach,
- logic-based approach, and
- knowledge-level approach.

Here, we will be using set-cover based approach. In set-cover-approach, we construct a set of most plausible hypotheses H given some observations O. Afterwards, we find the best possible explanation E based on H. In other words, *A domain for hypothesis assembly is defined by the triple ϕ, σ, ϵ), where ϕ is a finite set of hypotheses, σ is a set of observations and ϵ is a mapping from subsets of ϕ to subsets of σ. $\epsilon(\phi)$ is called the explanatory power of the set of hypotheses ϕ and determines the set of observations σ accounts for. An assembly problem is given by a set $\sigma' \subseteq \sigma$ of observations that have to be explained.* (Paul, 1993)

One core difference between abductive and the other two types of reasoning is, deductive reasoning is formulation of results based on rule and observation and inductive reasoning is formulation of rule based on result and observation. Whereas, abductive reasoning is formulation of an observation based on rule and result. For example (Flach and Kakas, 2000),

Rule : All the beans from this bag are white.

Result : These beans are white..

Conclusion : These beans are from this bag..

Difference between Inductive and Abductive Reasoning:

Analogical Reasoning: Analogical reasoning is the reasoning process concerned with comparison between two or more objects, arguments, entities etc. Formally we can define it as,

Premise : α is equivalent to ζ, κ, ϕ, and ω.

Premise : β is equivalent to ζ, κ, and ϕ.

Conclusion : β is probably equivalent to ω.

Difference between Inductive and Abductive Reasoning: The main differences between inductive and abductive reasoning are the completeness of premises and hypothesis generation. Abductive reasoning generates hypothesis in the preliminary phase in order to produce a conclusion whereas inductive reasoning generates a generalizable hypothesis as conclusion. Furthermore, abductive reasoning needs to generate a set of hypothesis due to it's partial complete premises whereas inductive reasoning generates generalizable hypothesis due to more complete premises. For example,

Premise (abductive+inductive) : My front lawn was wet.

Premise abductive+inductive : I found water under my car.

Premise (inductive only) : I found my radiator pipe broken.

Premise (inductive only) : It is currently autumn.

Conclusion abductive : It is likely rained last night.

Conclusion inductive : My lawn was due to dew as it is autumn now/Dew falls in autumn.

B Extended Related Work

In their work, Kawabata and Sugawara (2023) found that InstructGPT lacked critical reasoning capabilities when it came to logical reading comprehension. It failed to answer subquestions even if it was able to answer a given main questions correctly. Thalken et al. (2023) found that compared to PLMs, GPT-4 performed poorly on classification of legal reasoning as per jurisprudential philosophy, even when given instructions (i.e. prompts) equal to the instructions presented to human annotators. Kang et al. (2023) also found similar results where they showed that ChatGPT was not able to perform legal reasoning based on IRAC method and even if LLMs could produce reasonable answers, they mostly failed to yield correct reasoning paths that aligned with legal experts. Pan et al. (2023) showed that LLMs could be used to reformulate logical reasoning problems into symbolic formulation and then uses a deterministic symbolic solver to solve different forms of logical reasoning problems. Wu et al. (2023a) used LLMs to play Crafter (Hafner, 2022). Here the LLM parsed the Crafter paper to generate question answers related to game mechanics and afters it used to visual descriptor to take in game inputs, which was then transformed into Directed Acyclic Graph (DAG) with game mechanic Q/A. The LLM treated this DAG as the reasoning module and solved the DAG to take in game actions.

Apart from the above mentioned works, there have also been multiple works to evaluate the the reasoning capabilities of LLM through benchmarking them on multiple datasets. López Espejel et al. (2023) showed that even though GPT-4 performed better than GPT-3.5, overall ChatGPT had problems with inductive, mathematical, multi-hop and commonsense reasoning tasks. Bang et al. (2023) performed multilingual and multimodal reasoning evaluation of ChatGPT and reported that ChatGPT had variable performance on 10 different reasoning tasks. They also reported that ChatGPT was an unreliable reasoner, showed more issues with inductive reasoning than deductive and abductive reasoning. They also reported that ChatGPT was better at analogical reasoning compared to multi-hop reasoning and it also had hallucination problems. Furthermore, contrary to Gao et al. (2023a)'s finding, Bang et al. (2023) reported that ChatGPT was good at causal reasoning. However, Gao et al. (2023a) evaluated larger datasets compared to Bang et al. (2023). Qin et al. (2023)'s evaluation also showed that while GPT-4 was good at mathematical reasoning, it's performance suffered with logical and commonsense reasoning even with CoT. Moreover,Arkoudas (2023) posit that LLMs can not reason per se due to its inconsistent and often incompetent output, an argument we further by comparing ChatGPT's error pattern to that which might otherwise be expected. Borji (2023) also arrived at a similar conclusion for specific classes of reasoning.

Gurnee and Tegmark (2023) showed that location data encoded in LLMs can be cast onto a world map with reasonable accuracy, and argue that this reflects true understanding of the data. However, decade-old word embeddings can likewise be mapped with similar accuracy (Konkol et al., 2017), without the need for understanding. Fu and Khot (2022) suggest that models trained on code are better at reasoning and CoT, although admit the evidence is weak (due to a lack of non-code-trained non-reasoning models). They hypothesise that training on code may improve LLMs' abilities to handle long term dependencies and hierar-

chical structures. There are also works that show that different strategies can be employed to improve reasoning abilities of LLMs. Both Hao et al. (2023) and Zhao et al. (2023a) showed that adding a reward based planning component to LLMs improve their reasoning capabilities. Whereas, Si et al. (2023) proposed that, instead of using a single LLM, using multiple LLMs as a mixture of reasoning expert improved LLM reasoning performance.

Whether or not LLMs have the ability to reason consistently is still the subject of much debate. A distinction must be drawn between having the ability to reason per se, and being able to produce some otherwise constructed reasoning-like output (Ullman, 2023). To illustrate the point, an examiner marking the work of a student who has written a series of logical statements may assume they have sound reasoning ability, whereas in reality the student may have merely memorised the facts and written them verbatim without the need for reasoning.

C Explanations

Tables A1-A3 contain examples of prompts and ChatGPT's responses to them.

D Full Prompts for Manual CoT

Tables A4-A5 contain the prompts used for Manual CoT. For 2-class experiments, the *Unverified* prompts were removed and instructions modified appropriately. For non-mathematical experiments, the prompts in Table A5 were also removed.

E Additional Results

Tables A6-A9 contain results that did not fit into the main text due to space constraints.

	Prompt	Explanation
gpt-3.5-turbo	ZS	Based on the evidence provided, the rumour is false. The evidence states that the cause of the crash is not known and the plane did not send a distress signal.
	ZS CoT	As an expert journalist, it is important to analyze the evidence provided and make an informed classification of the rumor as true or false. The rumor states that France's transport minister says a distress signal was sent from the Germanwings plane. However, the evidence provided contradicts this claim. It states that the cause of the crash is not known and that the plane sent no distress signal during an eight-minute descent. Based on this evidence, it can be concluded that the rumor is false. The provided evidence clearly states that no distress signal was sent from the plane, which contradicts the claim made by the transport minister.
	Manual CoT	The evidence states that the plane did not send a distress signal, therefore the rumour is false.
gpt-4	ZS	-
	ZS CoT	The rumour is false.
	Manual CoT	The plane did not send a distress signal, therefore the rumour is false.
Rumour/Evidence	**Rumour:** ALERT France's transport minister says distress signal was sent from Germanwings plane - @patthomas , @europe1 **Evidence:** The cause of the crash is not known and the plane sent no distress signal during an eight-minute descent.	

Table A1: Example of generated explanation quality and verbosity.

	Prompt	Explanation
gpt-3.5-turbo	ZS	Based on the evidence provided, the rumour that there were 'numerous gunmen' at the Canada War Memorial shooting is false. The evidence states that a gunman shot a soldier at the Canadian War Memorial and then proceeded to storm the Canadian parliament. There is no mention of multiple gunmen involved in the incident.
	ZS CoT	Based on the evidence provided, it is true that there was a shooting at the Canadian War Memorial and the Canadian Parliament. It is also true that a Canadian soldier was shot and a shooter was seen running towards the parliament buildings. However, there is no specific mention of "numerous gunmen" in the evidence. Therefore, we cannot classify the rumour as true or false based on the provided evidence alone.
	Manual CoT	The evidence confirms that there was a shooting at the Canadian War Memorial and the Canadian parliament, therefore the rumour is true.
gpt-4	ZS	-
	ZS CoT	The rumour is partially true. The evidence confirms that there was a shooting at the Canada War Memorial and a person (a Canadian soldier) was shot. However, the evidence only mentions one gunman, not 'numerous gunmen'.
	Manual CoT	The evidence confirms that there was a shooting at the Canada War Memorial and the Parliament, but it only mentions one gunman. Therefore, the rumour is false.
Rumour/Evidence	**Rumour:** Ottawa Police Service : There were ' numerous gunmen ' at the Canada War Memorial shooting . One person was shot . **Evidence:** OTTAWA - Gunfire echoed through the Gothic halls of the Canadian parliament Wednesday as police swarmed in to tackle a gunman who had shot a soldier outside before storming the Canadian parliament. 151; – A gunman fatally shot a Canadian soldier in an attack near the Parliament in what the prime minster called a ' terrorist ' act. A Canadian soldier was shot at the Canadian War Memorial and a shooter was seen running towards the nearby parliament buildings, where more shots were fired..	

Table A2: Example of ZS CoT producing classification based on rumour decomposition.

	Prompt	Explanation
gpt-3.5-turbo	ZS	Based on the evidence provided, the rumour is true. The evidence states that the police officer, Ahmed Merabet, was killed in the Charlie Hebdo attacks, and a colleague confirmed that he was a Muslim.
	ZS CoT	Based on the evidence provided, it is true that the first person killed in the Charlie Hebdo attacks was a Muslim police officer named Ahmed Merabet. The evidence states that he was the second police officer to be killed that day, and it is also mentioned that he was a Muslim according to a colleague. Therefore, the rumour is true.
	Manual CoT	The evidence confirms that the first person killed in the Charlie Hebdo attacks was not a Muslim police officer, but another officer assigned to protect Charlie Hebdo's editorial director. Therefore, the rumour is false.
gpt-4	ZS	-
	ZS CoT	The rumour is partially true. The evidence confirms that a Muslim police officer, Ahmed Merabet, was indeed killed in the Charlie Hebdo attacks. However, he was not the first person killed, but the second police officer to be killed that day.
	Manual CoT	The evidence confirms that a Muslim police officer, Ahmed Merabet, was killed in the Charlie Hebdo attacks, but he was not the first person killed that day. Therefore, the rumour is false.
Rumour/Evidence	**Rumour:** The first person killed in Charlie Hebdo attacks was Muslim police officer **Evidence:** The officer, Ahmed Merabet, was the second police officer to be killed that day; another officer assigned to protect Charlie Hebdo's editorial director was also killed. PARIS The French police officer whose point-blank murder by suspected Islamist extremists on a Paris street was caught on video was a Muslim, according to a colleague..	

Table A3: Self-contradiction in ZS.

Dataset	Mode	Label	Prompts
Wiki	Deductive	T	**Claim:** Emmental cheese typically contains holes. **Evidence:** Bacteria within Emmental produce carbon dioxide, creating its charateristic spherical voids. **Explanation:** The spherical voids of carbon dioxide are holes, so the claim is true.
		F	**Claim:** The best selling game of all time is Tetris. **Evidence:** Minecraft has sold more copies than Tetris. **Explanation:** Since another game has sold more copies than Tetris, this claim is false.
	Abductive	T	**Claim:** Cows may cause serious injury when provoked. **Evidence:** The size and weight of a cow makes it difficult to topple by hand. **Explanation:** Due to being large animals, cows are probably able to cause serious injury, so the claim is likely to be true.
		F	**Claim:** Krakow does not have a substantial history. **Evidence:** Krakow, the capital city of Poland, was the first UNESCO World Heritage Site. **Explanation:** Its status as a world heritage site suggests that Krakow has a notable history, so this claim is probably false.
	-	U	**Claim:** The most common favourite color of Chicago residents is blue. **Evidence:** Chicago is a large American city. **Explanation:** The evidence regarding the size of Chicago does not contain any information about the residents' favourite color, therefore the claim can be classified as unknown.
Pheme	Deductive	T	**Claim:** Go outside sheeple! Low vitamin d levels linked to COVID-19 mortality rates #wakeup #touchgrass **Evidence:** A new meta-analysis suggests that low vitamin d levels are linked to COVID-19 survival rates. **Explanation:** The meta-analysis confirms that COVID survival rates are linked to vitamin D levels, therefore the rumour is true.
		F	**Claim:** 7 NEWS an eagle has been filmed carrying away a small dog: #natureismetal **Evidence:** The video apparently showing an eagle carrying away a dog is revealed to be a hoax generated by AI. **Explanation:** The video was fabricated by an AI, so the alleged incident did not occur and the rumour is false.
	Abductive	T	**Claim:** BREAKING NEWS: More hostages have escaped the rapidly escalating situation in Miami. The lockdown continues #MiamiLockdown **Evidence:** People were seen running out of the building, according to eyewitness reports **Explanation:** People running from the building in the context of this situation are probably hostages, therefore the rumour is likely to be true.
		F	**Claim:** Donald trump has died by following his own instructions on how to cure coronavirus #oops **Evidence:** More legal trouble on its way for former US President Donald Trump. **Explanation:** It is unlikely that Donald Trump has died if he is actively being litigated against, so this rumour is probably false.
	-	U	**Claim:** omg.. everyone knows that terrorists get double points on fridays.. of course it happened like that. **Evidence:** The gunman was shot shortly afterwards, although the town center remains in lockdown as police swarm the scene. **Explanation:** The evidence does not correspond to the outlandish claim that terrorists prefer to act on Fridays, therefore the rumour can be classified as unknown.

Table A4: Prompts used for Manual CoT. Here we provide the five examples that were mode specific only and were structured around rumours.

Mode	Process	Label	Prompts
Deductive	Mathematical	T	**Claim:** The book has over 200 pages. **Evidence:** The book has 264 pages. **Explanation:** 264 is greater than 200, therefore the claim is true.
	Temporal	F	**Claim:** Alice's birthday is in the last half of the year. **Evidence:** Alice was born on 10th March 1996. **Explanation:** March is not in the last half of the year, therefore the claim is false.

Table A5: Last two prompts used for Manual CoT. These prompts were targeted towards mathematical and temporal reasoning. Therefore, they were created as more generic examples than rumour specific.

	Deductive				Abductive			
	gpt-3.5-turbo		gpt-4		gpt-3.5-turbo		gpt-4	
	✓	x	✓	x	✓	x	✓	x
ZS	23	1	23	1	2	1	3	0
ZS CoT	22	2	24	0	3	0	3	0
Manual CoT	21	3	24	0	3	0	3	0

Table A6: 3-class classification results of GPT-3.5-Turbo and GPT-4 on WIKI under zero-shot (ZS) and chain-of-thought (CoT) paradigms, stratified by reasoning mode, and excluding mathematical claims. Abductive claims/rumours include any with an abductive step in their reasoning path.

	Deductive				Abductive			
	gpt-3.5-turbo		gpt-4		gpt-3.5-turbo		gpt-4	
	✓	x	✓	x	✓	x	✓	x
ZS	21	6	20	7	2	5	1	6
ZS CoT	16	11	17	10	1	6	2	5
Manual CoT	18	9	19	8	2	5	1	6

Table A7: 3-class classification results of GPT-3.5-Turbo and GPT-4 on PHEME under zero-shot (ZS) and chain-of-thought (CoT) paradigms, stratified by reasoning mode, and excluding mathematical rumours. Abductive claims/rumours include any with an abductive step in their reasoning path.

	F	T	U
F	19	0	1
T	0	37	3
U	3	1	16

Zero Shot

	F	T	U
F	18	0	2
T	0	38	2
U	5	1	14

ZS CoT (Kojima et al., 2023)

	F	T	U
F	19	0	1
T	1	36	3
U	1	0	19

Manual CoT (Wei et al., 2023)

Table A8: 3-class classification results of GPT-4 on the Wikipedia-based dataset under the Zero Shot (ZS) and Chain of Thought (CoT) paradigms. F = *False*, T = *True*, U = *Unverified*. Bold letters indicate the ground-truth.

	F	T	U
F	5	9	11
T	0	26	4
U	1	4	29

Zero Shot

	F	T	U
F	7	6	12
T	2	21	7
U	1	4	29

ZS CoT (Kojima et al., 2023)

	F	T	U
F	9	3	13
T	3	23	4
U	1	4	29

Manual CoT (Wei et al., 2023)

Table A9: 3-class classification results of GPT-4 on the PHEME-based dataset under the Zero Shot (ZS) and Chain of Thought (CoT) paradigms. F = *False*, T = *True*, U = *Unverified*. Bold letters indicate the ground-truth.