Assessing the Reasoning Capabilities of LLMs in the context of Evidence-based Claim Verification

Anonymous ACL submission

Abstract

Although LLMs have shown great performance on Mathematics and Coding related reasoning tasks, the reasoning capabilities of LLMs regarding other forms of reasoning are still an open problem. Here, we examine the issue of reasoning from the perspective of claim verification. We propose a framework designed to break down any claim paired with evidence into atomic reasoning types that are necessary for verification. We use this framework to create RECV, the first claim verification benchmark, incorporating real-world claims, to assess the deductive and abductive reasoning capabilities of LLMs. The benchmark comprises of three datasets, covering reasoning problems of increasing complexity. We evaluate three state-of-the-art proprietary LLMs under multiple prompt settings. Our results show that while LLMs can address deductive reasoning problems, they consistently fail in cases of abductive reasoning. Moreover, we observe that enhancing LLMs with rationale generation is not always beneficial. Nonetheless, we find that generated rationales are semantically similar to those provided by humans, especially in deductive reasoning cases.

1 Introduction

Large Language Models (LLMs) have shown remarkable proficiency in complex tasks where reasoning capabilities, such as logical deduction and semantic comparison, are paramount. Notable examples include solving MBA exams (Terwiesch, 2023), passing professional medical tests (Kung et al., 2023; Nori et al., 2023), performing quantitative reasoning (Lewkowycz et al., 2022), and communication games (Bakhtin et al., 2022; Xu et al., 2023; Gandhi et al., 2023). However, there is ongoing debate about whether such proficiency is due to LLMs manifesting reasoning capabilities or rather pattern matching and semantic similarity via memorization. For example, earlier claims that LLMs posses Theory of Mind (ToM) capabilities (Bubeck et al., 2023; Kosinski, 2023) were shown to be inaccurate (Ullman, 2023; Sileo and Lernould, 2023). In particular, despite appearing to manifest some form of ToM capabilities, LLMs mostly rely on shallow heuristics and spurious correlations (Shapira et al., 2023). Additionally, preliminary observations of emergent reasoning capabilities (Wei et al., 2022) were subsequently attributed to metric choice (Schaeffer et al., 2023), in-context learning (Lu et al., 2023b), and shortcuts (Kavumba et al., 2019).

These findings motivate the need for further research on the reasoning capabilities of LLMs, especially in high-stake real-world applications, where research on this topic is in its infancy. A notable example is fact-checking, where LLMs are considered to hold great potential for increased productivity overshadowed by the ease with which bad actors can proliferate misinformation (Guo et al., 2023). Verifying information is challenging since models require both accurate classification and strong rationale generation to be effective (Schlichtkrull et al., 2023). It is thus essential to understand the reasoning capabilities and limitations of LLMs in the context of fact-checking. In particular, we extend the current discussion around the reasoning abilities of LLMs, focusing on their ability to verify real-world claims.

In this work, we first propose a framework for breaking down complex claims into atomic reasoning steps. The motivation behind this is the lack of uniform terminology around reasoning evaluation. Most prominent evaluation datasets for reasoning are based on mathematics and coding, which involves deductive reasoning but is not treated as such (Sprague et al., 2024).

Our framework is rooted in existing philosophy literature concerning logical reasoning that aligns well with NLP (Wason and Johnson-Laird, 1972; Galotti, 1989). We use our framework to create

Reasoning in **E**vidence-based **C**laim **V**erification (RECV), the first reasoning benchmark for claim-verification. The benchmark comprises three datasets, curated from existing resources targeting different domains: VitaminC (Schuster et al., 2021) from Wikipedia, CLIMATE-FEVER (Diggelmann et al., 2020) from online claims and Wikipedia, and PHEMEPlus (Dougrez-Lewis et al., 2022) from rumours circulating on social media and associated evidence from news articles. The claims involve increasing levels of complexity as we move from VitaminC to PHEMEPlus, often requiring deductive and/or abductive reasoning.

We use RECV to evaluate three state-of-the-art proprietary LLMs that have shown impressive performance on various reasoning and language benchmarks (Huang and Chang, 2023; DeepSeek-AI et al., 2025). These models are: Claude V3 Sonnet (Anthropic, 2023), GPT-4 (OpenAI, 2023), and GPT-4o (OpenAI et al., 2024). In particular, we prompt models with and without Chain-of-Thought (CoT) (Wei et al., 2023) rationale generation to assess if and how the latter influences reasoning. In alignment with previous work (Saparov et al., 2023; Akyürek et al., 2024; Li et al., 2024) we find that LLMs are capable of deductive reasoning. However, they consistently fail at claim verification when presented with evidence that requires abductive reasoning. Furthermore, we observe conflicting results when prompting LLMs with CoT strategies. In particular, CoT leads to performance improvements for simple claim verification as in VitaminC, but detrimental in the case of complex claims such as those found in CLIMATE-FEVER and PHEMEPlus. Lastly, we carry out a qualitative analysis of generated rationales and observe high semantic similarity with human explanations, especially in deductive reasoning cases. In summary, we make the following contributions:

- We propose a framework for decomposing claim-evidence pairs into atomic reasoning types for verification, covering deductive and abductive reasoning (§3).
- We create the first reasoning benchmark for claim verification comprising three datasets of increasing complexity (§4).
- We extensively evaluate the reasoning capabilities of three state-of-the-art LLMs, showing that models fail when it comes to abductive reasoning and CoT's effectiveness is task-dependent (§5).
- We show that generated rationales are consistent with human reasoning for correct predictions, but the model is often unable to leverage such rationales for claim verification (§6).

2 RECV Logical Framework

Reasoning is often used interchangeably to denote critical thinking, decision-making, and logical reasoning. Following Wason and Johnson-Laird (1972) and Galotti (1989), we define reasoning as the process of logical steps that result in some form of decision-making or conclusion. Thus we define reasoning as a series of inference steps linking claims and evidence to reach a conclusion.

In particular we consider that reasoning consists of the interplay of three interrelated components: *types*, *processes*, and *tasks*. This is the basis of our RECV framework. Reasoning *types* are different forms of logical inference that we can use to reach a conclusion from a set of observations or premises. We distinguish between atomic and compound reasoning types. Atomic types denote basic forms of logical inference and include deductive, abductive, inductive, and analogical reasoning. A reasoning *task* is any task that requires multiple reasoning types, often in complex interaction with each other. For example claim verification is a composite reasoning task. A reasoning *process* is the method of interaction between reasoning types or even tasks within complex reasoning tasks. Notable examples of reasoning processes are multi-hop or multi-step inference, where individual steps or hops can be of different reasoning types. In this paper we focus particularly on atomic reasoning types.

2.1 Atomic Reasoning Types

Deduction: A conclusion is drawn directly from evidence. In the context of a claim, if the evidence supports the claim then the claim is deduced to be true (if P then Q, where P is the claim and Q is the evidence). For example,

- P: *Schools closed, Dammartin-en-Goele residents told to stay indoors, town 'like warzone'.* **[Claim]**
- Q: *Schools went into lockdown and the town appealed to residents to stay inside residents' houses.* **[Evidence]**
- C: Here, $P \implies Q$. The schools have been closed and citizens have been told to stay home. Thus, the town is like in a warzone situation. **[Conclusion]**

Equally if the evidence contradicts the claim then the claim is deduced to be false.

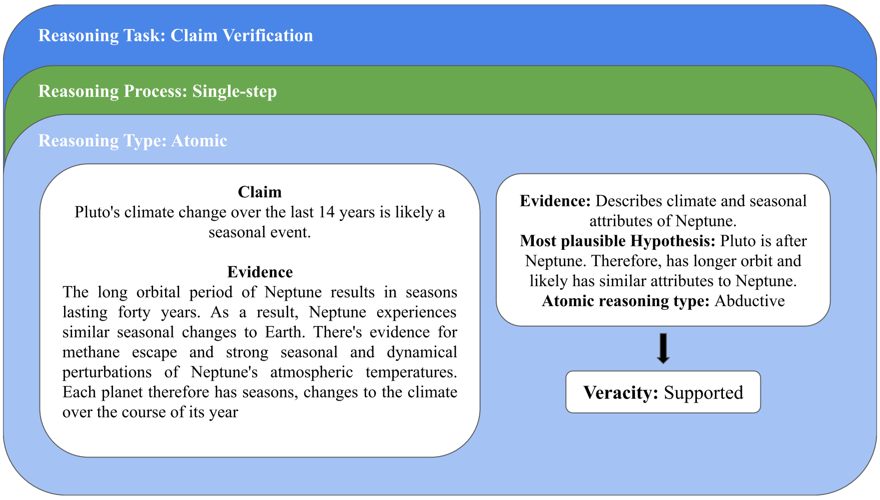

Figure 1: Resolution of claim verification via a single-step abductive reasoning type using RECV framework.

- P: *Heart goes out to 148 passengers and crew of Germanwings Airbus A320 that has crashed in French Alps, Southern France.* **[Claim]**
- Q: *German jetliner carrying 144 passengers and six crew en route from Barcelona, Spain, to Düsseldorf, Germany, has crashed in the French Alps, killing all 150 people on board.* **[Evidence]**
- C: Here, Q contradicts P. The evidence directly states the death toll is 150 which refutes the claim. **[Conclusion]**

Abduction: The most plausible conclusion is drawn from a set of candidate hypotheses, based on partial evidence. Abduction could lead to false conclusions.

- **Claim:** *Pluto's climate change over the last 14 years is likely a seasonal event.*
- **Evidence:** *The long orbital period of Neptune results in seasons lasting forty years. As a result, Neptune experiences similar seasonal changes to Earth. There's evidence for methane escape and strong seasonal and dynamical perturbations of Neptune's atmospheric temperatures. Each planet therefore has seasons, changes to the climate over the course of its year.*
- **Conclusion:** *The evidence only mentions Neptune. However, the claim is regarding Pluto. Given the partial evidence, the claim is supported based on the plausible hypothesis that Pluto is near Neptune and it is likely to have similar attributes when it comes to seasons and climate change.*

Induction: An inference is drawn from complete evidence (in a specific domain) and then a generalization (a rule that can be used beyond the initial domain) is derived from it. As per Flach and Kakas (2000), for inductive reasoning, the evidence can be true whilst only providing partial support for the conclusion, which typically generalizes beyond the evidence itself. Such generalization indicates there is no guarantee that the conclusion is true elsewhere.

Analogical reasoning: Conclusions are drawn based on the similarities between entities. While we do not provide examples for inductive and analogical reasoning in this section, they are still part of our framework. The focus on deduction and abduction is justified in (§3). We provide more formal definitions of atomic reasoning types in Appendix A with additional examples.

3 Methodology

We discuss our methodology for reasoning in claim verification. We first showcase the application of our RECV framework and then motivate our focus on deduction and abduction via a preliminary study.

RECV Logical Framework Application: The application of the RECV framework can be seen in Figure 1. The *reasoning task* here is claim verification and the *reasoning type* is composite. The claim is resolved using a *single-step process*, that consists of abductive *type atomic reasoning*. Here we only highlight the most plausible hypothesis that resolves the claim as *true*. However, in practice, we would generate multiple hypotheses before coming to the most plausible one.

Preliminary Study: Our objective here was to determine the atomic reasoning types necessary for accomplishing claim verification from evidence. We first collected a small dataset by manually selecting 90 claims and associated evidence from VitaminC (Schuster et al., 2021), CLIMATE-FEVER (Diggelmann et al., 2020), and PHEMEPlus (Dougrez-Lewis et al., 2022). We focus on these resources as they are widely used in claim and rumour verification and differ in complexity. Two annotators with expertise in Computer Science and native English proficiency assigned reasoning type labels to claim-evidence pairs following (§2). Disagreements encountered were resolved via a discussion stage with an independent expert. The Inter-Annotator Agreement (IAA) measured as Bennett's S score (Bennet et al., 1954) to account for label imbalance of reasoning types is 0.90, denoting almost perfect agreement. We observe that all examples either require deductive or abductive reasoning types.

Deductive and Abductive Reasoning: Our preliminary investigation suggests that inductive and analogical reasoning are rarely employed in claim verification. This is presumably because inductive reasoning relies on complete evidence, which is rarely available in real-world domain-specific settings. Generalisations from one domain to another, relevant to inductive reasoning, may only occur in scenarios that share common background knowledge, as in the medical domain. Similarly, analogical reasoning may be more suitable for other fact-checking related tasks like profiling and motive analysis where comparing information bits frequently occurs to reach a conclusion. By contrast, deductive and abductive reasoning types are often required in fact-checking (Pan et al., 2023; Tan et al., 2024). For these reasons, here we focus on deduction and abduction. We show that they represent a challenging setting for claim verification (§4), and model evaluation with LLMs (§5).

4 RECV Benchmark

We discuss the creation of RECV, in particular, our sample selection strategy and data annotation. See Appendix B for details regarding the three datasets.

Data Sampling Strategy. We build a heuristic-based sampling strategy to mitigate the anticipated data imbalance between deductive and abductive samples, as it was important to ensure both are represented in the annotated data. We used a combination of three embedding-based text similarity metrics to compute the average claim similarity between deductive samples. Likewise for abductive samples. The metrics are: cosine similarity, BERTScore (Zhang et al., 2020) and BLEURT score (Pu et al., 2021). We used the data collected during our preliminary study (§3) to set a similarity threshold for each reasoning type. We used each threshold to sample claims likely to be resolved via deductive and abductive reasoning, respectively. In particular, we exclude instances labeled as 'unverified' since such claims are always associated with deductive reasoning, either due to lack of proper evidence or contradictory evidence. In total, we sampled 500 claim-evidence pairs from each dataset. See Appendix C for more details on data sampling.

VitaminC	**Supported**	**Refuted**	**Total**
Deductive	272	199	471
Abductive	11	18	29
Total	283	217	500
CLIMATE-FEVER	**Supported**	**Refuted**	**Total**
Deductive	269	129	398
Abductive	88	14	102
Total	357	143	500
PHEMEPlus	**Supported**	**Refuted**	**Total**
Deductive	336	128	464
Abductive	22	14	36
Total	358	142	500

Table 1: RECV statistics.

Data Annotation We recruited 9 PhD students in Computer Science, fluent in English and grouped them in triples, one for each dataset. We evenly distributed dataset samples to annotators in a triple, so that 100 samples were annotated by all. In total, each annotator in a triple labeled 233 samples. Annotation guidelines per dataset are in Appendix D. We computed IAA as Bennett's S score to account for label imbalance (see Appendix E for pairwise agreement scores). The IAA is 0.75 for VitaminC, 0.56 for CLIMATE-FEVER, and 0.67 for PHEMEPlus. Table 1 reports our RECV statistics. In particular, we observe that the rate of abductive reasoning samples is relatively low compared to deductive ones: 5.8% in VitaminC, 20.4% in CLIMATE-FEVER, and 7.2% in PHEMEPlus. This imbalance is expected given the nature of collected evidence; most evidence provided, either in the form of Wikipedia articles as in VitaminC and CLIMATE-FEVER or news articles as in PHEMEPlus, contains detailed information to deductively verify the claim. In total, RECV consists of 1500 claim-evidence pairs with associated veracity and reasoning labels. The average sentence length for evidence in VitaminC is 1.084, 7.562 for PHEMEPlus and 7.828 for CLIMATE-FEVER. This highlights the varying complexity of the datasets and RECV.

5 Claim Verification with LLMs

Our objective here is to assess the capabilities of LLMs in performing deductive and abductive reasoning to determine the veracity of a claim.

Setup: We formulate claim verification as a prediction task. Given a claim-evidence pair, we prompt LLMs to predict whether the evidence supports or refutes the claim (Figure A2 (bottom)). We consider two settings: ***No-Exp*** and ***Exp***. In ***No-Exp***, we prompt LLMs to predict the claim veracity without any rationale generation. In ***Exp***, we first prompt LLMs to produce a rationale and then use the generated information to predict claim veracity. For each setting, we consider two different prompt strategies: Zero-Shot (ZS), and Manual Chain-of-Thought (M-CoT) (Wei et al., 2023). In addition, in ***Exp***, we also consider Zero-Shot Chain-of-Thought (ZS CoT) (Kojima et al., 2023). ZS CoT was applied only under ***Exp*** as rationale generation is integral to ZS CoT prompting. In all the prompts, we provide dataset specific personas and instructions in the system prompt and CoT examples in the user prompt. We report the prompts in Appendix F.

Metrics: We compute macro F1 score for veracity of claims given the evidence and the error rate of claim-evidence pairs concerning deductive and abductive reasoning types, respectively. F1 was chosen due to the class imbalance in CLIMATE-FEVER and PHEMEPlus (they have a 70/30 ratio between support and refute labels). We use annotators' reasoning type labels for claim-evidence pairs to identify errors in verification per category (cases of abduction vs deduction) and express it via error rate.

Models: We consider three state-of-the-art proprietary LLMs with remarkable proficiency in a wide range of tasks: Claude V3 Sonnet (Anthropic, 2023), GPT-4 (OpenAI, 2023), and GPT-4o (OpenAI et al., 2024). We conducted our experiments using OpenAI and Anthropic's official API.

5.1 Results

Table 2 reports classification performance and error rates per reasoning type for claim verification on RECV. We discuss dataset-specific results in detail.

VitaminC: Among prompting strategies, M-CoT leads to the highest increase in performance across all models. The average error rate across all models and settings is 10.31% for deductive reasoning and 32% for abductive reasoning. *This shows that all models struggle with abductive reasoning, even in less challenging settings like VitaminC.* Regarding model settings, we observe conflicting results. In particular, generating rationales improves veracity classification performance for deductive samples in all models, except for GPT-4 M-CoT. By contrast, only Claude ZS and GPT-4o ZS show improvements in *Exp* compared to *No-Exp* when targeting abductive reasoning. Overall, when moving to the *Exp* settings, we observe a 7.5% average performance drop, with GPT-4 reporting the highest degradation (-10%). Lastly, regarding prompting strategies, we observe that M-CoT outperforms CoT in deductive cases, while reporting comparable results in abductive ones.

CLIMATE-FEVER: The average error rate across all models and settings is 15.58% for deductive reasoning and 48.58% for abductive reasoning. Similar to VitaminC, these results denote that LLMs fail at predicting claim veracity when dealing with abductive reasoning. In particular, abductive reasoning samples are on average three times more challenging than deductive ones. Regarding model settings, we observe that rationale generation leads to performance degradation in all scenarios. Overall, we observe a 4.36% average performance drop for deductive cases and 14.76% for abductive ones. Regarding prompting strategies, we observe similar results to VitaminC where M-CoT outperforms CoT. In particular, the average error rate for M-CoT is 14.24% on deductive cases ($+2.76$) and 45.1% on abductive ones ($+8.17$).

PHEMEPlus: The results suggest that there is no model or prompting strategy that consistently outperforms others. The average error rate across all models and settings is 20.06% for deductive reasoning and 44.68% for abductive reasoning. Compared to VitaminC and CLIMATE-FEVER, PHEMEPlus represents a more challenging setting for deductive reasoning, while it is comparable in complexity with CLIMATE-FEVER when assessing

Model	VitaminC F1 ↑	VitaminC Deductive ↓	VitaminC Abductive ↓	CLIMATE-FEVER F1 ↑	CLIMATE-FEVER Deductive ↓	CLIMATE-FEVER Abductive ↓	PHEMEPlus F1 ↑	PHEMEPlus Deductive ↓	PHEMEPlus Abductive ↓
Claude ZS *No-Exp*	0.85	13.62	33.33	0.80	12.81	40.20	0.73	19.40	**38.89**
Claude M-CoT *No-Exp*	0.87	12.77	**23.33**	0.80	12.81	41.84	**0.76**	18.53	**38.89**
GPT-4 ZS *No-Exp*	0.86	12.13	30.00	0.87	**8.79**	**20.59**	0.69	20.69	**38.89**
GPT-4 M-CoT *No-Exp*	**0.90**	**8.30**	26.67	0.85	10.05	27.45	0.70	22.41	52.78
GPT-4o ZS *No-Exp*	0.88	10.43	40.00	0.84	9.55	33.33	0.72	20.04	41.67
GPT-4o M-CoT *No-Exp*	0.88	10.43	30.00	**0.92**	9.05	25.49	0.74	19.40	47.22
Claude ZS *Exp*	0.89	9.79 $^{(+3.83)}$	30.00 $^{(+3.33)}$	0.74	17.34 $^{(-4.53)}$	52.94 $^{(-12.75)}$	0.74	20.04 $^{(-0.65)}$	41.67 $^{(-2.78)}$
Claude ZS CoT *Exp*	0.88	11.06	30.00	0.70	22.61	56.86	0.73	21.34	41.67
Claude M-CoT *Exp*	0.90	8.72 $^{(+4.04)}$	30.00 $^{(-6.67)}$	0.73	17.59 $^{(-4.77)}$	57.84 $^{(-16.01)}$	0.73	23.49 $^{(-4.96)}$	41.67 $^{(-2.78)}$
GPT-4 ZS *Exp*	0.88	10.64 $^{(+1.49)}$	36.67 $^{(-6.67)}$	0.78	15.08 $^{(-6.28)}$	47.06 $^{(-26.47)}$	0.73	20.04 $^{(+0.65)}$	52.78 $^{(-13.89)}$
GPT-4 ZS CoT *Exp*	0.88	10.00	36.67	0.77	14.32	57.84	0.71	21.12	50.00
GPT-4 M-CoT *Exp*	0.89	8.72 $^{(-0.42)}$	36.67 $^{(-10.00)}$	0.82	11.31 $^{(-1.26)}$	35.29 $^{(-7.84)}$	0.73	19.61 $^{(+2.80)}$	50.00 $^{(+2.78)}$
GPT-4o ZS *Exp*	0.89	9.15 $^{(+1.28)}$	30.00 $^{(+10.00)}$	0.79	14.07 $^{(-4.52)}$	42.16 $^{(-8.82)}$	0.74	**18.32** $^{(+1.72)}$	44.44 $^{(-2.78)}$
GPT-4o ZS CoT *Exp*	0.89	9.36	30.00	0.78	14.07	55.88	0.74	18.97	44.44
GPT-4o M-CoT *Exp*	0.89	8.94 $^{(+9.96)}$	36.67 $^{(-6.67)}$	0.78	13.82 $^{(-4.77)}$	42.16 $^{(-16.67)}$	0.75	18.75 $^{(+0.65)}$	47.22 $^{(+0.00)}$

Table 2: Claim verification performance on RECV. Best results are in **bold**, second-best results are underlined. We report error rate delta performance between *No-Exp* and *Exp* settings in brackets. Negative delta indicates that rationale generation degrades perforamnce.

LLMs for claim verification. Regarding model settings, we observe minor performance improvements when prompting LLMs to generate rationales in deductive cases, with a 1.46% average gain. Claude is the only exception with a 2.81% average performance drop when moving to *Exp*. By contrast, we observe notable performance degradation in abductive reasoning cases, with GPT-4 ZS *Exp* being the worst (−13.89%). Lastly, regarding prompting strategies, we observe no performance difference between ZS CoT and M-CoT, highlighting the higher task complexity in PHEMEPlus.

6 Explanation evaluation

Providing reasonable explanations to support predicted veracity labels is a crucial aspect of claim verification systems. In particular, an automated system needs to be both convincing and trustworthy to convince users in practice (Schlichtkrull et al., 2023). Therefore, we evaluate the LLMs generated rationales in the *Exp* setting. This is paramount considering that LLMs tend to hallucinate (Bouyamourn, 2023; Rawte et al., 2023) and be self-contradictory at times (Mündler et al., 2023). We randomly selected 100 samples from each dataset in RECV and compared generated rationales against those provided by human annotators. We restricted sample selection to those where at least three models predicted wrong veracity labels. We follow Song et al. (2024) and compute *Factual Consistency* (FC), *Evidence Appropriateness* (EA), *BARTScore* (Yuan et al., 2021), and *Perplexity* (PPL) to assess the quality of the generated explanations. We provide additional details about metrics in Appendix G.

6.1 Results

Table 3 reports the results concerning explanation evaluation. We observe that all models achieve comparable results on appropriateness (EA), consistency (FC), and coherence measured via BARTScore (BART), while showing notable discrepancies regarding perplexity (PPL). In particular, GPT-4o ZS CoT has the most faithful rationales across all datasets. Moreover, prompting strategies like ZS CoT and M-CoT do not lead to consistent improvements over ZS, suggesting that their effectiveness may be problem- and model-dependent.

Additionally, we assess generated rationales regarding correct and wrong model predictions in Appendix G. Our results show that rationales from correct predictions better align with ground-truth explanations, suggesting that wrong predictions are usually the by-product of incorrect reasoning (the model is unable to leverage the explanation).

Lastly, we analyze how similar the generated rationales were between the models. To do so, we perform a permutation test using sentence-level contradiction scores from Fact_Score (see Appendix G). We find that Claude ZS has the most unique rationales on all datasets.

We discuss properties of rationales generated in the case of abductive and deductive errors per

	VitaminC				CLIMATE-FEVER				PHEMEPlus			
Model	**EA ↑**	**FC ↑**	**BART ↑**	**PPL ↓**	**EA ↑**	**FC ↑**	**BART ↑**	**PPL ↓**	**EA ↑**	**FC ↑**	**BART ↑**	**PPL ↓**
Claude ZS	0.85	0.85	-4.16	99.63	0.87	0.88	-4.26	31.08	0.82	0.81	-4.31	39.82
Claude ZS CoT	0.82	0.83	-4.38	52.85	0.82	0.83	-4.28	29.57	0.85	0.84	-4.44	38.81
Claude M-CoT	0.85	0.86	-4.05	68.53	0.89	0.90	-3.42	25.52	**0.89**	0.88	-4.17	37.83
GPT-4 ZS	0.87	0.87	-3.83	66.84	0.91	0.91	-3.67	27.93	0.87	0.86	-3.89	40.83
GPT-4 ZS CoT	0.87	0.86	-3.78	59.01	0.90	0.91	-3.65	**20.65**	0.87	0.87	-3.89	**28.83**
GPT-4 M-CoT	0.89	0.88	**-2.98**	45.84	**0.93**	**0.94**	**-2.90**	28.13	0.85	0.85	**-3.40**	47.15
GPT-4o ZS	0.90	0.88	-3.63	52.96	0.92	0.93	-3.64	58.42	0.85	0.86	-4.01	99.63
GPT-4o ZS CoT	**0.91**	**0.89**	-3.45	**35.82**	**0.93**	**0.94**	-3.39	46.92	**0.89**	**0.90**	-3.74	52.85
GPT-4o M-CoT	0.90	0.88	-3.63	50.10	0.90	0.91	-3.57	57.65	0.87	0.86	-4.08	68.53

Table 3: Qualitative evaluation on RECV in the *Exp* setting. Best results are in **bold**, second-best results are underlined.

dataset.

VitaminC We observe that LLMs struggle to generate faithful rationales in abductive cases. In particular, models tend to generate assertions rather than hedged information. This has implications for claim verification, where models predominantly refute or misclassify the veracity of the claim based on the generated explanations. Regarding deductive reasoning, we observe that the majority of errors are due to internal biases of LLMs, heavily influencing rationale generation, and to semantic faults in attending to only some parts of the claim and evidence.

CLIMATE-FEVER Regarding abductive cases, we observe the same issue reported in VitaminC. Regarding deductive reasoning, the majority of failures are due to implicit reasoning where relevant evidence information is implicit or where temporal relations between factual content must be understood to reach the correct conclusion.

PHEMEPlus Contrary to VitaminC and CLIMATE-FEVER, abductive and deductive reasoning errors are mainly due to semantic interpretation issues where models focus only on specific information in the claim and evidence. This limits models in assessing claim-evidence pairs in their entirety, thus, hindering them in capturing relations between the evidence and the claim. As in VitaminC, this issue affects the claim verification performance, often leading to misclassification.

7 Findings

We discuss the main findings of our work, including task complexity, the effectiveness of prompting strategies, and rationale generation.

Reasoning and Task complexity. Our results show that abductive reasoning is consistently more challenging than deductive reasoning. In particular, the performance gap between the two cases is around three times on average. This is mainly motivated by LLMs failing in performing uncertainty reasoning, often leading to erroneous assertive conclusions. Nonetheless, this is not the only issue that makes RECV challenging; task complexity plays a crucial role in reasoning performance. For instance, PHEMEPlus represents a more complex setting than VitaminC where news articles can contain extensive amount of information compared to Wikipedia pages. As shown in our qualitative analysis, LLMs tend to focus only on specific parts of input claim-evidence pairs, leading to suboptimal performance. For example, in the following claim from VitaminC,

Claim: Peking University is in a unitary sovereign state that's located in East Asia.

Evidence: Peking University abbreviated PKU is a major Chinese research university located in Beijing and a member of the C9 League.

the veracity of is deductively refuted. However, all the models labelled this pair as evidence supporting the claim. The rationale provided was that China was a sovereign country, ignoring the claim completely. This shows the over-reliance on specific parts of the evidence by the models while ignoring others. Overall, our findings suggest that LLMs' reasoning capabilities are domain and task dependent. Thus, we believe RECV represents a valuable resource to assess reasoning capabilities since it covers a wide spectrum of settings concerning claim verification.

Prompting Settings and Strategies Our experiments show that prompting strategies like ZS CoT and M-CoT do not lead to systematic performance improvements, but are rather specific to datasets (e.g., VitaminC) and models. These results align with recent findings about CoT being beneficial mainly for math- and code-related tasks (Sprague et al., 2024). This is likely derived by divergent reasoning paths within the models during inference that lead to reduction in performance (Chollet, 2023; Todd et al., 2023; Dutta et al., 2024). Furthermore, we also observe that internal alignment can hinder reasoning capabilities when it comes to abductive reasoning. Models are averse to provide predictions when evidence is incomplete. Yet abductive reasoning is often required for more complex tasks such as legal reasoning, just in time fact-checking, and other diverse forms of composite reasoning tasks. Hence, in order to achieve good results on these type of reasoning tasks, LLMs need to improve in the direction of abductive reasoning.

Explanation Quality Our evaluation of generated explanations shows that these are on average consistent with human rationales. In particular, ZS CoT rationales are more convincing due to their verbosity, whereas M-CoT rationales are more concise. Moreover, we observe that rationales generated for abductive reasoning cases resemble assertions as models disprefer generating uncertain rationales. Nonetheless, considering that our results are limited to macro performance results and given the limited number of abductive cases, we believe our estimates to decrease as dataset size increases. We leave a fine-grained analysis on generated rationales concerning an extended version of RECV as future work.

8 Related work

LLMs for Reasoning Several contributions have evaluated different reasoning capabilities in LLMs, including atomic and compounds types. For instance, LLMs can perform **abductive** reasoning for event prediction (Shi et al., 2023), but struggle with common sense reasoning (Zhao et al., 2023). Similarly, **deductive** reasoning in LLMs is beneficial to theorem-proving (Saparov et al., 2023), factual content generation (Akyürek et al., 2024), and question-answering (Li et al., 2024). Nonetheless, the observed improvements are often attributable to how prompts are designed rather than an emergent deductive capability (Chen et al., 2024). Moreover, LLMs perform out-of-context **inductive** reasoning (Treutlein et al., 2024), but fail in lexical tasks (Ye et al., 2023). Regarding **analogical** reasoning, LLMs address a wide variety of tasks, including nonverbal tests (Webb et al., 2023; Hu et al., 2023), question-answering (Yu et al., 2023), mathematical problem solving (Yasunaga et al., 2024), and planning (Yu et al., 2024), but present shortcomings in as many others (Ye et al., 2024; Sourati et al., 2024; Stevenson et al., 2024; Ahrabian et al., 2024; Lewis and Mitchell, 2024). Likewise, despite promising results in **compound** reasoning tasks, such as counterfactual (Wu et al., 2023), and compositional reasoning (Lu et al., 2023a), LLMs are notably unreliable (Gao et al., 2023; Zhang et al., 2024), sensitive to context (Hosseini et al., 2024; Chang and Bergen, 2024), and rely on shortcuts (Yang et al., 2024).

LLMs for Claim Verification Early work with LLMs focused on verifying simple facts (Lee et al., 2020). More recently, LLMs for claim verification have been augmented with external knowledge (Li et al., 2023a; Cheung and Lam, 2023), prompt-based reasoning (Cao, 2023; Li et al., 2023b; Lin et al., 2023), claim decomposition for fine-grained search into text chunks (Li et al., 2023a) or first-order logic terms (Wang and Shu, 2023), and data-augmentation (Alhindi et al., 2023). While LLMs have been extensively applied in fact-checking, the question of which reasoning capabilities are needed to verify claims remains unaddressed. Thus, we are the first to propose a reasoning benchmark for claim verification.

9 Conclusion

We propose a novel extendable logical reasoning framework for deconstructing claim-evidence pairs into reasoning steps, required to determine the veracity of a claim. We use our framework to create RECV, the first reasoning benchmark for claim verification focussed on deductive and abductive reasoning. Our results show that LLMs notably struggle with abductive reasoning, while performing better in deductive cases. Our findings show that LLMs reasoning capabilities are domain and task dependent. In particular, no specific prompting strategy, including rationale generation, is systematically beneficial across all datasets and models. Nevertheless, rationales generated by LLMs for deductive reasoning are on average consistent with human ones. Overall, these results highlight that RECV represents a challenging reasoning setting for LLMs and further research is required to reach satisfying performance.

Limitations

Dataset Selection and Reasoning Types Our focus on deductive and abductive reasoning types is dictated by our findings in the preliminary study. Nonetheless, other resources could be investigated to expand our approach to include other reasoning types. An example domain is biomedicine where datasets like COVID-Fact (Saakyan et al., 2021) could include examples where inductive reasoning is required to infer claim veracity.

Models We analyse three widely adopted proprietary LLMs. However, other models, including open-source ones, are also widely assessed in reasoning tasks. For a broader evaluation of LLMs, our study could include other models although these are currently unlikely to outperform the most established proprietary models.

Rationale Generation When LLMs generate 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.

Ethics Statement

The PHEMEPlus dataset is a pre-existing dataset of rumours, for which ethical approval was obtained by the original research team. The rest of the datasets were sampled from pre-existing datasets for which no ethical approval was required.

References

Kian Ahrabian, Zhivar Sourati, Kexuan Sun, Jiarui Zhang, Yifan Jiang, Fred Morstatter, and Jay Pujara. 2024. The curious case of nonverbal abstract reasoning with multi-modal large language models. *CoRR*, abs/2401.12117.

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.

Anthropic. 2023. The claude 3 model family: Opus, sonnet, haiku.

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.

E. M. Bennet, R. Alpert, and A. C. Goldstein. 1954. Communications through limited-response questioning*. *Public Opinion Quarterly*, 18(3):303–308.

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.

Tyler A. Chang and Benjamin K. Bergen. 2024. Language model behavior: A comprehensive survey. *Computational Linguistics*, 50(1):293–350.

Xinyun Chen, Ryan A. Chi, Xuezhi Wang, and Denny Zhou. 2024. Premise order matters in reasoning with large language models. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

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

François Chollet. 2023. How I think about LLM prompt engineering — fchollet.substack.com. `https://fchollet.substack.com/p/how-i-think-about-llm-prompt-engineering`. [Accessed 11-02-2025].

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Shengfeng Ye, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanjia Zhao, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning.

Thomas Diggelmann, Jordan Boyd-Graber, Jannis Bulian, Massimiliano Ciaramita, and Markus Leippold. 2020. Climate-fever: A dataset for verification of real-world climate claims.

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.

Subhabrata Dutta, Joykirat Singh, Soumen Chakrabarti, and Tanmoy Chakraborty. 2024. How to think step-by-step: A mechanistic understanding of chain-of-thought reasoning. *ArXiv*, abs/2402.18312.

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

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.

Jinglong Gao, Xiao Ding, Bing Qin, and Ting Liu. 2023. 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.

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.

Arian Hosseini, Alessandro Sordoni, Daniel Kenji Toyama, Aaron Courville, and Rishabh Agarwal. 2024. Not all LLM reasoners are created equal. In *The 4th Workshop on Mathematical Reasoning and AI at NeurIPS'24*.

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. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1049–1065, Toronto, Canada. 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.

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

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?

Martha Lewis and Melanie Mitchell. 2024. Using counterfactual tasks to evaluate the generality of analogical reasoning in large language models. *CoRR*, abs/2402.08955.

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.

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

Yian Li, Wentao Tian, Yang Jiao, Jingjing Chen, Na Zhao, and Yu-Gang Jiang. 2024. Look before you decide: Prompting active deduction of mllms for assumptive reasoning. *arXiv preprint arXiv:2404.12966*.

Yifei Li, Zeqi Lin, Shizhuo Zhang, Qiang Fu, Bei Chen, Jian-Guang Lou, and Weizhu Chen. 2023b. 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, 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?

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.

OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, Alex Nichol, Alex Paino, Alex Renzin, Alex Tachard Passos, Alexander Kirillov, Alexi Christakis, Alexis Conneau, Ali Kamali, Allan Jabri, Allison Moyer, Allison Tam, Amadou Crookes, Amin Tootoochian, Amin Tootoonchian, Ananya Kumar, Andrea Vallone, Andrej Karpathy, Andrew Braunstein, Andrew Cann, Andrew Codispoti, Andrew Galu, Andrew Kondrich, Andrew Tulloch, Andrey Mishchenko, Angela Baek, Angela Jiang, Antoine Pelisse, Antonia Woodford, Anuj Gosalia, Arka Dhar, Ashley Pantuliano, Avi Nayak, Avital Oliver, Barret Zoph, Behrooz Ghorbani, Ben Leimberger, Ben Rossen, Ben Sokolowsky, Ben Wang, Benjamin Zweig, Beth Hoover, Blake Samic, Bob McGrew, Bobby Spero, Bogo Giertler, Bowen Cheng, Brad Lightcap, Brandon Walkin, Brendan Quinn, Brian Guarraci, Brian Hsu, Bright Kellogg, Brydon Eastman, Camillo Lugaresi, Carroll Wainwright, Cary Bassin, Cary Hudson, Casey Chu, Chad Nelson, Chak Li, Chan Jun Shern, Channing Conger, Charlotte Barette, Chelsea Voss, Chen Ding, Cheng Lu, Chong Zhang, Chris Beaumont, Chris Hallacy, Chris Koch, Christian Gibson, Christina Kim, Christine Choi, Christine McLeavey, Christopher Hesse, Claudia Fischer, Clemens Winter, Coley Czarnecki, Colin Jarvis, Colin Wei, Constantin Koumouzelis, Dane Sherburn, Daniel Kappler, Daniel Levin, Daniel Levy, David Carr, David Farhi, David Mely, David Robinson, David Sasaki, Denny Jin, Dev Valladares, Dimitris Tsipras, Doug Li, Duc Phong Nguyen, Duncan Findlay, Edede Oiwoh, Edmund Wong, Ehsan Asdar, Elizabeth Proehl, Elizabeth Yang, Eric Antonow, Eric Kramer, Eric Peterson, Eric Sigler, Eric Wallace, Eugene Brevdo, Evan Mays, Farzad Khorasani, Felipe Petroski Such, Filippo Raso, Francis Zhang, Fred von Lohmann, Freddie Sulit, Gabriel Goh, Gene Oden, Geoff Salmon, Giulio Starace, Greg Brockman, Hadi Salman, Haiming Bao, Haitang Hu, Hannah Wong, Haoyu Wang, Heather Schmidt, Heather Whitney, Heewoo Jun, Hendrik Kirchner, Henrique Ponde de Oliveira Pinto, Hongyu Ren, Huiwen Chang, Hyung Won Chung, Ian Kivlichan, Ian O'Connell, Ian O'Connell, Ian Osband, Ian Silber, Ian Sohl, Ibrahim Okuyucu, Ikai Lan, Ilya Kostrikov, Ilya Sutskever, Ingmar Kanitscheider, Ishaan Gulrajani, Jacob Coxon, Jacob Menick, Jakub Pachocki, James Aung, James Betker, James Crooks, James Lennon, Jamie Kiros, Jan Leike, Jane Park, Jason Kwon, Jason Phang, Jason Teplitz, Jason Wei, Jason Wolfe, Jay Chen, Jeff Harris, Jenia Varavva, Jessica Gan Lee, Jessica Shieh, Ji Lin, Jiahui Yu, Jiayi Weng, Jie Tang, Jieqi Yu, Joanne Jang, Joaquin Quinonero Candela, Joe Beutler, Joe Landers, Joel Parish, Johannes Heidecke, John Schulman, Jonathan Lachman, Jonathan McKay, Jonathan Uesato, Jonathan Ward, Jong Wook Kim, Joost Huizinga, Jordan Sitkin, Jos Kraaijeveld, Josh Gross, Josh Kaplan, Josh Snyder, Joshua Achiam, Joy Jiao, Joyce Lee, Juntang Zhuang, Justyn Harriman, Kai

Fricke, Kai Hayashi, Karan Singhal, Katy Shi, Kavin Karthik, Kayla Wood, Kendra Rimbach, Kenny Hsu, Kenny Nguyen, Keren Gu-Lemberg, Kevin Button, Kevin Liu, Kiel Howe, Krithika Muthukumar, Kyle Luther, Lama Ahmad, Larry Kai, Lauren Itow, Lauren Workman, Leher Pathak, Leo Chen, Li Jing, Lia Guy, Liam Fedus, Liang Zhou, Lien Mamitsuka, Lilian Weng, Lindsay McCallum, Lindsey Held, Long Ouyang, Louis Feuvrier, Lu Zhang, Lukas Kondraciuk, Lukasz Kaiser, Luke Hewitt, Luke Metz, Lyric Doshi, Mada Aflak, Maddie Simens, Madelaine Boyd, Madeleine Thompson, Marat Dukhan, Mark Chen, Mark Gray, Mark Hudnall, Marvin Zhang, Marwan Aljubeh, Mateusz Litwin, Matthew Zeng, Max Johnson, Maya Shetty, Mayank Gupta, Meghan Shah, Mehmet Yatbaz, Meng Jia Yang, Mengchao Zhong, Mia Glaese, Mianna Chen, Michael Janner, Michael Lampe, Michael Petrov, Michael Wu, Michele Wang, Michelle Fradin, Michelle Pokrass, Miguel Castro, Miguel Oom Temudo de Castro, Mikhail Pavlov, Miles Brundage, Miles Wang, Minal Khan, Mira Murati, Mo Bavarian, Molly Lin, Murat Yesildal, Nacho Soto, Natalia Gimelshein, Natalie Cone, Natalie Staudacher, Natalie Summers, Natan LaFontaine, Neil Chowdhury, Nick Ryder, Nick Stathas, Nick Turley, Nik Tezak, Niko Felix, Nithanth Kudige, Nitish Keskar, Noah Deutsch, Noel Bundick, Nora Puckett, Ofir Nachum, Ola Okelola, Oleg Boiko, Oleg Murk, Oliver Jaffe, Olivia Watkins, Olivier Godement, Owen Campbell-Moore, Patrick Chao, Paul McMillan, Pavel Belov, Peng Su, Peter Bak, Peter Bakkum, Peter Deng, Peter Dolan, Peter Hoeschele, Peter Welinder, Phil Tillet, Philip Pronin, Philippe Tillet, Prafulla Dhariwal, Qiming Yuan, Rachel Dias, Rachel Lim, Rahul Arora, Rajan Troll, Randall Lin, Rapha Gontijo Lopes, Raul Puri, Reah Miyara, Reimar Leike, Renaud Gaubert, Reza Zamani, Ricky Wang, Rob Donnelly, Rob Honsby, Rocky Smith, Rohan Sahai, Rohit Ramchandani, Romain Huet, Rory Carmichael, Rowan Zellers, Roy Chen, Ruby Chen, Ruslan Nigmatullin, Ryan Cheu, Saachi Jain, Sam Altman, Sam Schoenholz, Sam Toizer, Samuel Miserendino, Sandhini Agarwal, Sara Culver, Scott Ethersmith, Scott Gray, Sean Grove, Sean Metzger, Shamez Hermani, Shantanu Jain, Shengjia Zhao, Sherwin Wu, Shino Jomoto, Shirong Wu, Shuaiqi, Xia, Sonia Phene, Spencer Papay, Srinivas Narayanan, Steve Coffey, Steve Lee, Stewart Hall, Suchir Balaji, Tal Broda, Tal Stramer, Tao Xu, Tarun Gogineni, Taya Christianson, Ted Sanders, Tejal Patwardhan, Thomas Cunninghman, Thomas Degry, Thomas Dimson, Thomas Raoux, Thomas Shadwell, Tianhao Zheng, Todd Underwood, Todor Markov, Toki Sherbakov, Tom Rubin, Tom Stasi, Tomer Kaftan, Tristan Heywood, Troy Peterson, Tyce Walters, Tyna Eloundou, Valerie Qi, Veit Moeller, Vinnie Monaco, Vishal Kuo, Vlad Fomenko, Wayne Chang, Weiyi Zheng, Wenda Zhou, Wesam Manassra, Will Sheu, Wojciech Zaremba, Yash Patil, Yilei Qian, Yongjik Kim, Youlong Cheng, Yu Zhang, Yuchen He, Yuchen Zhang, Yujia Jin, Yunxing Dai, and Yury Malkov. 2024. Gpt-4o system card.

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

Liangming Pan, Xiaobao Wu, Xinyuan Lu, Anh Tuan Luu, William Yang Wang, Min-Yen Kan, and Preslav Nakov. 2023. Fact-checking complex claims with program-guided reasoning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6981–7004, Toronto, Canada. Association for Computational Linguistics.

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

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.

Amy Pu, Hyung Won Chung, Ankur P Parikh, Sebastian Gehrmann, and Thibault Sellam. 2021. Learning compact metrics for mt. In *Proceedings of EMNLP*.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

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.

Arkadiy Saakyan, Tuhin Chakrabarty, and Smaranda Muresan. 2021. COVID-fact: Fact extraction and verification of real-world claims on COVID-19 pandemic. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2116–2129, Online. 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, Nedjma Ousidhoum, and Andreas Vlachos. 2023. The intended uses of automated fact-checking artefacts: Why, how and who. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8618–8642, Singapore. Association for Computational Linguistics.

Tal Schuster, Adam Fisch, and Regina Barzilay. 2021. Get your vitamin C! robust fact verification with contrastive evidence. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 624–643, Online. Association for Computational Linguistics.

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.

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.

Jiayu Song, Jenny Chim, Adam Tsakalidis, Julia Ive, Dana Atzil-Slonim, and Maria Liakata. 2024. Combining hierachical VAEs with LLMs for clinically meaningful timeline summarisation in social media. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14651–14672, Bangkok, Thailand. Association for Computational Linguistics.

Zhivar Sourati, Filip Ilievski, Pia Sommerauer, and Yifan Jiang. 2024. ARN: Analogical reasoning on narratives. *Transactions of the Association for Computational Linguistics*, 12:1063–1086.

Zayne Sprague, Fangcong Yin, Juan Diego Rodriguez, Dongwei Jiang, Manya Wadhwa, Prasann Singhal, Xinyu Zhao, Xi Ye, Kyle Mahowald, and Greg Durrett. 2024. To cot or not to cot? chain-of-thought helps mainly on math and symbolic reasoning.

Claire E. Stevenson, Alexandra Pafford, Han L. J. van der Maas, and Melanie Mitchell. 2024. Can large language models generalize analogy solving like people can? *CoRR*, abs/2411.02348.

Fiona Anting Tan, Jay Desai, and Srinivasan H. Sengamedu. 2024. Enhancing fact verification with causal knowledge graphs and transformer-based retrieval for deductive reasoning. In *Proceedings of the Seventh Fact Extraction and VERification Workshop (FEVER)*, pages 151–169, Miami, Florida, USA. Association for Computational Linguistics.

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.

Eric Todd, Millicent Li, Arnab Sen Sharma, Aaron Mueller, Byron C. Wallace, and David Bau. 2023. Function vectors in large language models. *ArXiv*, abs/2310.15213.

Johannes Treutlein, Dami Choi, Jan Betley, Samuel Marks, Cem Anil, Roger Baker Grosse, and Owain Evans. 2024. Connecting the dots: LLMs can infer and verbalize latent structure from disparate training data. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

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.

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.

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

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.

Sohee Yang, Elena Gribovskaya, Nora Kassner, Mor Geva, and Sebastian Riedel. 2024. Do large language models latently perform multi-hop reasoning? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10210–10229, Bangkok, Thailand. Association for Computational Linguistics.

Michihiro Yasunaga, Xinyun Chen, Yujia Li, Panupong Pasupat, Jure Leskovec, Percy Liang, Ed H. Chi, and Denny Zhou. 2024. Large language models as analogical reasoners. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

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.

Xiao Ye, Andrew Wang, Jacob Choi, Yining Lu, Shreya Sharma, Lingfeng Shen, Vijay Murari Tiyyala, Nicholas Andrews, and Daniel Khashabi. 2024. AnaloBench: Benchmarking the identification of abstract and long-context analogies. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 13060–13082, Miami, Florida, USA. Association for Computational Linguistics.

Junchi Yu, Ran He, and Zhitao Ying. 2024. Thought propagation: an analogical approach to complex reasoning with large language models. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Mengxia Yu, Zhihan Zhang, Wenhao Yu, and Meng Jiang. 2023. 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.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. Bartscore: Evaluating generated text as text generation. In *Advances in Neural Information Processing Systems*, volume 34, pages 27263–27277. Curran Associates, Inc.

Chenyang Zhang, Haibo Tong, Bin Zhang, and Dongyu Zhang. 2024. Probing causality manipulation of large language models. *CoRR*, abs/2408.14380.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Wenting Zhao, Justin Chiu, Claire Cardie, and Alexander Rush. 2023. 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.

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

A Logical Reasoning Examples

We hereby provide a formal description of atomic reasoning types, including examples on claim verification whenever possible.

Deductive. 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 hypotheses. A core element of deductive inference is that if the premises are true, then the conclusion is true. In formal logic, the rules of deduction are infinite (Morishita et al., 2023), where the most common ones are modus ponens, syllogism, and elimination. The reader is referred to the works of Morishita et al. (2023) and Saparov et al. (2023) for a more in-depth discussion of deduction rules.

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 references the school closing down and residents being told to shelter at home. Therefore, we deductively infer that the rumour is true as the conclusion logically follows the evidence.*

Abductive. There is much debate regarding definition of abductive reasoning (Plutynski, 2011). We follow the work of Paul (1993), which provides three different approaches towards defining abductive reasoning as:

- A set-cover-based approach;
- A logic-based approach;
- A knowledge-level approach.

In this work, we use the set-cover-based approach, in which we construct the 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).

Additionally, the key difference between abductive reasoning and the other forms of reasoning types is that, unlike the other types, abdctive reasoning works "backwards" towards the most plausible hypothesis from a given set of rules and happenings. 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 from 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.

Inductive. 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

$\exists x, observations(x) \implies conclusion$

amongst many other forms. A conclusion reached by inductive reasoning is not necessarily true. As per Flach and Kakas (2000), if the premises for any stated argument only provide partial support for its conclusion, then that argument is inductive supposing the premises are true.

Example 1.

Claim: *Injecting or consuming bleach is good for killing the virus (Covid-19).*

Evidence 1: *Applying bleach 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 the claim is false, as bleach causes harm to the body and would not kill any viruses within.*

Example 2.

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

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

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.

It is clear that each conclusion is true if we make a closed-world assumption regarding the premises. However, in reality, it is false as there exist flightless birds including Penguins and wingless birds such as Kiwi.

Analogical. Analogical reasoning is the reasoning process concerned with comparison between two or more objects, arguments, or entities.

Example.

Claim: entity α is equivalent to entities ζ, κ, ϕ, and ω.

Evidence: entity β is equivalent to entities ζ, κ, and ϕ.

Conclusion: entity β is probably equivalent to entity ω.

B Claim Verification Datasets

We select three popular resources for claim verification, covering different domains and increasing task complexity.

VitaminC (Schuster et al., 2021). A multi-task fact-checking dataset based on manual and synthetic English revisions to Wikipedia pages. The dataset comprises ∼450k claim-evidence pairs. For the claim verification task, claim-evidence pairs are annotated with veracity labels: *supports*, *refutes*, and *not-enough-information*. VitaminC is licensed under MIT License.

CLIMATE-FEVER (Diggelmann et al., 2020) . A claim verification dataset that consists of ∼1.5k real-world claims concerning climate change. The claims are retrieved from Google while the evidence is Wikipedia-based. The claim-evidence pairs are annotated with veracity labels: *supports*, *refutes*, and *not-enough-information*.

PHEMEPlus (Dougrez-Lewis et al., 2022). A rumour verification dataset comprising social media claims about real-world events. The dataset contains five different events where associated claim-evidence pairs are annotated with veracity labels: *true*, *false*, *not-enough-information*. PHEMEPlus is an extension of the PHEME (Zubiaga et al., 2016) dataset, where web-retrieved news articles are used as evidence in place of Twitter threads as done in PHEME.

C Sampling details

Figure A1 shows the distribution of cosine similarity (denoted as Sim Score), BERTScore and BLEURT score. We used this distribution to set up two different thresholds for deductive and abductive samples. We found that the Bertscores and Sim Scores for abductive sample did not seem to overlap. However, the deductive score had overlap between them. From this observation, we derived the following thresholds. For abductive samples, the threshold is: BERTScore $\leq 0.25 \wedge$ Sim Score ≥ 0.35. For deductive samples, the threshold is: Sim Score $\geq 0.36 \wedge$ BLEURT > 0.15.

D Annotation Guidelines

Figure A2 summarizes our annotation pipeline for RECV. In data annotation (Figure A2, **Top**), we provide a human annotator with claim-evidence pairs with corresponding veracity label. The annotator determines the reasoning type required to infer the claim veracity and provides a rationale in free-text format as motivation. Table A1 reports the annotation guidelines we used to instruct annotators in creating RECV.

E Data Annotation

Table A2 reports pairwise agreement scores for each dataset in RECV.

F Prompts

Tables A3, A4, and A5 report the prompts we used for VitaminC, CLIMATE-FEVER, and PHEMEPlus, respectively. We follow standard prompt construction strategies and provide dataset specific personas and instructions. Additionally, in M-CoT with provide examples to guide the model.

G Qualitative Analysis

Tables A9, A10, and A11 report pairwise permutation tests on RECV datasets. Moreover, Tables A6, A7, and A8 report qualitative analysis metrics on RECV datasets. In particular, we compute qualitative metrics on two sets of examples: those for which models correctly predicted the corresponding claim veracity (**Correct**), and those where models made wrong predictions (**Wrong**).

The metrics used for qualitative analysis are as following,

Factual consistency. We assess the consistency of LLM generated rationales R with human-written ones H, where consistency is the absence of contradiction. We define C to be a function that quantifies the consistency of text B based on text A:

$$C(A, B) = \frac{1}{|A| \cdot |B|} \sum_{a \in A} \sum_{b \in B} (1 - \text{NLI}(\text{Contradict}|a, b))$$

We calculate the consistency of LLM rationales to human rationales as,

$$\text{FC} = 1 - \frac{1}{N} \sum_{i=1}^{N} C_i \qquad (1)$$

where N is the total number of sentence pairs compared,C_i is the consistency score of the i-th comparison.

Evidence appropriateness. For evidence appropriateness, we use the same consistency score C as Fact_Expert.

$$\text{EA} = \frac{1}{M} \sum_{j=1}^{M} \left(\frac{1}{N_j} \sum_{i=1}^{N_j} (1 - c_{ij}) \right) \qquad (2)$$

Here, M is the total number of generated rationales, N_j is the number of sentences in the j-th generated rationale and C_ij is the consistency score for the i-th sentence in the j-th rationale. Evidence appropriateness can be considered as the mean factual consistency whereas Fact_Expert is the granular sentence level consistency.

Coherence. We estimate how easy it is to follow the rationales and how effectively it integrates information from the evidence using BARTScore.

Fluency. We estimate fluency for rationales using perplexity (PPL) under GPT-2-XL (Radford et al., 2019).

Read each claim and evidence pair samples along with their associated veracity labels. Afterwards you will label them with the type of reasoning you think was necessary for inferring the veracity label of the claim given the evidence. The reasoning types are abductive and deductive. Also, provide rationale for your labels.

The goal is to identify what type of reasoning is necessary to infer the veracity of the claim given the associated evidence, for each of the given pairs.

Example 1.

Claim: Climate change isn't increasing extreme weather damage costs.

Evidence: 1. Many analyses, such as that of the Stern Review presented to the British Government, have predicted reductions by several percent of world gross domestic product due to climate related costs such as dealing with increased extreme weather events and stresses to low-lying areas due to sea level rises. 2. Global losses reveal rapidly rising costs due to extreme weather-related events since the 1970s. 3. Global warming boosts the probability of extreme weather events, like heat waves, far more than it boosts more moderate events. 4. "Impacts [of climate change] will very likely increase due to increased frequencies and intensities of some extreme weather events".

Veracity: Refutes

Reasoning: Deductive

Rationale: The evidence deductively refutes the claim. We find explicit mention of increased damage cost in the second line of the evidence. While the last two lines of evidence provide explicit evidence of global causing more adverse weather events.

Example 2.

Claim: Pluto's climate change over the last 14 years is likely a seasonal event.

Evidence: The long orbital period of Neptune results in seasons lasting forty years. 2. As a result, Neptune experiences similar seasonal changes to Earth. 3. "Evidence for methane escape and strong seasonal and dynamical perturbations of Neptune's atmospheric temperatures". 4. Each planet therefore has seasons, changes to the climate over the course of its year.

Veracity: Supports

Reasoning: Abductive

Rationale: The claim is abductively supported. Given Pluto used to be a planet and now is labeled as a dwarf planet, we can hypothesize that it likely has the same attribute as neptune. Given pluto has the biggest orbital period, it is very much likely pluto seasons last over 10 years.

Table A1: Annotation guidelines used to create RECV. This specific example was for CLIMATE-FEVER. This partial representation of the guideline as we provided 8 examples for each dataset with a mix of supports and refutes label.

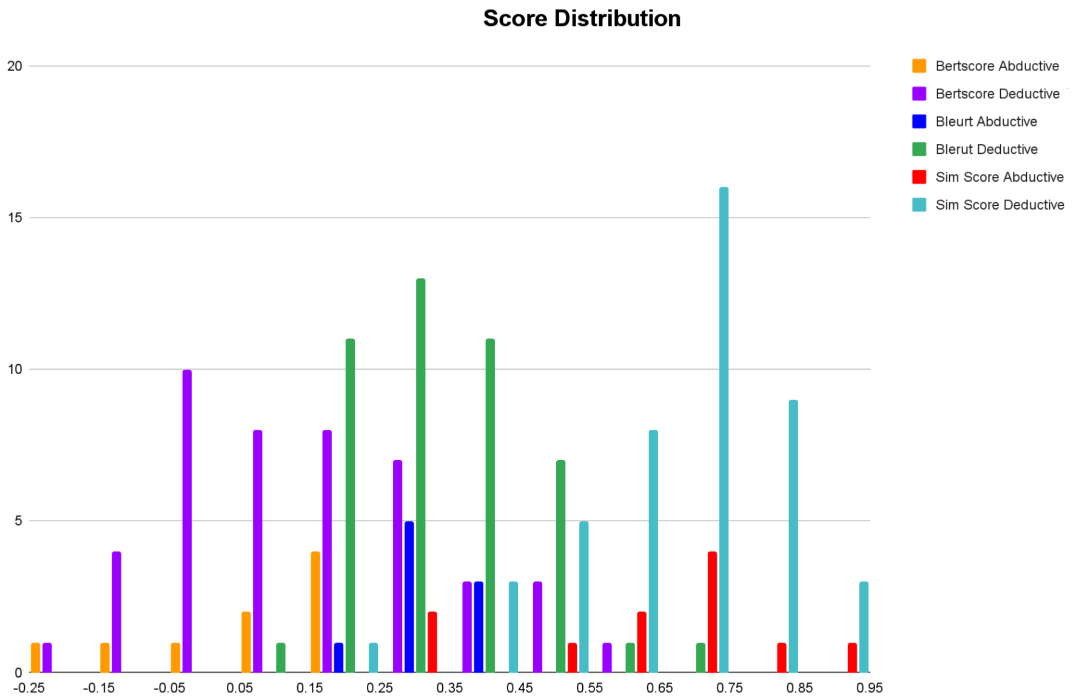

Figure A1: Distribution of the sampling metrics.

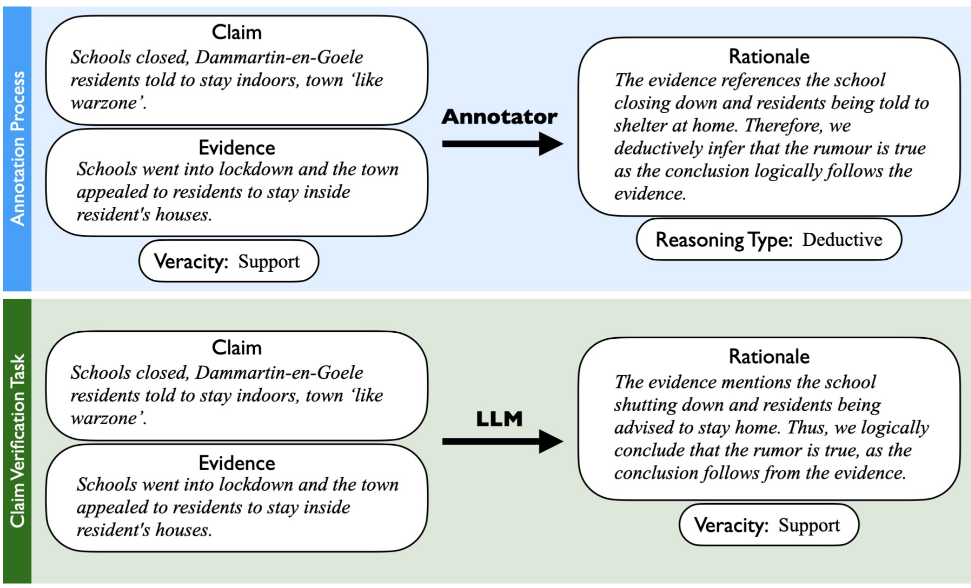

Figure A2: (**Top**) Our annotation process for reasoning-based claim verification. An annotator provides reasoning type required to infer the claim veracity and a rationale as motivation. (**Bottom**) The claim verification task where a LLM has to predict the claim veracity and generate a rationale as support.

	VitaminC			CLIMATE-FEVER			PHEMEPlus		
	Annotator A	Annotator B	Annotator C	Annotator A	Annotator B	Annotator C	Annotator A	Annotator B	Annotator C
Annotator A	-	0.72	0.74	-	0.56	0.56	-	0.64	0.68
Annotator B	0.72	-	0.78	0.56	-	0.56	0.64	-	0.68

Table A2: Pairwise Bennett's S Score across different datasets.

ZS.
You are an expert fact checker. As an expert fact checker, you will be helping us verify some claims. For your task, you will be provided with claims and evidence in this format Q:[<!C> Claim: ...<C!> \n <!E> Evidence: ...<E!>]. You will use the provided evidence to decide whether the associated claim is supported or refuted. You will first briefly explain your reasoning in one sentence, and then make the final judgment by writing LABEL: followed by a single word SUPPORTS or REFUTES.

CoT.
You are an expert fact checker. As an expert fact checker, you will be helping us verify some claims. For your task, you will be provided with claims and evidence in this format Q:[<!C> Claim: ...<C!> \n <!E> Evidence: ...<E!>]. You will use the provided evidence to decide whether the associated claim is supported or refuted. You will first briefly explain your reasoning in one sentence, and then make the final judgment by writing LABEL: followed by a single word SUPPORTS or REFUTES. Let's think step by step.

M-CoT.
You are an expert fact checker. As an expert fact checker, you will be helping us verify some claims. You will be provided with tuples of claim, evidence and answer as examples first. The example claims will be inside <!eC>...<eC!> tokens, evidence will be inside <!eE>...<eE!> tokens and the answer/reasoning will be inside <!eA>...<eA!> tokens. The answer is based on the evidence and it verifies whether the evidence supports or refutes the claim. For your task, you will be provided with claims and evidence in this format Q:[<!C> Claim: ...<C!> \n <!E> Evidence: ...<E!>]. You will use the provided evidence to decide whether the associated claim is supported or refuted. You will first briefly explain your reasoning in one sentence, and then make the final judgement by writing LABEL: followed by a single word SUPPORTS or REFUTES.

Here are some examples:
`{examples}`

Table A3: Prompts used in VitaminC.

ZS.
You are an expert climate scientist. As an expert climate scientist, you will be helping us verify some climate-related claims. For your task, you will be provided with climate-related claims and evidence in this format Q:[<!C> Claim: ...<C!> \n <!E> Evidence: ...<E!>]. You will use the provided evidence to decide whether the associated claim is supported or refuted. You will first briefly explain your reasoning in one sentence, and then make the final judgement by writing LABEL: followed by a single word SUPPORTS or REFUTES.

CoT.
You are an expert climate scientist. As an expert climate scientist, you will be helping us verify some climate-related claims. For your task, you will be provided with climate-related claims and evidence in this format Q:[<!C> Claim: ...<C!> \n <!E> Evidence: ...<E!>]. You will use the provided evidence to decide whether the associated claim is supported or refuted. You will first briefly explain your reasoning in one sentence, and then make the final judgement by writing LABEL: followed by a single word SUPPORTS or REFUTES. Let's think step by step.

M-CoT.
You are an expert climate scientist. As an expert climate scientist, you will be helping us verify some climate-related claims. You will be provided with tuples of claim, evidence and answer as examples first. The example claims will be inside <!eC>...<eC!> tokens, evidence will be inside <!eE>...<eE!> tokens and the answer/reasoning will be inside <!eA>...<eA!> tokens. The answer is based on the evidence and it verifies whether the evidence supports or refutes the claim. For your task, you will be provided with climate-related claims and evidence in this format Q:[<!C> Claim: ...<C!> \n <!E> Evidence: ...<E!>]. You will use the provided evidence to decide whether the associated claim is supported or refuted. You will first briefly explain your reasoning in one sentence, and then make the final judgement by writing LABEL: followed by a single word SUPPORTS or REFUTES.

Here are some examples:
`{examples}`

Table A4: Prompts used in CLIMATE-FEVER.

ZS.

You are an expert journalist. As an expert journalist, you will be helping us verify some rumours. For your task, you will be provided with rumours and evidence in this format Q:[<!R> Rumour: ...<R!> \n <!E> Evidence: ...<E!>]. You will use the provided evidence to decide whether the associated rumour is supported or refuted. You will first briefly explain your reasoning in one sentence, and then make the final judgement by writing LABEL: followed by a single word SUPPORTS or REFUTES.

CoT.

You are an expert journalist. As an expert journalist, you will be helping us verify some rumours. For your task, you will be provided with rumours and evidence in this format Q:[<!R> Rumour: ...<R!> \n <!E> Evidence: ...<E!>]. You will use the provided evidence to decide whether the associated rumour is supported or refuted. You will first briefly explain your reasoning in one sentence, and then make the final judgement by writing LABEL: followed by a single word SUPPORTS or REFUTES. Let's think step by step.

M-CoT.

You are an expert journalist. As an expert journalist, you will be helping us verify some rumours. You will be provided with tuples of rumour, evidence and answer as examples first. The example rumours will be inside <!eR>...<eR!> tokens, evidence will be inside <!eE>...<eE!> tokens and the answer/reasoning will be inside <!eA>...<eA!> tokens. The answer is based on the evidence and it verifies whether the evidence supports or refutes the rumour. For your task, you will be provided with rumours and evidence in this format Q:[<!R> Rumour: ...<R!> \n <!E> Evidence: ...<E!>]. You will use the provided evidence to decide whether the associated rumour is supported or refuted. You will first briefly explain your reasoning in one sentence, and then make the final judgement by writing LABEL: followed by a single word SUPPORTS or REFUTES.

Here are some examples:
`{examples}`

Table A5: Prompts used in PHEMEPlus.

	EA ↑		**FC ↑**		**BART ↑**		**PPL ↓**	
Model	**Correct**	**Wrong**	**Correct**	**Wrong**	**Correct**	**Wrong**	**Correct**	**Wrong**
Claude ZS	0.86	0.75	0.86	0.77	-4.15	-4.21	103.49	75.94
Claude ZS CoT	0.86	0.61	0.86	0.75	-4.42	-4.19	55.44	39.25
Claude M-CoT	0.87	0.82	0.86	0.84	-4.08	-3.90	72.02	47.03
GPT-4 ZS	0.89	0.76	0.88	0.82	-3.82	-3.87	69.82	51.17
GPT-4 ZS CoT	0.89	0.76	0.88	0.80	-3.77	-3.8	61.13	47.85
GPT-4 M-CoT	0.90	0.77	0.90	0.75	-2.98	-3.01	46.39	40.95
GPT-4o ZS	0.91	0.79	0.90	0.81	-3.67	-3.41	57.60	24.46
GPT-4o ZS CoT	0.93	0.75	0.92	0.76	-3.47	-3.29	37.77	21.57
GPT-4o M-CoT	0.92	0.77	0.90	0.78	-3.68	-3.35	53.63	26.52

Table A6: Qualitative evaluation on VitaminC. We distinguish between correct and wrong claim veracity predictions.

	EA ↑		FC ↑		BART ↑		PPL ↓	
Model	**Correct**	**Wrong**	**Correct**	**Wrong**	**Correct**	**Wrong**	**Correct**	**Wrong**
Claude ZS	0.87	0.88	0.87	0.89	-4.19	-4.60	61.26	79.73
Claude ZS CoT	0.81	0.86	0.91	0.86	-4.29	-4.26	29.27	25.08
Claude M-CoT	0.90	0.82	0.91	0.85	-3.36	-3.68	33.51	33.33
GPT-4 ZS	0.93	0.79	0.93	0.84	-3.65	-3.79	30.17	36.24
GPT-4 ZS CoT	0.92	0.81	0.92	0.87	-3.63	-3.76	29.23	31.51
GPT-4 M-CoT	0.96	0.70	0.96	0.75	-2.88	-3.07	25.50	25.65
GPT-4o ZS	0.93	0.89	0.93	0.90	-3.63	-3.71	28.09	26.62
GPT-4o ZS CoT	0.95	0.85	0.95	0.90	-3.37	-3.48	20.77	19.95
GPT-4o M-CoT	0.90	0.88	0.91	0.86	-3.56	-3.72	28.15	28.04

Table A7: Qualitative evaluation on CLIMATE-FEVER. We distinguish between correct and wrong claim veracity predictions.

	EA ↑		FC ↑		BART ↑		PPL ↓	
Model	**Correct**	**Wrong**	**Correct**	**Wrong**	**Correct**	**Wrong**	**Correct**	**Wrong**
Claude ZS	0.85	0.76	0.84	0.74	-4.29	-4.38	60.15	53.23
Claude ZS CoT	0.86	0.81	0.85	0.81	-4.43	-4.49	48.25	42.95
Claude M-CoT	0.89	0.87	0.89	0.86	-4.11	-4.30	57.30	58.42
GPT-4 ZS	0.88	0.81	0.89	0.81	-3.90	-3.87	41.33	35.52
GPT-4 ZS CoT	0.88	0.85	0.88	0.85	-3.90	-3.85	39.65	36.18
GPT-4 M-CoT	0.91	0.70	0.89	0.72	-3.40	-3.39	40.59	30.73
GPT-4o ZS	0.90	0.74	0.90	0.75	-4.02	-4.00	43.29	34.20
GPT-4o ZS CoT	0.92	0.82	0.92	0.84	-3.76	-3.70	30.33	24.55
GPT-4o M-CoT	0.89	0.79	0.89	0.79	-4.08	-4.05	49.46	40.59

Table A8: Qualitative evaluation on PHEMEPlus. We distinguish between correct and wrong claim veracity predictions.

	Claude ZS	**Claude ZS CoT**	**Claude M-CoT**	**GPT-4 ZS**	**GPT-4 ZS CoT**	**GPT-4 M-CoT**	**GPT-4o ZS**	**GPT-4o ZS CoT**
Claude ZS CoT	0.4061	-	-	-	-	-	-	-
Claude M-CoT	0.5985	0.1768	-	-	-	-	-	-
GPT-4 ZS	0.3461	0.0639	0.6466	-	-	-	-	-
GPT-4 ZS CoT	0.3303	0.0600	0.6636	0.9703	-	-	-	-
GPT-4 M-CoT	0.0830	0.0066	0.1943	0.3769	0.3489	-	-	-
GPT-4o ZS	0.0732	0.0067	0.1897	0.4086	0.3739	0.8896	-	-
GPT-4o ZS CoT	0.0158	0.0002	0.0558	0.1511	0.1269	0.6856	0.5457	-
GPT-4o M-CoT	0.0866	0.0064	0.2211	0.4259	0.3837	0.8893	0.9988	0.5626

Table A9: Pairwise Permutation Test on 100 evaluation samples from VitaminC.

	Claude ZS	**Claude ZS CoT**	**Claude M-CoT**	**GPT-4 ZS**	**GPT-4 ZS CoT**	**GPT-4 M-CoT**	**GPT-4o ZS**	**GPT-4o ZS CoT**
Claude ZS CoT	0.0480	-	-	-	-	-	-	-
Claude M-CoT	0.4062	0.0027	-	-	-	-	-	-
GPT-4 ZS	0.0811	0.0001	0.5795	-	-	-	-	-
GPT-4 ZS CoT	0.0889	0.0002	0.6516	0.8770	-	-	-	-
GPT-4 M-CoT	0.0022	0.0001	0.0802	0.1529	0.1005	-	-	-
GPT-4o ZS	0.0108	0.0001	0.2442	0.4047	0.3009	0.5190	-	-
GPT-4o ZS CoT	0.0006	0.0001	0.0528	0.0932	0.0536	0.9319	0.4036	-
GPT-4o M-CoT	0.1325	0.0001	0.7312	0.7618	0.8751	0.0741	0.2653	0.0369

Table A10: Pairwise Permutation Test on 100 evaluation samples from CLIMATE-FEVER.

	Claude ZS	Claude ZS CoT	Claude M-CoT	GPT-4 ZS	GPT-4 ZS CoT	GPT-4 M-CoT	GPT-4o ZS	GPT-4o ZS CoT
Claude ZS CoT	0.1872	-	-	-	-	-	-	-
Claude M-CoT	0.0068	0.0953	-	-	-	-	-	-
GPT-4 ZS	0.0276	0.3199	0.4636	-	-	-	-	-
GPT-4 ZS CoT	0.0082	0.1534	0.6888	0.7094	-	-	-	-
GPT-4 M-CoT	0.2363	0.8827	0.2256	0.4959	0.3116	-	-	-
GPT-4o ZS	0.0534	0.4611	0.3528	0.8139	0.5373	0.6487	-	-
GPT-4o ZS CoT	0.0002	0.0033	0.4062	0.0791	0.1584	0.0263	0.0518	-
GPT-4o M-CoT	0.0270	0.3188	0.4208	0.9707	0.6609	0.5043	0.8387	0.0681

Table A11: Pairwise Permutation Test on 100 evaluation samples from PHEMEPlus.