Inference to the Best Explanation in Large Language Models

Anonymous ACL submission

Abstract

How do Large Language Models (LLMs) generate explanations? While LLMs are increasingly adopted in real-world applications, the principles and properties behind their explanatory process are still poorly understood. This paper proposes an interpretability and evaluation framework for LLMs' explanatory reasoning inspired by philosophical accounts on Inference to the Best Explanation (IBE). In particular, the framework aims to estimate the quality of natural language explanations through a combination of criteria computed on linguistic and logical features, including consistency, parsimony, coherence, and uncertainty. We conduct extensive experiments on Causal Question Answering (CQA), instantiating our framework to select among competing explanations generated by LLMs (i.e., ChatGPT and LLama 2). The results reveal that the proposed methodology can successfully identify the best explanation supporting the correct answers with up to 77% accuracy ($\approx$ 27% above random) suggesting that LLMs indeed conform to features of IBE. At the same time, we found notable differences across LLMs, with ChatGPT significantly outperforming LLama 2. Finally, we analyze the degree to which different criteria can predict the correct answer, suggesting potential implications for external verification methods for LLM-generated output.

1 Introduction

Large Language Models (LLMs) such as OpenAIs ChatGPT (Brown et al., 2020) and Llama 2 (Touvron et al., 2023) have been highly effective across a diverse range of language understanding and reasoning tasks (Liang et al., 2023). While LLM performances have been thoroughly investigated across various benchmarks (Wang et al., 2019; Srivastava et al., 2023; Gao et al., 2023; Touvron et al., 2023), the principles and properties behind their step-wise reasoning process are still poorly formalized and understood. LLMs are notoriously considered black-box models that are difficult to interpret (Chakraborty et al., 2017; Danilevsky et al., 2020). Further, the commercialization of LLMs has led to strategic secrecy around model architectures and training details (Xiang, 2023; Knight, 2023). Finally, neural models are susceptible to hallucinations and adversarial perturbations (Geirhos et al., 2020; Camburu et al., 2020), often producing plausible but factually incorrect answers (Ji et al., 2023; Huang et al., 2023). As the size and complexity of LLM architectures increase, the systematic study of the generated explanations become crucial since it can provide efficient and pragmatic mechanisms to better interpret and validate the internal inference processes (Wei et al., 2022b; Lampinen et al., 2022; Huang and Chang, 2022).

Currently, there is no systematic way to automatically evaluate explanations (Valentino et al., 2021). Without resource-intensive annotation methodologies (Wiegreffe and Marasovic, 2021; Thayaparan et al., 2020; Dalvi et al., 2021; Camburu et al., 2018), explanation quality methods tend to rely on weak supervision scenarios, where arriving at the correct answer is taken as evidence of good explanation quality. In this paper, we seek to better understand LLM explanatory process through the investigation of explicit linguistic and logical properties. While explanations are hard to formalize due to their open-ended nature, we hypothesize that they can be analyzed as linguistic objects, with observable and measurable features that can serve to define criteria for assessing their quality. These criteria can potentially lead to model selection and safety mechanisms and serve as a critical qualitative understanding device – i.e., to determine inference phenomena that are not fully addressed by a given model (e.g., logical consistency).

Specifically, this paper investigates the following overarching research question: *"Can linguistic and logical properties associated with LLMs' generated explanations be used to qualify the models'*

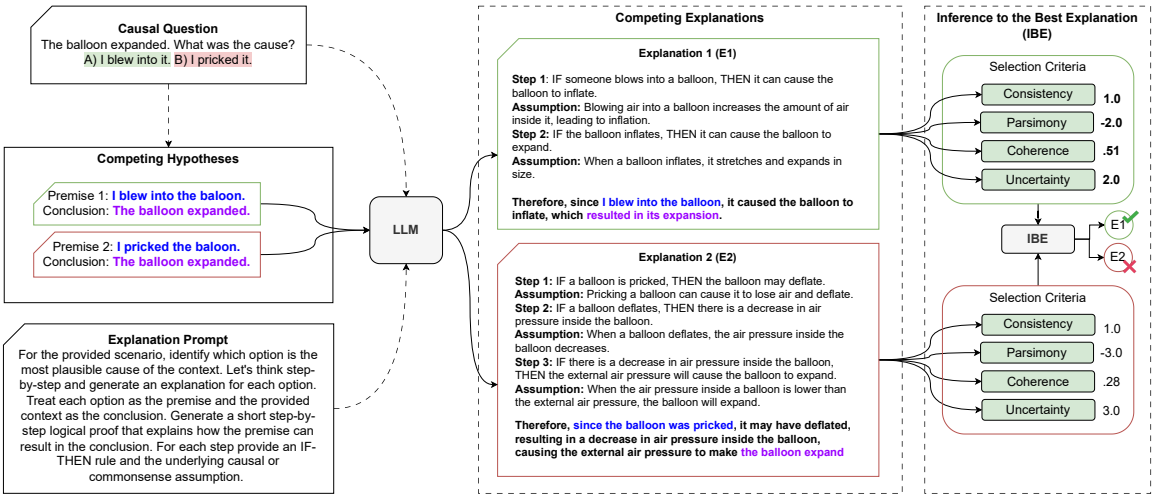

Figure 1: The IBE framework qualifies LLM-generated explanations with a set of logical and linguistic selection criteria to identify the most plausible hypothesis.

reasoning process?". To this end, we propose an interpretable framework inspired by philosophical accounts on *Inference to the Best Explanation (IBE)* – i.e., the process of selecting among competing explanatory theories (Lipton, 2017). In particular, the framework is designed to estimate the quality of natural language explanations according to a combination of metrics computed on a set of interpretable features, namely logical consistency, parsimony, coherence, and linguistic uncertainty.

To evaluate the proposed framework, we conduct extensive experiments on Causal Question Answering (CQA) tasks, implementing each metric with external models to select among competing explanations generated by LLMs (i.e., ChatGPT and LLama 2). The results reveal that the proposed methodology can successfully identify the best explanation supporting the correct answers with up to 77% accuracy ($\approx$ 27% above random) confirming that the proposed criteria possess complementary features. At the same time, however, we found notable differences across metrics and LLMs, with ChatGPT significantly outperforming LLama.

In summary, this paper provides the following contributions:

1. To the best of our knowledge, we are the first to propose an interpretable framework inspired by philosophical accounts on Inference to the Best Explanation (IBE) to automatically assess the quality of natural language explanations.

2. We demonstrate how the framework can be instantiated for Large Language Models (LLMs) with the use of external tools, performing an extensive empirical evaluation of LLM explanations on CQA tasks.

3. We found that the IBE criteria are predictors of the correct answer with degrees of statistical significance, with uncertainty, parsimony and coherence being the best predictors. This is evidence that LLMs indeed conform to features of IBE. At the same time, however, we found that LLMs are strong rationalizers and can produce consistent explanations even for less plausible candidates, making the consistency metric less effective in practice.

The entire experimental code is available online[1] to encourage future research in the field.

2 Inference to the Best Explanation (IBE)

Explanatory reasoning is a distinctive feature of human rationality underpinning problem-solving and knowledge creation in both science and everyday scenarios (Lombrozo, 2012). Accepted epistemological accounts characterize the creation of an explanation as composed of two distinct phases: conjecturing and criticism (Popper, 2014; Deutsch, 2011). According to this view, the explanatory process always involves a conflict between plausible explanations, which is typically resolved through the criticism phase via a selection process, where competing explanations are assessed according to a set of criteria. The criticism phase is often referred

[1] anonymous_link

to as *Inference to the Best Explanation (IBE)* (Lipton, 2017; Harman, 1965). While the conjecturing phase is hard to formalize due to its open-ended nature, the criteria according to which an explanation has to be preferred among competing ones are more easily definable. Therefore, philosophers have attempted to identify what characterizes a good explanation, defining criteria such as parsimony, coherence, unification power, and hardness to variation (Mackonis, 2013; Thagard, 1978, 1989; Kitcher, 1989; Valentino and Freitas, 2022).

As LLMs become interfaces for natural language explanations, epistemological frameworks offer an opportunity for developing criticism mechanisms to better understand the explanatory process underlying state-of-the-art models. To this end, this paper considers an LLM as a conjecture device producing linguistic objects that can be subject to criticism. In particular, we focus on a subset of criteria that can be automatically computed on explicit linguistic and logical features, namely: consistency, parsimony, coherence, and uncertainty.

To assess LLMs' alignment to such criteria, we focus on the task of selecting among competing explanations in a Multiple Choice Question Answering (MCQA) setting (Figure 1). Specifically, given a set of competing hypotheses $H = \{h_1, h_2, \ldots, h_n\}$ (each corresponding to a possible cause-effect relation between a premise and a conclusion), we prompt an LLM to generate plausible explanations supporting each hypothesis (Section 3). Subsequently, we adopt the selection criteria to assess the quality of the generated explanations (Section 4). The explanation with the highest quality score is then selected to predict the final answer and assessed as the extent to which observable explanatory features are correlated with QA accuracy. Specifically, we hypothesize that the quality of LLMs' explanations for the correct answers should be higher than the ones generated for alternative hypotheses and that higher-quality explanations can be explicitly identified via the IBE selection criteria.

3 Explanation Prompting

The first stage in our methodology consists of prompting LLMs to generate competing explanations for different hypotheses. To this end, we employ a variation of Chain-of-Thoughts (CoT) prompting (Wei et al., 2022a). Specifically, the original CoT method is modified to instruct the LLM to produce an explanation for each hypothesis (see Figure 1). To this end, we adopt a methodology similar to Valentino et al. (2021) to constrain the generated explanations into an entailment form for the downstream IBE evaluation. In line with Valentino et al. (2021), we posit that a valid explanation should demonstrate an entailment relationship between the premise and conclusion.

To elicit logical connections between statements and facilitate subsequent analysis, each generated explanation is constrained to use weak syllogisms expressed as IF-THEN statements for each explanation step. Additionally, the LLM is instructed to produce the associated causal or commonsense assumption underlying each explanation step. This output is then post-processed to extract the relevant supporting knowledge for evaluation via the selection criteria. Additional details and examples of prompts are reported in Appendix A.1.

4 Selection Criteria

To perform IBE, we investigate a set of criteria that can be automatically computed on explicit linguistic and logical features, namely: consistency, parsimony, coherence, and uncertainty.

4.1 Consistency

The first criterion adopted for assessing explanation quality is logical consistency. Given a hypothesis, composed of a premise p_i, a conclusion c_i, and an explanation consisting of a set of statements $E = s_1, ..., s_i$, we define E to be logically consistent if $p_i \cup E \models c_i$. An explanation, therefore, is logically consistent if it is possible to build a complete deductive proof linking premise and conclusion. To implement the logical consistency metric, we leverage external symbolic solvers along with autoformalization – i.e., the translation of natural language into a formal language (Wu et al., 2022). Specifically, hypotheses and explanations are formalized into a Prolog program which will attempt to generate a deductive proof via backward chaining (Weber et al., 2019).

To perform autoformalization, we leverage the translation capabilities of ChatGPT-3.5-Turbo. Specifically, we instruct the model to convert each IF-Then implication from the generated explanation steps into an implication rule, while the premise statement is converted into grounding atoms. On the other end, the entailment condition and the conclusion are used to create a Prolog

query. Further details about the autoformalization process can be found in Appendix A.2.

After autoformalization, we adopt an external Prolog solver for entailment verification. The explanation is considered logically consistent if the Prolog solver can satisfy the query and successfully build a deductive proof. Additional technical on proof generation can be found in Appendix A.3.

4.2 Parsimony

The parsimony principle, often referred to as Ockham's razor, is considered an important criterion for choosing among competing explanations. In particular, parsimony favors the selection of the simplest explanation consisting of the fewest elements and assumptions (Sober, 1981). This is because an explanation with fewer assumptions tends to leave fewer statements unexplained, improving specificity and alleviating the infinite regress (Thagard, 1978). Further, parsimony is an essential feature of causal interpretability, as only parsimonious solutions are guaranteed to reflect causation in comparative analysis (Baumgartner, 2015). In this paper, we adopt two metrics as a proxy of parsimony, namely *proof depth*, and *concept drift*.

Proof depth, denoted as $Depth$, is defined as the cardinality of the set of rules, R, required by the Prolog solver to connect the conclusion to the premise via backward chaining. Let h be a hypothesis candidate composed of a premise p and a conclusion c, and let E be a formalized explanation represented as a set of rules R'. The proof depth is the number of rules $|R|$, with $R \subseteq R'$, traversed during backward chaining to connect the conclusion c to the premise p:

$$Depth(h) = |R|$$

Concept drift, denoted as $Drift$, is defined as the number of additional concepts and entities, outside the ones appearing in the hypothesis (i.e., premise and conclusion), that are introduced by the LLM to support the entailment. For simplicity, we consider nouns as concepts. Let $N = \{Noun_p, Noun_c, Noun_E\}$ be the unique nouns found in the premise, conclusion, and explanation steps. Concept drift is the cardinality of the set difference between the nouns found in the explanation and the nouns in the hypothesis:

$$Drift(h) = |Noun_E - (Noun_p \cup Noun_c)|$$

Intuitively, the parsimony principle would predict the most plausible hypothesis as the one supported by the explanation with the smallest observed proof depth and concept drift. Implementation details can be found in Appendix A.4.

4.3 Coherence

An explanation can be formally consistent on the surface while still including implausible or ungrounded intermediate assumptions or implication steps. As these assumptions cannot be reliably identified via external logical solvers, we introduce an additional metric named *coherence*. Coherence aims to evaluate the quality of each If-then implication in the generated explanation measuring the entailment strength between the clauses. To this end, we employ a fine-tuned Natural Language Inference (NLI) model. Formally, let S be a set of explanation steps, where each step s consists of an If-then statement, $s = (If_s, Then_s)$. For a given step s_i, let $ES(s_i)$ denote the entailment score obtained via an NLI model between If_s and $Then_s$ clauses. The step-wise entailment score $SWE(S)$ is then calculated as the averaged sum of the entailment scores across all explanation steps $|S|$:

$$\text{SWE}(S) = \frac{1}{|S|} \sum_{i=1}^{|S|} \text{ES}(s_i)$$

We hypothesize that the LLMs should generate a higher step-wise entailment score for the most plausible candidate hypotheses, as such explanations should include stronger entailment relations between the If-then clauses. Additional details can be found in Appendix A.5.

4.4 Uncertainty

Finally, we consider the observed linguistic certainty expressed in the generated explanation as a proxy for plausibility. Hedging words such as *probably*, *might be*, *could be*, etc typically signal ambiguity and are often used when the truth condition of a statement is unknown or improbable. For instance, Pei and Jurgens (2021) found that the strength of scientific claims in research papers is strongly correlated with the use of direct language. In contrast, the use of hedging language suggested that the veracity of the claim was weaker or highly contextualized.

To measure the linguistic certainty (LC) of an explanation, we consider the explanation's underlying assumptions (A_i) and the overall explanation

summary (S), calculating the linguistic certainty score using the fine-tuned sentence-level RoBERTa model from Pei and Jurgens (2021). The overall linguistic certainty score (LC_{overall}) is the sum of the assumption and explanation summary scores:

$$LC_{\text{overall}} = LC(A) + LC(S)$$

Where $LC(A)$ is the sum of the linguistic certainty scores ($LC(A)$) across all the assumptions $|A|$ associated with each explanation step i:

$$LC(A) = \sum_{i=1}^{|A|} LC(a_i)$$

and linguistic certainty of the explanation summary $LC(S)$. We hypothesize that the LLM will use more hedging language when explaining the weaker hypothesis reflecting in a lower linguistic certainty score. Further details can be found in Appendix A.6.

4.5 Inference to Best Explanation

To perform IBE, we define a vanilla linear regression model $\theta(\cdot)$ fitted on the features extracted from the selection criteria to predict the probability that an explanation E_i corresponds to the correct answer. Specifically, we employ the linear model to score the explanations generated for each hypothesis independently and then select the final answer a via argmax:

$$a = \underset{i}{\text{argmax}}[\theta(E_1), \ldots, \theta(E_n)]$$

Additional details can be found in Appendix A.7.

5 Experimental Setting

Causal Question-Answering (CQA) requires reasoning about the causes and effects of a hypothetical or observed event. For our experiments, we specifically consider the task of cause and effect prediction in a Multiple-Choice Question Answering (MCQA) setting, where given a question and two candidate answers, the LLM must decide which is the most plausible cause or effect. Causal reasoning is a challenging task as the model must both possess commonsense knowledge about the plausibility of a causal relationship and consider the chain of events and context which would make one option more plausible than the other. For our experiments, we use the Choice of Plausible Alternatives (COPA) (Gordon et al., 2012) and E-CARE (Du et al., 2022) datasets.

COPA. COPA is a multiple choice commonsense causal QA dataset consisting of 500 train and test examples that were manually generated. Each multiple-choice example consists of a question premise and a set of answer candidates which are potential causes or effects of the premise. COPA is a well-established causal reasoning benchmark that is both a part of SuperGlue (Wang et al., 2019) and the CALM-Bench (Dalal et al., 2023).

E-CARE. E-CARE is a large-scale multiple-choice causal QA dataset consisting of crowd-sourced 15K train and 2k test examples. Each example is further annotated with a conceptual explanation describing the underlying concepts required for reasoning. E-CARE follows the same task format as COPA and can be considered an extension of COPA. We randomly sample 500 examples from the E-CARE test set for our experiments and do not use the associated explanations.

LLMs. We consider ChatGPT-Turbo-3.5, LLaMA 2 13B and LLaMA 2 7B for all experiments. ChatGPT is a proprietary model based on GPT-3 (Brown et al., 2020) highly effective across a wide range of natural language reasoning tasks (Laskar et al., 2023). We additionally evaluate the open-source LLaMA 2 model (Touvron et al., 2023). Here, we consider both the 13B and 7B variants of LLaMa 2 as both are seen as viable commodity ChatGPT alternatives and have been widely adopted by the research community for LLMs benchmarking and evaluation.

6 Results

To assess LLMs' alignment with the proposed IBE framework, we run a regression analysis and conduct a set of ablation studies evaluating the relationship between the selection criteria and question-answering accuracy. The main results are presented in Figure 2 and Table 1.

Overall, our empirical evaluation reveals that **while feature importance varies across LLMs, the analyzed explanatory features tend to conform to IBE expectations**. This is mostly apparent in ChatGPT, where all criteria are found to be statistically significant. At the same time, we found that **LLMs can generate plausible explanations for the less plausible answers, making some of the criteria less effective**, especially in LLaMa models. A comparison across metrics reveals that **linguistic uncertainty is the best indicator across LLMs**

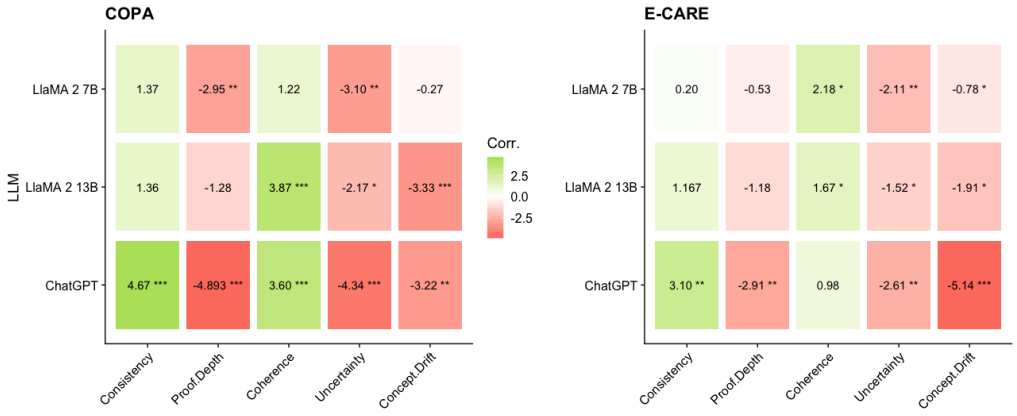

Figure 2: The results show that proof depth, concept drift, and linguistic uncertainty have varying levels of statistical significance in predicting question-answering accuracy. Across all LLMs and datasets, linguistic certainty is the strongest predictor. '***' p>0.001, '**' p>0.01 '*' p>0.05.

	COPA			E-CARE		
	ChatGPT	**LlaMA 2 13B**	**LlaAMA 2 7B**	**ChatGPT**	**LlaMA 2 13B**	**LlaAMA 2 7B**
Consistency	.51	.52	.55	.54	.54	.54
Depth (Parsimony)	.67	.53	.63	.66	.56	.54
Drift (Parsimony)	.67	.63	.58	.66	.57	.57
Coherence	.66	.66	.56	.56	.57	.59
Linguistic Uncertainty	.70	.65	.61	.59	.56	.60
Random	*.50*	*.50*	*.50*	*.50*	*.50*	*.50*
+ Consistency	.51	.52	.55	.54	.54	.54
+ Depth	.67	.53	.63	.66	.56	.56
+ Drift	.70	.65	.65	.72	.66	.65
+ Coherence	.73	.71	.69	.73	.68	.69
+ Uncertainty	**.77**	**.74**	**.70**	**.74**	**.70**	**.73**

Table 1: Ablation study of IBE features on question-answering accuracy. While IBE feature importance differs across LLMs, generated explanations tend to conform to IBE expectations. In the best case, IBE can achieve up to 77% and 74% accuracy considering all the criteria on COPA and E-CARE.

and the feature that is mostly correlated with answer accuracy. Next, we explore each explanation feature in further detail to better understand the variances across criteria and LLMs.

6.1 Consistency

We found that all the LLMs are surprisingly strong conjecture models, being able to generate logically consistent explanations across all hypotheses (Figure 3). This is confirmed by the **high consistency scores for both correct and wrong hypotheses, with the consistency criteria being able to improve accuracy only by 6pp over a random baseline**. Moreover, we observe that consistency tends to be statistically insignificant for the Llama models. Therefore, we conclude that **evidence of logical consistency provides a limited signal for plausibility and is better understood in the context of other IBE features**. For the incorrect candidate explanations, we find that LLMs over-rationalize and introduce additional premises to demonstrate entailment in their explanations. We further explore this phenomenon in Section 6.2.

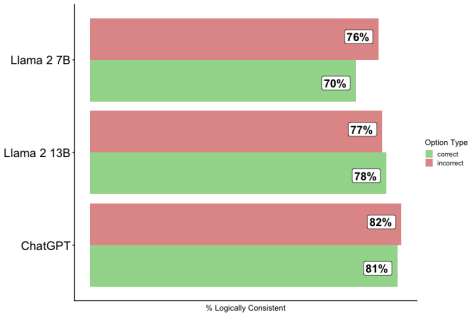

Figure 3: Evaluation of explanation consistency. LLMs are strong rationalizers and can generate logically consistent explanations at equal rates for both correct and incorrect answers.

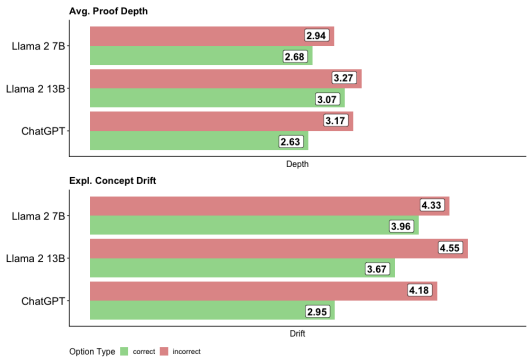

Figure 4: Explanation parsimony is evaluated using proof depth and concept drift. Both metrics are consistently lower for explanations supporting the correct answers, implying higher parsimony.

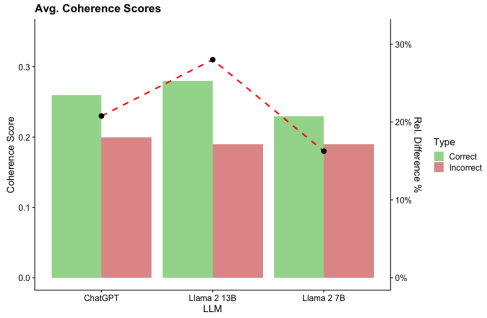

Figure 5: The average coherence score is consistently higher for explanations corresponding to the correct hypotheses.

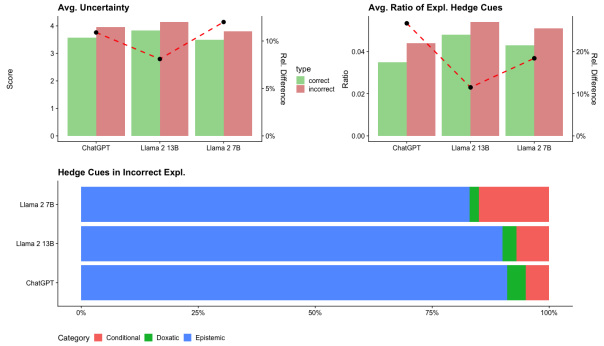

Figure 6: LLMs tend to use more hedging language in explanations supporting less plausible hypotheses. This language is mostly classified as *epistemic*.

6.2 Parsimony

The results suggest that parsimony has a more consistent effect and represents a better predictor of answer accuracy. **We observe a negative correlation between proof depth, concept drift, and question-answering accuracy, suggesting that LLMs tend to introduce more concepts and explanation steps when explaining less plausible hypotheses**. On average, we found the proof depth and concept drift to be 6% and 10% greater for the incorrect option across all LMMs (see Figure 4). Moreover, the results suggest that **as the LLM size grows, the ability to over-rationalize tends to grow linearly**. This is attested by the fact that the average difference in proof depth and concept drift is the greatest in ChatGPT, suggesting that the model tends to find the most efficient explanations for stronger hypotheses and introduce articulated explanation steps for supporting the weaker candidates. Finally, we found that **Llama models tend to generate more complex explanations**, with Llama 2 13B exhibiting the largest concept drift for less plausible hypotheses. **The parsimony criterion supports the IBE predictive power with an average of 14% improvement over consistency**.

6.3 Coherence

Similarly to parsimony, **we found coherence to be a better indicator of explanation quality when compared to consistency, being statistically significant for both ChatGPT and Llama 2 13B on COPA and both Llama 2 models on E-Care**. In the ablation studies, coherence tends to improve accuracy by 10pp for COPA and 2.5pp for E-Care. We found that the average coherence score is consistently greater for the stronger hypothesis across all LLMs and datasets (see Figure 5). **Both ChatGPT and Llama 2 13B exhibit a higher relative difference between the correct and incorrect hypotheses in contrast to Llama 2 7B**.

6.4 Uncertainty

Finally, the results reveal that **linguistic uncertainty is the strongest predictor of answer accuracy and is statistically significant across all LLMs**. This suggests that LLMs use more qualifying and hedging words when explaining weaker hypotheses (see Figure 6). We found that **uncertainty can improve accuracy by 13pp on COPA and 4pp on E-CARE**. As a further analysis, we examine the uncertainty cues expressed by LLMs by analyzing both the frequency of hedge words and categorizing the observed uncertainty cues. To this end, we use a fine-tuned BERT-based token classification model to classify all the words in the generated explanation with uncertainty categories introduced in the 2010 CoNLL shared task on Hedge Detection (Farkas et al., 2010). Farkas et al. (2010) classify hedge cues into three cate-

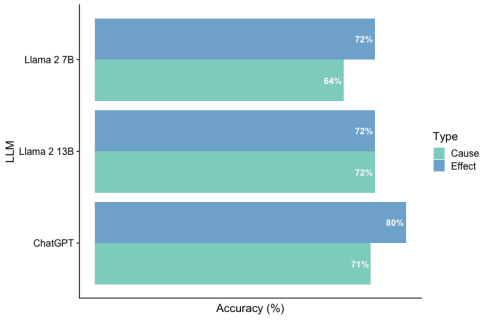

Figure 7: Accuracy in predicting the most plausible causes vs effects on COPA.

gories: epistemic, doxatic, and conditional. Epistemic cues refer to hedging scenarios where the truth value of a proposition can be determined but is unknown in the present (e.g. the blocks *may* fall). Doxatic cues refer to beliefs and hypotheses that can be held to be true or false by others (e.g. the child *believed* the blocks *would* fall). Finally, conditional cues refer to propositions whose truth value is dependent on another proposition's truth value (e.g. *if* the balloon is pricked it may deflate). The results show that **the distribution of hedge words across LLMs tends to be similar, with only minor differences between LLMs** (see Figure 6). Epistemic cues were most frequently used by all three models, with Llama 2 7B being more likely to use conditional cues.

6.5 Causal Directionality

When considering the causal directionality (i.e. cause vs effect), we observed that accuracy tended to differ between LLMs on COPA. In particular, **we found both ChatGPT and Llama 2 7B to be more accurate in predicting the effects in causal scenarios** (see Figure 7). We hypothesize that **LLMs may suffer the challenge of causal sufficiency as the space of potential causal explanations can be far greater than the range of effects once an event has been observed**. This hypothesis is partly supported by the fact that ChatGPT and Llama 2 7B express greater linguistic uncertainty and produce more complex explanations when predicting causes rather than effects.

7 Related Works

Explorations of LLM reasoning capabilities across various domains (e.g. arithmetic, commonsense, planning, symbolic, etc) are an emerging area of interest (Xu et al., 2023; Huang and Chang, 2023). Prompt-based methods (Wei et al., 2022b; Zhou et al., 2023; Wang et al., 2023), such as CoT, investigate strategies to elicit specific types of reasoning behavior through direct LLM interaction. Olausson et al. (2023) investigate automatic proof generation and propose a neurosymbolic framework with an LLM semantic parser and external solver. Creswell et al. (2022) propose an inference framework where the LLM acts as both a selection and inference module to produce explanations consisting of causal reasoning steps in entailment tasks. This paper primarily draws inspiration from recent work on the evaluation of natural language explanations (Valentino et al., 2021; Wiegreffe and Marasovic, 2021; Thayaparan et al., 2020; Dalvi et al., 2021; Camburu et al., 2018). However, differently from previous methods that require extensive human annotations, we are the first to propose a set of criteria that can be automatically computed on explicit linguistic and logical features.

8 Conclusion

This paper proposed an interpretable framework for LLM explanation evaluation inspired by philosophical accounts of IBE. Utilizing a range of selection criteria, including logical consistency, parsimony, coherence, and linguistic uncertainty, the framework can identify the best explanation with up to 77% accuracy in a CQA scenario. Across all LLMs, the IBE features were found to be statistically significant in general with varying importance. Linguistic uncertainty, in particular, represents the best predictor across different LLMs. Our results suggest that LLMs tend to be strong conjecture models being able to generate logically consistent explanations for less plausible hypotheses. Regarding different models, ChatGPT was found to produce the most parsimonious explanations for correct hypotheses but also tended to over-rationalize for less plausible examples. In contrast, the Llama 2 models tend to produce more complex explanations, with Llama 2 13B exhibiting the greatest average proof depth and concept drift. In conclusion, we found that the proposed selection criteria represent strong predictors of question-answering accuracy when applied in combination, suggesting that LLMs do often conform to IBE expectations. We believe our findings can open new lines of research on external evaluation methods for LLMs-generated output, as well as the development of new AI safety mechanisms.

9 Limitations

IBE offers an interpretable explanation evaluation framework utilizing logical and linguistic features. Our current instantiation of the framework is primarily limited in that it does not consider grounded truth for factuality. We observe that the model can generate factually incorrect but logically consistent explanations. In some cases, the coherence metric can identify those factual errors when the step-wise entailment score is comparatively lower. However, our reliance on aggregated metrics can hide weaker entailment especially when the explanation is longer or the entailment strength of the surrounding steps is stronger. Future work can introduce metrics to evaluate grounded knowledge or perform more granular evaluations of explanations to better weight factual inaccuracies.

Finally, the list of criteria considered in this work is not exhaustive and can be extended in future work. However, additional criteria for IBE might not be straightforward to implement (e.g., unification power, hardness to variation) and would probably require further progress in both epistemological accounts and existing NLP technology.

References

Michael Baumgartner. 2015. Parsimony and causality. *Quality & Quantity*, 49:839–856.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

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

Lars Buitinck, Gilles Louppe, Mathieu Blondel, Fabian Pedregosa, Andreas Mueller, Olivier Grisel, Vlad Niculae, Peter Prettenhofer, Alexandre Gramfort, Jaques Grobler, Robert Layton, Jake VanderPlas, Arnaud Joly, Brian Holt, and Gaël Varoquaux. 2013. API design for machine learning software: experiences from the scikit-learn project. In *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, pages 108–122.

Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. 2018. e-snli: Natural language inference with natural language explanations. *Advances in Neural Information Processing Systems*, 31.

OM Camburu, B Shillingford, P Minervini, T Lukasiewicz, and P Blunsom. 2020. Make up your mind! adversarial generation of inconsistent natural language explanations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020. ACL Anthology.

Supriyo Chakraborty, Richard Tomsett, Ramya Raghavendra, Daniel Harborne, Moustafa Alzantot, Federico Cerutti, Mani Srivastava, Alun Preece, Simon Julier, Raghuveer M. Rao, Troy D. Kelley, Dave Braines, Murat Sensoy, Christopher J. Willis, and Prudhvi Gurram. 2017. Interpretability of deep learning models: A survey of results. In *2017 IEEE SmartWorld, Ubiquitous Intelligence Computing, Advanced Trusted Computed, Scalable Computing Communications, Cloud Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI)*, pages 1–6.

Antonia Creswell, Murray Shanahan, and Irina Higgins. 2022. Selection-inference: Exploiting large language models for interpretable logical reasoning.

Dhairya Dalal, Paul Buitelaar, and Mihael Arcan. 2023. CALM-bench: A multi-task benchmark for evaluating causality-aware language models. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 296–311, Dubrovnik, Croatia. Association for Computational Linguistics.

Bhavana Dalvi, Peter Jansen, Oyvind Tafjord, Zhengnan Xie, Hannah Smith, Leighanna Pipatanangkura, and Peter Clark. 2021. Explaining answers with entailment trees. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7358–7370.

Marina Danilevsky, Kun Qian, Ranit Aharonov, Yannis Katsis, Ban Kawas, and Prithviraj Sen. 2020. A survey of the state of explainable AI for natural language processing. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 447–459, Suzhou, China. Association for Computational Linguistics.

David Deutsch. 2011. *The beginning of infinity: Explanations that transform the world.* penguin uK.

Li Du, Xiao Ding, Kai Xiong, Ting Liu, and Bing Qin. 2022. e-CARE: a new dataset for exploring explainable causal reasoning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 432–446, Dublin, Ireland. Association for Computational Linguistics.

Richárd Farkas, Veronika Vincze, György Móra, János Csirik, and György Szarvas. 2010. The CoNLL-2010 shared task: Learning to detect hedges and their scope in natural language text. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning – Shared Task*, pages 1–12, Uppsala, Sweden. Association for Computational Linguistics.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. 2023. A framework for few-shot language model evaluation.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. 2020. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673.

Andrew Gordon, Zornitsa Kozareva, and Melissa Roemmele. 2012. SemEval-2012 task 7: Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 394–398, Montréal, Canada. Association for Computational Linguistics.

Gilbert H Harman. 1965. The inference to the best explanation. *The philosophical review*, 74(1):88–95.

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

Jie Huang and Kevin Chen-Chuan Chang. 2022. Towards reasoning in large language models: A survey. *arXiv preprint arXiv:2212.10403*.

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

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2023. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Comput. Surv.*, 55(12).

Philip Kitcher. 1989. Explanatory unification and the causal structure of the world.

Will Knight. 2023. Ai is becoming more powerful-but also more secretive.

Andrew Lampinen, Ishita Dasgupta, Stephanie Chan, Kory Mathewson, Mh Tessler, Antonia Creswell, James McClelland, Jane Wang, and Felix Hill. 2022. Can language models learn from explanations in context? In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 537–563.

Md Tahmid Rahman Laskar, M Saiful Bari, Mizanur Rahman, Md Amran Hossen Bhuiyan, Shafiq Joty, and Jimmy Xiangji Huang. 2023. A systematic study and comprehensive evaluation of chatgpt on benchmark datasets.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D. Manning, Christopher Ré, Diana Acosta-Navas, Drew A. Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue Wang, Keshav Santhanam, Laurel Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. 2023. Holistic evaluation of language models.

Peter Lipton. 2017. Inference to the best explanation. *A Companion to the Philosophy of Science*, pages 184–193.

Tania Lombrozo. 2012. Explanation and abductive inference. *Oxford handbook of thinking and reasoning*, pages 260–276.

Adolfas Mackonis. 2013. Inference to the best explanation, coherence and other explanatory virtues. *Synthese*, 190(6):975–995.

Yixin Nie, Haonan Chen, and Mohit Bansal. 2019. Combining fact extraction and verification with neural semantic matching networks. In *Association for the Advancement of Artificial Intelligence (AAAI)*.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. Adversarial NLI: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Theo X. Olausson, Alex Gu, Benjamin Lipkin, Cedegao E. Zhang, Armando Solar-Lezama, Joshua B. Tenenbaum, and Roger Levy. 2023. Linc: A neurosymbolic approach for logical reasoning by combining language models with first-order logic provers.

Jiaxin Pei and David Jurgens. 2021. Measuring sentence-level and aspect-level (un)certainty in science communications. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9959–10011, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Karl Popper. 2014. *Conjectures and refutations: The growth of scientific knowledge*. routledge.

R Core Team. 2013. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Elliott Sober. 1981. The principle of parsimony. *The British Journal for the Philosophy of Science*, 32(2):145–156.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power, Alex Ray, Alex Warstadt, Alexander W. Kocurek, Ali Safaya, Ali Tazarv, Alice Xiang, Alicia Parrish, Allen Nie, Aman Hussain, Amanda Askell, Amanda Dsouza, Ambrose Slone, Ameet Rahane, Anantharaman S. Iyer, Anders Andreassen, Andrea Madotto, Andrea Santilli, Andreas Stuhlmüller, Andrew Dai, Andrew La, Andrew Lampinen, Andy Zou, Angela Jiang, Angelica Chen, Anh Vuong, Animesh Gupta, Anna Gottardi, Antonio Norelli, Anu Venkatesh, Arash Gholamidavoodi, Arfa Tabassum, Arul Menezes, Arun Kirubarajan, Asher Mullokandov, Ashish Sabharwal, Austin Herrick, Avia Efrat, Aykut Erdem, Ayla Karakaş, B. Ryan Roberts, Bao Sheng Loe, Barret Zoph, Bartłomiej Bojanowski, Batuhan Özyurt, Behnam Hedayatnia, Behnam Neyshabur, Benjamin Inden, Benno Stein, Berk Ekmekci, Bill Yuchen Lin, Blake Howald, Bryan Orinion, Cameron Diao, Cameron Dour, Catherine Stinson, Cedrick Argueta, César Ferri Ramírez, Chandan Singh, Charles Rathkopf, Chenlin Meng, Chitta Baral, Chiyu Wu, Chris Callison-Burch, Chris Waites, Christian Voigt, Christopher D. Manning, Christopher Potts, Cindy Ramirez, Clara E. Rivera, Clemencia Siro, Colin Raffel, Courtney Ashcraft, Cristina Garbacea, Damien Sileo, Dan Garrette, Dan Hendrycks, Dan Kilman, Dan Roth, Daniel Freeman, Daniel Khashabi, Daniel Levy, Daniel Moseguí González, Danielle Perszyk, Danny Hernandez, Danqi Chen, Daphne Ippolito, Dar Gilboa, David Dohan, David Drakard, David Jurgens, Debajyoti Datta, Deep Ganguli, Denis Emelin, Denis Kleyko, Deniz Yuret, Derek Chen, Derek Tam, Dieuwke Hupkes, Diganta Misra, Dilyar Buzan, Dimitri Coelho Mollo, Diyi Yang, Dong-Ho Lee, Dylan Schrader, Ekaterina Shutova, Ekin Dogus Cubuk, Elad Segal, Eleanor Hagerman, Elizabeth Barnes, Elizabeth Donoway, Ellie Pavlick, Emanuele Rodola, Emma Lam, Eric Chu, Eric Tang, Erkut Erdem, Ernie Chang, Ethan A. Chi, Ethan Dyer, Ethan Jerzak, Ethan Kim, Eunice Engefu Manyasi, Evgenii Zheltonozhskii, Fanyue Xia, Fatemeh Siar, Fernando Martínez-Plumed, Francesca Happé, Francois Chollet, Frieda Rong, Gaurav Mishra, Genta Indra Winata, Gerard de Melo, Germán Kruszewski, Giambattista Parascandolo, Giorgio Mariani, Gloria Wang, Gonzalo Jaimovitch-López, Gregor Betz, Guy Gur-Ari, Hana Galijasevic, Hannah Kim, Hannah Rashkin, Hannaneh Hajishirzi, Harsh Mehta, Hayden Bogar, Henry Shevlin, Hinrich Schütze, Hiromu Yakura, Hongming Zhang, Hugh Mee Wong, Ian Ng, Isaac Noble, Jaap Jumelet, Jack Geissinger, Jackson Kernion, Jacob Hilton, Jaehoon Lee, Jaime Fernández Fisac, James B. Simon, James Koppel, James Zheng, James Zou, Jan Kocoń, Jana Thompson, Janelle Wingfield, Jared Kaplan, Jarema Radom, Jascha Sohl-Dickstein, Jason Phang, Jason Wei, Jason Yosinski, Jekaterina Novikova, Jelle Bosscher, Jennifer Marsh, Jeremy Kim, Jeroen Taal, Jesse Engel, Jesujoba Alabi, Jiacheng Xu, Jiaming Song, Jillian Tang, Joan Waweru, John Burden, John Miller, John U. Balis, Jonathan Batchelder, Jonathan Berant, Jörg Frohberg, Jos Rozen, Jose Hernandez-Orallo, Joseph Boudeman, Joseph Guerr, Joseph Jones, Joshua B. Tenenbaum, Joshua S. Rule, Joyce Chua, Kamil Kanclerz, Karen Livescu, Karl Krauth, Karthik Gopalakrishnan, Katerina Ignatyeva, Katja Markert, Kaustubh D. Dhole, Kevin Gimpel, Kevin Omondi, Kory Mathewson, Kristen Chiafullo, Ksenia Shkaruta, Kumar Shridhar, Kyle McDonell, Kyle Richardson, Laria Reynolds, Leo Gao, Li Zhang, Liam Dugan, Lianhui Qin, Lidia Contreras-Ochando, Louis-Philippe Morency, Luca Moschella, Lucas Lam, Lucy Noble, Ludwig Schmidt, Luheng He, Luis Oliveros Colón, Luke Metz, Lütfi Kerem Şenel, Maarten Bosma, Maarten Sap, Maartje ter Hoeve, Maheen Farooqi, Manaal Faruqui, Mantas Mazeika, Marco Baturan, Marco Marelli, Marco Maru, Maria Jose Ramírez Quintana, Marie Tolkiehn, Mario Giulianelli, Martha Lewis, Martin Potthast, Matthew L. Leavitt, Matthias Hagen, Mátyás Schubert, Medina Orduna Baitemirova, Melody Arnaud, Melvin McElrath, Michael A. Yee, Michael Cohen, Michael Gu, Michael Ivanitskiy, Michael Starritt, Michael Strube, Michał Swędrowski, Michele Bevilacqua, Michihiro Yasunaga, Mihir Kale, Mike Cain, Mimee Xu, Mirac Suzgun, Mitch Walker, Mo Tiwari, Mohit Bansal, Moin Aminnaseri, Mor Geva, Mozhdeh Gheini, Mukund Varma T, Nanyun Peng, Nathan A. Chi, Nayeon Lee, Neta Gur-Ari Krakover, Nicholas Cameron, Nicholas Roberts, Nick Doiron, Nicole Martinez, Nikita Nangia, Niklas Deckers, Niklas Muennighoff, Nitish Shirish Keskar, Niveditha S. Iyer, Noah Constant, Noah Fiedel, Nuan Wen, Oliver Zhang, Omar Agha, Omar Elbaghdadi, Omer Levy, Owain Evans, Pablo Antonio Moreno Casares, Parth Doshi, Pascale Fung, Paul Pu Liang, Paul Vicol, Pegah Alipoormolabashi, Peiyuan Liao, Percy Liang, Peter Chang, Peter Eckersley, Phu Mon Htut, Pinyu Hwang, Piotr Miłkowski, Piyush Patil, Pouya Pezeshkpour, Priti Oli, Qiaozhu Mei, Qing Lyu, Qinlang Chen, Rabin Banjade, Rachel Etta Rudolph, Raefer Gabriel, Rahel Habacker, Ramon Risco, Raphaël Millière, Rhythm Garg, Richard Barnes, Rif A. Saurous, Riku Arakawa, Robbe

Raymaekers, Robert Frank, Rohan Sikand, Roman Novak, Roman Sitelew, Ronan LeBras, Rosanne Liu, Rowan Jacobs, Rui Zhang, Ruslan Salakhutdinov, Ryan Chi, Ryan Lee, Ryan Stovall, Ryan Teehan, Rylan Yang, Sahib Singh, Saif M. Mohammad, Sajant Anand, Sam Dillavou, Sam Shleifer, Sam Wiseman, Samuel Gruetter, Samuel R. Bowman, Samuel S. Schoenholz, Sanghyun Han, Sanjeev Kwatra, Sarah A. Rous, Sarik Ghazarian, Sayan Ghosh, Sean Casey, Sebastian Bischoff, Sebastian Gehrmann, Sebastian Schuster, Sepideh Sadeghi, Shadi Hamdan, Sharon Zhou, Shashank Srivastava, Sherry Shi, Shikhar Singh, Shima Asaadi, Shixiang Shane Gu, Shubh Pachchigar, Shubham Toshniwal, Shyam Upadhyay, Shyamolima, Debnath, Siamak Shakeri, Simon Thormeyer, Simone Melzi, Siva Reddy, Sneha Priscilla Makini, Soo-Hwan Lee, Spencer Torene, Sriharsha Hatwar, Stanislas Dehaene, Stefan Divic, Stefano Ermon, Stella Biderman, Stephanie Lin, Stephen Prasad, Steven T. Piantadosi, Stuart M. Shieber, Summer Misherghi, Svetlana Kiritchenko, Swaroop Mishra, Tal Linzen, Tal Schuster, Tao Li, Tao Yu, Tariq Ali, Tatsu Hashimoto, Te-Lin Wu, Théo Desbordes, Theodore Rothschild, Thomas Phan, Tianle Wang, Tiberius Nkinyili, Timo Schick, Timofei Kornev, Titus Tunduny, Tobias Gerstenberg, Trenton Chang, Trishala Neeraj, Tushar Khot, Tyler Shultz, Uri Shaham, Vedant Misra, Vera Demberg, Victoria Nyamai, Vikas Raunak, Vinay Ramasesh, Vinay Uday Prabhu, Vishakh Padmakumar, Vivek Srikumar, William Fedus, William Saunders, William Zhang, Wout Vossen, Xiang Ren, Xiaoyu Tong, Xinran Zhao, Xinyi Wu, Xudong Shen, Yadollah Yaghoobzadeh, Yair Lakretz, Yangqiu Song, Yasaman Bahri, Yejin Choi, Yichi Yang, Yiding Hao, Yifu Chen, Yonatan Belinkov, Yu Hou, Yufang Hou, Yuntao Bai, Zachary Seid, Zhuoye Zhao, Zijian Wang, Zijie J. Wang, Zirui Wang, and Ziyi Wu. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models.

Paul Thagard. 1989. Explanatory coherence. *Behavioral and brain sciences*, 12(3):435–467.

Paul R Thagard. 1978. The best explanation: Criteria for theory choice. *The journal of philosophy*, 75(2):76–92.

Mokanarangan Thayaparan, Marco Valentino, and André Freitas. 2020. A survey on explainability in machine reading comprehension. *arXiv preprint arXiv:2010.00389*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models.

Marco Valentino and André Freitas. 2022. Scientific explanation and natural language: A unified epistemological-linguistic perspective for explainable ai. *arXiv preprint arXiv:2205.01809*.

Marco Valentino, Ian Pratt-Hartmann, and André Freitas. 2021. Do natural language explanations represent valid logical arguments? verifying entailment in explainable NLI gold standards. In *Proceedings of the 14th International Conference on Computational Semantics (IWCS)*, pages 76–86, Groningen, The Netherlands (online). Association for Computational Linguistics.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. *CoRR*, abs/1905.00537.

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

Leon Weber, Pasquale Minervini, Jannes Münchmeyer, Ulf Leser, and Tim Rocktäschel. 2019. Nlprolog: Reasoning with weak unification for question answering in natural language.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed H. Chi, Quoc Le, and Denny Zhou. 2022a. Chain of thought prompting elicits reasoning in large language models. *CoRR*, abs/2201.11903.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022b. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Sarah Wiegreffe and Ana Marasovic. 2021. Teach me to explain: A review of datasets for explainable natural language processing. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. 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 1112–1122. Association for Computational Linguistics.

Yuhuai Wu, Albert Q. Jiang, Wenda Li, Markus N. Rabe, Charles Staats, Mateja Jamnik, and Christian Szegedy. 2022. Autoformalization with large language models.

Chloe Xiang. 2023. Openai's gpt-4 is closed source and shrouded in secrecy.

Fangzhi Xu, Qika Lin, Jiawei Han, Tianzhe Zhao, Jun Liu, and Erik Cambria. 2023. Are large language models really good logical reasoners? a comprehensive evaluation and beyond.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc Le, and Ed Chi. 2023. Least-to-most prompting enables complex reasoning in large language models.

A Appendix

A.1 Explanation Prompting

Type	Example	Entailment Forms
Cause Prediction	Context: The balloon expanded. Question: What was the cause? A) I blew into it. B) I pricked it.	Premise: I blew into it. Conclusion: The balloon expanded. Premise: I pricked it. Conclusion: The balloon expanded.
Effect Prediction	Context: The child punched the stack of blocks. Question: What was the effect? A) The stack towered over the boys head. B) The blocks scattered all over the rug.	Premise: The child punched the stack of blocks. Conclusion: The stack towered over the boys head. Premise: The child punched the stack of blocks. Conclusion: The blocks scattered all over the rug.

Figure 8: To perform IBE we convert the CQA context and answer candidates into an entailment form (i.e., EEV) (Valentino et al., 2021).

A modified CoT prompt is used to instruct the LLM to generate explanations. The prompt includes a set of instructions for explanation generation and an in-context example. Appended to the end of the prompt are the CQA context, causal question, and answer candidates. The LLM is instructed to first convert the options into the EEV format consisting of a premise and conclusion. The EEV format will differ depending on the directionality of the causal question (see Figure 8). Cause prediction questions will treat the answer candidate as the premise and the context as the conclusion. In contrast, effect prediction reverses the relationship treating the context as the premise and the answer options as the conclusion. After the EEV conversion, the model is instructed to generate a step-by-step explanation consisting of IF-THEN statements and the associated causal or commonsense assumptions. For ease of post-processing, the LLM is instructed to use headers and enumerate steps using the *Step #* format. A full example of the prompt template is exhibited in Figure 9.

```
For the provided scenario, identify which option is the most plausible cause of
the context. Let's think step-by-step and generate an explanation for each
option. Treat each option as the premise and the provided context as the conclus
ion.  Generate a short step-by-step logical proof that explains how the premise
can result in the conclusion. For each step provide an IF-THEN rule and the
underlying causal or commonsense assumption. After generating the explanations,
select which option is the most plausible cause. Ensure your response contains
the following sections: Option 1 Explanation, Option 2 Explanation, and Answer.
Your final answer should be the option which is the most plausible. You must
select only one option as the final answer. Incomplete answers will be
penalized.

Example:
Context: The woman banished the children from her property.
Question: What was the cause?
Options:
(a) the children trampled through her garden
(b) the children hit a ball into her yard

Option 1 Explanation:
Premise: the children trampled through her garden.
Conclusion: The woman banished the children from her property.
Step 1: IF children trample through someone's garden, THEN it can cause damage
to the garden.
Assumption: Trampling through a garden can result in damage to the garden.
[...]
Step 5: Therefore, since the children trampled through her garden, causing damag
e, the woman may have felt upset or angry and decided to banish the children fro
m her property as a way to prevent further damage.

Option 2 Explanation
Premise: the children hit a ball into her yard.
Conclusion: The woman banished the children from her property.

Step 1: IF children hit a ball into her yard, THEN the woman may feel her proper
ty is being invaded.
Assumption: Having objects thrown into one's yard can be seen as an invasion of
privacy.
[...]
Step 5: Therefore, since the children hit a ball into her yard, the woman may ha
ve felt her property was being invaded, which could have led to her becoming ang
ry and ultimately banishing the children from her property.

Answer: (a) the children trampled through her garden
--------------------------
Context:
Question:
Options: |
```

Figure 9: An example of the modified CoT prompt template for explanation generation.

```
Convert the provided premise, conclusion, and explanation into the Prolog
syntax. Generate the goal from the Conclusion. Generate the facts from the Premi
se. Generate the rules from the Explanation. Ensure there is only one variable p
er predicate. Do not generate rules or facts with more than one variable. For ex
ample 'intoxicated(X, main)'' is not allowed. 'intoxicated(X,Y)' is not allowed.
Do not generate goals with multiple constants. For example 'leaking(water_pipe,
frozen)' is not allowed. Ensure that the goal and facts refer to the same consta
nt.

Example 1:
Premise: Tom's pancreas was injured.
Conclusion: He has a high blood sugar level.
Explanation:
- IF pancreas are injured, THEN pancreas may be dysfunctional.
- IF pancreas are dysfunctional, THEN pancreas have a reduced capacity for insul
in production.
- IF there is a reduced capacity for insulin production, THEN there there is hig
h levels of blood sugar.
- Therefore, since Tom's pancreas was injured, he may have a reduced capacity fo
r insulin production, leading to insufficient insulin and high blood sugar level
s.

Goal:
- has_high_blood_sugar(tom).
Formal Goal:
- has_high_blood_sugar(X) :- tom(X).
Facts:
- injured_pancreas(tom)
- tom(tom)
Rules:
- dysfunctional_pancreas(X) :- injured_pancreas(X).
- reduced_insulin_production(X) :- dysfunctional_pancreas(X)
- has_high_blood_sugar(X) :- reduced_insulin_production(X)

Example 2:
[....]
--------------------------
Premise:
Conclusion:
Explanation: |
```

Figure 10: An example of the autoformalization prompt.

A.2 Autoformalisation

Autoformalisation is the process of translating natural language descriptions into formal specifications (Wu et al., 2022). We adopt the translational capabilities of ChatGPT-3.5-Turbo to convert the explanation into a formal entailment hypothesis. The IF-THEN explanation steps are converted into a set of Prolog rules, the entailment description is used to generate Prolog atoms, and the conclusion statement is translated into a Prolog query. We provide an example of the autoformalization prompt in Figure 10 and an example of the formalized output in Figure 11. After autoformalization, we deploy a post-processing script to extract the formalized rules, atoms, and query and generate a Prolog program for entailment verification.

A.3 Logical Consistency

1. Explanation	2. Formalized Output	3. Generated Proof
Premise: **I blew into it.** Conclusion: **The balloon expanded** **Step 1**: IF someone blows into a balloon, THEN it can cause the balloon to inflate. **Assumption:** Blowing air into a balloon increases the amount of air inside it, leading to inflation. **Step 2:** IF the balloon inflates, THEN it can cause the balloon to expand. **Assumption:** When a balloon inflates, it stretches and expands in size. Therefore, since I blew into the balloon, it caused the balloon to inflate, which resulted in its expansion.	**Prolog Query** expanded_balloon(me). **Program** % Atoms blew_into_balloon(me). me(me). % Rules inflated_balloon(X) :- blew_into_balloon(X). expanded_balloon(X) :- inflated_balloon(X).	expanded_balloon(me) –> expanded_balloon(X) :- inflated_balloon(X) –> inflated_balloon(X) :- blew_into_balloon(X) –> **blew_into_balloon(me)**

Figure 11: An example of the autoformalization prompt.

An explanation hypothesis is considered logically consistent if the external solver can build a deductive proof connecting the conclusion to the premise. We use NLProlog (Weber et al., 2019), a neuro-symbolic Prolog solver integrating backward chaining with word embedding models via a weak unification mechanism. NLProlog allows for a level of flexibility and robustness that is necessary for NLP use cases (e.g. unification applied to synonyms). We provide the autoformalized query, atoms, and rules to NLProlog. If NLProlog can satisfy the entailment query, it will return the proof consisting of the set of rules traversed, the weak unification score, and the proof depth. For simplicity, we assign a score of one if the entailment query is satisfied and zero if it is not. The proof depth score is evaluated as part of the parsimony analysis. An end-to-end example of consistency evaluation can be found in Figure 11.

A.4 Parsimony

Parsimony measures the complexity of an explanation and is represented by the proof depth and concept drift metrics. Proof depth is automatically calculated by NLProlog and reflects the number

Algorithm 1: Neuro-symbolic Solver

Input : Symbolic KB kb, Goal $goal$, Glove embedding model $e(\cdot)$
Output : proof chain $chain$, proof depth $depth$

threshold $\leftarrow$ 0.13;
$depth \leftarrow 1$;
$chain \leftarrow emptylist$;
foreach *step* t **in** *backward_chaining(kb, goal)* **do**
  foreach $max\_unification(q, q_t)$ **do**
    $unification\_score \leftarrow CosineSimilarity(e(q, m_s), e(q_t, m_s))$;
    $depth \leftarrow depth \times unification\_score$;
  end
  $chain \leftarrow$ backward_chaining(kb, $goal$);
end
if *chain is not empty* **and** $depth >$ *threshold* **then**
  $chain \leftarrow$ current_proof_chain[0];
end
else
  $depth \leftarrow 0$;
end
return *chain, depth*;

of rules traversed by the solver to satisfy the entailment query. If the hypothesis is not logically consistent, depth is set to zero. The concept drift metric measures the entropy of novel concepts introduced to bridge the premise and conclusion. To compute the drift of an explanation, we consider the nouns found in the premise, conclusion, and explanation steps. We use Spacy (Honnibal and Montani, 2017) to tokenize and extract part-of-speech (POS) tags. All tokens with the 'NOUN' POS tag extracted. For normalization purposes, we consider the lemma of the tokens. Concept drift then is calculated as the set difference between the unique nouns found across all explanation steps and those found in the premise and conclusion.

A.5 Coherence

Coherence evaluates the plausibility of the intermediate explanation. We propose stepwise entailment as a metric to measure the entailment strength of the If-then implications. We employ a RoBERTa-

Algorithm 2: Concept Drift

Input : Premise, Conclusion, Explanation, Spacy model $spacy(\cdot)$
Output : Drift Score $drift$

$Noun_p \leftarrow spacy(Premise)$;
$Noun_c \leftarrow spacy(Conclusion)$;
$Noun_E \leftarrow spacy(Explanation)$;
$N \leftarrow \{Noun_p, Noun_c, Noun_E\}$;
$drift \leftarrow length(set(Noun_E) - set(Noun_p \cup Noun_c))$;
return $drift$;

based NLI model (Nie et al., 2020) that has been finetuned on a range of NLI and fact verification datasets consisting of SNLI (Bowman et al., 2015), aNLI (Nie et al., 2020), multilingual NLI (Williams et al., 2018)), and FEVER-NLI (Nie et al., 2019). To compute the stepwise entailment score, we first measure the entailment strength between the If and Then propositions. For example, to calculate the score of the statement *"IF a balloon is pricked, THEN the balloon may deflate"* we consider *"a balloon is pricked"* and *"the balloon may deflate"* as input sentences for the NLI model. The NLI will produce independent scores for the entailment and contradiction labels. We compute the entailment strength by subtracting the contraction label score from the entailment label score. An entailment strength of one indicates the If-then implication is strongly plausible whereas a score of zero suggests that it is likely implausible. The overall stepwise entailment score is the average of entailment strength measures across all explanation steps.

A.6 Linguistic Uncertainty

Linguistic uncertainty measures the confidence of a statement where hedging cues and indirect language suggest ambiguity around the proposition. To measure sentence-level uncertainty, we employ a finetuned RoBERTa model provided by (Pei and Jurgens, 2021). The model was trained on a sentence-level dataset consisting of findings and statements extracted from new articles and scientific publications and human annotated evaluation of sentence certainty. A scale from one to six was used to annotate sentences where one corresponds to the lowest degree of certainty and six is the highest expressed by the sentence. We invert the scale to retrieve the uncertainty scores. To compute the overall linguistic uncertainty of an explanation, we

Algorithm 3: Stepwise Entailment

Input : Explanation E, NLI Model $nli(\cdot)$
Output : Average Entailment Strength $strength$

$EntailmentStrengthScores \leftarrow$ empty list;
foreach *Step* $(If_s, Then_s)$ *in* E **do**
  $EntailmentScore \leftarrow nli(If_s, Then_s)$;
  $ContradictionScore \leftarrow nli(If_s, Then_s)$;
  $EntailmentStrength \leftarrow EntailmentScore - ContradictionScore$;
  Append $EntailmentStrength$ to $EntailmentStrengthScores$;
end
$strength \leftarrow Avg(EntailmentStrengthScores)$;
return $strength$;

first compute the uncertainty for each assumption and the explanation summary and then average all the scores.

A.7 Inference to Best Explanation

To perform IBE, we first fit a linear regression model over the extracted explanation features from the COPA train set and 500 random sample train examples from the E-CARE train set. We consider all explanations independently and annotate each explanation with a 1 if it corresponds to a correct answer or 0 if corresponds to an incorrect answer. After the linear model is fitted, we evaluate the COPA and E-CARE test sets. For each example, we use the trained linear model to score each answer candidate explanation and then select a candidate with the highest score. We use the linear regression implementation from scikit-learn (Buitinck et al., 2013) for the IBE model. We additionally use the R stats package (R Core Team, 2013) for conducting our regression analysis.

A.8 E-CARE Results

A.8.1 E-CARE Consistency

See Figure 12.

A.8.2 E-CARE Proof Depth

See Figure 13.

Algorithm 4: Linguistic Uncertainty

Input : Assumptions, Explanation Summary, Uncertainty Estimator Model $uc(\cdot)$

Output : Overall Uncertainty

$AssumptionUncertaintyList \leftarrow$ empty list;

foreach *Assumption in Assumptions* **do**

  $UncertaintyScore \leftarrow uc(UncertaintyModel, Assumption)$;

  Append $UncertaintyScore$ to $AssumptionUncertaintyList$;

end

$AverageAssumptionUncertainty \leftarrow Avg(AssumptionUncertaintyList)$;

$ExplanationUncertainty \leftarrow uc(UncertaintyModel, ExplanationSummary)$;

$OverallExplanationUncertainty \leftarrow AverageAssumptionUncertainty + ExplanationUncertainty$;

return $OverallExplanationUncertainty$;

A.8.3 E-CARE Concept Drift

See Figure 13.

A.8.4 E-CARE Coherence

See Figure 15.

A.8.5 E-CARE Uncertainty

See Figure 16.

A.8.6 E-CARE Hedge Ratio

See Figure 17.

A.8.7 E-CARE Hedge Distribution

See Figure 18.

A.9 Dataset Details

COPA is released under a BSD-2 license and made available for broad research usage with copyright notification restrictions [2]. We do not modify or use COPA outside of its intended use which is primarily open-domain commonsense causal reasoning. E-CARE is released under the MIT license and can be used for broad purposes with copyright notification restrictions [3]. We do not modify or use E-CARE outside of its intended use which is causal reasoning evaluation of language models.

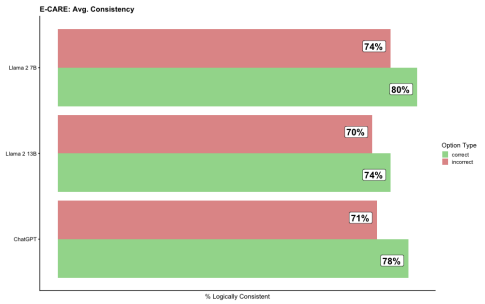

Figure 12: Average consistency comparison between correct and incorrect options for the E-CARE dataset.

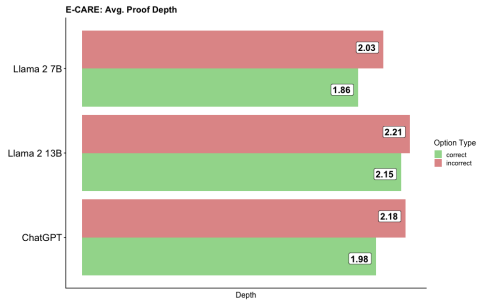

Figure 13: Comparison of average proof depth between correct and incorrect options.

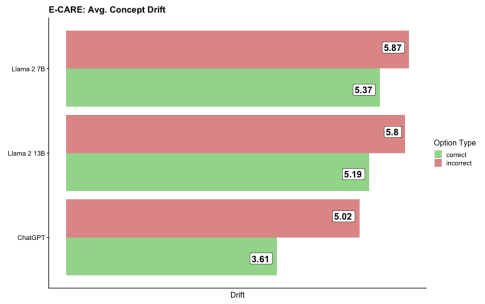

Figure 14: Comparison of average concept drift between correct and incorrect options.

[2] people.ict.usc.edu/ gordon/copa.html

[3] github.com/Waste-Wood/e-CARE?tab=MIT-1-ov-file#readme

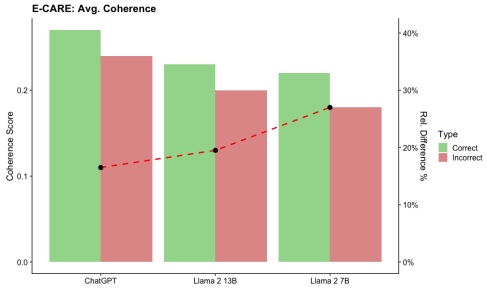

Figure 15: Comparison of average coherence scores between correct and incorrect options.

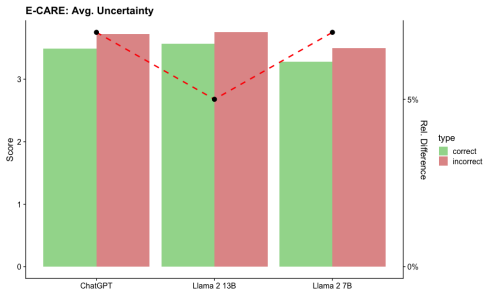

Figure 16: Comparison of average uncertainty scores between correct and incorrect options.

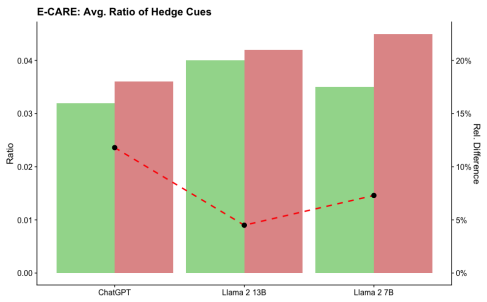

Figure 17: Comparison of the average ratio of hedge cues between correct and incorrect options.

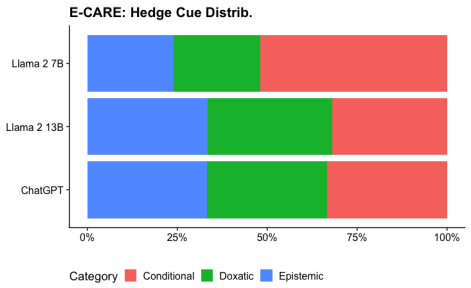

Figure 18: Distribution of hedge cues across incorrect explanations.