

# Inference to the Best Explanation in Large Language Models

Dhairya Dalal<sup>1</sup>, Marco Valentino<sup>2</sup>, André Freitas<sup>2,3,4</sup>, Paul Buitelaar<sup>1,5</sup>

<sup>1</sup> SFI Centre for Research and Training in Artificial Intelligence, University of Galway, Ireland

<sup>2</sup> Idiap Research Institute, Switzerland

<sup>3</sup> Department of Computer Science, University of Manchester, UK

<sup>4</sup> National Biomarker Centre, CRUK-MI, University of Manchester, UK

<sup>5</sup> Insight SFI Research Centre for Data Analytics, University of Galway, Ireland

Email: d.dalal1@universityofgalway.ie

## Abstract

While Large Language Models (LLMs) have found success in real-world applications, their underlying explanatory process is still poorly understood. This paper proposes *IBE-Eval*, a framework inspired by philosophical accounts on *Inference to the Best Explanation (IBE)* to advance the interpretation and evaluation of LLM explanations. *IBE-Eval* estimates the plausibility of natural language explanations through a combination of explicit logical and linguistic features including: *consistency*, *parsimony*, *coherence*, and *uncertainty*. Extensive experiments are conducted on *Causal Question Answering (CQA)*, where *IBE-Eval* is tasked to select the most plausible causal explanation amongst competing ones generated by the LLM (e.g. GPT 3.5 or LLaMA 2). The experiments reveal that *IBE-Eval* can successfully identify the best explanation with up to 77% accuracy ( $\approx 27\%$  above random), improving upon a GPT 3.5-as-a-judge baseline ( $\approx +17\%$ ) while being intrinsically more efficient and interpretable. Additional analysis suggests that, despite LLM-specific variances, generated explanations tend to conform to IBE criteria and that *IBE-Eval* is significantly correlated with human judgment, opening up opportunities for future development of automated explanation verification tools.

## 1 Introduction

Large Language Models (LLMs) such as OpenAI’s GPT (Brown et al., 2020) and LLaMA (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 understood (Valentino et al., 2021). LLMs are no-

toriously black-box and can be difficult to interpret (Chakraborty et al., 2017; Danilevsky et al., 2020). Moreover, the commercialization of LLMs has led to strategic secrecy around model architectures and training details (Xiang, 2023; Knight, 2023). Finally, LLMs 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 generated explanations becomes crucial to better interpret and validate the LLM’s internal inference and reasoning processes (Wei et al., 2022b; Lampinen et al., 2022; Huang and Chang, 2022).

The automatic evaluation of natural language explanations presents several challenges (Atanasova et al., 2023; Camburu et al., 2020). Without resource-intensive annotation (Wiegrefe and Marasovic, 2021; Thayaparan et al., 2020; Dalvi et al., 2021; Camburu et al., 2018), explanation quality methods tend to rely on either weak supervision, where the identification of the correct answer is taken as evidence of explanation quality, or require the injection of domain-specific knowledge (Quan et al., 2024). In this paper, we seek to better understand the 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 measurable features that can serve to define criteria for assessing their quality.

Specifically, this paper investigates the following overarching research question: “*Can the linguistic and logical properties associated with LLM-generated explanations be used to qualify the models’ reasoning process?*”. To this end, we propose an interpretable framework inspired by philosophical accounts of abductive inference, also known as *Inference to the Best Explanation (IBE)* - i.e. the

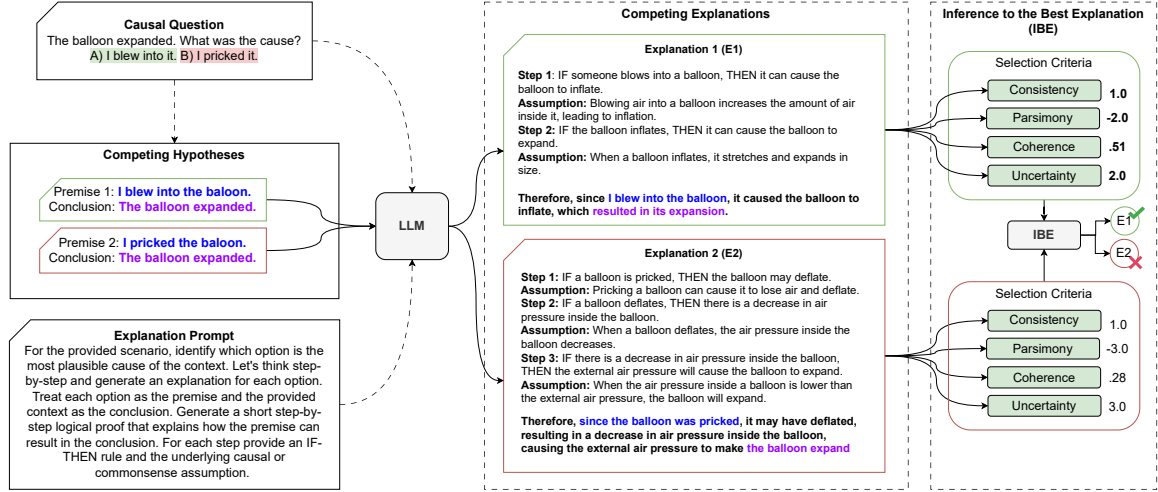


Figure 1: *IBE-Eval* qualifies LLM-generated explanations with a set of logical and linguistic selection criteria to identify the most plausible hypothesis. The corresponding explanation for each hypothesis is evaluated across the IBE criteria of logical consistency, parsimony, internal coherence, and linguistic uncertainty. A final plausibility score is computed across those features and the hypothesis with highest score is identified as the best explanation.

process of selecting among competing explanatory theories (Lipton, 2017). In particular, we aim to measure the extent to which LLM-generated explanations conform to IBE expectations when attempting to identify the most plausible explanation. To this end, we present *IBE-Eval*, a framework designed to estimate the plausibility of natural language explanations through a set of explicit logical and linguistic features, namely: *logical consistency*, *parsimony*, *coherence*, and *linguistic uncertainty*.

To evaluate the efficacy of *IBE-Eval*, we conduct extensive experiments in the multiple-choice Causal Question Answering (CQA) setting. The overall results and contributions of the paper can be summarized as follows:

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 propose *IBE-Eval*, a framework that can be instantiated with external tools for the automatic evaluation of LLM-generated explanations and the identification of the best explanation in a multiple-choice CQA setting.
3. We provide empirical evidence that LLM-generated explanations tend to conform to IBE expectations with varying levels of statistical significance correlated to the LLM’s size.

4. We additionally find that uncertainty, parsimony, and coherence are the best predictors of plausibility and explanation quality across all LLMs. However, we also find that the LLMs tend to be strong rationalizers and can produce logically consistent explanations even for less plausible candidates, making the consistency metric less effective in practice.
5. *IBE-Eval* can successfully identify the best explanation supporting the correct answers with up to 77% accuracy ( $\approx 27\%$  above random and  $\approx 17\%$  over GPT 3.5-as-a-Judge baselines)
6. *IBE-Eval* is significantly correlated with human judgment, outperforming a GPT3.5-as-a-Judge baseline in terms of alignment with human preferences.

For reproducibility, our code is made available on Github<sup>1</sup> 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; Deutsch, 2011). Accepted epistemological accounts characterize the creation of an explanation as composed of two distinct phases: conjecturing and criticism (Popper,

<sup>1</sup><https://github.com/dhairyalal/IBE-eval>

2014). 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 such as parsimony, coherence, unification power, and hardness to variation (Lipton, 2017; Harman, 1965; 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 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 computed on explicit linguistic and logical features, namely: *consistency*, *parsimony*, *coherence*, and *uncertainty*.

To assess the LLM’s alignment to such criteria, we focus on the task of selecting among competing explanations in a multiple-choice CQA setting (Figure 1). Specifically, given a set of competing hypotheses (i.e. the multiple-choice options),  $H = \{h_1, h_2, \dots, h_n\}$ , we prompt the LLM to generate plausible explanations supporting each hypothesis (Section 3). Subsequently, we adopt the proposed IBE selection criteria to assess the quality of the generated explanations (Section 4). *IBE-Eval* computes an explanation plausibility score derived from the linear combination of the computed selection criteria. The explanation with the highest score is selected as the predicted answer and additionally assessed as to the extent to which the observable IBE features are correlated with QA accuracy. We hypothesize that IBE-Eval will produce higher scores for the explanation associated with the correct answer and that the IBE criteria should meaningfully differentiate between competing explanations.

### 3 Explanation Generation

For the first stage, the LLM is prompted to generate competing explanations for the hypotheses using a modified Chain-of-Thought (CoT) prompt (Wei et al., 2022a). Specifically, the COT prompt is modified to instruct the LLM to produce an explanation for each competing hypothesis (see Figure 1). We adopt a methodology similar to Valentino et al. (2021), where the generated explanation is

constrained into an entailment form for the downstream IBE evaluation. In particular, we posit that a valid explanation should demonstrate an entailment relationship between the premise and conclusion which are derived from the question-answer pair.

To elicit logical connections between explanation steps and facilitate subsequent analysis, the LLM is constrained to use weak syllogisms expressed as If-Then statements. 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 explanation steps and supporting knowledge for evaluation via the IBE selection criteria. Additional details and examples of prompts are reported in Appendix A.2.

### 4 Linguistic & Inference Criteria

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

**Consistency.** Consistency aims to verify whether the explanation is logically valid. Given a hypothesis, comprised of a premise  $p_i$ , a conclusion  $c_i$ , and an explanation consisting of a set of If-Then statements  $E = s_1, \dots, s_i$ , we define  $E$  to be logically consistent if  $p_i \cup E \models c_i$ . Specifically, an explanation is logically consistent if it is possible to build a deductive proof linking premise and conclusion.

To evaluate logical consistency, we leverage external symbolic solvers along with autoformalization - i.e., the translation of natural language into a formal language (Wu et al., 2022). Specifically, the 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 GPT 3.5. Specifically, we instruct GPT 3.5 to convert each IF-Then explanation step from the generated explanation into an implication rule and the premise statement into grounding atoms. On the other end, the entailment condition and the conclusion are used to create a Prolog query. The query instructs the Prolog solver to attempt to find a path through the implications rules such that the conclusion be directly connected to the premise. Further details about the autoformalization process can be found in Appendix A.3.

After autoformalization, following recent work on neuro-symbolic integration for LLM explanations (Quan et al., 2024), we adopt an external Prolog solver for entailment verification<sup>2</sup>. The explanation is considered consistent if the Prolog solver can satisfy the query and successfully build a deductive proof. Technical details can be found in Appendix A.5.

**Parsimony.** The parsimony principle, also known as Ockham’s razor, favors the selection of the simplest explanation consisting of the fewest elements and assumptions (Sober, 1981). Epistemological accounts posit that 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 an explanation with the smallest observed proof depth and concept drift. Implementation details can be found in Appendix A.6.

**Coherence.** *Coherence* attempts to measure the logical validity at the level of the specific explanation steps. An explanation can be formally consistent on the surface while still including implausible or ungrounded intermediate assumptions. Coherence evaluates the quality of each intermediate If-Then implication by measuring the entailment strength between the If and Then 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 the 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  $|S|$ :

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

We hypothesize that the LLM should generate a higher coherence score for more plausible hypotheses, as such explanations should exhibit stronger step-wise entailment. Additional details can be found in Appendix A.7.

**Uncertainty.** Finally, we consider the 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. 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, they found that the use of hedging language suggested that the veracity of the claim was weaker or highly contextualized.

To measure the linguistic uncertainty ( $UC$ ) of an explanation, we consider the explanation’s underlying assumptions ( $A_i$ ) and the overall explanation summary ( $S$ ). The linguistic uncertainty score is extracted using the fine-tuned sentence-level RoBERTa model from Pei and Jurgens (2021). The overall linguistic uncertainty score ( $UC_{\text{overall}}$ ) is the sum of the assumption and explanation summary scores:

<sup>2</sup>[https://github.com/neuro-symbolic-ai/explanation\\_based\\_ethical\\_reasoning](https://github.com/neuro-symbolic-ai/explanation_based_ethical_reasoning)



$$UC_{\text{overall}} = UC(A) + UC(S)$$

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

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

and linguistic uncertainty of the explanation summary  $UC(S)$ . We hypothesize that the LLM will use more hedging language when explaining the weaker hypothesis resulting in a higher uncertainty score. Further details can be found in Appendix A.8.

#### 4.1 Inference to Best Explanation

After the IBE criteria are computed for each competing hypothesis, they are used to generate the final explanation plausibility score. We define a simple linear regression model  $\theta(\cdot)$ , which was fitted on a small set of training examples consisting of extracted IBE features to predict the probability that an explanation  $E_i$  corresponds to the correct answer. Specifically, we employ *IBE-Eval* to score each generated explanation independently and then select the final answer  $a$  via argmax:

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

Additional details can be found in Appendix A.9.

## 5 Experimental Setting

Causal Question-Answering (CQA) requires reasoning about the causes and effects given an event description. We specifically consider the task of cause and effect prediction in a multiple-choice 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 causal relationships and consider the event 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 crowd-sourced QA dataset consisting of 15K train and 2k test examples. Similar to COPA, the task requires the selection of the most likely cause or effect provided an event description. We randomly sample 500 examples from the E-CARE test set for our experiments.

**LLMs.** We consider GPT-Turbo-3.5, LLaMA 2 13B, and LLaMA 2 7B for all experiments. GPT 3.5 is a proprietary model (Brown et al., 2020) and is 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). We consider both the 13B and 7B variants of Llama 2 as both are seen as viable commodity GPT alternatives and have been widely adopted by the research community for LLM benchmarking and evaluation.

**Baselines.** We employ LLM-as-a-Judge (Zheng et al., 2023) and human evaluators as baseline methods for the selection of the best explanation in the CQA setting. (Zheng et al., 2023) found LLMs can align with human judgment and be utilized for automated evaluation and judgment. We specifically uses GPT 3.5 as the LLM judge. For each CQA example, we present the judges with two competing explanations generated by the target LLM. The judge is asked to identify the best and most plausible explanation. Additional details about the baselines can be found in Appendix A.4.

## 6 Preliminary Analysis

We conduct a preliminary analysis as a sanity check to measure the extent to which LLMs generate *self-evident or tautological* explanations - i.e., explanations that simply restate the premises and conclusions. Tautological explanations present a risk for *IBE-Eval* as the metrics would be theoretically uninformative if the LLM adopts the tested causal relation as the explanation itself (e.g.  $A \rightarrow B$ ) without providing additional supporting statements.

We consider the *parsimony* metric to compute the percentage of explanations with *proof depth* equal to 1 (i.e, explanations containing only one inference step) and *concept drift* equal to 0 (i.e. no

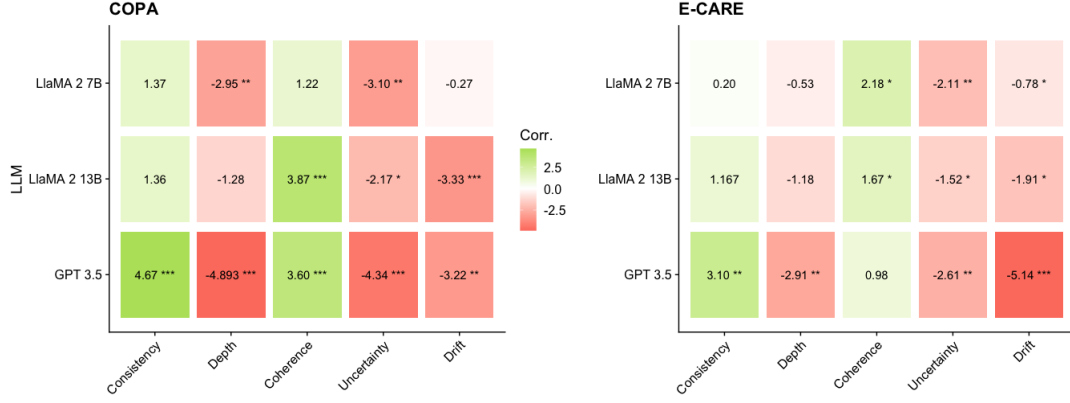


Figure 2: A regression analysis measuring the correlation between IBE criteria and question accuracy. All the LLMs tend to conform to IBE expectations with GPT 3.5 exhibiting the most consistent and significant alignment. Linguistic uncertainty is the strongest IBE predictor for explanation quality, where higher uncertainty is negatively correlated with question accuracy. Statistical significance is noted as: ‘\*\*\*’  $p < 0.001$ , ‘\*\*’  $p < 0.01$  ‘\*’  $p < 0.05$ .

	COPA			E-CARE		
	GPT 3.5	LLaMA 2 13B	LLaMA 2 7B	GPT 3.5	LLaMA 2 13B	LLaMA 2 7B
<b>Baselines</b>						
GPT3.5 Judge	.59	.47	.63	.43	.61	.52
<b>Human</b>	<b>.95</b>	<b>1.0</b>	<b>.91</b>	<b>.90</b>	<b>.91</b>	<b>.92</b>
<b>IBE Features</b>						
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
<b>Composed Model</b>						
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
+ Linguistic Uncertainty	<b>.77</b>	<b>.74</b>	<b>.70</b>	<b>.74</b>	<b>.70</b>	<b>.73</b>

Table 1: An ablation study and evaluation of the IBE criteria and the composed *IBE-Eval* model. *IBE-Eval* outperforms the GPT 3.5 Judge baseline by an average of +17.5% across all all models and tasks.

additional concepts other than the ones stated in premises and conclusions appear in the explanation). In such cases, the LLM is effectively generating a self-evident or tautological explanation.

**We found that about 2% of the cases consist of self-evident explanations.** For GPT 3.5, LLaMA 2 13B, and LLaMA 2 7B, 2% of the generated explanations exhibit a concept drift of 0, and on average 1.5% of the explanations have a proof depth of 1. We then conducted an error analysis to evaluate the cases where *IBE-Eval* selected a self-evident explanation as the best one. **Across all LLMs, less than 0.1% of the errors were caused by the selection of such explanations.** Our analysis suggests that the impact of self-evident explanations is not sig-

nificant and that the IBE framework can be robustly applied to identify such cases.

## 7 Results

To assess the LLM’s alignment with the proposed IBE framework and evaluate the efficacy of *IBE-Eval*, we run a regression analysis and conduct a set of ablation studies to evaluate the relationship between IBE and question accuracy. The main results are presented in Figure 2 and Table 1.

Our regression analysis finds that the IBE criteria are generally consistent across the LLMs as demonstrated by similar correlation patterns found on both the COPA and E-CARE tasks (Figure 2). GPT 3.5 exhibits the strongest alignment with IBE

expectations as we observe nearly all the IBE criteria have statistically significant and directionally aligned correlations across both tasks. **Thus our proposed IBE criteria can serve as promising build blocks for future work on automated explanation evaluation.**

In Table 1 we evaluate the accuracy of the IBE criteria and *IBE-Eval* in selecting the most plausible explanation in the CQA setting. We find that though independently the IBE criteria are generally limited in their ability to identify the more plausible explanation - they still outperform the GPT-3.5-as-a-judge baseline. ***IBE-Eval*, which considers all IBE criteria, improves the ability to select the best explanation by 17% over both the GPT 3.5-as-a-judge and random baselines.** We can achieve up to 77% accuracy utilizing just the extracted IBE criteria demonstrating IBE’s potential value for automatic explanation evaluation.

Next, we explore each explanation feature in further detail to better understand the variances across the IBE criteria and LLMs.

**Consistency.** We find that the LLMs are surprisingly strong conjecture models. The LLMs can generate logically consistent explanations for any hypothesis as observed by similar consistency scores for correct and incorrect (Figure 3) explanations. Moreover, we observe that consistency tends to be a statistically insignificant predictor 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 criteria.** For the incorrect candidate explanations, we find that LLMs over-rationalize and introduce additional premises to demonstrate entailment in their explanations.

**Parsimony.** The results suggest that parsimony has a more consistent effect and is a better predictor of explanation quality. **We observe negative correlations 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 depth and drift to be 6% and 10% greater for the incorrect option across all LLMs (Figure 4). Moreover, the results suggest that **as the LLM parameter size increases, the tendency to over-rationalize increases as well.** This is attested by the fact that the average difference in depth and drift is the great-

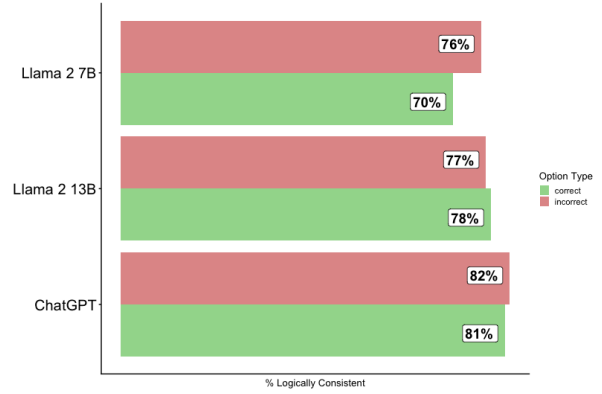


Figure 3: An evaluation of explanation consistency. LLMs are strong rationalizers and can generate logically consistent explanations at equal rates for explanations associated with both correct and incorrect answers options.

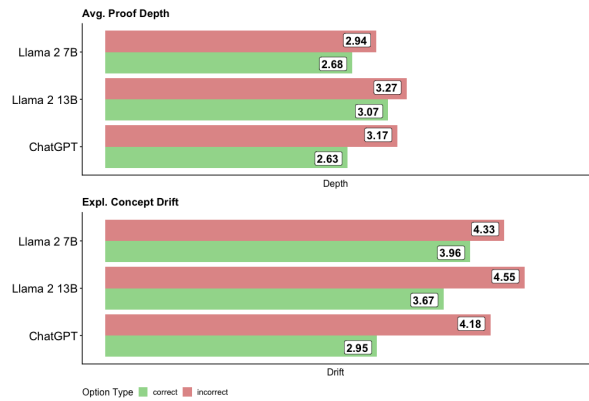


Figure 4: Explanation parsimony is evaluated using proof depth and concept drift. Both metrics are consistently lower for explanations supporting the correct answers suggesting that LLMs are able to generate efficient explanations for the more plausible hypothesis.

est in GPT 3.5, suggesting that the model tends to find the most efficient explanations for stronger hypotheses and articulates explanations for weaker candidates. Finally, we found that the **LLaMA models tend to generate more complex explanations overall**, 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.**

**Coherence.** Similarly to parsimony, we found coherence to be a better indicator of explanation quality being statistically significant for both GPT 3.5 and Llama 2 13B on COPA and both Llama 2 models on E-Care. We found that the average coherence score is consistently greater for the

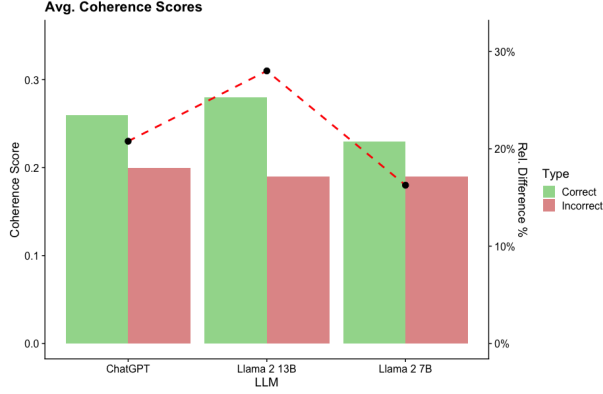


Figure 5: An evaluation of the explanation coherence and question accuracy. The average coherence score is consistently higher for explanations corresponding to the correct hypotheses across the LLMs.

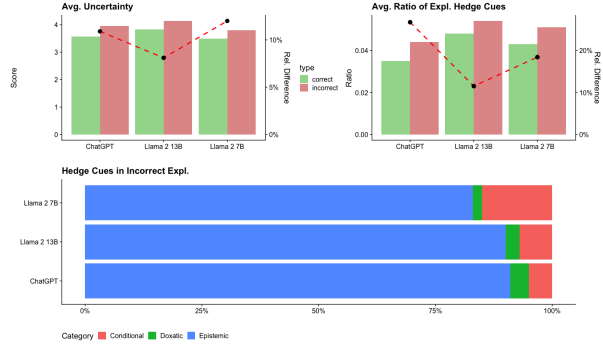


Figure 6: Evaluation of linguistic uncertainty in LLM-generated explanations. LLMs tend to use more hedging language in explanations supporting less plausible hypotheses. Across the LLMs, the hedging language is found to be predominantly *epistemic* A.8.

stronger hypothesis across all LLMs and datasets (see Figure 5). **Both GPT and Llama 2 13B exhibit a higher relative difference between the correct and incorrect hypotheses in contrast to Llama 2 7B.**

**Uncertainty.** The results reveal that **linguistic uncertainty is the strongest predictor of explanation quality and is a statistically significant feature for all LLMs.** This suggests that LLMs use more qualifying language when explaining weaker hypotheses (see Figure 6). We found that **uncertainty can improve accuracy by 13pp on COPA and 4pp on E-CARE.** We also examine the uncertainty cues expressed by LLMs by analyzing both the frequency of hedge words and the types of hedge cues employed in incorrect explanations. We find the distribution of hedge cues across LLMs tends to be similar, with only minor differences be-

tween LLMs (Figure 6). Epistemic cues were most frequently used by all three models, with LLaMA 2 7B being more likely to use conditional cues. See Appendix A.8 for further details..

## 7.1 Correlation with Human Judgement.

We first sample 100 generated explanation pairs across both the COPA and E-CARE tasks and evaluated LLMs. Two human evaluators are instructed to evaluate the pair of explanations and to select which explanation is most plausible. No additional information about the original question nor the correct answer is provided to prevent biasing the judge.

The human evaluators on average were able to identify the explanation associated with the correct answer 96% (COPA) and 91% (E-Care) of time. We compute the inter-evaluator agreement score between two human evaluators and find that there is Cohen’s Kappa score of .68 suggesting there is a strong agreement between the two evaluators.

To evaluate if *IBE-Eval* is correlated with human judgement, we compute the Spearman’s rank correlation between GPT-3.5-as-a-judge, *IBE-Eval* and human judgment. We find that GPT-3.5-as-a-judge exhibits a weak and statistically insignificant correlation with human judgment (0.31). **In contrast, we find that the *IBE-Eval* is significantly aligned with human preferences (with a Spearman’s correlation of 0.64 and  $p < 0.01$ ) further suggesting the IBE’s potential for automatic explanation evaluation.**

## 8 Related Work

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. Research on LLM faithfulness (Atanasova et al., 2023) investigates if LLM explanations are robust to spurious input alterations. Parcalabescu and Frank (2024) propose



a self-consistency measure CC-SHAP which measures how specific alterations to a model’s input contribute to the generated explanation. This paper primarily draws inspiration from recent work on the evaluation of natural language explanations (Quan et al., 2024; Valentino et al., 2021; Wiegrefe 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 or specific domain knowledge, we are the first to propose a set of criteria that can be automatically computed on explicit linguistic and logical features.

## 9 Conclusion

This paper proposed *IBE-Eval*, an interpretable framework for LLM explanation evaluation inspired by philosophical accounts of Inference to the Best Explanation (IBE). *IBE-Eval* can identify the best explanation supporting the correct answer with up to 77% accuracy in CQA scenarios, improving upon a GPT 3.5 Judge baselines by +17%. Our regression study suggests that LLM explanations tend to conform to IBE expectations and that *IBE-Eval* is strongly correlated with human judgment. Linguistic uncertainty is the stronger IBE predictor for explanation quality closely followed by parsimony and coherence. However, we also found that LLMs tend to be strong conjecture models able to generate logically consistent explanations for less plausible hypotheses, suggesting limited applicability for the logical consistency criterion in isolation. We believe our findings can open new lines of research on external evaluation methods for LLMs as well as interpretability tools for understanding the LLM’s underlying explanatory process.

## 10 Limitations

*IBE-Eval* 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 knowledge for factuality. We observe that LLMs 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 internal entailment especially when the explanation is longer or the entailment strength of the surrounding explanation

steps is stronger. Future work can introduce metrics to evaluate grounded knowledge or perform more granular evaluations of explanations to better weight factual inaccuracies.

Additionally, *IBE-Eval* currently does not support single natural language explanations and was evaluated in the limited domain of causal common-sense reasoning. Future work will explore globally calibrating *IBE-Eval* plausibility scores to extend evaluation to more diverse explanation generation and QA settings. Calibration efforts would allow for *IBE-Eval* to generate comparable scores across unrelated explanations and could be used to produce global thresholds explanation classification.

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.

## 11 Ethics Statement

The human annotators for computing the human judgment baseline are all authors of the papers and as such were not further compensated for the annotation task.

## Acknowledgements

This work was partially funded by the Swiss National Science Foundation (SNSF) project NeuMath (200021\_204617), by the EPSRC grant EP/T026995/1 entitled “EnnCore: End-to-End Conceptual Guarding of Neural Architectures” under Security for all in an AI-enabled society, by the CRUK National Biomarker Centre, and supported by the Manchester Experimental Cancer Medicine Centre, the Science Foundation Ireland under grants SFI/18/CRT/6223 (Centre for Research Training in Artificial Intelligence), SFI/12/RC/2289\_P2 (Insight), co-funded by the European Regional Development Fund, and the NIHR Manchester Biomedical Research Centre.

## References

Pepa Atanasova, Oana-Maria Camburu, Christina Lioma, Thomas Lukasiewicz, Jakob Grue Simonsen, and Isabelle Augenstein. 2023. [Faithfulness tests for natural language explanations](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*,

- pages 283–294, Toronto, Canada. Association for Computational Linguistics.
- 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, Yanis 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 Yuksekogonul, 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](#).
- Letitia Parcalabescu and Anette Frank. 2024. [On measuring faithfulness or self-consistency of natural language explanations](#).
- Jiaxin Pei and David Jurgens. 2021. [Measuring sentence-level and aspect-level \(un\)certainly 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.
- Xin Quan, Marco Valentino, Louise A Dennis, and André Freitas. 2024. Enhancing ethical explanations of large language models through iterative symbolic refinement. *arXiv preprint arXiv:2402.00745*.
- 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 Askill, 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 Tabasum, Arul Menezes, Arun Kirubakaran, 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 Engelfu 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 Chifullo, 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, Lütfti 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 Śwędrowski, Michele Bevilacqua, Michihiro Yasunaga, Mihir Kale, Mike Cain, Mimeo Xu, Mirac Suzgun, Mitch Walker, Mo Tiwari, Mohit Bansal, Moin Aminnaseri, Mor Geva, Mozdeh 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 Asadi, 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 Mishnerghi, 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 Wiegrefe 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](#).
- Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Haoteng Zhang, Joseph Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *ArXiv*, abs/2306.05685.

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 Reproducibility

All experimental code is available online<sup>3</sup> to encourage future research in the field. We additionally summarize all the model implementations and technical resources used for the computation of the proposed IBE criteria below:

- We adopt the Prolog solver for neuro-symbolic integration from (Quan et al., 2024).
- We use Spacy (Honnibal and Montani, 2017) to tokenize and extract part-of-speech (POS) tags.
- To compute coherence, we employ the RoBERTa-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 measure sentence-level uncertainty, we employ a finetuned RoBERTa model provided by (Pei and Jurgens, 2021).
- 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).

### A.2 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 7: To perform IBE we convert the CQA context and answer candidates into an entailment form (i.e., EEV) (Valentino et al., 2021).

<sup>3</sup><https://github.com/dhairyalal/IBE-eval>

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 conclusion. 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 damage, the woman may have felt upset or angry and decided to banish the children from 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 property 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 have felt her property was being invaded, which could have led to her becoming angry and ultimately banishing the children from her property.

Answer: (a) the children trampled through her garden

-----  
Context:  
Question:  
Options: |

Figure 8: 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 is high lev
els 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 9: An example of the autoformalization prompt.

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 7). 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 common-sense 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 8.

### A.3 Autoformalization

Autoformalization is the process of translating natural language descriptions into formal specifications (Wu et al., 2022). We adopt the translational capabilities of GPT-3.5-Turbo to convert the expla-

nation 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 9 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.4 LLM-as-a-Judge Baseline

Given the two explanations below (explanation 1 and explanation 2) which explanation is more plausible. A good explanation should be logically consistent and arrive at the correct conclusion. Respond with either explanation 1 or explanation 2 as your final answer.

Explanation 1:  
...

Explanation 2:  
...

Answer:

Figure 10: An example of prompt used by the LLM-as-a-Judge model for evaluating competing explanations.

GPT 3.5 is used as the LLM for the LLM-as-a-Judge baseline. Similar to the human evaluators, GPT is presented with both generated explanations and asked to identify which explanation is more plausible.

### A.5 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.

---

**Algorithm 1: Neuro-symbolic Solver**


---

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

```

1 threshold  $\leftarrow 0.13$ ;
2  $depth \leftarrow 1$ ;
3  $chain \leftarrow emptylist$ ;
4 foreach step  $t$  in  $backward\_chaining(kb, goal)$  do
5   foreach  $max\_unification(q, q_t)$  do
6      $unification\_score \leftarrow$   

        $CosineSimilarity(e(q, m_s), e(q_t, m_s))$ ;
7      $depth \leftarrow depth \times$   

        $unification\_score$ ;
8   end
9    $chain \leftarrow backward\_chaining(kb, goal)$ ;
10 end
11 if  $chain$  is not empty and  $depth >$   

     $threshold$  then
12    $chain \leftarrow current\_proof\_chain[0]$ ;
13 end
14 else
15    $depth \leftarrow 0$ ;
16 end
17 return  $chain, depth$ ;

```

---

### A.6 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 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 Mon-

tani, 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.

---

**Algorithm 2: Concept Drift**


---

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

```

1  $Noun_p \leftarrow spacy(Premise)$ ;
2  $Noun_c \leftarrow spacy(Conclusion)$ ;
3  $Noun_E \leftarrow spacy(Explanation)$ ;
4  $N \leftarrow \{Noun_p, Noun_c, Noun_E\}$ ;
5  $drift \leftarrow length(set(Noun_E) -$   

    $set(Noun_p \cup Noun_c))$ ;
6 return  $drift$ ;

```

---

### A.7 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-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 contradiction 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.8 Linguistic Uncertainty

Linguistic uncertainty measures the confidence of a statement where hedging cues and indirect



---

**Algorithm 3: Stepwise Entailment**

---

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

```

1  $EntailmentStrengthScores \leftarrow$  empty
  list;
2 foreach  $Step(If_s, Then_s)$  in  $E$  do
3    $EntailmentScore \leftarrow nli(If_s,$ 
      $Then_s)$ ;
4    $ContradictionScore \leftarrow nli(If_s,$ 
      $Then_s)$ ;
5    $EntailmentStrength \leftarrow$ 
      $EntailmentScore -$ 
      $ContradictionScore$ ;
6   Append  $EntailmentStrength$  to
      $EntailmentStrengthScores$ ;
7 end
8  $strength \leftarrow$ 
    $Avg(EntailmentStrengthScores)$ ;
9 return  $strength$ ;
```

---

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

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 categories: 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).

---

**Algorithm 4: Linguistic Uncertainty**

---

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

```

1  $AssumptionUncertaintyList \leftarrow$  empty
  list;
2 foreach  $Assumption$  in  $Assumptions$  do
3    $UncertaintyScore \leftarrow$ 
      $uc(UncertaintyModel,$ 
        $Assumption)$ ;
4   Append  $UncertaintyScore$  to
      $AssumptionUncertaintyList$ ;
5 end
6  $AverageAssumptionUncertainty \leftarrow$ 
    $Avg(AssumptionUncertaintyList)$ ;
7  $ExplanationUncertainty \leftarrow$ 
    $uc(UncertaintyModel,$ 
      $ExplanationSummary)$ ;
8  $OverallExplanationUncertainty \leftarrow$ 
    $AverageAssumptionUncertainty +$ 
    $ExplanationUncertainty$ ;
9 return
    $OverallExplanationUncertainty$ ;
```

---

**A.9 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.

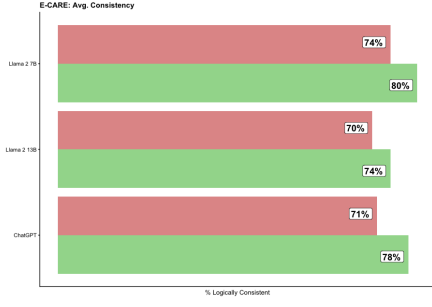


Figure 12: Average consistency comparison between correct and incorrect options for the E-CARE dataset.

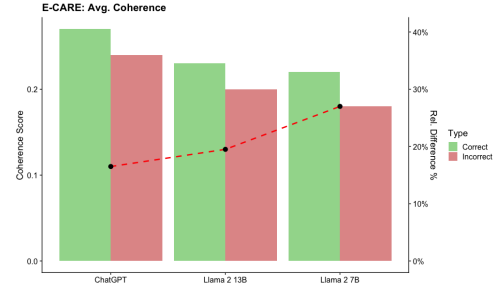


Figure 15: Comparison of average coherence scores between correct and incorrect options.

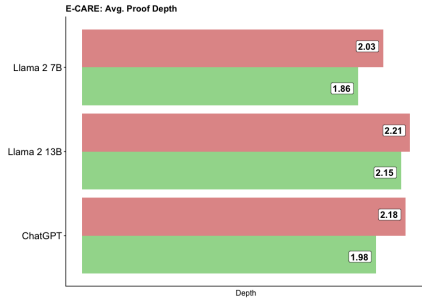


Figure 13: Comparison of average proof depth between correct and incorrect options.



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

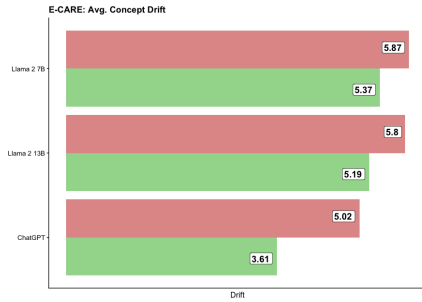


Figure 14: Comparison of average concept drift between correct and incorrect options.

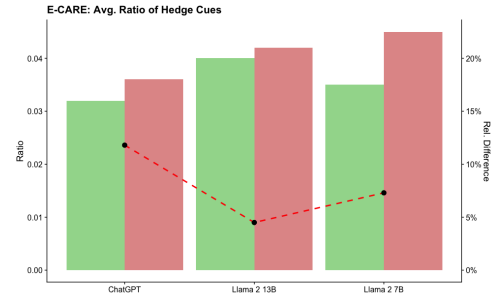


Figure 17: Comparison of the average ratio of hedge cues between correct and incorrect options.

## A.10 E-CARE Results

### A.10.1 E-CARE Consistency

See Figure 12.

### A.10.2 E-CARE Proof Depth

See Figure 13.

### A.10.3 E-CARE Concept Drift

See Figure 13.

### A.10.4 E-CARE Coherence

See Figure 15.

### A.10.5 E-CARE Uncertainty

See Figure 16.

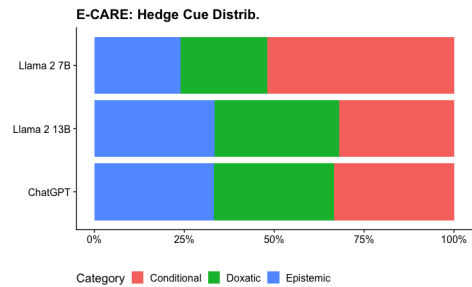


Figure 18: Distribution of hedge cues across incorrect explanations.

### A.10.6 E-CARE Hedge Ratio

See Figure 17.

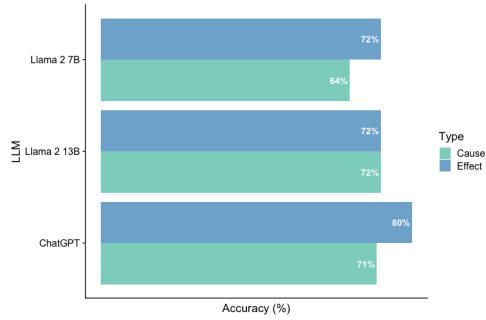


Figure 19: Accuracy in predicting the most plausible causes vs effects on COPA.

### A.10.7 E-CARE Hedge Distribution

See Figure 18.

### A.11 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 GPT and LLaMA 2 7B to be more accurate in predicting the effects in causal scenarios** (see Figure 19). 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 GPT and LLaMA 2 7B express greater linguistic uncertainty and produce more complex explanations when predicting causes rather than effects.

### A.12 Dataset Details

COPA is released under a BSD-2 license and made available for broad research usage with copyright notification restrictions<sup>4</sup>. 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<sup>5</sup>. We do not modify or use E-CARE outside of its intended use which is causal reasoning evaluation of language models.

<sup>4</sup>[people.ict.usc.edu/~gordon/copa.html](https://people.ict.usc.edu/~gordon/copa.html)

<sup>5</sup>[github.com/Waste-Wood/e-CARE?tab=MIT-1-ov-file#readme](https://github.com/Waste-Wood/e-CARE?tab=MIT-1-ov-file#readme)