Explanations *explained*. Influence of free-text explanations on LLMs and the role of implicit knowledge.

Anonymous ACL submission

Abstract

Despite their remarkable performance, LLMs' ability to provide transparent and faithful explanations for their predictions remains a challenge. We investigate the influence of different types of natural language explanations on LLM predictions, focusing on four different datasets presenting tasks that involve leveraging implicit knowledge. We conduct experiments on three SOTA LLMs on 8 types of explanations, both written by humans or machine-generated, through three generation methods: label-agnostic, label-aware, and counterfactual (label-contradicting) explanation generation. Our results consistently demonstrate that providing explanations significantly improves the accuracy of LLM predictions, even when the models are not explicitly trained to generate explanations, and propose a method to study the relationship between implicitness and explanation effectiveness.[1]

1 Introduction

Large Language Models (LLMs) excel at various natural language processing tasks, including text generation, translation, and question answering (Touvron et al., 2023; OpenAI, 2023). However, understanding their reasoning remains challenging, hindering trust and adoption in high-stakes domains (Hase et al., 2020; Kaneko and Okazaki, 2023; Kotonya and Toni, 2020; Atanasova et al., 2020). One approach is to train LLMs to generate explanations for their predictions. Existing methods, like pipeline models (Wiegreffe et al., 2020) and self-rationalizing models (Lei et al., 2016), often focus on extractive rationales suitable for information extraction tasks (Jacovi et al., 2021). However, complex reasoning tasks require free-text explanations, especially when implicit knowledge is involved (Wiegreffe et al., 2021).

[1] Code and data will be publicly released upon acceptance.

Generating explanations raises concerns about faithfulness, as LLMs might produce plausible-sounding explanations without genuine connection to their reasoning (Narang et al., 2020). This is particularly problematic for implicit knowledge, which relies on the model's internal representations of the world (McClelland et al., 2020).

This study investigates the impact of different natural language explanations on LLM predictions, focusing on the role of implicit knowledge. We analyze human-written and LLM-generated explanations across three experimental setups (label-aware, label-agnostic, and label-contradicting) (Sections 3 and 2) and four tasks requiring implicit knowledge (Section 4).

We hypothesize that the effectiveness of explanations, measured by downstream task performance, correlates with the degree of *implicitness*, i.e. novel, yet relevant, information they provide. Section 7 explores this hypothesis by examining the relationship between explanation effectiveness and metrics approximating novelty and relatedness.

The main contributions of this paper are the following:

- We categorize types of explanations and propose a methodology to test their impact on LLM predictions across tasks and languages.
- We demonstrate that providing explanations can boost prediction accuracy, even without explicit training.
- We propose a method to measure the correlation between explanation effectiveness and the conveyed implicit knowledge, presenting preliminary metrics and results.

2 Methodology

2.1 Problem definition

We address the problem of explaining the semantic relationship between two textual fragments, under the assumption that the relationship involves implicit or world knowledge, and the hypothesis that explanations eliciting more implicit knowledge represent higher quality explanations.

For our study we define an *explanatory task* in the following way. Consider a pair of sentences $< s_1, s_2 >$, and a semantic relation r holding between s_1 and s_2 (e.g., s_1 temporally precedes s_2, s_1 is caused by s_2, s_1 contradicts s_2, etc.). The task consists in a model M_1 generating an explanation e_i given the relation r, and then in a model M_2 that uses the explanation e_i to predict the relation r for the same sentence pair, when r is not given. Given this setting, the goal is to support the hypothesis that using explanations results in better predictions, and to investigate the correlation between explanation quality, implicit information elicitation, and relation prediction.

We consider different semantic relations, explanation types and generation modalities, as well as different large language models.

2.2 Explanatory pipeline

More in detail, to investigate explanation quality, we propose a three-step methodology, described in the following.

Step 1: explanation generation. First, given an explanatory task, we ask a model M_1 to generate a set of possible explanations E for the semantic relation r_c for the sentence pair $< s_1, s_2 >$. We assume ground truth relations R_c from human annotators, as they guarantee explanations are consistent with the actual semantic relations of the sentence pair.

$$M_1(s_1, s_2, r_c) \Rightarrow E$$

To keep under control our experimental setting, we assume that there is only one semantic relation r_c for a given sentence pair.

As we are interested in comparing different explanations $E = \{e_1, e_2, \dots e_n\}$ for the same sentence pair and the same relation r_c (e.g., a counterfactual explanation vs. a why-explanation) each explanation e_i is generated independently, prompting a generative model for each specific explanation type. In Section 3 we define in detail the set E of explanation types.

Step 2: model prediction. Here, model M_2 is asked to predict a semantic relation r_p between s_1 and s_2 given one individual explanation e_i in E, injected into the input along with the sentence pair. Adding one explanation e_i is meant to potentially add new information, implicit in s_1 and s_2, that can help the model M_2 to predict the correct relation r_c.

$$M_2(s_1, s_2, e_i) \Rightarrow r_p$$

The two models used in step 1 and step 2, M_1 and M_2, might be the same model, in which case the goal is to assess the self consistency of the model (generate the explanation and then use it for prediction), or two different models, in which case the goal is to have an independent assessment of the explanation quality. M_1 has to be a generative model, as it has to produce the set of explanations E, while the model M_2 is typically a classification model, or a generative one performing a classification task (as in our experimental setup).

Step 3: quality assessment. At this step we assess the quality of the explanations in E generated by M_1. Intuitively, the quality of an explanation e_i depends on its ability to provide useful content to solve a relation prediction task: the more e_i is useful to the model M_2 to predict the correct relation r_c, the better its *effectiveness*, taken as a proxy of the quality of e_i. Accordingly, here we assume that the M_2 performance is an indicator of the explanation effectiveness, such that better explanations are those that contribute to better prediction accuracy. Given an explanation e_i in the set E, its effectiveness relative to a model M_2 is given by the ability of the model to predict a relation r_p that approximates the correct relation r_c for a given sentence pair.

$$Effectivness(e_i, M_2) = r_p \approx r_c$$

Practically, the accuracy of the model M_2 on a relation prediction task is used as the main metric for explanation *effectiveness*. There are two interesting aspects to be considered. First, the delta between the relation prediction of the M_2 model without and with

e_i: this is an indicator of the absolute effectiveness of a certain explanation. Second, the relative ranking of all explanations in E given by the M_2 accuracy: this will give us a metric to assess if one explanation *type* is better (i.e., more effective) than another.

2.3 Measuring implcitness

While effectiveness is relative to a certain model, explanation type or relation, we want to explore whether better explanations are those that are able to introduce highly relevant implicit knowledge, i.e., not present in the sentence pair $< s_1, s_2 >$, that the M_2 model can use of for predicting r_p. Intuitively, a good explanation for an implicit knowledge-based relationship should maximize both its *novelty*, i.e., it has to bring new, implicit content with respect to $< s_1, s_2 >$, and its *relevance* with respect to $< s_1, s_2 >$, i.e. it has to be grounded to entities and events mentioned in the sentences.

As a preliminary step towards validating this hypothesis, we define the amount of implicitness of an explanation e_i as the combination of the *relevance* and the *novelty* of e_i with respect to a sentence pair $< s_1, s_2 >$.

$$Impl(s_1, s_2, e_i) = Rel(e_i, s_1, s_2) * Nov(e_i, s_1, s_2)$$

In Section 7, we propose some preliminary metrics to estimate these measures and assess them using implicitness as a direct evaluation measure for explanations assessed against effectiveness as computed in the first set of experiments.

3 Types of Explanations

In this section we present the types of explanations used by model M_2 with different characteristics (for a characterization of explanations in NLP, see (Jansen et al., 2016)). The explanations are free-text and can be generated either by a human or by a model M_1, so that E is representative both of how humans provide explanations in real contexts, and of the generative capacities and prompting techniques of current Large Language Models. To exemplify the various types, suppose the following working example:

s_1 = *The sky is cloudy today.*
s_2 = *I'll take an umbrella.*
$r_c = s_1$ *causes* s_2

Human explanations. These (called `human` in our experiments) are explanations directly generated or manually checked by humans, given the correct relation r_c and can virtually take any of the type described in the later sections. While the quality of human generated explanations can be considered high (e.g., we expect that they point out relevant and implicit information), there is no guarantee that, when used by a model M_2, they perform better than model generated explanations. For the purposes of this paper, we carefully select datasets that provide reference human-generated or human-edited explanations.

Why explanations. This kind of explanation (`why`) is the most typical way to provide an explanation, i.e., as an answer to a why question (). In our setting, a why explanation is an answer to *Why is r_c the relation holding between s_1 and s_2?*. Then, a common why explanation would be:

> *Cloudy skies indicate it might be raining, and the umbrella prevents one from getting wet.*

Why-not explanations. This type of explanation (`why-not`) argues that the alternative relation(s) *cannot* hold as correct between s_1 and s_1. The rationale is based on *reasoning by exclusion*, a common strategy in argumentation. In our setting, where we have binary or three-way relations, a why-not explanation is a why explanation for the relation $\neg r_c$: *Why is $\neg r_c$ a relation not holding between s_1 and s_2?*. Suppose the same example as above. Then, a common why-not explanation would be:

> *Cloudy skies indicate it might be raining, and not taking an umbrella could result in getting wet, which is undesirable.*

Example-based explanation. This kind of explanation (`ex-exp`) asks for a supplementary or equivalent example, or a specific instance of s_1 and s_2 holding relation r_c. The rationale is that making examples is considered a useful communicative instrument to improve understanding (Kim et al., 2016). In our setting, an example-based explanation would be

Provide a supplementary, equivalent or specific example of s_3 and s_4 where the relation r_c still holds.. For our case, a common example-based explanation would be:

> *Yesterday it was raining in Rome and I went to work with an umbrella to not get wet.*

Self-rationalizing explanations. This kind of explanation does not assume knowledge of r_c, and asks to either (i) explain the reasoning then predict r_c (`pre-hoc`), or (ii) first predict r_c then explain the prediction (`post-hoc`). These explanations are respectively inspired by "explain-then-predict" strategies in NLP, using techniques such as as chain-of-thought in-context learning (Wei et al., 2022), and "predict-then-explain" strategy , using post-hoc self-rationalisations(Lei et al., 2016).

Counterfactual explanations. This kind of explanation, in its classical formulation, asks for what (minimal) changes are needed to be made on s_1 and s_2 in order to falsify the relation r_c. Then, in a counterfactual situation, the negation of a binary relation r_c holds between the modified s_1 and s_2. The rationale for a counterfactual explanation is that forcing changes on s_1 and s_2 forces to change r_c into $\neg r_c$ (Wachter et al., 2017; Verma et al., 2022). In our setting, a counterfactual (`c-factual`) explanation originates from the following question: *What are the conditions in which relation r_c may not hold for s_1 and s_2?*. Let's use again our example, for which a common counterfactual explanation would be:

> *If I were deciding whether to take an umbrella for a trip to the desert, a cloudy sky would not be a reason to take an umbrella in my backpack.*

For the sake of our experiments, we also consider a more shallow interpretation of counterfactual (`not(why-exp)`) that simply represents the falsification of a why question.

> *Cloudy skies indicate it might be raining, but an umbrella does not prevent one from getting wet.*

4 Experiments on explanation effectiveness

4.1 Models

For Step 1, explanation generation, we used GPT-3.5, a proprietary large language model from OpenAI (OpenAI, 2023), known for its high performance in text generation and reasoning tasks. For Step 2, model prediction, as M_2 we use another instance of GPT-3.5, to assess the effect of generated explanations on the same model,and two different models to which we apply Step 1's output: Llama-2 13B (Touvron et al., 2023), a large language model from Meta AI, distinguished by its open-source nature and wide training data, and Mixtral 8x7B (Jiang et al., 2024), a recently released open-source mixture-of-expert architecture from Mistral AI, notable for its strong performance on benchmarks while being smaller than other competing models.

4.2 Datasets

We use 4 datasets that propose tasks involving different kinds of reasoning and eliciting implicit or external knowledge to different extents. All the datasets provide either human-generated or human-collected and curated explanations explanations, which we use as the `human` explanation type.

- **e-RTE-3-it** (Zaninello et al., 2023): a dataset in Italian for Recognizing Textual Entailment (RTE), featuring pairs of texts-hypotheses and human-written explanations for the entailment relation. The dataset consists of 1,600 sentence pairs and is annotated for three entailment classes: entailment (YES), contradiction (NO), and neutrality (UNKNOWN).

- **e-SNLI** (Camburu et al., 2018), a version of the Stanford Natural Language Inference (SNLI) corpus enriched with human-written natural language explanations. The dataset includes 570k sentence pairs labeled for the same three entailment classes as e-RTE-3-ITA.

- **e-CARE** (Du et al., 2022): a dataset focused on causal reasoning questions, featuring human-annotated explanations for the causal questions, The dataset consists of 21k causal reasoning questions with both correct and incorrect answers. We accommodate this dataset into our experimental setup by pairing each question (s_1) with either the correct answer (s_1, label: YES) or the incorrect answer (s_1, label: NO).

- **(e-)StrategyQA** (Geva et al., 2021): A question-answering dataset designed to require multiple steps strategic reasoning or implicit knowledge to answer. The dataset comprises 2,780 strategy question (which we use as s_2) with answer "YES" or "NO" (labels), its decomposition into multi-step reasoning paths (which we use as explanation) and evidence paragraphs giving the context of the question (which we use as s_2).

4.3 Generation and inference setups

In this section we describe how the explanation types presented in Section 3 can practically be produced and introduced in our explanatory and prediction pipelines (Section 2.2). We then present our experimental setup and results grouping explanations by whether they are produced by assuming either 1. knowledge of the correct relation label marked as correct (4.5); 2. no knowledge of the correct relation label (4.6); 3. knowledge of the correct label, marked as incorrect (4.7).

To ensure that the explanations do not simply suggest the right answer but are not informative, we "anonimize" them by substituting each explicit reference to the labels or other obvious suggestions with a placeholder. To include the explanations in Step 2, we prompt M_2 to use a "hint" to give its answer, represented by the explanation[2].

4.4 Baseline Generation

We use two baselines in our experiments: **no-explanation** (`no-exp`), where the model M_2 performs 0-shot relation r_p prediction; **dummy explanation** (`dummy`), where we use a copy of s_2 as the explanation, to ensure virtually zero new information given, and that results may not be due simply to data augmentation/larger contexts.

4.5 Relation Aware Explanations

In this setup, we are assuming a *relation aware* approach, where the generation process is driven by the correct relation r_c holding between s_1 and s_2. In this setup we include both human generated (`human`) and model generated explanations (`why`, `why-not`, `ex-exp`) (see Section 3). To generate the explanations, we prompted GPT-3.5 differently for each explanation type, providing it with the golden label during explanation generation. We prompt the model to return some structure in the output, and parse it with regular expressions to collect the explanations. Similarly, we parse the output of M_2 with regular expressions to extract the label, and resolve manually conflicting cases. Results for this setup, as the accuracy on test sets, are reported in Table 1.

4.6 Relation Agnostic Explanations

In Section 4.5 we have assumed that most explanations are generated knowing the correct relation r_c holding between s_1 and s_2, i.e., referred as *relation-aware*. However, we are also interested in experimenting on a *relation-agnostic*, self-supervised generation, where a model M_1 generates an explanation while contextually being asked to predict the relation. We call this modality *label agnostic generation*, which makes use of the `pre-hoc` and `post-hoc` explanation types.

In Table 2 we report accuracy for this setup. For the sake of comparison, we also report (in brackets) the results that consider the label that M_1 contextually outputted in Step 1. Note that, being predicted contextually with the explanation generation in Step 1, the relation r_p explained can be either correct or wrong, with potential error propagation in Step 2.

4.7 Relation Contradicting Explanations

In this final setup, we use counterfactual explanations, i.e. explanations that are explicitly contradicting the golden label `c-factual` or that are falsifying the explanation for a correct label $\neg$`(why-exp)` to test the robustness of models to potentially false or misleading information, as well as highlight how different model may be differently sensitive to explanation injection. For this setup, we also report accuracy, but we interpret higher accuracy as an indicator of less effectiveness of this special type of explanations (Table 3).

5 Results and Discussion

This study reveals that explanations, even without explicit training, enhance LLMs' semantic relation

[2]All the code, prompts and the data will be made publicly available in the camera-ready version.

prediction accuracy, as shown in Tables 1, 2, and 3. Across various models, datasets, and explanation types, label-aware explanations consistently yield the greatest improvement, while even relation-agnostic explanations surpass baseline performance.

Label-Aware Explanations significantly boost LLM accuracy. Models with access to explanations, particularly "why" explanations, perform best, demonstrating the utility of providing detailed reasoning steps. "Why-not" explanations also effectively refine decision-making processes, typically ranking second in performance.

Relation-Agnostic Explanations enhance accuracy even without targeting specific relations, underscoring the value of generic explanations. Pre-hoc explanations (generated before predictions) tend to outperform post-hoc ones (generated afterward). The accuracy of these explanations varies with the contextually predicted label, emphasizing the risk of error propagation from incorrect predictions.

Relation-Contradicting Explanations show that LLMs struggle with misleading information, as seen with lower performance from "c-factual" and "¬(why-exp)" explanations compared to baselines. This indicates the need for accuracy and validation in explanation content to aid LLMs effectively.

Model sensitivity to explanation types varies; for instance, GPT-3.5 excels with "why" explanations, while Llama 13b prefers "why-not." Dataset characteristics also influence explanation effectiveness, with StrategyQA showing lower gains compared to e-CARE, highlighting the impact of the complexity and type of reasoning required.

In summary, explanations significantly enhance LLM performance, though their effectiveness varies with the explanation type, model architecture, and dataset complexity. Further research is essential to optimize explanation use and improve LLM reasoning capabilities.

6 Related work

Explainable AI (XAI) aims to make complex models more understandable, with various ***types of explanations*** contributing to this goal. The concept of "explanation" has been interpreted differently in XAI literature, each serving distinct purposes and applying to different aspects of model interpretation. A comprehensive review of these techniques can be found in Molnar et al. (2020). **Local explanations** focus on providing insights into the decision-making process for individual predictions, with techniques like LIME (Ribeiro et al., 2016) and SHAP (Lundberg and Lee, 2017) being prominent examples. **Feature importance** explanations aim to identify which features are most influential in a model's decision-making process, while global explanations seek to convey an understanding of the model's behavior across all predictions. Friedman (2001) provides significant contributions to understanding ensemble models' global behavior. **Counterfactual explanations** offer a different perspective by illustrating how changes in the input text could alter the prediction (Wachter et al., 2017; Verma et al., 2022; Tolkachev et al., 2022). Example-based explanations utilize specific instances from the dataset to explain how the model behaves under certain conditions (Kim et al., 2016). **Attention** mechanisms, since the introduction of the Transformer model (Vaswani et al., 2017), have been utilized for model interpretation. However, there is debate on whether attention can effectively serve as a proxy for explanations, with some arguing for its limitations (Jain and Wallace, 2019), while others challenge this claim (Wiegreffe and Pinter, 2019). An overview of this topic can be found in Bibal et al. (2022). **Causal inference** methods, as detailed by Pearl (2009), offer a deeper level of explanation by understanding the causal relationships within the data that the model leverages for predictions.

The ***role of explanations in NLP models*** has been explored by various researchers. Paranjape et al. (2021) focuses on template-based contrastive explanations, while our work delves into different types of explanations and their connection to implicit knowledge in language models. Lampinen et al. (2022) and Ye and Durrett (2022) demonstrate the benefits of in-context explanations for large models in challenging reasoning tasks. Similar to our approach, Pruthi et al. (2022) measure explanation quality based on downstream performance. Their methodology involves training a student model on explanations generated by a teacher, resembling our generation-and-evaluation setup. However, they utilize automatic explanation generation techniques and train the student for the end task. Finally, Cambria et al. (2023)

LABEL-AWARE EXPLANATIONS						
MODEL	**no-exp**	**dummy**	**human**	**why**	**why-not**	**ex-exp**
e-RTE-3-ITA						
GPT-3.5	65	57	69	**80**	75	72
LLama 13b	57	45	75	**91**	84	82
Mistral 8x7b	76	64	88	**90**	85	86
e-SNLI						
GPT-3.5	65	64	69	**88**	86	**88**
LLama 13b	53	44	75	84	79	87
Mistral 8x7b	74	69	89	87	84	**93**
e-CARE						
GPT-3.5	52	65	81	**90**	88	**93**
LLama 13b	30	48	62	**95**	89	90
Mistral 8x7b	62	63	77	91	78	**93**
e-Strategy-QA						
GPT-3.5	45	47	50	71	**72**	44
LLama 13b	45	26	57	**74**	64	68
Mistral 8x7b	39	43	**57**	49	44	41

Table 1: Accuracy of LLMs on test sets of the selected datasets with label-aware explanations. We boldface the best scoring type of explanation for each model.

provides a comprehensive survey of approaches for generating natural language explanations, while Hartmann and Sonntag (2022) examines the benefits of explanations for NLP models.

7 Experiments on measuring implicitness

In Section 7 we have defined the amount of implicitness of an explanation e_i as the combination of *relatedness* and *novelty* of e_i with respect to a sentence pair $< s_1, s_2 >$.

This set of experiments aims to propose a preliminary study to quantify the degree of implicit information brought up by the explanation, and how it correlates with explanation effectiveness.

We define two simple metrics to capture different degrees of explanation relatedness and one for novelty, measuring implicitness as the product of relatedness and novelty.

7.1 Relatedness

Semantic Similarity. We leverage cosine similarity between the sentence embeddings of the combined text-hypothesis pair and the explanation. Given an input sentence s, the model outputs a fixed-dimensional vector $\mathbf{e}_s$ representing its contextualized embedding. The *sentence-transformers/all-mpnet-base-v2* model (Reimers and Gurevych, 2019) was used to generate semantically rich sentence representations.

Entailment. We use a pre-trained NLI model to determine the degree to which the explanation is implied by the combined text-hypothesis pair. A sigmoid function was applied to the entailment score p_{ent} output by the NLI model. Higher scores indicate stronger entailment relation between combined text-hypothesis pairs (t and h, respectively) and their corresponding explanations (e), suggesting that the explanation is likely to be related to the input. For calculations, we use the *roberta-large-mnli* model (Liu et al., 2019), fine-tuned on the Multi-Genre NLI dataset (Williams et al., 2018).

7.2 Novelty

Novelty: This metric, inspired by classical work on surprisal in information theory (Shannon, 1948), captures the unexpectedness of words in the explanation given the combined text-hypothesis context. We calculated the average word surprisal of an explanation as:

$$\text{Novelty}(t, h, e) = \frac{1}{|E_s|} \sum_{w \in E_s} - \log P(w|t, h) \quad (1)$$

LABEL-AGNOSTIC EXPLANATIONS				
MODEL	**no-exp**	**dummy**	**pre-hoc**	**post-hoc**
e-RTE-3-ITA				
GPT-3.5	65	57	55 (56)	55 (63)
LLama 13b	57	45	61	60
Mistral 8x7b	76	64	66	66
e-SNLI				
GPT-3.5	65	63	64	66
LLama 13b	53	44	58	65
Mistral 8x7b	73	69	74	71
e-CARE				
GPT-3.5	51	64	67	**69**
LLama 13b	29	48	62	**63**
Mistral 8x7b	61	63	63	**65**
Strategy-QA				
GPT-3.5	45	46	**48**	47
LLama 13b	44	26	45	**46**
Mistral 8x7b	39	43	42	**43**

Table 2: Accuracy on test sets for the setup using label agnostic explanations.

LABEL-CONTRADICTING EXPLANATIONS				
MODEL	**no-exp**	**dummy**	**c-factual**	**¬(why-exp)**
e-RTE-3-ITA				
GPT-3.5	65	57	15	30
LLama 13b	57	45	18	10
Mistral 8x7b	76	64	36	33
e-SNLI				
GPT-3.5	65	64	28	42
LLama 13b	53	44	13	12
Mistral 8x7b	74	69	42	52
e-CARE				
GPT-3.5	52	65	6	27
LLama 13b	30	48	1	16
Mistral 8x7b	62	63	4	26
Strategy-QA				
GPT-3.5	45	47	37	37
LLama 13b	45	26	37	27
Mistral 8x7b	39	43	39	31

Table 3: Accuracy on test sets of the tested models for the label-contradicting explanations.

where $P(w|t,h)$ is the probability of a word w in the explanation to occur in the input, estimated using the word frequencies in the combined text-hypothesis context. We define an empirical smoothing parameter $alpha = 0.1$ as the frequency of words non occurring in the input.

7.3 Preliminary results

The analysis of the correlation among implicitness measures and the prediction outcomes in the datasets highlights some common trends, which we report in detail for the Mixtral model results on the e-CARE dataset (Table 4).

The correlation coefficient between similarity and prediction and entailment and prediction are moderately strong ($r = 0.53$, $r = 0.49$), indicating that higher relatedness often correlates with a higher likelihood of a correct prediction. Novelty alone exhibits a negative correlation with prediction ($r = 0.36$), indicating that higher novelty often may lead to incorrect predictions.

However, considering feature interaction, the interaction between similarity and novelty shows a positive correlation with predictions (r=0.55), suggesting that the interaction between the two has a potential predictive power that needs to be further investigated. The interaction of entailment with novelty correlates positively with prediction (r=0.51), confirming the potential influence of implicitness in the prediction phase. These findings encourage us to further explore the dimension of implicitness in explanations.

8 Conclusion

In this study, we tested the effects of explanations on LLMs, showing that they can significantly improve their accuracy in predicting relations between sentences. This improvement is consistent across different models, datasets, and explanation types. Our experiments also show a correlation between explanation effectiveness and the degree of implicit knowledge conveyed by the explanations, suggesting that explanations that introduce novel and relevant information are more likely to be helpful to LLMs. Furthermore, our analysis reveals that different LLMs exhibit varying sensitivity to different explanation types. Our findings contribute to research on the role of explanations in enhancing LLM performance. By understanding the nuances of model sensitivity to different explanation types and the ways in which explanations contribute to implicit knowledge acquisition, we can develop more effective techniques for explaining and improving the reasoning capabilities of LLMs.

Limitations

This study has several limitations that should be considered.

Limited Scope: We focus on a specific type of NLP task involving implicit knowledge and investigate the impact of explanations on relation prediction. Further research is needed to extend these findings to a broader range of NLP tasks and model architectures.

Artificial Setting: We utilize a controlled experimental setup, where explanations are provided in a specific format and injected into the model during inference. Real-world applications might involve more complex scenarios with less controlled input and output formats.

Simplification of Implicitness: Our measurement of implicitness relies on basic metrics like cosine similarity and novelty, which may not fully capture the nuanced nature of implicit knowledge in language. More sophisticated techniques are needed for a comprehensive evaluation of implicitness. Data Dependence: Our results are based on specific datasets with curated explanations. Further exploration with different datasets is required to assess the generalizability of our findings.

References

Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, and Isabelle Augenstein. 2020. Generating fact checking explanations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7352–7364, Online. Association for Computational Linguistics.

Adrien Bibal, Rémi Cardon, David Alfter, Rodrigo Wilkens, Xiaoou Wang, Thomas François, and Patrick Watrin. 2022. Is attention explanation? an introduction to the debate. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3889–3900, Dublin, Ireland. Association for Computational Linguistics.

Erik Cambria, Lorenzo Malandri, Fabio Mercorio, Mario Mezzanzanica, and Navid Nobani. 2023. A survey on xai and natural language explanations. *Information Processing Management*, 60(1):103111.

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

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.

Jerome H. Friedman. 2001. Greedy function approximation: A gradient boosting machine. *Annals of statistics*, pages 1189–1232.

Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. 2021. Did aristotle use a laptop? a question answering benchmark with implicit reasoning strategies. *Transactions of the Association for Computational Linguistics*, 9:346–361.

Mareike Hartmann and Daniel Sonntag. 2022. A survey on improving NLP models with human explanations. In *Proceedings of the First Workshop on Learning with Natural Language Supervision*, pages 40–47, Dublin, Ireland. Association for Computational Linguistics.

Peter Hase, Shiyue Zhang, Harry Xie, and Mohit Bansal. 2020. Leakage-adjusted simulatability: Can models generate non-trivial explanations of their behavior in natural language? In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4351–4367, Online. Association for Computational Linguistics.

Alon Jacovi, Swabha Swayamdipta, Shauli Ravfogel, Yanai Elazar, Yejin Choi, and Yoav Goldberg. 2021. Contrastive explanations for model interpretability. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1597–1611, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Sarthak Jain and Byron C. Wallace. 2019. Attention is not Explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3543–3556, Minneapolis, Minnesota. Association for Computational Linguistics.

Peter Alexander Jansen, Niranjan Balasubramanian, Mihai Surdeanu, and Peter Clark. 2016. What's in an explanation? characterizing knowledge and inference requirements for elementary science exams. In *International Conference on Computational Linguistics*.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2024. Mixtral of experts.

Masahiro Kaneko and Naoaki Okazaki. 2023. Controlled generation with prompt insertion for natural language explanations in grammatical error correction.

Been Kim, Rajiv Khanna, and Oluwasanmi O. Koyejo. 2016. Examples are not enough, learn to criticize! criticism for interpretability. In *Advances in Neural Information Processing Systems*, volume 29.

Neema Kotonya and Francesca Toni. 2020. Explainable automated fact-checking for public health claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7740–7754, Online. Association for Computational Linguistics.

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, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Tao Lei, Regina Barzilay, and T. Jaakkola. 2016. Rationalizing neural predictions. *ArXiv*, abs/1606.04155.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Scott M. Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Advances in neural information processing systems*, volume 30.

James L. McClelland, Felix Hill, Maja Rudolph, Jason Baldridge, and Hinrich Schütze. 2020. Placing language in an integrated understanding system: Next steps toward human-level performance in neural language models. *Proceedings of the National Academy of Sciences*, 117(42):25966–25974.

Christoph Molnar, Giuseppe Casalicchio, and Bernd Bischl. 2020. *Interpretable Machine Learning – A Brief History, State-of-the-Art and Challenges*, pages 417–431.

Sharan Narang, Colin Raffel, Katherine Lee, Adam Roberts, Noah Fiedel, and Karishma Malkan. 2020. Wt5?! training text-to-text models to explain their predictions.

OpenAI. 2023. Gpt-4 technical report.

Bhargavi Paranjape, Julian Michael, Marjan Ghazvininejad, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2021. Prompting contrastive explanations for commonsense reasoning tasks. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4179–4192, Online. Association for Computational Linguistics.

Judea Pearl. 2009. *Causality*. Cambridge university press.

Danish Pruthi, Rachit Bansal, Bhuwan Dhingra, Livio Baldini Soares, Michael Collins, Zachary C. Lipton, Graham Neubig, and William W. Cohen. 2022. Evaluating explanations: How much do explanations from the teacher aid students? *Transactions of the Association for Computational Linguistics*, 10:359–375.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, pages 3997–4007.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. Why should i trust you? explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144.

Claude E Shannon. 1948. A mathematical theory of communication. *Bell System Technical Journal*, 27(3):379–423.

George Tolkachev, Stephen Mell, Stephan Zdancewic, and Osbert Bastani. 2022. Counterfactual explanations for natural language interfaces. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 113–118, Dublin, Ireland. Association for Computational Linguistics.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models.

Ashish Vaswani et al. 2017. Attention is all you need. In *Advances in neural information processing systems*, volume 30.

Sahil Verma, Varich Boonsanong, Minh Hoang, Keegan E. Hines, John P. Dickerson, and Chirag Shah. 2022. Counterfactual explanations and algorithmic recourses for machine learning: A review.

Sandra Wachter, Brent Mittelstadt, and Chris Russell. 2017. Counterfactual explanations without opening the black box: Automated decisions and the gdpr. *Harvard Journal of Law & Technology*, 31(2).

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed H. Chi, Quoc V Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*.

Sarah Wiegreffe, Ana Marasović, and Noah A. Smith. 2020. Measuring association between labels and free-text rationales. In *Conference on Empirical Methods in Natural Language Processing*.

Sarah Wiegreffe, Ana Marasović, and Noah A. Smith. 2021. Measuring association between labels and free-text rationales. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10266–10284, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Sarah Wiegreffe and Yuval Pinter. 2019. Attention is not not explanation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 11–20, Hong Kong, China. Association for Computational Linguistics.

Adina Williams, Nikita Nangia, Samuel R. Bowman, Martin Abadi, and Antoine Bordes. 2018. A broad-coverage challenge corpus for sentence understanding through inference. *Transactions of the Association for Computational Linguistics*, 6:309–324.

Xi Ye and Greg Durrett. 2022. The unreliability of explanations in few-shot prompting for textual reasoning. In *Advances in Neural Information Processing Systems*, volume 35, pages 30378–30392. Curran Associates, Inc.

Andrea Zaninello, Sofia Brenna, and Bernardo Magnini. 2023. Textual entailment with natural language explanations: The italian e-rte-3 dataset.

A Appendix

Table 4: Correlation Matrix for Features from the e-CARE dataset based on the Mistral label-aware predictions.

Feature	Similarity	Entailment	Novelty	Sim x Nov	Ent x Nov	Prediction
Similarity	1.00	0.45	-0.20	0.86	0.42	0.53
Entailment	0.45	1.00	-0.25	0.40	0.95	0.49
Novelty	-0.20	-0.25	1.00	-0.18	-0.22	-0.36
Sim x Nov	0.86	0.40	-0.18	1.00	0.38	0.55
Ent x Nov	0.42	0.95	-0.22	0.38	1.00	0.51
Prediction	0.53	0.49	-0.36	0.55	0.51	1.00