Explanations *explained*. Influence of Free-text Explanations on LLMs and the Role of Implicit Knowledge

Andrea Zaninello[1,2], Roberto Dessì[3], Malvina Nissim[4], Bernardo Magnini[2]

[1]Free University of Bolzano, Italy

[2]Fondazione Bruno Kessler, Italy

[3]Not Diamond, San Francisco, USA

[4]University of Groningen, Netherlands

azaninello@fbk.eu, m.nissim@rug.nl, magnini@fbk.eu

Abstract

In this work, we investigate the relationship between the quality of explanations produced by different models and the amount of implicit knowledge they are able to provide beyond the input. We approximate explanation quality through accuracy on a downstream task with a standardized pipeline (GEISER) and study its correlation with three different association measures, each capturing different aspects of implicitness, defined as a combination of relevance and novelty. We conduct experiments with three SOTA LLMs on four tasks involving implicit knowledge, with explanations either confirming or contradicting the correct label. Our results demonstrate that providing quality explanations consistently improves the accuracy of LLM predictions, even when the models are not explicitly trained to take explanations as input, and underline the correlation between implicit content delivered by the explanation and its effectiveness.[1]

1 Introduction

Large Language Models (LLMs) excel at numerous language processing tasks, including text generation, translation, and question answering (Touvron et al., 2023; OpenAI, 2023). Still, understanding their reasoning is 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 towards "intrinsic explainability" is to have LLMs 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 (Jacovi et al., 2021). However, complex reasoning tasks require free-text explanations, especially when implicit knowledge is involved (Wiegreffe et al., 2021). Also, generating explanations raises concerns about their faithfulness, as LLMs might produce plausible-sounding explanations with no 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).

With the rise of retrieval-augmented generation (RAG, Lewis et al. (2020)), language models are increasingly supplemented with external information, such as explanations, retrieved from knowledge bases or provided via in-context learning (ICL). The effectiveness of these approaches depends on the quality of the retrieved or injected text, which serves as additional context for the model's reasoning. While traditional RAG studies focus on improving retrieval mechanisms (e.g., optimizing factual correctness), less attention has been paid to evaluating the quality of explanations used in these frameworks. Recent work by He et al. (2024) shows that augmenting ICL with natural language explanations (NLEs) improves model robustness. However, their study focuses on performance benefits rather than the quality of different explanation types, and their evaluation is limited to downstream accuracy without assessing what makes an explanation effective in guiding a model's decision.

Our work addresses this gap by providing a principled evaluation of explanation quality, particularly in sentence pair reasoning tasks, measured by downstream task performance. Moreover, we show that explanation effectiveness correlates with the degree of *implicit content*, i.e., novel yet relevant information they provide. We test this hypothesis by examining the relationship between explanation effectiveness and three metrics approximating novelty and relevance, and show that they have high, yet different correlation with explanation quality according to the examined task.

The main contributions of this paper are:

[1]Code and data available here github.com/andreazaninello/geiser.

*Proceedings of the 14th Joint Conference on Lexical and Computational Semantics (*SEM 2025)*, pages 212–224

- we propose GEISER, a standardized pipeline to evaluate the effectiveness of different types of explanations using LLM relation predictions on tasks involving varying degrees of implicit reasoning and external knowledge;

- using the proposed pipeline, we report experimental results on different kinds of explanations (human- and machine-generated), across three LLMs, four tasks and two languages;

- through our analysis, we introduce "implicit knowledge" as a key factor of explanation quality, and study different metrics to estimate it, showing its correlation with explanation effectiveness.

2 Related Work

The role of explanations in NLP has been extensively studied. For instance, Cambria et al. (2023) provide a comprehensive survey of natural language **explanation generation** approaches, and Hartmann and Sonntag (2022) examine the benefits of explanations for improving NLP models. Paranjape et al. (2021) focus on template-based explanations, while Lampinen et al. (2022) and Ye and Durrett (2022) highlight the advantages of in-context explanations for complex reasoning tasks. Jansen et al. (2016) provide a comprehensive characterization of different kinds of explanations, each one with different insight into model behavior.

Traditionally, **explanation quality** has been assessed using automated metrics like BLEU (Papineni et al., 2002), ROUGE (ROUGE, 2004), or BERT-Score (Zhang et al., 2019), which compare outputs to human-written references. However, these metrics may not fully capture explanation quality or align with human judgment, and collecting human references is often costly. More recently, (human) *simulatability scores* have emerged as an alternative to overlap metrics, based on the idea that explanation quality can be defined as the "utility to an end-user" (Kim et al., 2016). This approach evaluates how explanations improve predictive performance on downstream tasks rather than overlap with ground truth explanations and, while humans were initially the predictors (Wiegreffe et al., 2021), trained models now automate this process, showing strong correlations with human judgments (Hase et al., 2020). For example, Pruthi et al. (2022) measures explanation quality by training a student model on teacher-generated explanations for downstream tasks.

Prior work has largely focused on eliciting explanations from models or evaluating them based on task performance, our work shifts the focus toward understanding how explanations can reveal implicit knowledge, offering a novel perspective on explanation quality assessment. While, to the best of our knowledge, there are no previous works addressing **implicit content measures** directly, in the context of information retrieval, *relevance* and *novelty* have been recognized as key aspects of *novelty detection tasks* (Ghosal et al., 2022, 2018), and similarly to us exploit Textual Entailment (Bentivogli et al., 2011) for sentence level novelty mining. Metrics such as the cosine similarity between high-dimensional embeddings has been traditionally used to quantify semantic similarity of texts, but has also been recently questioned as a faithful representation (Steck et al., 2024). Other works, on the other hand, have focused on estimating the *causal strength* between textual fragments, and proposed learned metrics such as CEQ (Du et al., 2022) or CESAR (Cui et al., 2024), in the attempt to improve more simplistic yet effective metrics such as Pointwise Mutual Information (PMI).

3 Methodology

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

3.1 Explanatory task

Given a pair of sentences $< s_1, s_2 >$, and a semantic relation r 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 for the relation r and then in a model M_2 using the explanation e_i to predict the relation r for the same sentence pair, when r is not given. The goal is to support the hypothesis that using explanations results in better predictions, and that an increase in prediction accuracy corresponds to higher explanation effectiveness, as well as investigate the correlation between explanation quality, implicit information elicitation, and relation prediction.

3.2 The GEISER Pipeline

To estimate the quality of the explanations, we propose GEISER (Generation and evaluation of Explanations for Implicit SEmantic Relations) a three-step methodology inspired by work on human simulatability scores.

Step 1: Generate Explanations with M1. 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 consistent with the actual semantic relations of the sentence pair.

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

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 5 we define in detail the set E of explanation types.[2]

Step 2: Predict Relation with M2. 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 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 must be a generative model, as it has to produce the set of explanations E, while M_2 is a generative model performing a classification task.

Step 3: Evaluate M1's Explanations through M2's performance. Our final goal is to 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.

$$\textit{Effectiveness}(e_i, M_2) = r_p \approx r_c$$

In practice, overall accuracy of a model M_2 on a relation prediction task is used as a proxy metric for explanation *effectiveness*. There are two interesting aspects to be considered. First, the difference 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, as an aggregation metric, the relative ranking of all explanations in $E_t \in E$ given by the M_2 accuracy according to their type and how they were generated: this will give us an indication of whether an explanation type or a generative model is better (i.e., more effective) than another.

3.3 Measuring Implicit Content via Explanation–Input Association Measures

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 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 (Ghosal et al., 2018).

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

We operationalise the implicit content calculations comparing three different association measures between the input sentences $< s_1, s_2 >$ and the explanation e_i: *Causal Strength (CS)*, *Entailment Probability*, and *Cosine Similarity*. These are

[2] To keep under control our experimental setting, we assume only one semantic relation r_c for a given sentence pair.

intended to variously reflect how well an explanation relates to the input sentences s_1 and s_2, while bringing new, potentially useful information.

Causal Strength

Firstly, we consider the metric of CAUSAL STRENGTH (CS) as proposed by Cui et al. (2024), which—in its original formulation—aggregates token-level associations between a *cause* sequence C (length n) and an *effect* sequence E (length m). However, following Du et al. (2022), in our setting the cause sequence C is obtained by taking the maximum value obtained by calculating the causal strength either between a. the concatenation of s_1 and the explanation and the effect sequence E (corresponding to s_2) or 2. s_1 as C and the concatenation of the explanation and s_2 as E. The method uses causal token embeddings from a BERT model pre-trained on a cause–effect corpus, and attention weights to focus on the most relevant token pairs.

Formally, the score is defined as:

$$\mathrm{CS}(C, E) \;=\; \sum_{i=1}^{n}\sum_{j=1}^{m} a_{ij} \left| \frac{\mathbf{c}_i^\top \mathbf{e}_j}{\|\mathbf{c}_i\|\|\mathbf{e}_j\|} \right|$$

where $\mathbf{c}_i$ and $\mathbf{e}_j$ are the token embeddings of C and E, and a_{ij} are normalized attention weights over token pairs ($\sum_{i,j} a_{ij} = 1$). This attention-weighted absolute cosine similarity captures both semantic alignment and token-level causal informativeness.

Intuitively, this metric emphasizes token pairs that are both semantically aligned (via cosine similarity) and deemed important by the attention mechanism—highlighting explanatory tokens that are both novel and causally relevant. This implementation, known as CESAR (Cui et al., 2024), yields more robust predictions of causal strength changes when additional information (e.g., supporters or defeaters) is introduced.

Entailment Probability

Secondly, we consider the probability of entailment, also referred to as NATURAL LANGUAGE INFERENCE (NLI), calculated via a pre-trained NLI model. Given an explanation exp, an input statement s_1, and a target statement s_2, we define:

$$\mathrm{NLI}(\mathrm{exp}, s_1, s_2) \;=\; P_{\mathrm{NLI}}\big(s_1 \wedge \mathrm{exp} \models s_2\big).$$

This directional measure, while theoretically capturing both relevance and novelty, may in practice favour relatedness over new information, and its reliability is limited by the accuracy of the underlying NLI model.

Cosine Similarity

Thirdly, we consider simply comparing the COSINE SIMILARITY (COS) of the embedding vectors $\mathbf{e}_{\mathrm{exp}}$ and $\mathbf{e}_{s_1,s_2}$:

$$\mathrm{Cos}(\mathbf{e}_{\mathrm{exp}}, \mathbf{e}_{s_1,s_2}) \;=\; \frac{\mathbf{e}_{\mathrm{exp}} \cdot \mathbf{e}_{s_1,s_2}}{\|\mathbf{e}_{\mathrm{exp}}\|\,\|\mathbf{e}_{s_1,s_2}\|}.$$

This measure captures semantic relatedness but not novelty, and can be sensitive to embedding behaviors (e.g., scale-invariance may obscure frequency effects).

In summary (Table 1), CS offers an interpretable balance of novelty and relevance; NLI aligns closely with the conceptual role of explanations, though reliability is tied to model strength and may not fully reflect novelty; COS, on the other hand, is easy to compute but lacks novelty sensitivity. For this reason, we hypothesize that CS should better align with explanation effectiveness, as defined above, and thus positively correlate with accuracies at the system level.

Measure	Relevance	Novelty	Reliability
CS	yes	yes	Corpus-based, robust
NLI	yes	no	Theoretical, model-limited
COS	yes	no	Fast, but surface-level

Table 1: Overview of implicit content measures and their features.

4 Tasks and Datasets

We use four datasets that propose tasks involving different kinds of reasoning and eliciting implicit or external knowledge to various extents. All datasets provide either human-generated or human-collected and curated explanations (which we use as the `gold` baseline, see Section 5)[3]

e-SNLI (Natural Language Inference). A version of the Stanford Natural Language Inference (SNLI) corpus, includes 570k sentence pairs (which we use as s_1 and s_2) labeled for three entailment classes: "entailment", "contradiction", and "neutrality"; each pair is enriched with 3 human-written, natural language explanations (Camburu et al., 2018), which we use in concatenation as our "gold" explanations.

[3]An example of how each dataset is preprocessed in the GEISER pipeline is provided in the Appendix.

StrategyQA (Multi-hop Question Answering). A question-answering dataset designed to require multiple-step strategic reasoning and/or implicit knowledge to answer a question. The dataset (Geva et al., 2021) comprises 2,780 strategy questions (which we use as s_2) with answer "yes" or "no" (labels), its decomposition into multi-step reasoning paths (which we use in combination as gold explanations) and evidence paragraphs giving the context of the question (which we use as s_1).

e-CARE (Causality). A dataset focused on causal reasoning, featuring human-annotated explanations for the causal questions, The dataset consists of 21k causal reasoning questions with both correct and incorrect answers (Du et al., 2022). We accommodate this dataset into our experimental setup by pairing both input sentences as s_1 and, for each pair, ask the question represented by s_2, focusing on whether the first sentence is the cause of the second (label "yes") or not (label "no").

e-RTE-3-it (Recognizing Textual Entailment in Italian). A dataset in Italian for Recognizing Textual Entailment (RTE), featuring pairs of texts-hypotheses and human-written, manually curated explanations for the entailment relation (Zaninello et al., 2023). It consists of 1,600 sentence pairs (which we use as s_1 and s_2, respectively) and is annotated with the same labels as e-SNLI.

5 Explanation Types

We test two different modes of explanation generation: explanations that *confirm* the given relationship between s_1 and s_2, explaining why it holds (*why*, *gold*) and explanations that potentially *contradict* the relationship between s_1 and s_2 explaining the circumstances when the relationship may not hold (*counterfactual*).

Why explanations. This kind of machine-generated explanation (`why`) is the most typical way to provide an explanation, i.e., the answer to a "why" question. In our setting, a why explanation is an answer to the question "*Why is r_c the relation holding between s_1 and s_2?*".

Gold explanations. These explanations (`gold`) are the explanations provided in the original dataset, either directly written or manually checked by humans given the correct relation r_c, thus falling into the label-confirming explanation type like *why* explanations. While the quality of human-generated explanations is generally 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 will perform better than model-generated ones. Therefore, for the purposes of this study, we evaluate them along with the generated ones and take them as a strong baseline, rather than consider them a target or reference explanation.

Counterfactual explanations. In our setting, a counterfactual (`cf`) explanation (Wachter et al., 2017; Verma et al., 2022) explicitly contradicts the golden label. It originates from the following question: "*What are the conditions in which relation r_c may not hold for s_1 and s_2?*". The aim of these explanations is to test the robustness of models to potentially false or misleading information, as well as highlight how different models may be differently sensitive to explanation injection[4].

6 Experiments

6.1 Experimental Setup

Models. We utilize three open-access language models of comparable size, which we combine as both M_1 and M_2: Llama-3-8B-Instruct (Team Llama et al., 2024), Gemma-7b-it (Gemma et al., 2024) and DeepSeek-R1-Distill-Qwen-7B (DeepSeek-AI et al., 2025; Qwen et al., 2025).

To compute Cosine Similarity (Section 3.3), we use *sentence-transformers/all-MiniLM-L6-v2* (Wang et al., 2020). For Entailment, we use the pre-trained NLI model *deberta-large* (Liu et al., 2019), fine-tuned on the Multi-Genre NLI dataset (Williams et al., 2018).

Prompting and Inference Details. Our implementation leverages the HuggingFace's `lm_eval harness` library to ensure consistent and reproducible evaluation across tasks. For M_1 generation, we use the output type *generate_until*. We employ greedy decoding for all experiments, and all prompts are constructed in English (so that all explanations are returned in English, regardless of input)[5]. For M_2 prediction, we use the *multiple_choice* output type, which calculates logits for a given set of labels.

[4] See prompts and example explanations in the Appendix.

[5] Due to computational constraints, we used the first 800 examples from the test sets of each dataset to keep generation within our capacity limits. This approach allowed us to maintain a balance between comprehensive evaluation and practical feasibility.

	noexp	gold	cf-llama-m1	cf-gemma-m1	cf-deepsk-m1	why-llama-m1	why-gemma-m1	why-deepsk-m1
ESNLI (3 labels)								
llama-m2	0.54	0.71	0.34	0.44	0.59	0.72	0.89	**0.95**
gemma-m2	0.61	0.79	0.53	0.46	0.61	0.76	0.90	**0.95**
deepseek-m2	0.34	0.34	0.40	0.36	0.61	0.45	0.70	**0.96**
all-m2	0.50	0.61	0.42	0.42	0.60	0.64	0.83	**0.95**
SQA (2 labels)								
llama-m2	0.64	0.78	0.62	0.66	0.46	0.80	0.75	**0.91**
gemma-m2	0.62	0.68	0.55	0.58	0.50	0.81	0.75	**0.88**
deepseek-m2	0.45	0.45	0.43	0.38	0.52	0.45	0.45	**0.77**
all-m2	0.57	0.64	0.53	0.54	0.49	0.69	0.65	**0.85**
ECARE (2 labels)								
llama-m2	0.53	0.56	0.51	0.54	0.61	0.81	0.76	**0.91**
gemma-m2	0.48	0.71	0.54	0.51	0.62	0.92	0.75	**0.94**
deepseek-m2	0.48	0.49	0.50	0.48	0.59	0.53	0.54	**0.83**
all-m2	0.50	0.59	0.52	0.51	0.61	0.75	0.68	**0.89**
ERTEIT (3 labels)								
llama-m2	0.48	0.53	0.29	0.26	0.20	**0.70**	0.62	0.67
gemma-m2	0.44	0.49	0.23	0.19	0.20	**0.70**	0.59	0.67
deepseek-m2	0.48	0.48	0.50	0.38	0.20	0.58	0.53	**0.67**
all-m2	0.47	0.50	0.34	0.28	0.20	0.66	0.58	**0.67**

Table 2: Accuracy of M_2 models across the four datasets and explanation types, using explanations generated by M_1. Explanations marked as *noexp* and *gold* represent the baselines. Values are reported as accuracy scores of each M_2 model and as mean across all M_2 models (*all-m2*), with standard errors omitted for brevity. The best-performing explanation type for each M_2 is boldfaced.

To make generated explanations comparable to gold explanations, we ask M_1 to explain in approx. 3 sentences. To include the explanations in Step 2, we prompt M_2 to use a "hint" to give its answer, represented by the explanation.

Anonymization to Prevent Label Leakage. To ensure that the explanations do not simply suggest the right answer without genuinely being informative, we "anonymize" them by substituting each explicit reference to the labels with a placeholder (`XXX`) using regular expressions to fetch either the label (e.g. "YES" and "NO") or words directly connected to the relation (e.g. "contradict", "contradiction" etc.). Moreover, we explicitly ask the M_1 model to avoid stating the answer directly when generating the explanation.

6.2 Evaluation

Baselines. We select two baselines: no explanation given (**`noexp`**), where the model M_2 performs 0-shot relation r_p prediction; human explanation (**`gold`**), where we use the explanation provided in the original dataset as the hint, providing a strong baseline. Gold explanations too, like the generated ones, underwent the process of anonymisation.

Explanation quality. For explanation quality through GEISER, we calculate the average accuracy (**`acc`**) of each $M2$ model separately using either the explanations generated by the same model ($M_1 = M_2$), or by another model ($M_1 \neq M_2$). We report average accuracy for each explanation type/M_1, and both separately and ensembling by M_2, along with the accuracy obtained by with gold and noexp baselines (Table 2).

Correlation with Implicitness. To study the correlation of the selected implicitness measures (CS, COS, NLI) with explanation quality, we report the average score separately for each explanation type/M_1, each ranging from 0 to 1 (Table 3. Then, we calculate the Pearson correlation coefficient (r) ($p = 0.05$), to assess the linear relationship between each M_2 accuracy for all explanation types in each dataset, and the association measure for the same system for that dataset.

7 Results and Discussion

Accuracy Trends. In Table 2 we report the performances on the GEISER experiments for the four datasets under different explanation types, both with $M_1 = M_2$ and $M_1 \neq M_2$. Results are re-

	noexp	gold	cf-llama-m1	cf-gemma-m1	cf-deepsk-m1	why-llama-m1	why-gemma-m1	why-deepsk-m1
ESNLI (3 labels)								
Causal Strength	0.45	0.69	0.48	0.51	0.65	0.61	0.67	0.81
Cosine Similarity	0.05	0.70	0.48	0.46	0.48	0.44	0.56	0.47
Entailment prob.	0.34	0.42	0.12	0.13	0.82	0.22	0.23	0.77
SQA (2 labels)								
Causal Strength	0.55	0.70	0.53	0.58	0.71	0.64	0.69	0.77
Cosine Similarity	0.05	0.58	0.75	0.78	0.69	0.75	0.83	0.76
Entailment prob.	0.08	0.03	0.07	0.08	0.62	0.10	0.17	0.89
ECARE (2 labels)								
Causal Strength	0.50	0.85	0.52	0.53	0.61	0.65	0.64	0.73
Cosine Similarity	0.06	0.46	0.55	0.51	0.44	0.61	0.63	0.45
Entailment prob.	0.44	0.20	0.17	0.13	0.81	0.29	0.19	0.88
ERTEIT (3 labels)								
Causal Strength	0.60	0.71	0.47	0.53	0.53	0.72	0.70	0.80
Cosine Similarity	0.05	0.75	0.44	0.48	0.39	0.45	0.63	0.46
Entailment prob.	0.58	0.38	0.29	0.27	0.87	0.50	0.45	0.93

Table 3: Mean association measures (Causal Strength, Cosine Similarity, Entailment probability) across datasets, models, and explanation types. The *noexp* and *gold* columns indicate the baselines results using no-explanation and the human-generated ones. The remaining columns indicate results obtained by counterfactual *cf* or *why* explanations generated by the three LLMs as M_1.

ported both separately for the different M_2 models, as well as the average accuracy across all models (*all-m2*).

The best scoring task was ESNLI with *why* explanations written by Deepseek, which also presents the largest gain over the *noexp* baseline (from 0.34 to 0.96 with Deepseek both as $M1$ and M_2). Despite presenting a very similar task, the lower scoring dataset and the smallest gains were with ERTEIT (min. 0.20 with *cf*, max. 0.70 with LLama's *why*). This seems to indicate that the models still struggle with languages other than English, or are possibly mislead by the language shift between the input (Italian) and the explanations (English).

The preferred explanations were those of Deepseek with all M_2s in all tasks, with the exception of ERTEIT, where Llama-m2 and Gemma-m2 scored higher with Llama-m1's *why* explanations.

Label-confirming explanations (*why*) consistently led to the highest accuracy across all datasets and models, confirming that explanations aligned with the gold label can meaningfully support the m2 model's decision-making. On the other hand, label-contradicting explanations generally scored lower than the *noexp* baseline, as was expected, indicating that "bad" explanations can indeed be detrimental to the model's accuracy. However, there are a few cases where *cf* explanations improved over the *noexp* baseline, specifically on ESNLI and ECARE and mainly with Deepseek-m2's explanations. We manually inspected a sample of the *cf* explanations that led to a correct prediction, and noticed a common trend: in fact, in these cases either the model "refused" to support the opposite label, or it produced a long "chain of thought" style explanation which was truncated, and therefore did not contain the section of the explanation supporting the opposite label. This was especially the case with Deepseek, which produced an initial "reasoning" independently of the supported relationship, which was helpful for the downstream model to predict the correct label. Finally, it is worth noticing that machine-generated *why* explanations consistently outperform the human-generated ones (*gold*), which nonetheless are beneficial to prediction accuracy compared to *noexp*.

Association Measures. We computed three association measures between the explanation and the input: causal strength, cosine similarity, and entailment probability (Table 3). Label-confirming explanations showed higher values on all three measures compared to other conditions. For example, in SQA, the average causal strength and cosine similarity were highest (0.64 and 0.75, respectively) for label-confirming explanations, indicating a stronger semantic and causal link to the

input. This confirms that these metrics are not only sensitive to novelty but also to relevance (*cf* being new information, which is not relevant for the correct label). On the other hand, entailment probability showed more variation across datasets, likely due to inherent differences in how entailment is interpreted in each task.

Correlations with Accuracy. To identify which measures best predict explanation utility, we correlated the association measures with accuracy across all explanation types (*gold*, *cf*, *why*) and generative models (Table 4). While all measures generally show a positive correlation with accuracy, Causal strength showed the highest and most significant correlation across all datasets (e.g., 0.86 for llama-m2 in ESNLI, 0.90 for all-m2 in ERTEIT). This suggests that the extent to which an explanation causally supports the input is a reliable predictor of its usefulness. Cosine similarity and entailment probability were weaker and less consistent predictors, though entailment reached high correlations in specific cases (e.g., 0.91 for deepseek-m2 on SQA).

Implications. These results suggest that injecting label-confirming explanations improves model performance, particularly when the explanations exhibit strong causal links to the input. Among the evaluated association measures, causal strength emerges as the most promising indicator of explanation quality. This highlights its potential as a diagnostic tool for filtering or scoring explanations before injection. Cosine similarity and entailment probability offer additional, though less robust, signals.

	CS	COS	NLI
ESNLI			
llama-m2	**0.86**	0.29	0.45
gemma-m2	0.33	0.33	0.33
deepseek-m2	0.63	0.19	0.63
all-m2	0.68	0.28	0.53
SQA			
llama-m2	0.42	0.21	0.14
gemma-m2	0.54	0.21	0.21
deepseek-m2	**0.71**	0.13	**0.91**
all-m2	0.62	0.21	0.48
ECARE			
llama-m2	0.43	0.37	0.46
gemma-m2	0.66	0.47	0.47
deepseek-m2	0.36	0.08	**0.80**
all-m2	0.54	0.36	0.53
ERTEIT			
llama-m2	**0.92**	0.20	0.18
gemma-m2	**0.93**	0.18	0.18
deepseek-m2	**0.70**	0.13	-0.02
all-m2	**0.90**	0.18	0.17

Table 4: Pearson correlation coefficient (r), between M_2 accuracy across the four datasets and the three association measures: Causal Strength (CS), Cosine Similarity (COS) and Entailment probability (NLI). The boldfaced figures indicate statistical significance according to a t-test with $n-2$ degrees of freedom, and $p = 0.05$.

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. Future work should explore how to automatically generate or filter explanations with high causal alignment to further boost downstream model performance.

Limitations

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.

Our measurement of implicitness relies on metrics like cosine nli and casual strength, which do not distinguish between relevance and novelty, and may not fully capture the nuanced nature of implicit

knowledge in language. Finer-grained techniques are needed for a comprehensive evaluation of implicitness. Future work should explore additional features, such as explanation length and syntactic complexity, to better understand their interplay with model performance.

We utilized 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.

Also, while our study focused on sentence-pair tasks, the GEISER pipeline can in principle be extended to multi-hop reasoning chains and other explanation-rich settings by iteratively injecting intermediate explanations, which we plan to explore in future work.

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.

Luisa Bentivogli, Peter Clark, Ido Dagan, and Danilo Giampiccolo. 2011. The seventh pascal recognizing textual entailment challenge. *Theory and Applications of Categories*.

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.

Shaobo Cui, Lazar Milikic, Yiyang Feng, Mete Ismayilzada, Debjit Paul, Antoine Bosselut, and Boi Faltings. 2024. Exploring defeasibility in causal reasoning. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 6433–6452.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning.

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.

Gemma, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, Léonard Hussenot, Pier Giuseppe Sessa, Aakanksha Chowdhery, Adam Roberts, Aditya Barua, Alex Botev, Alex Castro-Ros, Ambrose Slone, Amélie Héliou, et al. 2024. Team gemma and : Open models based on gemini research and technology.

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.

Tirthankar Ghosal, Tanik Saikh, Tameesh Biswas, Asif Ekbal, and Pushpak Bhattacharyya. 2022. Novelty detection: A perspective from natural language processing. *Computational Linguistics*, 48(1):77–117.

Tirthankar Ghosal, Amitra Salam, Swati Tiwari, Asif Ekbal, and Pushpak Bhattacharyya. 2018. TAP-DLND 1.0 : A corpus for document level novelty detection. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

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.

Xuanli He, Yuxiang Wu, Oana-Maria Camburu, Pasquale Minervini, and Pontus Stenetorp. 2024. Using natural language explanations to improve robustness of in-context learning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13477–13499, Bangkok, Thailand. 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.

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*.

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.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880.

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*.

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.

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.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

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.

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.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2025. Qwen2.5 technical report.

Lin CY ROUGE. 2004. A package for automatic evaluation of summaries. In *Proceedings of Workshop on Text Summarization of ACL, Spain*.

Harald Steck, Chaitanya Ekanadham, and Nathan Kallus. 2024. Is cosine-similarity of embeddings really about similarity? In *Companion Proceedings of the ACM Web Conference 2024*, WWW '24, page 887–890. ACM.

AI@Meta Team Llama, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, et al. 2024. The llama 3 herd of models.

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.

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).

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Advances in neural information processing systems*, 33:5776–5788.

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.

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.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Appendix

Illustrative Examples from the Four Datasets

Example 1 — ESNLI (with `gold` explanation)

Task Input

Premise: This church choir sings to the masses as they sing joyous songs from the book at a church.
Hypothesis: The church is filled with song.
Label: entailment

Prediction Prompt (with `gold` explanation)

Your task is to predict the entailment relationship (entailment, neutral, contradiction) between a premise and a hypothesis given a hint.
Premise: This church choir sings to the masses as they sing joyous songs from the book at a church. Hypothesis: The church is filled with song. Hint: "Filled with song" is a rephrasing of "choir sings to the masses." Hearing song brings joyousness in the church. If the choir sings, then the church is filled with song. Answer:

Example 2 — SQA (with `noexp` explanation)

Task Input

Context: The Police = English rock band.
Question: Could the members of The Police perform lawful arrests?
Label: no

Prediction Prompt (with `noexp` explanation)

Your task is to answer a question with 'yes' or 'no' given a context and a hint.
Context: The Police = English rock band. Question: Could the members of The Police perform lawful arrests? Hint: None. Answer:

Example 3 — ECARE (with `why` explanation)

Task Input

Sentence 1: The woman gave birth to a child.
Sentence 2: The child brought psycho-physical phenomena on a new life.
Label: yes

Generation Prompt

Your task is to explain in approx. 3 sentences the answer to a question given a pair of sentences, without stating the answer directly.
Sentence 1: The woman gave birth to a child. Sentence 2: The child brought psycho-physical phenomena on a new life. Question: Is sentence 1 the cause of sentence 2? Answer = yes Explanation (max 3 sentences, avoid stating the answer directly):

Prediction Prompt (with `why` explanation by *DeepSeek*)

Your task is to answer a question with 'yes' or 'no' given a pair of sentences and a hint.
Sentence 1: The woman gave birth to a child. Sentence 2: The child brought psycho-physical phenomena on a new life. Question: Is sentence 1 the cause of sentence 2? Hint: Sentence 1 describes an event that happened before Sentence 2. In many cases, the XXX of an event is the immediate preceding event. Additionally, if Sentence 1 directly leads to Sentence 2, such as in the scenario where the woman gave birth to a child, it can be inferred that Sentence 1 is the XXX of Sentence 2. Answer:

Example 4 — ERTEIT (with `cf` explanation)

Task Input

Premise: Claude Chabrol (nato il 24 giugno 1930) è un regista francese diventato famoso negli ultimi 40 anni sin dal suo primo film, *Le Beau Serge*, per le sue agghiaccianti storie di omicidi, tra cui *Le Boucher*.
Hypothesis: *Le Beau Serge* è stato diretto da Chabrol.
Label: entailment

Generation Prompt

Your task is to convince me in approx. 3 sentences that the entailment relationship between a premise and a hypothesis is WRONG, without stating the relationship directly.
Premise: Claude Chabrol [...] *Le Beau Serge* [...] Hypothesis: *Le Beau Serge* è stato diretto da Chabrol. Entailment relationship: entailment. Explanation (max 3 sentences, avoid stating the relationship directly):

Prediction Prompt (with `cf` explanation by *Gemma*)

Your task is to predict the entailment relationship (entailment, neutral, contradiction) between a premise and a hypothesis given a hint.
Premise: Claude Chabrol [...] *Le Beau Serge* [...] Hypothesis: *Le Beau Serge* è stato diretto da Chabrol. Hint: The premise describes a renowned French director, Claude Chabrol, known for his intense murder stories. The hypothesis assumes that Chabrol directed the film *Le Beau Serge*. However, the premise does not necessarily XXX the hypothesis, as it does not provide any information about the film's authorship. Therefore, the XXX relationship between the premise and the hypothesis is incorrect. Answer: