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

Anonymous ACL submission

Abstract

In this work, we investigate the influence of different types of natural language explanations on LLMs' predictions, focusing on four different datasets presenting tasks that involve leveraging implicit knowledge. We conduct experiments with three SOTA LLMs on five types of explanations, either written by humans or machine-generated, through three generation methods: explain given the correct label (label-aware), explain and predict the label contextually (label-agnostic), and support the falseness of the correct label (label-contradicting). Our results demonstrate that providing explanations consistently improves the accuracy of LLM predictions, even when the models are not explicitly trained to take explanations as input, and pave the way to a study of the relationship 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. We investigate 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 generation modes (label-aware, label-agnostic, and label-contradicting) and four tasks requiring implicit knowledge. We hypothesize that explanation effec-

[1] Code and data will be distributed upon acceptance.

tiveness, measured by downstream task performance, correlates with their degree of *implicit content*, i.e., novel yet relevant information they provide. Section 4 explores this hypothesis by examining the relationship between explanation effectiveness and metrics approximating novelty and relevance. This insight is crucial for RAG settings, where explanations serve as intermediate reasoning steps to enhance factual accuracy and robustness. If explanations merely rephrase retrieved evidence or fail to introduce new insights, they may be redundant or misleading rather than helpful.

The main contributions of this paper are the following: (i) 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; (ii) using the proposed pipeline, we report extensive experimental results on different kinds of explanations (human- and machine-generated), across three LLMs, four tasks and two languages; (iii) through our analysis, we introduce "implicit knowledge" as a key factor of explanation quality, and propose a metric to estimate it showing its correlation with explanation effectiveness.

2 Related Work

The **role of explanations** in NLP has been extensively studied. Cambria et al. (2023), for instance, surveys natural language explanation generation, while Hartmann and Sonntag (2022) explores their benefits for NLP models. Paranjape et al. (2021) focuses 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.

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.

As for the **types of explanations** used in NLP, a comprehensive characterization of explanations is provided by Jansen et al. (2016), each with different insights into model behavior from different perspectives. For instance, *local explanations* focus on individual predictions (Ribeiro et al., 2016; Lundberg and Lee, 2017) to estimate feature importance. These methods help understand model decisions at the instance level but may not fully capture the overall implicit model knowledge. *Feature importance explanations* generalize this idea by identifying which input features contribute most to a model's predictions. In contrast, *global explanations* aim to describe the model's overall decision-making behavior across all inputs, with early foundational work by Friedman (2001) providing key insights into ensemble models, while *attention-based explanations* have gained popularity since the introduction of the Transformer model (Vaswani et al., 2017). However, the effectiveness of attention as a faithful explanation, and its correlation with model decisions, is debated (Jain and Wallace, 2019; Wiegreffe and Pinter, 2019). In our work, we focus on natural language explanations and their impact on downstream performance rather than inspecting model behavior by analyzing its inner computations.

Finally, to the best of our knowledge, there are no previous works addressing **implicit content measures** directly. However, 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.

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 $E = \{e_1, e_2, \ldots e_n\}$ 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 6 we define in detail the set E of explanation types.[2]

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

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.

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

Therefore, accuracy of the model M_2 on a relation prediction task is used as a proxy metric of 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, 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.

4 Measuring Implicit Content

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 $< s_1, s_2 >$.

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

We define four metrics to assess explanation *relevance* and *novelty*:

Relevance (REL)

1. **Semantic Similarity (A)**: Measures cosine similarity between sentence embeddings of the input (text + hypothesis) and the explanation.

2. **NLI-based Relevance (B)**: Uses a pre-trained NLI model to determine whether the explanation entails the input ($s_1 + s_2$), assuming stronger entailment indicates higher relevance.

Novelty (NOV)

1. **Probability of Not-Entailment (A)**: Measures $(1 - \textit{prob\_entailment})$ between input ($s_1 + s_2$) and explanation, assuming higher values indicate novelty.

2. **Probability of Neutral (B)**: Uses a 3-label NLI model to detect whether the explanation is neutral (neither entailed nor contradictory) with respect to the input, suggesting the presence of new, non-redundant information.

Both aspects should be balanced since novelty does not necessarily imply relevance.

5 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` explanation type, see Section 6.1).

e-RTE-3-it (Recognizing Textual Entailment) A dataset in Italian for Recognizing Textual Entailment (RTE), featuring pairs of texts-hypotheses and human-written 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 for three entailment classes: "entailment", "contradiction", and ""neutrality.

e-SNLI (Natural Language Inference) A version of the Stanford Natural Language Inference (SNLI) corpus, includes 570k sentence pairs labeled for the same three entailment classes as e-RTE-3-it enriched with 3 human-written, natural language explanations (Camburu et al., 2018), which we use in concatenation as our "gold" explanation.

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 (s_2) whether the first sentence is the cause of the second (label "yes") or not (label "no").

StrategyQA (Multi-hop Question Answering) A question-answering dataset designed to require multiple-step strategic reasoningandor 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).

6 Generation Modes and Explanation types

In this section we present the generation strategies and the types of explanations generated by model M_1

and used by model M_2 with different characteristics.

To reproduce a real-world scenario, we group different types of explanations based on whether, when they are generated, the model is given knowledge of the true relation between the two sentences. We consider three different modes:

- the correct relationship between s_1 and s_2 is known at generation explanation time (**label-aware**)
- the correct relationship is not known at the time of generation, and has to be predicted and explained contextually (**label-agnostic**)
- the correct relationship is known but is said to be incorrect at the time of generation, so a counterfactual explanation is required (**label-contradicting**).

The latter type of explanation has the aim of testing the consistency of a model to inputs that can potentially mislead the correct prediction.

6.1 Label-aware Explanations

In the *label-aware* approach, the generation process is driven by the correct relation r_c holding between s_1 and s_2. We include both human generated (**`gold`**) and model generated explanations (**`why`**) in this setup.

Gold explanations. These explanations (called `gold` in our experiments) are the explanations provided in the original dataset, either directly generated or manually checked by humans given the correct relation r_c.

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 perform better than model generated explanations. Therefore, for the purposes of this study, we evaluate them along with the generated ones rather than consider them a target or reference explanation.

Why explanations. This kind of 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 *Why is r_c the relation holding between s_1 and s_2?*.

6.2 Label-agnostic Explanations

In Section 6.1 we have assumed that explanations are generated knowing the correct relation r_c holding between s_1 and s_2, i.e., referred as *label-aware*. However, to simulate a more realistic world scenario, we are also interested in experimenting on *relation-agnostic* explanations, where a model M_1 generates an explanation contextually predicts the relation. We call this modality *label agnostic generation*.

This kind of explanation does not assume knowledge of r_c, and asks to either (i) explain the reasoning then predict r_c (**`cot`**), or (ii) first predict r_c then explain the prediction (**`phr`**).

Chain-of-Thought Explanations. This kind of explanation, inspired by "explain-then-predict" strategies such as chain-of-thought in-context learning (Wei et al., 2022), does not assume knowledge of r_c, and asks to first provide the reasoning to get to the final answer, then predict the correct relation (`cot`).

Post-hoc Rationalizations. inspired by the "predict-then-explain" strategies using post-hoc self-rationalizations (Lei et al., 2016) asks the model to first predict the correct relation, then explain its prediction (`phr`).

6.3 Label-contradicting Explanations

In this final setup (*label-contradicting*), we use counterfactual explanations (**`c-factual`**)(Wachter et al., 2017; Verma et al., 2022), explicitly contradicting the golden label.

Counterfactual explanations. 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?*. 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.

7 Experiments

7.1 Models

We utilized three open-access language models of comparable size to assess the quality of explanations: 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). Llama 3-8B-Instruct, developed by Meta, is an 8 billion-parameter model designed for instruction-following tasks. It features a context window of 8,000 tokens and has demonstrated strong performance across various benchmarks, including a 68.4% accuracy on MMLU (Hendricks et al., 2016). Gemma-7b-it is a 7-billion-parameter model fine-tuned for instruction tasks. Built upon the research and technology of Google's Gemini models (Team, 2023), Gemma models have shown strong performance across academic benchmarks for language understanding, reasoning, and safety. DeepSeek-R1-Distill-Qwen-7B is a 7-billion-parameter model distilled from the larger DeepSeek-R1, focusing on enhancing reasoning capabilities. It has shown competitive performance on benchmarks such as the American Invitational Mathematics Examination 2024, achieving a pass@1 score of 55.5.

To compute inference scores for novelty and relevance (Section 4), we use a pre-trained NLI model. A sigmoid function is applied to the entailment score p_{ent} of the NLI model. In its classical formulation, higher scores indicate stronger entailment relation between combined a text and a hypothesis, while in our setting we take it as a proxy of the degree of relatedness between the concatenation of sentence s_1 and s_2 and their corresponding explanation e, suggesting that the explanation is likely to be relevant to the input. For calculations, we use the *deberta-large* model (Liu et al., 2019), fine-tuned on the Multi-Genre NLI dataset (Williams et al., 2018).

7.2 Experiment setups

Prompting and Inference Details Our implementation leveraged the HuggingFace's `lm_eval harness` library to ensure consistent and reproducible evaluation across tasks, with *output type generate_until* and *multiple_choice* for M_1 and M_2, respectively. 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. We employed greedy decoding for all experiments, and all prompts were constructed in English (so all explanations were returned in English, regardless of input). 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 "anonimize" them by substituting each explicit reference to the labels with a placeholder using regular expressions. Moreover, we explicitly ask the M_1 model to avoid stating the answer directly when generating the explanation.

Baselines We use three 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; we also set the hint given to the $M2$ model as a copy of the right label, to set an upperbound baseline (**`obvious`**) to check whether the model is sensitive to label leakage regardless of the explanatory form of the hint.

7.3 Performance Measures

GEISER We calculate the accuracy (**`acc`**) of the $M2$ models using either the explanations generated by the same model (Table 1), or by another model (Table 2), which we report along with the accuracy obtained by the gold and the baseline explanations.

Implicitness Here, we analyze the correlation both with the accuracy obtained by M_2 using the explanations (**`acc`**), as well as their potential to change a prediction from wrong to right (**`acc_change`**), which we set = 0 if the same label is predicted with and without explanation, 1 if the prediction becomes right using the explanation, -1 if it becomes wrong.

8 Results and Discussion

8.1 GEISER results

In Table 1 we report the performance on the GEISER experiments with $M_1 = M_2$ and Table 2 for $M_1 \neq M_2$ of the three models the across four datasets under different explanation types.

The figures show that, providing LLMs with explanations, even if they have not been explicitly trained

for this, can significantly boost their accuracy in predicting semantic relations between sentences. The improvement is consistent across different models, datasets, and explanation types, with label-aware explanations with the most significant gains.

The performance of models varies significantly across datasets. The e-RTE-3-it dataset has lower accuracy scores across all explanation types, while ESNLI and ECARE show higher accuracy, particularly with *why* explanations. The StrategyQA dataset exhibits mixed results, with *why* and *cot* explanations performing well in different scenarios. This variability suggests that the effectiveness of explanation types may depend on the specific characteristics and language of the dataset (Italian), even though in some cases (M1: Llama - M2: Gemma on e-RTE-3-it) gold explanations, written in Italian, outperform why explanations (written in English).

As for Same-Model vs. Cross-Model scenarios, models generally achieve higher accuracy when generating and using their own explanations (M1 = M2), indicating better alignment between explanation style and internal reasoning. However, certain cross-model combinations (e.g., M1: Qwen - M2: Llama on ECARE) outperform same-model scenarios, highlighting the potential for leveraging complementary strengths in cross-model setups.

Label-aware explanations, particularly *why*, consistently outperform other types. Label-agnostic explanations (*cot*, *phr*) generally underperform but show occasional utility in cross-model scenarios on the StrategyQA dataset. Label-contradicting explanations (cf) consistently yield the lowest accuracy, emphasizing the detrimental impact of misleading information on model performance. However, it is interesting to notice that in a few cases, for example the ECARE dataset with M1=Qwen, *cf* explanations are still outperforming the *noexp* and *dummy* baselines. Another interesting observation is that in some cases (e.g. on ESNLI with LLama and Qwen as M_2) the *obvious* (upper bound), expected to outperform all types as it is a direct suggestion of the correct label, is lower than the best performing explanation type. These facts seem to indicate that input in an *explanatory* form is indeed influencing the "reasoning" of the model, leading it to better predictions.

8.2 Implicitness Results

Implicitness measures show limited predictive power across datasets, with the highest correlation at 0.574 for `anon-gold` in *Qwen + Gemma* on **ERTEIT**. Dataset-specific trends reveal weak correlations in **SQA** ($R^2 < 0.02$) but stronger effects in **ERTEIT** and **ESNLI**, particularly for entailment-based features. For example, `REL (2)` achieves 0.434 for `gold` in *Qwen + Gemma* on **ERTEIT**, and 0.530 for `dummy` in **ESNLI**, highlighting the role of novelty and explicit entailment.

`Gold` explanations consistently show the strongest correlations, while `dummy` explanations occasionally influence model behavior. Label-agnostic (`cot`, `phr`) and label-contradicting (`cf`) explanations underperform, with `cf` showing negative or negligible correlations. *Qwen + Gemma* exhibits stronger sensitivity to implicitness features than *Qwen + Llama*, suggesting Gemma benefits more from structured explanations.

In summary, implicitness measures influence accuracy changes but are not definitive, with stronger effects in reasoning-heavy datasets like **ERTEIT** and **ESNLI**.

9 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.

GEISER Results (M1 = M2)								
MODEL	**noexp**	**dummy**	**obvious**	**gold**	**why**	**cot**	**phr**	**cf**
e-RTE-3-it (3 labels)								
M1: Llama - M2: Llama	0.4862	0.4987	0.5725	0.5362	**0.5637**	0.4837	0.4900	0.1725
M1: Gemma - M2: Gemma	0.4400	0.4700	0.5725	0.4962	**0.505**	0.4700	0.4550	0.16125
M1: Qwen - M2: Qwen	0.4850	0.4850	0.4950	0.4850	**0.5512**	0.4725	0.4787	0.1150
ESNLI (3 labels)								
M1: Llama - M2: Llama	0.5437	0.5975	0.6762	**0.7162**	0.7075	0.3563	0.3850	0.3450
M1: Gemma - M2: Gemma	0.6100	0.535	0.9962	0.7975	**0.8762**	0.4363	0.4275	0.4575
M1: Qwen - M2: Qwen	0.3412	0.3412	0.6250	0.3425	**0.9400**	0.4550	0.4087	0.6287
ECARE (2 labels)								
M1: Llama - M2: Llama	0.5350	0.5450	0.9062	0.5613	**0.7975**	0.5475	0.5525	0.5137
M1: Gemma - M2: Gemma	0.4887	0.5037	1.0000	0.7125	**0.8050**	0.5775	0.5375	0.5562
M1: Qwen - M2: Qwen	0.4887	0.4900	0.9500	0.4987	**0.8625**	0.5487	0.4925	0.5750
StrategyQA (2 labels)								
M1: Llama - M2: Llama	0.6450	0.6837	0.5660	**0.7870**	0.7587	0.6420	0.6462	0.5887
M1: Gemma - M2: Gemma	0.6275	0.6237	0.9812	0.6850	**0.7875**	0.5825	0.5937	0.5800
M1: Qwen - M2: Qwen	0.4575	0.4550	0.7575	0.4550	**0.7512**	0.5775	0.5612	0.5100

Table 1: Accuracy of models across the four datasets and explanation types, using explanations generated by the same model (M1 = M2). Explanations marked as *noexp* and *dummy* represent the baselines, *obvious* represents the upper bound, remaining columns represent label-aware (*gold*, *why*), label-agnostic (*cot*, *phr*) and label-contradicting (*cf*) explanations. Values are reported as accuracy scores of M_2 models, with standard errors omitted for brevity. The best-performing explanation type for each model-dataset combination is boldfaced.

GEISER Results (M1 $\neq$ M2)								
MODEL	**noexp**	**dummy**	**obvious**	**gold**	**why**	**cot**	**phr**	**cf**
e-RTE-3-it (3 labels)								
M1: Llama - M2: Gemma	0.4387	0.4700	0.5725	0.4950	**0.5575**	0.3375	0.4850	0.1462
M1: Llama - M2: Qwen	0.4850	0.4850	0.4950	0.4850	**0.5075**	0.4825	0.4975	0.4762
M1: Gemma - M2: Llama	0.4863	0.4987	0.5725	**0.5325**	0.5287	0.4637	0.4625	0.1837
M1: Gemma - M2: Qwen	0.4850	0.4850	0.4938	0.4850	**0.495**	0.4762	0.4675	0.3700
M1: Qwen - M2: Llama	0.4862	0.4987	0.5725	0.5362	**0.5487**	0.4525	0.4750	0.1025
M1: Qwen - M2: Gemma	0.4387	0.4700	0.5725	0.4950	**0.5462**	0.4150	0.4737	0.1112
ESNLI (3 labels)								
M1: Llama - M2: Gemma	0.6100	0.5350	0.9962	**0.7975**	0.7587	0.3688	0.3875	0.5213
M1: Llama - M2: Qwen	0.3412	0.3412	0.6250	0.3425	**0.4362**	0.3862	0.3850	0.3762
M1: Gemma - M2: Llama	0.5437	0.5975	0.6762	0.7162	**0.885**	0.4550	0.5200	0.4375
M1: Gemma - M2: Qwen	0.3412	0.3412	0.6250	0.3425	**0.6725**	0.4663	0.3775	0.3625
M1: Qwen - M2: Llama	0.5438	0.5975	0.6765	0.7162	**0.9550**	0.5487	0.4312	0.6150
M1: Qwen - M2: Gemma	0.6100	0.5350	0.9962	0.7975	**0.9575**	0.4987	0.4362	0.6287
ECARE (2 labels)								
M1: Llama - M2: Gemma	0.4887	0.5037	1.0000	0.7125	**0.8962**	0.5512	0.5700	0.5325
M1: Llama - M2: Qwen	0.4862	0.4987	0.5725	**0.5362**	0.5637	0.4837	0.4900	0.1725
M1: Gemma - M2: Llama	0.5350	0.5450	0.9062	0.5612	**0.7500**	0.5887	0.5875	0.5687
M1: Gemma - M2: Qwen	0.4887	0.4900	0.9500	0.4987	**0.5287**	0.5212	0.5150	0.4750
M1: Qwen - M2: Llama	0.5350	0.5450	0.9062	0.5613	**0.9337**	0.5750	0.5062	0.5825
M1: Qwen - M2: Gemma	0.4887	0.5037	1.0000	0.7125	**0.9450**	0.5662	0.4912	0.5850
StrategyQA (2 labels)								
M1: Llama - M2: Gemma	0.6275	0.62375	0.9812	0.6850	**0.8637**	0.6112	0.6787	0.5762
M1: Llama - M2: Qwen	0.4575	0.4550	0.7575	0.4550	0.4537	**0.5287**	0.4675	0.4500
M1: Gemma - M2: Llama	0.6450	0.6837	0.5663	**0.7875**	0.7662	0.6025	0.6162	0.6487
M1: Gemma - M2: Qwen	0.4575	0.4550	0.7575	0.4550	0.4775	**0.5562**	0.4650	0.4300
M1: Qwen - M2: Llama	0.6450	0.6837	0.5662	0.7875	**0.8762**	0.6375	0.5862	0.5150
M1: Qwen - M2: Gemma	0.6275	0.6237	0.9812	0.6850	**0.8487**	0.6037	0.5750	0.5050

Table 2: Accuracy of models across the four datasets and explanation types, using explanations generated by the another model ($M_1 \neq M_2$). The best-performing explanation type for each model-dataset combination is boldfaced.

Limitations

The limitations of our studies include the following.

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 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. Future work should explore additional features, such as explanation length and syntactic complexity, to better understand their interplay with model performance.

Finally, 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.

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.

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.

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

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.

Lisa Anne Hendricks et al. 2016. Generating visual explanations. In *European Conference on Computer Vision*, pages 3–19. Springer.

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*.

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*.

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.

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.

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.

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

Gemini Team. 2023. Gemini: A family of highly capable multimodal models.

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.

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.

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

A Appendix

Correlation of Implicitness measures

Implicitness Correlation Results with ACCURACY using Explanation								
Dataset	Model 1	Model 2	Explanation	Corr. REL (1)	Corr. REL (2)	Corr. NOV (1)	Corr. NOV (2)	R-squared
SQA	qwen	llama	dummy	0.106	-0.010	0.019	0.007	0.014
SQA	qwen	llama	gold	0.002	0.091	-0.040	-0.067	0.011
SQA	qwen	llama	why	0.057	0.040	-0.063	-0.065	0.005
SQA	qwen	llama	cot	0.027	0.019	0.035	0.039	0.004
SQA	qwen	llama	phr	-0.014	0.025	-0.043	-0.052	0.004
SQA	qwen	llama	cf	0.015	0.032	-0.034	-0.035	0.002
SQA	qwen	gemma	dummy	0.097	0.078	0.054	0.056	0.020
SQA	qwen	gemma	gold	0.018	-0.019	-0.089	0.012	0.011
SQA	qwen	gemma	why	0.083	0.035	-0.109	-0.105	0.014
SQA	qwen	gemma	cot	0.008	0.028	-0.030	-0.026	0.002
SQA	qwen	gemma	phr	0.017	0.087	-0.076	-0.077	0.010
SQA	qwen	gemma	cf	0.021	0.040	-0.043	-0.043	0.002
ERTEIT	qwen	llama	dummy	0.132	0.337	-0.139	-0.071	0.117
ERTEIT	qwen	llama	gold	0.298	0.339	-0.103	-0.666	0.462
ERTEIT	qwen	llama	why	0.143	0.121	-0.134	-0.127	0.033
ERTEIT	qwen	llama	cot	-0.037	0.161	-0.140	-0.181	0.035
ERTEIT	qwen	llama	phr	0.008	0.124	-0.022	-0.137	0.031
ERTEIT	qwen	llama	cf	0.071	-0.075	-0.028	0.096	0.042
ERTEIT	qwen	gemma	dummy	0.123	0.333	-0.158	-0.061	0.113
ERTEIT	qwen	gemma	gold	0.254	0.434	-0.051	-0.740	0.574
ERTEIT	qwen	gemma	why	0.148	0.130	-0.141	-0.133	0.036
ERTEIT	qwen	gemma	cot	-0.027	0.213	-0.184	-0.239	0.059
ERTEIT	qwen	gemma	phr	0.001	0.135	-0.019	-0.137	0.032
ERTEIT	qwen	gemma	cf	0.044	-0.038	-0.041	0.064	0.027
ESNLI	qwen	llama	dummy	0.119	0.456	-0.269	0.006	0.289
ESNLI	qwen	llama	gold	0.044	0.157	-0.299	-0.330	0.156
ESNLI	qwen	llama	why	0.095	0.037	-0.177	-0.125	0.041
ESNLI	qwen	llama	cot	-0.191	0.069	-0.109	-0.117	0.046
ESNLI	qwen	llama	phr	-0.128	0.179	-0.213	-0.252	0.093
ESNLI	qwen	llama	cf	0.088	-0.302	0.163	0.247	0.123
ESNLI	qwen	gemma	dummy	0.266	0.530	-0.164	-0.038	0.294
ESNLI	qwen	gemma	gold	0.185	0.209	-0.063	-0.262	0.084
ESNLI	qwen	gemma	why	0.059	0.057	-0.119	-0.085	0.016
ESNLI	qwen	gemma	cot	-0.166	0.126	-0.080	-0.147	0.053
ESNLI	qwen	gemma	phr	-0.115	0.185	-0.202	-0.220	0.079
ESNLI	qwen	gemma	cf	0.053	-0.274	0.212	0.268	0.107
ERTEIT	qwen	llama	dummy	0.132	0.337	-0.139	-0.071	0.117
ERTEIT	qwen	llama	gold	0.298	0.339	-0.103	-0.666	0.462
ERTEIT	qwen	llama	why	0.143	0.121	-0.134	-0.127	0.033
ERTEIT	qwen	llama	cot	-0.037	0.161	-0.140	-0.181	0.035
ERTEIT	qwen	llama	phr	0.008	0.124	-0.022	-0.137	0.031
ERTEIT	qwen	llama	cf	0.071	-0.075	-0.028	0.096	0.042
ERTEIT	qwen	gemma	dummy	0.123	0.333	-0.158	-0.061	0.113
ERTEIT	qwen	gemma	gold	0.254	0.434	-0.051	-0.740	0.574
ERTEIT	qwen	gemma	why	0.148	0.130	-0.141	-0.133	0.036
ERTEIT	qwen	gemma	cot	-0.027	0.213	-0.184	-0.239	0.059
ERTEIT	qwen	gemma	phr	0.001	0.135	-0.019	-0.137	0.032
ERTEIT	qwen	gemma	cf	0.044	-0.038	-0.041	0.064	0.027

Table 3: Correlation of implicit measures with accuracy change using the explanation across the four datasets and explanation types, using explanations generated by Gwen and predictions of all three models.

Dataset	Model 1	Model 2	Explanation	Corr. REL (1)	Corr. REL (2)	Corr. NOV (1)	Corr. NOV (2)	R-squared
SQA	qwen	llama	dummy	0.106	-0.010	0.019	0.007	0.015
SQA	qwen	llama	gold	0.002	0.091	-0.040	-0.067	0.009
SQA	qwen	llama	why	0.057	0.040	-0.063	-0.065	0.010
SQA	qwen	llama	cot	0.027	0.019	0.035	0.039	0.011
SQA	qwen	llama	phr	-0.014	0.025	-0.043	-0.052	0.003
SQA	qwen	llama	cf	0.015	0.032	-0.034	-0.035	0.006
SQA	qwen	gemma	dummy	0.097	0.078	0.054	0.056	0.004
SQA	qwen	gemma	gold	0.018	-0.019	-0.089	0.012	0.002
SQA	qwen	gemma	why	0.083	0.035	-0.109	-0.105	0.013
SQA	qwen	gemma	cot	0.008	0.028	-0.030	-0.026	0.003
SQA	qwen	gemma	phr	0.017	0.087	-0.076	-0.077	0.017
SQA	qwen	gemma	cf	0.021	0.040	-0.043	-0.043	0.005
ERTEIT	qwen	llama	dummy	0.132	0.337	-0.139	-0.071	0.009
ERTEIT	qwen	llama	gold	0.298	0.339	-0.103	-0.666	0.129
ERTEIT	qwen	llama	why	0.143	0.121	-0.134	-0.127	0.044
ERTEIT	qwen	llama	cot	-0.037	0.161	-0.140	-0.181	0.023
ERTEIT	qwen	llama	phr	0.008	0.124	-0.022	-0.137	0.009
ERTEIT	qwen	llama	cf	0.071	-0.075	-0.028	0.096	0.032
ERTEIT	qwen	gemma	dummy	0.123	0.333	-0.158	-0.061	0.004
ERTEIT	qwen	gemma	gold	0.254	0.434	-0.051	-0.740	0.078
ERTEIT	qwen	gemma	why	0.148	0.130	-0.141	-0.133	0.005
ERTEIT	qwen	gemma	cot	-0.027	0.213	-0.184	-0.239	0.037
ERTEIT	qwen	gemma	phr	0.001	0.135	-0.019	-0.137	0.009
ERTEIT	qwen	gemma	cf	0.044	-0.038	-0.041	0.064	0.022
ESNLI	qwen	llama	dummy	0.119	0.456	-0.269	0.006	0.032
ESNLI	qwen	llama	gold	0.044	0.157	-0.299	-0.330	0.079
ESNLI	qwen	llama	why	0.095	0.037	-0.177	-0.125	0.096
ESNLI	qwen	llama	cot	-0.191	0.069	-0.109	-0.117	0.031
ESNLI	qwen	llama	phr	-0.128	0.179	-0.213	-0.252	0.078
ESNLI	qwen	llama	cf	0.088	-0.302	0.163	0.247	0.110
ESNLI	qwen	gemma	dummy	0.266	0.530	-0.164	-0.038	0.103
ESNLI	qwen	gemma	gold	0.185	0.209	-0.063	-0.262	0.002
ESNLI	qwen	gemma	why	0.059	0.057	-0.119	-0.085	0.017
ESNLI	qwen	gemma	cot	-0.166	0.126	-0.080	-0.147	0.038
ESNLI	qwen	gemma	phr	-0.115	0.185	-0.202	-0.220	0.047
ESNLI	qwen	gemma	cf	0.053	-0.274	0.212	0.268	0.061
ERTEIT	qwen	llama	dummy	0.132	0.337	-0.139	-0.071	0.009
ERTEIT	qwen	llama	gold	0.298	0.339	-0.103	-0.666	0.129
ERTEIT	qwen	llama	why	0.143	0.121	-0.134	-0.127	0.044
ERTEIT	qwen	llama	cot	-0.037	0.161	-0.140	-0.181	0.023
ERTEIT	qwen	llama	phr	0.008	0.124	-0.022	-0.137	0.009
ERTEIT	qwen	llama	cf	0.071	-0.075	-0.028	0.096	0.032
ERTEIT	qwen	gemma	dummy	0.123	0.333	-0.158	-0.061	0.004
ERTEIT	qwen	gemma	gold	0.254	0.434	-0.051	-0.740	0.078
ERTEIT	qwen	gemma	why	0.148	0.130	-0.141	-0.133	0.005
ERTEIT	qwen	gemma	cot	-0.027	0.213	-0.184	-0.239	0.037
ERTEIT	qwen	gemma	phr	0.001	0.135	-0.019	-0.137	0.009
ERTEIT	qwen	gemma	cf	0.044	-0.038	-0.041	0.064	0.022

Table 4: Correlation of implicit measures with accuracy change (from acc. without using the explanation to acc. using the explanation) across the four datasets and explanation types, using explanations generated by Gwen and predictions of all three models.