Retrieving Versus Understanding Extractive Evidence in Few-Shot Learning

Karl Elbakian[1], **Samuel Carton**[1]

[1]University of New Hampshire
karl.elbakian@unh.edu, samuel.carton@unh.edu

Abstract

A key aspect of alignment is the proper use of within-document evidence to construct document-level decisions. We analyze the relationship between the retrieval and interpretation of within-document evidence for large language models in a few-shot setting. Specifically, we measure the extent to which model prediction errors are associated with evidence retrieval errors with respect to gold-standard human-annotated extractive evidence for five datasets, using two popular closed proprietary models. We perform two ablation studies to investigate when both label prediction and evidence retrieval errors can be attributed to qualities of the relevant evidence. We find that there is a strong empirical relationship between model prediction and evidence retrieval error, but that evidence retrieval error is mostly not associated with evidence interpretation error–a hopeful sign for downstream applications built on this mechanism.

Code — https://github.com/kelbakian/llm-rationale-fidelity

1 Introduction

AI alignment refers to the goal of ensuring that model output is aligned with human intents and values (Shen et al. 2023; Anwar et al. 2024; Shen et al. 2024). One key element of alignment is **verification**, the ability to confirm that a model's predictions have indeed accorded with those intents and values. When we use models to automate or assist human-affecting tasks such as moderation (Kumar, AbuHashem, and Durumeric 2024), resume screening (Gan, Zhang, and Mori 2024), grading (Pinto et al. 2023), or medical decision-making (Thirunavukarasu et al. 2023), we want a human auditor to be able to review their decisions for mistakes or pathologies of behavior such as bias or the use of spurious evidence. Verification has traditionally been one of the major goals of model interpretability (Fok and Weld 2023). Implicitly, the assumption underlying this function is that it is easier for a human auditor to catch model mistakes at the explanation level and propagate them upward to an appropriate skepticism about the model's overall prediction, than to inspect that prediction alone.

With the rise of large language models (LMs), the discourse on AI interpretability has turned towards methods that take advantage of their emergent capabilities. Free-text explanations can provide clear, comprehensible, and interactive descriptions of why an LM made a certain prediction (Singh et al. 2024), while work in LM reasoning such as Chain-of-Thought (Wei et al. 2022) and its descendants force the model to break its reasoning into discrete steps which can be individually inspected. What unites these approaches is that they are **abstractive**, synthesizing the raw input into a concise and human-comprehensible form.

By contrast, **extractive** approaches to interpretability such as rationale models (Lei, Barzilay, and Jaakkola 2016) or LIME (Ribeiro, Singh, and Guestrin 2016) have fallen somewhat out of vogue because of their inexpressiveness compared to abstractive approaches (Siegel et al. 2024; Hu et al. 2023), as well as the parametric and/or computational overhead they add to models which are already large, unwieldy, and in some cases accessible via API only. However, some recent work has examined the ability of LMs to generatively mimic extractive explanations, or "self-rationalize", via prompting, finding them of comparable quality to traditional methods (Huang et al. 2023).

However, when we want to verify an LM's prediction over an input document, there are basic questions that are more appropriately answered by extractive approaches than abstractive ones. Namely: **what evidence from the input document is the model using as the basis for its decision, and does that evidence support its predicted label?** Fully abstractive explanations don't directly elucidate the relationship between the within-text evidence and label, and can in fact obfuscate it. Figure 1 shows a simple example where abstractive explanations of a moderation decision are inadequate for refuting it.

The problem of verification is related to the problem of faithfulness (Jacovi and Goldberg 2020), the idea that a model's explanation should be coupled with, and thus display the true underlying logic of, its prediction. Prompt-based explanations, abstractive or extractive, are intrinsically unfaithful by this definition, as they are just a proximal generation by the model alongside the label (Lyu et al. 2023; Turpin et al. 2023). For the purpose of verification, however, we are less interested in answering "what does the explanation tell us about the model's prediction?" than "what does the correctness of the model's explanation tell us about the correctness of its prediction?" Analogous to the concept of

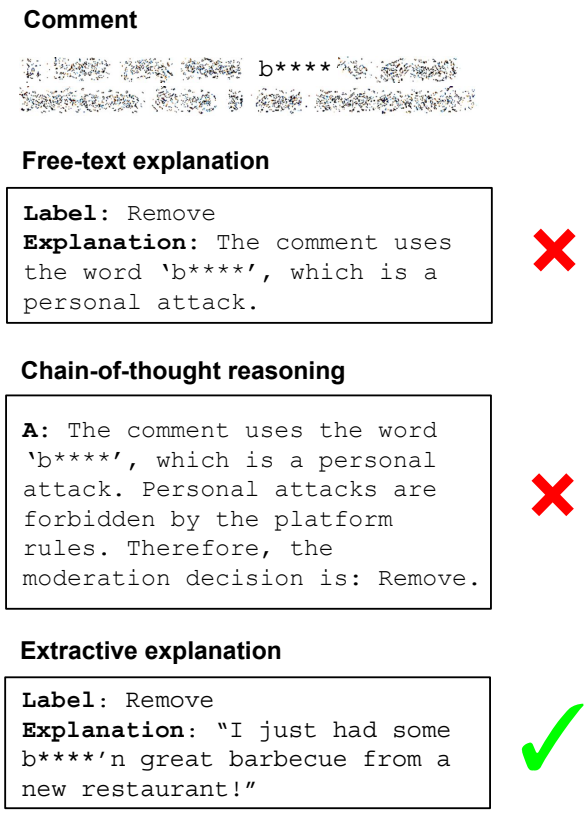

Figure 1: Artificial examples of abstractive and extractive explanations for an erroneous moderation prediction. Only the extractive explanation provides a basis for refuting it.

internal vs. external validity in experimental design (McDermott 2011), these two questions can be viewed as **internal vs. external faithfulness**, respectively.

In this paper we examine the external faithfulness of prompt-based self-rationalization in two prominent large proprietrary language models: GPT-4 (OpenAI 2023) and Gemini (Team et al. 2023). Our high-level research question is: **Can large language models reliably and meaningfully identify within-document evidence for their predictions?** With an eye towards developing downstream human-model interaction systems based on this mechanism, we investigate the following specific questions: (1) Can LMs reliably quote evidence snippets from an input document without mistakes?; (2) Does forcing self-rationalization impact model accuracy?; (3) Does operation order (explain-then-predict versus predict-then-explain) affect performance?; (4) Are label prediction errors correlated with evidence retrieval errors, relative to human gold-standard evidence?; (5) What types of evidence retrieval errors cause label prediction errors?; (6) Under what circumstances do LMs fail to retrieve evidence? These last two questions are of key importance for the design of any human-model collaborative system based on extractive evidence as an underlying mechanism. If the characteristic failure mode for the model is to identify the correct pertinent evidence and then simply to misinterpret it, this is much more correctable, via either prompt engineering approaches such as self-consistency (Wang et al. 2023) or human inspection, than if errors stem from missing key evidence entirely. We focus exclusively on **extractive** self-rationalization for the reasons outlined above: they are easier to assess for correctness (and thus faithfulness), and invite fewer pitfalls as a verification mechanism.

We experiment with five datasets for which gold standard human-annotated extractive evidence is available, covering a wide range of tasks: MultiRC (Khashabi et al. 2018), SciFact (Wadden et al. 2022), WikiAttack (Carton, Mei, and Resnick 2018), Evidence Inference (DeYoung et al. 2020), and HealthFC (Vladika, Schneider, and Matthes 2024). For a sample of each dataset, we run experiments prompting the model to self-rationalize label predictions under varying conditions, comparing and contrasting the result to the gold-standard evidence in each dataset. We find broadly that (1) these models **can** reliably quote within-text evidence; (2) self-rationalization mostly does **not** effect label accuracy, (3) operation order does **not** matter; (4) label prediction error **is** highly correlated with evidence error for most datasets; (5) label error is more commonly linked with capturing confounding evidence rather than missing key evidence; and finally (6) missing key evidence is mostly commonly linked to the presence of redundant evidence rather than more intractable interpretation issues. All of these are positive outcomes, suggesting the potential for LM extractive self-rationalization as a mechanism for powering downstream applications. All code and experimental results can be found in our github repository.

2 Related Work

Recent interpretability work has tended to focus on abstractive approaches such as posthoc free-text explanations (Singh et al. 2024; Zhu et al. 2024), or explanations as a byproduct of explicit reasoning processes like Chain-of-Thought (Wei et al. 2022; Lanham et al. 2023). However, one of the major goals of interpretability is verification (Fok and Weld 2023), the implicit assumption being that it is easier to recognize an erroneous explanation than an erroneous label. This mode requires explanations to be faithful (Jacovi and Goldberg 2020) to the overall prediction, but this quality is difficult to measure in abstractive approaches (Agarwal, Tanneru, and Lakkaraju 2024; Siegel et al. 2024).

Extractive approaches are less problematic in this regard because they directly attribute the prediction to evidence within the input. Even if this evidence is not technically faithful to the model's prediction, an observer can still assess whether it truly supports the label without being potentially beguiled by misleading abstractive generations of the model. Traditional approaches to identifying extractive evidence, such as the rationale model architecture (Lei, Barzilay, and Jaakkola 2016) or the LIME perturbation method (Ribeiro, Singh, and Guestrin 2016), add impractical levels of computational overhead to already large models, so recent work has investigated whether LMs can be prompted to produce such attributions as a generative output (Huang et al. 2023; Hu et al. 2023; Majumder et al. 2022). An especially relevant recent work is (Madsen, Chandar, and Reddy

2024), which investigates the internal faithfulness of several types of extractive explanations produced by three open LMs. Our approach follows this work in directly prompting LMs to produce extractive evidence for their predictions. To assess whether LMs can identify the "correct" evidence, we use datasets with gold-standard evidence annotations. The ERASER collection (DeYoung et al. 2019) includes a number of these datasets, and (Wiegreffe and Marasović 2021) surveys yet more.

Finally, our goal of retrieving relevant evidence buried in potentially long documents is similar to that of"needle-in-the-haystack" evaluations (Dhinakaran and Jolly 2024), which ask questions about evidence manually inserted into a long context. Where we differ is in applying this approach to naturally-occurring evidence that the model may not be able to properly interpret, rather than artificially-inserted evidence the model is assumed to be able to comprehend if it can find it.

3 Datasets

We analyze five datasets for which gold-standard human-annotated extractive evidence is included alongside ground truth labels: MultiRC (Khashabi et al. 2018), SciFact (Wadden et al. 2022), and WikiAttack (Carton, Mei, and Resnick 2018), Evidence Inference (DeYoung et al. 2020), and HealthFC (Vladika, Schneider, and Matthes 2024). To reduce API access costs, we perform our analysis on randomly-sampled 300-item subsets of the development set for each dataset.

SciFact (SF) (Wadden et al. 2022) is a scientific claim verification dataset, involving identifying whether abstracts from the research literature either support or refute a given scientific claim. The dataset contains 1,400 expert-written claims, paired with evidence-containing abstracts annotated with veracity labels and sentence-level rationales. Rationales are direct spans of text from the document which support the class label given. Each claim has a class label of *SUPPORT*, *NO INFO*, or *CONTRADICT*.

MultiRC (MRC) (Khashabi et al. 2018) is a reading comprehension dataset containing sets of short paragraphs and questions that depend on information from multiple sentences within the paragraph. Each paragraph-question pair contains five answers, with a variable number of correct answer-options. Additionally, answer-options do not have to be a span from the text. The dataset contains sentence-level rationales in the form of relevant sections of the paragraph, as well as *True* or *False* labels for each answer candidate.

WikiAttack (WA) (Carton, Mei, and Resnick 2018) contains excerpts of Wikipedia comment threads original included in the WikiToxic dataset of (Wulczyn, Thain, and Dixon 2017) which are scanned for instances of personal attacks or harassment. Class labels are given as either *attack* or *nonattack*, and rationales for comments labeled as attack are given in the form of spans from the comment.

Evidence Inference (EI) (DeYoung et al. 2020) dataset contains scientific articles and queries asking to compare

A) Predict-then-explain prompt

```
Based on the document between the <doc> tags,
classify the claim between the <claim> tags as
SUPPORT, CONTRADICT, or NOINFO, with respect to
the document. Respond in json format, with a
'label' field for the classification and an
'explanation' field listing the parts of the
document that support the label.
```

{{NOINFO, CONTRADICT, and SUPPORT exemplars}}

<doc>**Genetically identical cells sharing an environment can display markedly different phenotypes**... allows a trial commitment to multicellularity that external signals could extend.</doc>
<claim>**Gene expression does not vary appreciably across genetically identical cells.**</claim>

```
{"label":"NOINFO", "explanation":[]}
```

B) Evidence-given w/o document prompt

{{Prompt instructions}}

<doc>**Genetically identical cells sharing an environment can display markedly different phenotypes.**</doc>
<claim>**Gene expression does not vary appreciably across genetically identical cells.**</claim>

```
{"label": "CONTRADICT"}
```

Figure 2: Two examples of GPT-4 prompts on the same SciFact item. Model output highlighted in green. Human-annotated evidence underlined, claim bolded. The model misses the evidence and mislabels the document in the predict-then-explain setting, but correctly labels it when the evidence is provided, even without its surrounding context.

the relative effectiveness of two treatments with respect to a given outcome. Labels include *significantly increased*, *significantly decreased*, and *no significant difference*. Annotations are in the form of quotes from the text which support the label.

HealthFC (HFC) (Vladika, Schneider, and Matthes 2024) is a medical domain QA dataset. Given a health-related claim, verdicts are decided from a systematic review or clinical trial document, as one of *supported*, *refuted*, and *not enough information*. Additionally, annotations containing up to five quotes from the document are provided. After reviewing the dataset, we found discrepancies between the original, expert-annotated, German verdicts, and their English label mappings. Thus, we performed a re-labeling of the dataset, first by prompting GPT-4 to generate an English label given the query and German verdict, and then by manually labeling each item in the dataset, using a holistic review of the original English label, German verdict, cited evidence, and full document text. The adjusted labels are included in our repository.

Dataset	Failure Rate (%)	
	GPT-4	Gemini-1.5
MRC	2.00	2.33
SF	0.33	0.00
WA	0.00	4.33
EI	40.33	25.0
HFC	15.33	0.33

Table 1: Model evidence quoting failure rates in the explain-then-predict condition

4 Implementation Details

Model details. Because of their to-date superior few-shot performance compared to open models, we focus on two closed proprietary models: GPT-4 and Gemini. For GPT-4, the *gpt-4-0613* version is used for MultiRC, WikiAttack, and SciFact. Because of long input lengths, *gpt-4-turbo* is used for Evidence Inference and HealthFC. All GPT model temperatures are set to the default 0.7. For Gemini, the *gemini 1.5 pro* model version is used for all datasets, with temperature set to the default 1.0

Exemplar sampling. All few-shot conditions involve sampling two exemplars for each possible class, resulting in 4-shot learning for MultiRC and WikiAttack, and 6-shot learning for SciFact, Evidence Inference, and HealthFC. We randomly sample and shuffle exemplars independently for each item. For MultiRC, SciFact, Evidence Inference, and HealthFC, exemplars are sampled from the training set. For WikiAttack, where human rationales are unavailable for the training set, they are instead sampled from the development set, resulting in effectively 250-fold cross-validation.

Experimental conditions and prompting. We compare model performance across 7 different prompting conditions: (1) **zero-shot**, with neither exemplars nor self-rationalization; (2) standard **few-shot** with no rationalization; (3) **predict then explain**, where the model is asked to first make a label prediction then provide evidence to justify its answer; (4) **explain then predict**, where the model is asked the reverse order of the former condition; (5) **evidence given**, in which the human-annotated evidence is provided as part of the input prompt; (6) **evidence given without document**, where the human annotated evidence is provided without the context of the surrounding document; and (7) **evidence occluded**, where the human-annotated evidence is removed without replacement from the document.

Fig. 2 shows example prompts for the "predict-then-explain" and "evidence given without document" conditions. Model output is solicited in `JSON` format using `<claim>` and `<doc>` tags to denote different elements of the input. Prompt scaffolds can be found in our repository.

5 Results

5.1 Can language models reliably quote evidence from the input document?

Table 1 shows self-rationalization failure rates across the explain-then-predict condition, where the model fails to return any valid quote from an input text, using string substring matching. We observe mostly very low error rates, with the two high error rates for GPT-4 are associated with the `GPT-4 turbo` variant model, which is known to be weaker than the base model. Gemini also shows a high rate of error for Evidence Inference. The results show that, by and large, the strongest contemporary models can reliably quote evidence from the input document. Evidence Inference is the outlier, possibly due to long context lengths.

Dataset	Condition	Label accuracy	
		GPT-4	Gemini-1.5
MRC	Zero-shot	0.887	0.840
	Few-shot	0.893	0.892
	Predict then explain	0.897	0.885
	Explain then predict	0.861	0.874
	Evidence given	0.887	0.885
	Evidence given w/o document	0.827	0.797
	Evidence occluded	0.663	0.666
SF	Zero-shot	0.853	0.850
	Few-shot	0.843	0.820
	Predict then explain	0.870	0.837
	Explain then predict	0.833	0.841
	Evidence given	0.947	0.947
	Evidence given w/o document	0.907	0.890
	Evidence occluded	0.633	0.587
WA	Zero-shot	0.743	0.674
	Few-shot	0.777	0.703
	Predict then explain	0.790	0.728
	Explain then predict	0.743	0.690
	Evidence given	0.813	0.832
	Evidence given w/o document	0.917	0.901
	Evidence occluded	0.533	0.563
EI	Zero-shot	0.832	0.765
	Few-shot	0.809	0.767
	Predict then explain	0.870	0.839
	Explain then predict	0.888	0.831
	Evidence given	0.907	0.880
	Evidence given w/o document	0.920	0.893
	Evidence occluded	0.721	0.698
HFC	Zero-shot	0.763	0.803
	Few-shot	0.733	0.787
	Predict then explain	0.835	0.809
	Explain then predict	0.783	0.819
	Evidence given	0.947	0.857
	Evidence given w/o document	0.850	0.863
	Evidence occluded	0.677	0.687

Table 2: Label Accuracy for different prompting paradigms

Dataset	Model	Correct predictions				Incorrect predictions			
		Count	Evidence			Count	Evidence		
			F1	Recall	Prec.		F1	Recall	Prec.
MRC	GPT-4	253	0.71	0.68	0.84	41	0.59**	0.62	0.63***
	Gemini-1.5	256	0.67	0.63	0.86	37	0.67	0.53	0.67
SF	GPT-4	249	0.76	0.77	0.84	50	0.11***	0.75	0.12***
	Gemini-1.5	252	0.80	0.77	0.86	48	0.14***	0.60	0.45
WA	GPT-4	223	0.85	0.94	0.85	77	0.15***	0.92	0.18***
	Gemini-1.5	198	0.82	0.89	0.87	89	0.16***	0.90	0.23
EI	GPT-4	159	0.62	0.62	0.69	20	0.38***	0.33***	0.48**
	Gemini-1.5	187	0.83	0.61	0.63	38	0.81	0.41	0.39
HFC	GPT-4	199	0.37	0.38	0.44	55	0.29	0.29	0.36
	Gemini-1.5	245	0.53	0.38	0.48	54	0.53	0.35	0.36

Table 3: Evidence rationalization performance for correct vs incorrect label predictions in the *explain then predict* condition. Asterisks denote significant differences between values using a 2-sided t-test; *:p<0.1; **:p<0.05; ***p<0.005

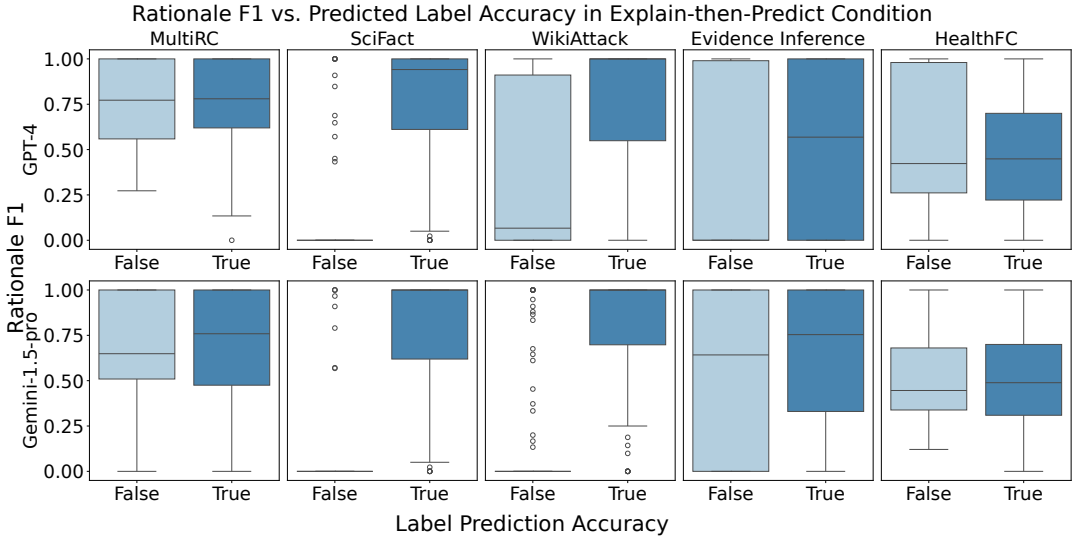

Figure 3: Box-and-whisker plots of label prediction error versus mean predicted evidence rationale F1 for all five datasets in the *explain then predict* condition.

5.2 Does self-rationalization impact model accuracy?

Table 2 summarizes the overall label prediction accuracy results for all conditions. For two datasets, Evidence Inference and HealthFC, forcing the model to support its prediction has a substantial positive effect (+7.9/5.0 for GPT-4 and +6.4/3.2 for Gemini-1.5, respectively). This effect is marginal for the other three datasets, even sometimes having a negative effect. This effect interestingly is the opposite of the failure rate results, suggesting that longer documents benefit more from this prompting requirement.

5.3 Effect of prompt operation order

Table 2 also provides a comparison between asking the model to first rationalize its prediction (*explain then predict*) and then provide a label, versus rationalizing after making its label prediction (*predict then explain*). Using GPT-4, for MultiRC, SciFact, and HealthFC, changing the request order caused a significant difference in the label prediction performance while Gemini-1.5 showed no significant differences. Using GPT-4, the predictive performance difference is mostly in favor of the *predict then explain* condition, suggesting that in most cases constraining the model to preemptively justify its label has a slightly negative impact on its label accuracy.

5.4 Is evidence retrieval error correlated with label prediction error?

Table 3 shows the mean self-rationalization performance (compared to human-annotated evidence) for correct and incorrect label predictions made by each model, in the *explain then predict* condition. Figure 3 illustrates the distribution of rationale F1s under this condition.

For GPT-4, there is a strong correlation between evidence F1 and label accuracy in SciFact, WikiAttack, and Evidence Inference, with respective evidence-F1 differences of 0.65,

Explain then predict		Low evidence recall		High evidence recall	
Evidence given		Correct label	Incorrect label	Correct label	Incorrect label
Dataset	Model	"Retrieval issue"	"Unexplainable"	"Confounding evidence"	"Misinterpretation"
MRC	GPT-4	0.12 (5)	0.24 (10)	0.22 (9)	**0.41 (17)**
	Gemini-1.5	0.16 (6)	0.30 (11)	0.14 (5)	**0.41 (15)**
SF	GPT-4*	0.10 (5)	0.14 (7)	**0.60 (30)**	0.16 (8)
	Gemini-1.5*	0.19 (9)	0.21 (10)	**0.50 (24)**	0.10 (5)
WA	GPT-4	0.00 (0)	0.06 (5)	0.34 (26)	**0.60 (46)**
	Gemini-1.5	0.06 (5)	0.05 (4)	**0.45 (40)**	0.44 (39)
HFC	GPT-4*	**0.73 (40)**	0.07 (4)	0.13 (7)	0.07 (4)
	Gemini-1.5	**0.44 (24)**	0.35 (19)	0.11 (6)	0.09 (5)
EI	GPT-4	0.26 (5)	**0.32 (6)**	0.16 (3)	0.26 (5)
	Gemini-1.5*	**0.39 (15)**	0.16 (6)	0.16 (6)	0.29 (11)

Table 4: Contingency table of label prediction errors in *explain then predict* condition; split by low vs. high evidence recall and correct vs. incorrect prediction in *explanation given* condition. Asterisks denote significant differences via the Fisher exact test with p-value 0.05. An interpretation of each contingency is presented in quotes.

0.70, and 0.24. It is likewise strong for Gemini-1.5 in SciFact and WikiAttack, with evidence-F1 differences of 0.66 for both. With respect to SciFact, hardly any incorrect predictions display any level of correct alignment. Again, this effect is consistent across both models.

This result suggests that for certain datasets (SciFact and WikiAttack), the primary challenge is in retrieving the correct evidence, and when this can be done the label can be predicted with high accuracy. There are others (MultiRC, Evidence Inference in the case of Gemini), where evidence retrieval is successful, and the challenge lies more in interpreting the correct answer from that evidence. HealthFC is an outlier, with both models performing poorly on evidence retrieval, for both correct and incorrect label predictions.

5.5 Why does the model make prediction errors?

When the model makes prediction errors, to what extent can we attribute these errors to missing evidence or misinterpretation of correct evidence? Is it more common for the model to miss key evidence entirely (which is difficult to correct via prompt engineering), or to misinterpret relevant evidence (which can be addressed by reinspection approaches)? Table 4 explores these questions by breaking down label prediction errors made in the *explain then predict* condition by (1) the evidence recall relative to human-annotated evidence in this condition, and (2) the model's label accuracy on those corresponding items on the *evidence given* condition. This division lets us ask: **given that the model was wrong, did it identify the correct evidence, and would it still have been wrong if it had done so?**

In this division, the (low-recall, correct label) condition represents cases where the model's label prediction mistake would have been overturned by presenting it with the human rationale, meaning that we can attribute the mistake to the model having failed to retrieve the correct evidence. The (low-recall, incorrect label) condition represents cases where the model's mistake persists with or without the human rationale, making it impossible to attribute. The (high-recall, correct label) condition implies that there was extra confounding evidence the model picked up in addition to recovering the human rationale, which caused its label prediction error. Finally, the (high-recall, incorrect label) condition represents scenarios where the model successfully recovered the human rationale, but made the incorrect prediction regardless (misinterpretation).

We find that in three of out five datasets (MultiRC, SciFact, and WikiAttack), label prediction errors are mostly associated with high rationale recall (misinterpretation and confounding evidence) rather than low recall (missing evidence). In MultiRC and WikiAttack, the model is most likely to recover the human-annotated evidence but be unable to interpret it properly, while for SciFact the model is likely to recover both the human rationale and extraneous evidence, which causes it to produce an incorrect prediction. For two datasets, HealthFC and Evidence Inference, a majority of prediction errors are associated with low evidence recall. In both models' cases, providing human evidence to the model would produce a correct label for HealthFC. For Evidence Inference, this remains true only for Gemini-1.5. This means that HealthFC is the only dataset for which we can attribute a majority of errors to missing key evidence that the model had the capacity to interpret. In other cases, the model is more likely to recover the relevant evidence and then misinterpret it.

5.6 Why does the model miss key evidence?

Similar to Section 5.5, we can use the results of different prompting paradigms to ask why the model misses key evidence. In particular, we can investigate the following hypothesis: **the model tends to miss evidence primarily when it is unable to interpret it correctly.**

Table 5 represents a contingency table of low-evidence-recall instances from the *explain then predict* condition, divided by label prediction error in the *evidence occluded* and *evidence given* conditions. The (occluded incorrect, given correct) contingency represents cases where the human evidence was both necessary and sufficient for the model to correctly predict the label, while (occluded incorrect, given incorrect label) represents cases where the human evidence was not unnecessary, but also not sufficient to predict the

Evidence occluded		Label incorrect		Label correct	
Evidence given		Label correct	Label incorrect	Label correct	Label incorrect
Note	Model	"Retrieval issue"	"Uninterpretable evidence"	"Other viable evidence present"	
MRC	GPT-4*	0.22 (19)	0.14 (12)	**0.60 (52)**	0.05 (4)
	Gemini-1.5	0.26 (29)	0.05 (6)	**0.49 (55)**	0.20 (22)
SF	GPT-4*	0.37 (25)	0.10 (7)	**0.51 (34)**	0.01 (1)
	Gemini-1.5*	**0.42 (32)**	0.09 (7)	0.41 (31)	0.08 (6)
WA	GPT-4	**0.47 (7)**	0.20 (3)	0.13 (2)	0.20 (3)
	Gemini-1.5	**0.46 (13)**	0.04 (1)	0.32 (9)	0.18 (5)
HFC	GPT-4*	0.30 (45)	0.01 (2)	**0.65 (97)**	0.04 (6)
	Gemini-1.5*	0.21 (45)	0.06 (13)	**0.63 (133)**	0.09 (20)
EI	GPT-4*	0.16 (8)	0.08 (4)	**0.76 (38)**	0.00 (0)
	Gemini-1.5*	0.18 (11)	0.03 (2)	**0.77 (48)**	0.02 (1)

Table 5: Contingency table of low human rationale recall instances in *explain then predict* condition; split by correct vs. incorrect label prediction in *evidence occluded* and *evidence given* conditions. Asterisks denote significant differences via the McNemar test with p-value 0.05. Intepretations provided in quotes.

label. Finally, either contingency where the label prediction was correct for the explanation-occluded condition is one where there was additional viable evidence beyond the human-annotated evidence, rendering it unnecessary, and low evidence recall less of a real error. Thus, the question being asked here is: **given that the model missed key evidence, was that evidence necessary and would it have been sufficient if it had been found?**

If misses of necessary human-annotated evidence were mostly associated with them being uninterpretable by the model, a majority of cases would fall in the (occluded incorrect, given incorrect) contingency. Instead, Table 5 shows that in a majority of cases, datasets fall into the (occluded correct, given correct) contingency, meaning that models mostly fail to retrieve human-annotated evidence when there is additional viable evidence that the model can use to successfully predict the label.

Even when no such additional viable evidence is possible (incorrect label in the occluded condition), a majority of cases for all datasets are ones where the label is correct in the evidence-given condition, meaning the model does know how to correctly interpret the key evidence, it simply fails to retrieve it in the *explain then predict* condition. In WikiAttack alone does this contingency represent a majority of overall cases. Hence, the hypothesis that retrieval failures are associated with evidence the model cannot interpret correctly is not supported in our analysis.

6 Discussion

Consider a human-model collaborative system such as an interpretable LM-driven moderation system, grounded in the mechanism of extractive self-rationalization. For verification to be possible in such a system, the LM must be able to consistently extract explanatory evidence from the input document, and that evidence should display external faithfulness, in that if it can be proven incorrect then the model's prediction should also be incorrect as well. As error modes go, it is better for the model to identify correct evidence and then misinterpret it than to miss evidence entirely, as this is more correctable via prompting or human inspection. And if the model is to miss evidence, it is better for it to be able to interpret it correctly if provided, as this could be supported by additional within-document retrieval support e.g. (Singhania, Razniewski, and Weikum 2024). The most irrecoverable outcome is the model missing evidence that it cannot interpret properly at all.

The results shown above are largely supportive of these requirements. We find that models can mostly reliably quote evidence from the input, and that for at least some datasets, evidence retrieval performance is correlated with label prediction performance. We find that label errors are mostly associated with either confounding evidence or with missing evidence that could at least be interpreted correctly if it had been retrieved. Finally, when the model misses key evidence, we find it mostly associated with the presence of other viable evidence, meaning that it is not truly missing evidence at all in most apparent cases.

This is a hopeful outcome for any verification system based on self-rationalization. While the properties (and average document lengths) of the data in question has a major impact on the viability of this mechanism, there are at least some datasets for which it will tend to work well and potentially serve as the basis for such a system. Further work might develop an evaluation protocol to determine whether or not a given dataset falls into this category.

7 Conclusion

In this work, we investigate the relationship between prediction and extractive rationalization in few-shot learning. We find that there is a strong correlation between the classification accuracy and agreement with human-annotated rationales. Furthermore, we find that significantly more model error is attributable to imprecise rationalization than incomplete rationalization, a positive sign for downstream methods based on this mechanism.

References

Agarwal, C.; Tanneru, S. H.; and Lakkaraju, H. 2024. Faithfulness vs. Plausibility: On the (Un)Reliability of Explanations from Large Language Models. ArXiv:2402.04614 [cs].

Anwar, U.; Saparov, A.; Rando, J.; Paleka, D.; Turpin, M.; Hase, P.; Lubana, E. S.; Jenner, E.; Casper, S.; Sourbut, O.; Edelman, B. L.; Zhang, Z.; Günther, M.; Korinek, A.; Hernandez-Orallo, J.; Hammond, L.; Bigelow, E.; Pan, A.; Langosco, L.; Korbak, T.; Zhang, H.; Zhong, R.; hÉigeartaigh, S. ; Recchia, G.; Corsi, G.; Chan, A.; Anderljung, M.; Edwards, L.; Bengio, Y.; Chen, D.; Albanie, S.; Maharaj, T.; Foerster, J.; Tramer, F.; He, H.; Kasirzadeh, A.; Choi, Y.; and Krueger, D. 2024. Foundational Challenges in Assuring Alignment and Safety of Large Language Models. ArXiv:2404.09932 [cs].

Carton, S.; Mei, Q.; and Resnick, P. 2018. Extractive Adversarial Networks: High-Recall Explanations for Identifying Personal Attacks in Social Media Posts. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 3497–3507. Brussels, Belgium: Association for Computational Linguistics.

DeYoung, J.; Jain, S.; Rajani, N. F.; Lehman, E.; Xiong, C.; Socher, R.; and Wallace, B. C. 2019. ERASER: A Benchmark to Evaluate Rationalized NLP Models. *arXiv preprint*. ArXiv: 1911.03429.

DeYoung, J.; Lehman, E.; Nye, B.; Marshall, I. J.; and Wallace, B. C. 2020. Evidence Inference 2.0: More Data, Better Models. ArXiv:2005.04177 [cs].

Dhinakaran, A.; and Jolly, E. 2024. The Needle In a Haystack Test: Evaluating the Performance of LLM RAG Systems.

Fok, R.; and Weld, D. S. 2023. In Search of Verifiability: Explanations Rarely Enable Complementary Performance in AI-Advised Decision Making. ArXiv:2305.07722 [cs].

Gan, C.; Zhang, Q.; and Mori, T. 2024. Application of LLM Agents in Recruitment: A Novel Framework for Resume Screening. ArXiv:2401.08315 [cs].

Hu, X.; Hong, Z.; Zhang, C.; King, I.; and Yu, P. 2023. Think Rationally about What You See: Continuous Rationale Extraction for Relation Extraction. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '23, 2436–2440. New York, NY, USA: Association for Computing Machinery. ISBN 978-1-4503-9408-6.

Huang, S.; Mamidanna, S.; Jangam, S.; Zhou, Y.; and Gilpin, L. H. 2023. Can Large Language Models Explain Themselves? A Study of LLM-Generated Self-Explanations. ArXiv:2310.11207 [cs].

Jacovi, A.; and Goldberg, Y. 2020. Towards Faithfully Interpretable NLP Systems: How should we define and evaluate faithfulness? *arXiv:2004.03685 [cs]*. ArXiv: 2004.03685.

Khashabi, D.; Chaturvedi, S.; Roth, M.; Upadhyay, S.; and Roth, D. 2018. Looking Beyond the Surface: A Challenge Set for Reading Comprehension over Multiple Sentences. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 252–262. New Orleans, Louisiana: Association for Computational Linguistics.

Kumar, D.; AbuHashem, Y.; and Durumeric, Z. 2024. Watch Your Language: Investigating Content Moderation with Large Language Models. ArXiv:2309.14517 [cs].

Lanham, T.; Chen, A.; Radhakrishnan, A.; Steiner, B.; Denison, C.; Hernandez, D.; Li, D.; Durmus, E.; Hubinger, E.; Kernion, J.; Lukošiūtė, K.; Nguyen, K.; Cheng, N.; Joseph, N.; Schiefer, N.; Rausch, O.; Larson, R.; McCandlish, S.; Kundu, S.; Kadavath, S.; Yang, S.; Henighan, T.; Maxwell, T.; Telleen-Lawton, T.; Hume, T.; Hatfield-Dodds, Z.; Kaplan, J.; Brauner, J.; Bowman, S. R.; and Perez, E. 2023. Measuring Faithfulness in Chain-of-Thought Reasoning. ArXiv:2307.13702 [cs].

Lei, T.; Barzilay, R.; and Jaakkola, T. 2016. Rationalizing Neural Predictions. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 107–117. 571 citations (Semantic Scholar/arXiv) [2022-08-26].

Lyu, Q.; Havaldar, S.; Stein, A.; Zhang, L.; Rao, D.; Wong, E.; Apidianaki, M.; and Callison-Burch, C. 2023. Faithful Chain-of-Thought Reasoning. ArXiv:2301.13379 [cs].

Madsen, A.; Chandar, S.; and Reddy, S. 2024. Are self-explanations from Large Language Models faithful? ArXiv:2401.07927.

Majumder, B. P.; Camburu, O.-M.; Lukasiewicz, T.; and McAuley, J. 2022. Knowledge-Grounded Self-Rationalization via Extractive and Natural Language Explanations. ArXiv:2106.13876 [cs].

McDermott, R. 2011. Internal and external validity. *Cambridge handbook of experimental political science*, 27.

OpenAI. 2023. GPT-4 Technical Report. ArXiv:2303.08774 [cs].

Pinto, G.; Cardoso-Pereira, I.; Monteiro, D.; Lucena, D.; Souza, A.; and Gama, K. 2023. Large Language Models for Education: Grading Open-Ended Questions Using ChatGPT. In *Proceedings of the XXXVII Brazilian Symposium on Software Engineering*, SBES '23, 293–302. New York, NY, USA: Association for Computing Machinery. ISBN 9798400707872.

Ribeiro, M. T.; Singh, S.; and Guestrin, C. 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*, 1135–1144. ACM. ISBN 978-1-4503-4232-2. 1999 citations (Semantic Scholar/DOI) [2022-08-26].

Shen, H.; Knearem, T.; Ghosh, R.; Alkiek, K.; Krishna, K.; Liu, Y.; Ma, Z.; Petridis, S.; Peng, Y.-H.; Qiwei, L.; Rakshit, S.; Si, C.; Xie, Y.; Bigham, J. P.; Bentley, F.; Chai, J.; Lipton, Z.; Mei, Q.; Mihalcea, R.; Terry, M.; Yang, D.; Morris, M. R.; Resnick, P.; and Jurgens, D. 2024. Towards Bidirectional Human-AI Alignment: A Systematic Review for Clarifications, Framework, and Future Directions. ArXiv:2406.09264 [cs].

Shen, T.; Jin, R.; Huang, Y.; Liu, C.; Dong, W.; Guo, Z.; Wu, X.; Liu, Y.; and Xiong, D. 2023. Large Language Model Alignment: A Survey. ArXiv:2309.15025 [cs].

Siegel, N. Y.; Camburu, O.-M.; Heess, N.; and Perez-Ortiz, M. 2024. The Probabilities Also Matter: A More Faithful Metric for Faithfulness of Free-Text Explanations in Large Language Models. ArXiv:2404.03189 [cs].

Singh, C.; Inala, J. P.; Galley, M.; Caruana, R.; and Gao, J. 2024. Rethinking Interpretability in the Era of Large Language Models. ArXiv:2402.01761 [cs].

Singhania, S.; Razniewski, S.; and Weikum, G. 2024. Recall Them All: Retrieval-Augmented Language Models for Long Object List Extraction from Long Documents. ArXiv:2405.02732 [cs].

Team, G.; Anil, R.; Borgeaud, S.; Wu, Y.; Alayrac, J.-B.; Yu, J.; Soricut, R.; Schalkwyk, J.; Dai, A. M.; Hauth, A.; and others. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Thirunavukarasu, A. J.; Ting, D. S. J.; Elangovan, K.; Gutierrez, L.; Tan, T. F.; and Ting, D. S. W. 2023. Large language models in medicine. *Nature Medicine*, 29(8): 1930–1940. Publisher: Nature Publishing Group.

Turpin, M.; Michael, J.; Perez, E.; and Bowman, S. R. 2023. Language Models Don't Always Say What They Think: Unfaithful Explanations in Chain-of-Thought Prompting. ArXiv:2305.04388 [cs].

Vladika, J.; Schneider, P.; and Matthes, F. 2024. HealthFC: Verifying Health Claims with Evidence-Based Medical Fact-Checking. In Calzolari, N.; Kan, M.-Y.; Hoste, V.; Lenci, A.; Sakti, S.; and Xue, N., eds., *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 8095–8107. Torino, Italia: ELRA and ICCL.

Wadden, D.; Lo, K.; Kuehl, B.; Cohan, A.; Beltagy, I.; Wang, L. L.; and Hajishirzi, H. 2022. SciFact-Open: Towards open-domain scientific claim verification. ArXiv:2210.13777 [cs].

Wang, X.; Wei, J.; Schuurmans, D.; Le, Q.; Chi, E.; Narang, S.; Chowdhery, A.; and Zhou, D. 2023. Self-Consistency Improves Chain of Thought Reasoning in Language Models. ArXiv:2203.11171 [cs].

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Ichter, B.; Xia, F.; Chi, E.; Le, Q.; and Zhou, D. 2022. Chain of Thought Prompting Elicits Reasoning in Large Language Models. ArXiv:2201.11903 [cs].

Wiegreffe, S.; and Marasović, A. 2021. Teach Me to Explain: A Review of Datasets for Explainable Natural Language Processing. ArXiv:2102.12060 [cs].

Wulczyn, E.; Thain, N.; and Dixon, L. 2017. Ex Machina: Personal Attacks Seen at Scale. In *Proceedings of the 26th International Conference on World Wide Web*, 1391–1399. ISBN 978-1-4503-4913-0.

Zhu, Z.; Chen, H.; Ye, X.; Lyu, Q.; Tan, C.; Marasovic, A.; and Wiegreffe, S. 2024. Explanation in the Era of Large Language Models. In Zhang, R.; Schneider, N.; and Chaturvedi, S., eds., *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 5: Tutorial Abstracts)*, 19–25. Mexico City, Mexico: Association for Computational Linguistics.