Evaluating Human Alignment and Model Faithfulness of LLM Rationale

Anonymous ACL submission

Abstract

We study how well large language models (LLMs) explain their generations through rationales – a set of tokens extracted from the input text that reflect the decision-making process of LLMs. Specifically, we systematically study rationales derived using two approaches: (1) popular prompting-based methods, where prompts are used to guide LLMs in generating rationales, and (2) technical attribution-based methods, which leverage attention or gradients to identify important tokens. Our analysis spans three classification datasets with annotated rationales, encompassing tasks with varying performance levels. While prompting-based self-explanations are widely used, our study reveals that these explanations are not always as "aligned" with the human rationale as attribution-based explanations. Even more so, fine-tuning LLMs to enhance classification task accuracy does not enhance the alignment of prompting-based rationales. Still, it does considerably improve the alignment of attribution-based methods (e.g., Input×Gradient). More importantly, we show that prompting-based self-explanation is also less "faithful" than attribution-based explanations, failing to provide a reliable account of the model's decision-making process. To evaluate faithfulness, unlike prior studies that excluded misclassified examples, we evaluate all instances and also examine the impact of fine-tuning and accuracy on alignment and faithfulness. Our findings suggest that inconclusive faithfulness results reported in earlier studies may stem from low classification accuracy. These findings underscore the importance of more rigorous and comprehensive evaluations of LLM rationales.[1]

1 Introduction

The rise of large language models (LLMs) has significantly transformed the field of natural language processing (NLP) (Touvron et al., 2023; Team et al., 2023; OpenAI et al., 2024), enabling a wide range of applications from web question answering to complex reasoning tasks. However, if they cannot clearly and reliably explain their outputs (Ji et al., 2023), it limits their deployment in high-stakes scenarios.

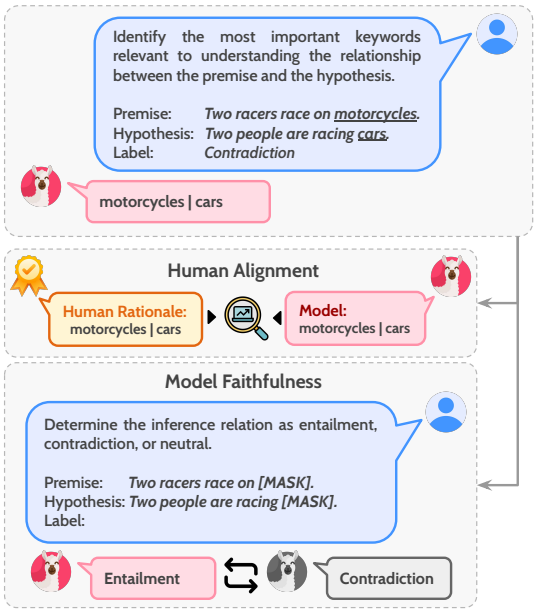

Figure 1: An example of our analysis methodology on the e-SNLI dataset. *Human alignment* compares model rationales with human-annotated rationale; *Model faithfulness* measures when model prediction changes (e.g. from `Contradiction` to `Entailment`) after masking the rationales identified by the model.[2]

Rationales[3], i.e., tokens of the input text that are most influential to the models' predictions, are widely studied in the NLP community prior to the era of LLMs to interpret model predictions (Lei et al., 2016; DeYoung et al., 2019; Wiegreffe and Pinter, 2019; Jacovi and Goldberg, 2020). For LLMs, besides attribution-based methods like at-

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

[2] In the first prompt, we use the true label for human alignment, and the predicted label for faithfulness experiments.

[3] Also called *self-explanation* or *extractive rationales* in previous work (Huang et al., 2023a; Madsen et al., 2024).

tention weights (Wiegreffe and Pinter, 2019) or gradients (Li et al., 2016), rationales can also be extracted by leveraging the instruction-following ability of LLMs and guiding them with explicit prompts to explain their predictions (Figure 1). We call these prompting-based rationales.

To evaluate different rationales, previous works on model interpretation establish two properties of rationales that are critical for successful interpretability: human alignment (DeYoung et al., 2019; Hase and Bansal, 2022) and faithfulness (Jacovi and Goldberg, 2020). *Human alignment* refers to the degree to which the rationales match or align with human-annotated rationales, while *faithfulness* assesses whether the rationales truly reflect the model's internal process. However, studies on LLM rationales either focus on the faithfulness of off-the-shelf LLMs (Huang et al., 2023a; Madsen et al., 2024), or their human alignment (Chen et al., 2023), but lack a comprehensive exploration of the two properties together. Moreover, they consider LLMs only as out-of-the-box models, without fine-tuning for specific tasks. The impact of fine-tuning to improve task accuracy on the alignment and faithfulness of LLM rationales remains underexplored.

In this paper, we conduct extensive experiments to comprehensively evaluate LLM rationales and bridge the gap in existing research. We consider five state-of-the-art LLMs, encompassing both open-source models (Llama2 (Touvron et al., 2023), Llama3, Mistral (Jiang et al., 2023)) and proprietary models (GPT-3.5-Turbo, GPT-4-Turbo (OpenAI et al., 2024)). Our study leverages three annotated natural language classification datasets, e-SNLI (Camburu et al., 2018a), FEVER (Thorne et al., 2018a) and Medical-Bios (Eberle et al., 2023), to evaluate and compare rationale extraction methods based on prompting strategies and feature attribution-based techniques such as Input×Gradient (Li et al., 2016).

Through our experiments, we find that attribution-based methods align more closely with humans in most cases, especially after fine-tuning. These methods also show more consistent improvements from fine-tuning compared to prompting. Additionally, we observe that low classification performance and collapsing predictions are linked to the limitations in evaluating the faithfulness of LLM rationales, shedding light on the task- and model-dependent findings of previous research. Most importantly we demonstrate that attribution-based methods offer a more faithful reflection of the model's inner workings than prompting.

In summary, our work provides a systematic framework, empirical evidence, and practical recommendations for extracting and evaluating LLM rationales. Our findings highlight critical limitations in prompting-based rationales, emphasizing the need for further efforts to improve the interpretability and trustworthiness of LLMs.

2 Related Work

Interpretability Recent literature in natural language processing (NLP) has seen a surge in interpretability methods aimed at making models more transparent and understandable. The traditional interpretability methods include 1) attention-based methods, which leverage the attention weights in models like transformers to identify which parts of the input the model focuses on when making a decision (Vaswani et al., 2023; Clark et al., 2019; Abnar and Zuidema, 2020), 2) Gradient-based methods, which provide explanations by identifying which input tokens most influence the model's output, often using techniques like gradient-based saliency maps (Simonyan et al., 2014a), or its extension by incorporating the input vector norms or integration (Sundararajan et al., 2017). 3) Vector-based methods that propagate the decomposed representations throughout the model achieving the best faithfulness results on encoder-based models (Kobayashi et al., 2020, 2021; Ferrando et al., 2022; Modarressi et al., 2022, 2023).

Rationales Rationales can be categorized as free-form or extractive. Free-form rationales use natural language to explain the model's reasoning, filling in commonsense knowledge gaps. They can improve model performance (Sun et al., 2022a) and user interpretability (Sun et al., 2022b). Extractive rationales highlight specific parts of the input text that provide sufficient evidence for a prediction, independent of the rest of the input (Lei et al., 2016; DeYoung et al., 2020). They can also enhance model performance (Huang et al., 2021; Carton et al., 2021) and improve human interpretability (Strout et al., 2019). Our work focuses on extractive rationales for interpretability evaluation. In this research area, Huang et al. (2023b) studied faithfulness in ChatGPT, comparing prompting and Lime (Ribeiro et al., 2016). Madsen et al. (2024) investigated LLM faithfulness on models like Llama2 (Touvron et al., 2023), Falcon (Penedo et al., 2023),

	Input	Human	Mistral-7B	LLama-3-8B	GPT-4 Turbo
e-SNLI	Premise: **Five** children playing soccer chase after a ball. Hypothesis: There are **ten** children playing. Label: Contradiction	**[Five, ten]**	[**Five** children, **ten** children]	[children, playing]	[children, **ten**]
FEVER	Claim: The Lincoln-Douglas debates happened in Alton, Illinois. Evidence: The Lincoln – Douglas Debates -LRB- also known as The Great Debates of 1858 -RRB- were a series of seven debates between Abraham Lincoln ... The debates were held in seven towns in the state of Illinois : Ottawa on August 21 ... **Alton, on, October, 15** ... Label: Supports	**[Alton, on, October, 15]**	[Lincoln, Douglas, Illinois, debates]	[**Alton**, **October**, Illinois, Debates]	[**Alton**, debates, **October**, **15**]
MedicalBios	His basic **surgical training** was undertaken in Newcastle. After gaining his **MRCS** and spending some time carrying out locum SpR work at Hurstwood Park, the Royal London Hospital Label: Surgeon	**[surgical, training, MRCS]**	[His, **MRCS**, **surgical**]	[**surgical**, **training**, **MRCS**]	[**surgical**, **MRCS**, SpR]

Table 1: Examples from the datasets are presented, showcasing rationales from human annotations alongside those generated by three models. Human rationales and the corresponding generated words are highlighted in **bold font**.

and Mistral (Jiang et al., 2023), noting its dependence on both model and dataset. Despite this, there is still a scarcity of analyses comparing LLM-generated rationales to interpretability methods. To the best of our knowledge, no studies seem to have assessed human alignment and model faithfulness jointly or used fine-tuning to overcome faithfulness evaluation limitations and explore its effects.

3 Experimental Setup

3.1 Datasets

We use three natural language classification datasets, each annotated with human rationales that highlight the key input words essential for determining the correct label. Table 1 shows examples of these datasets alongside the human rationale annotation and model-generated rationale.

e-SNLI This dataset (Camburu et al., 2018b; DeYoung et al., 2019) is a natural language inference task with three classes including Entailment, Contradiction, and Neutral, showing the relation between the premise and hypothesis sentences. We utilize 5,000 examples from the training set and 300 examples from the test set.

FEVER The Fact Extraction and Verification dataset (Thorne et al., 2018b; DeYoung et al., 2019) focuses on verifying claims based on provided evidence. Each claim is labeled as either "supports" or "refutes," with an annotated rationale indicating the specific portion of the evidence that underpins this classification. We use 5,000 samples from the training set and 300 from the test set.

MedicalBios (Eberle et al., 2023) consists of human rationale annotations for a subset of 100 samples (five medical classes) from the BIOS dataset (De-Arteaga et al., 2019) for the occupation classification task.

3.2 Models

We employ five of the latest LLMs. From the open-source models, we utilize Llama2 (Touvron et al., 2023), LLama3, and Mistral (Jiang et al., 2023). For proprietary models, we include GPT3.5-Turbo and GPT4-Turbo (OpenAI et al., 2024). All models are prompted without sampling during generation, leading to deterministic outputs (See Table A.1 for model information).

3.3 Methods

3.3.1 Prompting-Based Method

We employ different prompts for each stage of our experiments which are shown in Tables A.3, A.4, and A.5. Our prompts are engineered to convey the necessary information in a few sentences without confusing the models. We also experiment with two variations of each explanation prompt to manage the number of words the model generates.

Unbound Prompt In this method, the model autonomously determines the appropriate length of its generated text without word count restrictions.

Top-Var Prompt In this prompt, the model generates the top-k words, where k matches the exact number of words annotated in the human ratio-

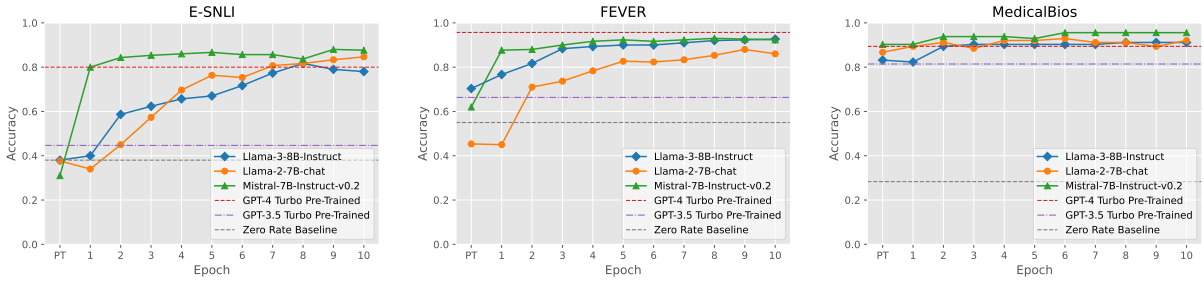

Figure 2: Classification accuracy throughout 10 epochs of fine-tuning. PT denotes the pre-trained model's accuracy. "Zero Rate Baseline" refers to the performance of a classifier that assigns all inputs to the majority class. Our tasks include E-SNLI and FEVER, where pre-trained models tend to underperform, as well as MedicalBios, where they demonstrate strong performance.

nales for each sentence. This method controls for word count in our experiments, allowing us to evaluate model alignment independently of its word importance threshold and ensuring a fair evaluation of faithfulness by keeping the number of masked words consistent across experiments.

3.3.2 Attribution-Based Methods

We employ the Inseq library (Sarti et al., 2023) to implement attribution-based methods for LLMs. Specifically, we select three available options: (i) Attention Weight Attribution, which utilizes the model's internal attention weights (Wiegreffe and Pinter, 2019); (ii) Simple Gradients (Saliency), which is based on the gradients of the output with respect to the inputs (Simonyan et al., 2014b); and (iii) Input×Gradient, which factors in both the input vector size and the gradient in its calculations (Li et al., 2016). We choose these methods because of their demonstrated faithfulness in previous work on NLP models (Atanasova et al., 2020; Modarressi et al., 2022, 2023), and their potential for efficient execution on large language models with limited computational resources. For each method, we focus on the output label word produced by the model in response to the classification prompt (Table A.3), calculating its attributions to the input tokens.

4 Results

In this section, we delve into utilizing both prompting-based and attribution-based approaches to extract rationale from the model, focusing on two aspects: human alignment and model faithfulness. Furthermore, we conduct fine-tuning experiments on open LLMs to examine how task performance influences alignment and faithfulness.

4.1 Task Performance

We begin our results by examining the accuracy of the models on the datasets. Figure 2 presents the pre-trained (PT) off-the-shelf accuracy of the models, along with their performance improvements during fine-tuning for 10 epochs.

As described by Wang et al. (2024) and Zhong et al. (2023), LLMs may underperform small fine-tuned models such as BERT, and smaller LLMs like LLaMA-2-7B might even collapse entirely. Madsen et al. (2024) reports similar behavior, with certain combinations of tasks and models performing poorly across different prompt variations. This pattern is evident in our results as well, where pre-trained open LLMs exhibit near-random accuracy on the e-SNLI and FEVER datasets (Figure 2).

Unlike previous work (Madsen et al., 2024) that simply excluded misclassified examples from their faithfulness evaluations—which may represent more than half of the dataset—we include all examples in our analysis. Additionally, we fine-tune[4] the LLMs using LoRA (Hu et al., 2022). As shown in Figure 2, the classification performance of LLaMA-2, LLaMA-3, and Mistral improves significantly after fine-tuning on the underperforming datasets, e-SNLI and FEVER, while also showing slight enhancements on the well-performing MedicalBios dataset. This fine-tuning narrows the performance gap with GPT-4-Turbo across all cases.

In the following sections, we will explore the alignment and faithfulness of pre-trained and fine-tuned models and investigate how task performance influences these aspects.

[4]The hyperparameters are provided in Table A.2.

Model	Method	Selection	Pre-trained Model F1			Fine-tuned Model F1		
			E-SNLI	FEVER	MedBios	E-SNLI	FEVER	MedicalBios
Mistral-7B Instruct-v0.2	PROMPTING	UNBOUND	36.36	24.68	38.14	33.03 (**−3.33**)	25.64 (+0.95)	38.87 (+0.72)
		TOP-VAR	40.08	26.24	44.11	35.29 (**−4.79**)	26.45 (+0.21)	45.02 (+0.92)
	ATTENTION	TOP-VAR	36.26	37.23	37.36	42.14 (**+5.87**)	38.64 (+1.41)	38.29 (+0.93)
	SALIENCY	TOP-VAR	46.46	37.96	40.76	49.69 (**+3.24**)	40.44 (**+2.47**)	46.91 (**+6.15**)
	INPUT×GRAD	TOP-VAR	40.45	36.36	40.10	42.70 (**+2.25**)	39.21 (**+2.85**)	45.38 (**+5.28**)
Llama-2-7b chat	PROMPTING	UNBOUND	38.57	14.69	32.64	38.44 (−0.13)	14.19 (−0.50)	33.87 (+1.22)
		TOP-VAR	45.65	20.89	49.77	44.33 (−1.31)	20.43 (−0.46)	50.48 (+0.71)
	ATTENTION	TOP-VAR	31.65	32.16	31.79	32.90 (+1.25)	32.94 (+0.78)	32.33 (+0.54)
	SALIENCY	TOP-VAR	34.92	30.52	36.49	46.56 (**+11.64**)	38.52 (**+8.00**)	34.12 (**−2.37**)
	INPUT×GRAD	TOP-VAR	35.43	31.01	37.84	46.38 (**+10.96**)	38.05 (**+7.03**)	35.35 (**−2.49**)
Llama-3-8B Instruct	PROMPTING	UNBOUND	37.68	23.79	46.52	28.14 (**−9.53**)	21.83 (**−1.96**)	47.94 (+1.42)
		TOP-VAR	43.07	28.76	59.99	30.03 (**−13.04**)	32.58 (**+3.82**)	59.76 (−0.23)
	ATTENTION	TOP-VAR	28.67	33.87	26.48	28.60 (−0.07)	34.80 (+0.93)	29.57 (**+3.10**)
	SALIENCY	TOP-VAR	41.42	34.29	39.68	44.62 (**+3.20**)	35.61 (+1.32)	41.67 (**+1.98**)
	INPUT×GRAD	TOP-VAR	44.59	35.27	45.61	49.22 (**+4.63**)	36.30 (+1.04)	46.68 (+1.07)
GPT-3.5 Turbo 1106	PROMPTING	UNBOUND	43.14	22.76	42.95	-	-	-
		TOP-VAR	46.06	34.27	53.96	-	-	-
GPT-4 Turbo 2024-04-09	PROMPTING	UNBOUND	51.44	29.48	53.25	-	-	-
		TOP-VAR	**52.53**	**43.23**	**60.77**	-	-	-
Random	RANDOM	TOP-VAR	26.54	33.90	21.97	26.54	33.90	21.97

Table 2: Human alignment F1↑ score. The difference in alignment between the fine-tuned (10 epochs) and pre-trained model is reported in the parentheses. Average random baseline (Selecting Top-Var random words) over 100 seeds is also reported. The highest alignment in each combination of model and dataset is underlined. Attribution-based methods are more aligned with humans in the majority of the cases, especially after fine-tuning. They also exhibit more consistent improvements by fine-tuning compared to prompting.

4.2 Human Alignment

The human-annotated rationales provide explanations for the ground truth label. Therefore, we first ask the model to generate its rationale for the true label (by prompting or attribution). Then to measure alignment, we compute the F1 score, which balances precision (the proportion of correctly generated words out of all generated words) and recall (the proportion of correctly generated words out of all relevant words). This score compares the model's generated rationales with human-annotated ones, assessing how well the model aligns with human explanations. Table 2 presents the F1 scores for the pre-trained and fine-tuned (epoch 10) models.

First, comparing the evaluated models reveals GPT-4-Turbo to be the most aligned with humans. Although the other models show varying levels of alignment depending on the task, GPT-3.5-Turbo and Llama3 often provide better prompting explanations than Llama2 and Mistral.

Second, we note that providing additional information about the number of words selected by humans in *Top-Var* settings enhances alignment, indicating disparities between model thresholds for word importance in *Unbound* prompting compared to human annotators.

Third, attribution-based methods tend to align more closely with human reasoning than prompting-based methods, especially after fine-tuning. This suggests that the model may be attending to words in a manner similar to human thought processes, although this cannot be fully explained through language modeling. Furthermore, among the attribution-based methods, Input×Gradient and Saliency appear to outperform raw attention weights, which is anticipated based on existing literature (Ferrando et al., 2022).

4.3 Effect of Classification Performance on Alignment

To analyze the effect of task performance on alignment more comprehensively, we rerun the alignment experiments for "Prompt Top-Var" and "Input×Gradient Top-Var" across all 10 epochs, as shown in Figure 3. Moreover, Table 3 shows

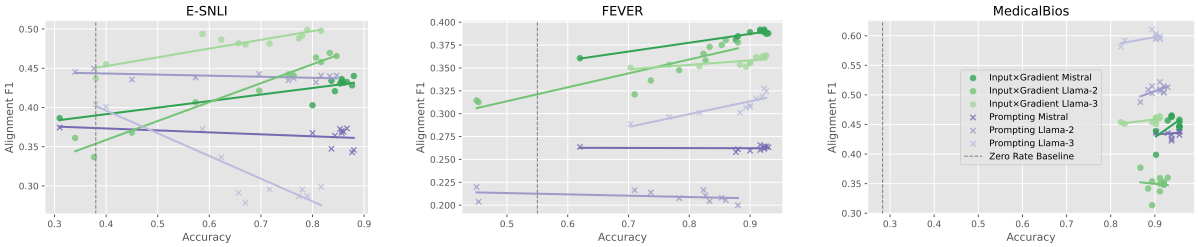

Figure 3: Accuracy and Human Alignment F1 score of 10 epochs of fine-tuning.

Dataset	Method Model	Input×Gradient	Prompting
E-SNLI	Mistral-7B	0.85 (0.00)	-0.35 (0.30)
	Llama-2-7b	0.98 (0.00)	-0.52 (0.10)
	Llama-3-8B	0.88 (0.00)	-0.90 (0.00)
FEVER	Mistral-7B	0.97 (0.00)	-0.06 (0.86)
	Llama-2-7b	0.90 (0.00)	-0.36 (0.28)
	Llama-3-8B	0.70 (0.02)	0.86 (0.00)
MedicalBios	Mistral-7B	0.58 (0.06)	0.16 (0.64)
	Llama-2-7b	-0.09 (0.79)	0.57 (0.07)
	Llama-3-8B	0.61 (0.04)	0.61 (0.05)

Table 3: The Pearson Correlation (p-value) of Human alignment F1↑ score and classification accuracy across 10 epochs of fine-tuning each model. Input×Gradient is highly correlated with accuracy while prompting doesn't show a clear trend.

the correlation between the Alignment F1 and the model's classification accuracy.

The results suggest a general trend of improved alignment for attribution-based methods with increasing accuracies throughout epochs.[5] This enhancement can stem from their reliance on the classification prompt and the classification capabilities of LLMs, which may fall short in off-the-shelf models. Consequently, fine-tuning enhances performance and affects the model's internal processes, potentially making them more accessible and transparent for detection via attention or gradient-based methods. Moreover, fine-tuning can guide the model's attention to the correct words, especially in datasets like e-SNLI where pre-trained classification accuracy was low.[6] As a result, these gradient-based methods can identify more human-aligned rationales by tracing back the attributions from the output label to the input sentence in fine-tuned models. Nonetheless, in prompting methods, this improvement in classification seems to function independently of the explanation task, as it does not demonstrate clear or significant enhancements.

[5]With the exception of MedicalBios, which does not show a significant range of accuracy improvement (Less than 10%), making it not a great candidate for this analysis.

[6]Figure A.3 illustrates qualitative examples of such cases.

4.4 Faithfulness to the Model

While human alignment provides a useful measure of the plausibility of LLM rationales, it is more important to consider the faithfulness of these rationales to the model's actual decision-making process. A word may be crucial for the model's decision even if it does not align with human rationale and vice versa. Therefore, we must ask: Are the self-explanations genuinely influential in the model's decision-making process?

To evaluate faithfulness, we employ a perturbation-based experiment similar to previous work (Madsen et al., 2024; Modarressi et al., 2023). In this experiment, we mask the important words identified by the prompting and attribution methods and measure the flip rate of the predicted label during classification. A higher flip rate indicates that the masked words are indeed important to the model, leading it to change its previous decision, and this suggests that the explanation is more faithful to the model's decision-making process.

4.4.1 Limitations of Faithfulness Evaluation before Fine-Tuning

Table 4 presents the faithfulness flip rate of the pre-trained and fine-tuned LLMs. A noteworthy finding is that in the e-SNLI and FEVER datasets, where classification accuracy was notably low (Figure 2), both attribution-based and prompting-based methods result in a very small (Less than 5%) flip rate in pre-trained models. Even more concerning, masking all the words in the input sentence led to less than a 2% flip rate for the Mistral and Llama-2 models (Mask ALL).

This issue arises from pre-trained models being heavily biased toward specific labels in poorly performing datasets (Figure A.1). For instance, in the e-SNLI dataset, Llama2, Llama3, and GPT-3.5 predominantly predict "entailment" in over 80% of the examples. Consequently, even masking the entire input does not change their biased predictions, resulting in extremely low faithfulness flip rates.

Model	Method	Pre-trained Faithfulness E-SNLI	Pre-trained Faithfulness FEVER	Pre-trained Faithfulness MedBios	Fine-tuned Faithfulness E-SNLI	Fine-tuned Faithfulness FEVER	Fine-tuned Faithfulness MedicalBios
Mistral-7B Instruct-v0.2	PROMPTING	**1.00**	**1.70**	19.27	16.33 (**+15.33**)	7.33 (+5.63)	8.85 (**−10.42**)
	ATTENTION	**0.00**	15.65	24.77	43.67 (**+43.67**)	11.33 (−4.31)	12.39 (**−12.38**)
	SALIENCY	**0.00**	14.97	30.28	50.00 (**+50.00**)	12.33 (−2.63)	15.93 (**−14.35**)
	INPUT×GRAD	**0.00**	14.29	28.44	41.33 (**+41.33**)	11.67 (−2.62)	16.81 (**−11.63**)
	RANDOM	**0.33**	12.24	5.50	27.00 (**+26.67**)	5.00 (−7.24)	2.65 (−2.85)
	HUMAN	**0.67**	**1.70**	44.04	50.00 (**+49.33**)	18.67 (**+16.97**)	14.16 (**−29.88**)
	ALL	**0.00**	**1.70**	100.00	68.67 (**+68.67**)	23.00 (**+21.30**)	84.96 (**−15.04**)
Llama-2-7b chat	PROMPTING	**1.34**	**0.00**	20.56	33.00 (**+31.66**)	8.33 (+8.33)	14.16 (−6.40)
	ATTENTION	**0.67**	**4.00**	25.23	31.33 (**+30.66**)	9.00 (+5.00)	14.16 (**−11.07**)
	SALIENCY	**1.68**	**4.33**	37.38	46.67 (**+44.99**)	12.33 (+8.00)	16.81 (**−20.57**)
	INPUT×GRAD	**2.01**	**4.33**	37.38	46.00 (**+43.99**)	11.33 (+7.00)	16.81 (**−20.57**)
	RANDOM	**1.34**	**3.33**	15.89	26.67 (**+25.32**)	7.00 (+3.67)	5.31 (**−10.58**)
	HUMAN	**1.01**	**0.00**	50.47	50.00 (**+48.99**)	14.33 (**+14.33**)	20.35 (**−30.11**)
	ALL	**0.00**	**0.00**	71.03	57.33 (**+57.33**)	19.33 (**+19.33**)	72.57 (+1.54)
Llama-3-8B Instruct	PROMPTING	15.33	**1.33**	36.94	22.33 (+7.00)	11.33 (+10.00)	16.07 (**−20.87**)
	ATTENTION	21.33	17.33	21.62	29.67 (+8.33)	11.33 (−6.00)	10.71 (**−10.91**)
	SALIENCY	22.00	17.67	27.03	45.00 (**+23.00**)	12.33 (−5.33)	15.18 (**−11.85**)
	INPUT×GRAD	22.67	16.33	28.83	46.00 (**+23.33**)	14.00 (−2.33)	18.75 (**−10.08**)
	RANDOM	18.00	12.67	**4.50**	28.33 (**+10.33**)	8.67 (−4.00)	0.89 (−3.61)
	HUMAN	24.00	**1.33**	40.54	51.00 (**+27.00**)	18.00 (**+16.67**)	22.32 (**−18.22**)
	ALL	79.00	**1.33**	67.57	81.33 (+2.33)	29.67 (**+28.33**)	73.21 (+5.65)

Table 4: Faithfulness FLIP RATE↑ percentage. The difference in faitfhulness between the fine-tuned (10 epochs) and pre-trained model is reported in the parentheses. The number of words to mask is enforced (TOP-VAR), and no method could mask more than the specified number for each sentence (except ALL). The highest faithfulness in each combination of model and dataset is underlined. Faithfulness evaluation before fine-tuning is unreliable due to collapsing predictions and biasing toward specific labels which causes Less than 5% flip rate in some cases. Attribution-based methods are more faithful in all of the cases after fine-tuning when the evaluation is more reliable.

Figure 4 illustrates the relationship between accuracy and faithfulness flip rate, based on 10 epochs of fine-tuning each model on each dataset. Additionally, Table 5 displays the correlation between model faithfulness and classification accuracy over the same 10 epochs, alongside the pre-trained model's accuracy. Both analyses indicate that models and datasets with lower pre-trained performance, relative to the zero-rate baseline, tend to show a stronger correlation between improvements in accuracy and increases in faithfulness. This phenomenon can also be attributed to label bias. When a pre-trained model has very low accuracy, it often indicates significant label bias, meaning that masking any number of words may not impact the biased decision. Fine-tuning the model helps mitigate this label bias (Figure A.1), leading to a higher faithfulness flip rate and, consequently, a correlation with accuracy improvements.

Although fine-tuning is not a definitive solution, it brings attention to a critical issue in previous faithfulness evaluations of LLMs, which were primarily developed with encoder-based models like BERT in mind. Encoder-based models are typically fine-tuned for specific classification tasks, whereas LLMs are pre-trained for broader tasks such as language modeling and instruction following. As a result, using the same faithfulness evaluations assesses the explanations over language modeling capabilities of LLMs rather than classification tasks, where word attributions can differ significantly. The assumption underlying these evaluations is that if a rationale is truly faithful, masking it should meaningfully change the model's decision. However, our research shows that for LLMs such as Llama2 and datasets like e-SNLI, the model's performance can be so poor (e.g., consistently predicting the same class) that even masking the entire input text has little to no impact on its predictions, leading to extremely low faithfulness. This problem is further exacerbated by the differences between BERT's straightforward classifier approach and the prompting used for LLMs, which introduces additional prompt engineering

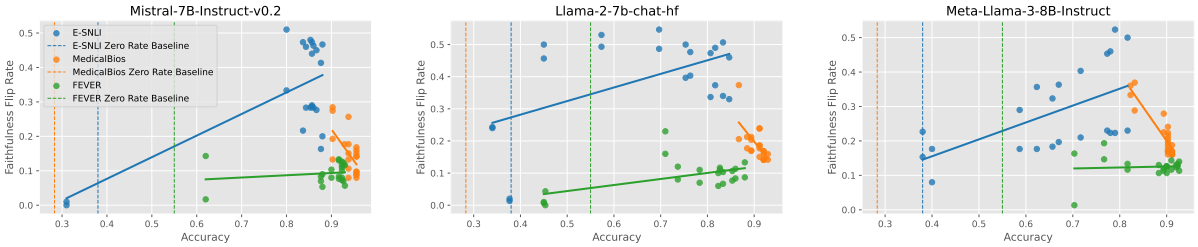

Figure 4: Accuracy and Faithfulness Flip Rate of 10 epochs of fine-tuning.

Dataset	Model	PT Acc.	Correlation
E-SNLI	Mistral-7B	31.00	0.68 (0.00)
	Llama-2-7b	37.67	0.53 (0.01)
	Llama-3-8B	38.00	0.58 (0.00)
FEVER	Mistral-7B	62.00	0.19 (0.41)
	Llama-2-7b	45.33	0.55 (0.01)
	Llama-3-8B	70.33	0.07 (0.77)
MedicalBios	Mistral-7B	90.27	-0.60 (0.00)
	Llama-2-7b	86.73	-0.65 (0.00)
	Llama-3-8B	83.19	-0.91 (0.00)

Table 5: The "Pearson Correlation (p-value)" between Faithfulness Flip Rate and Classification Accuracy over 10 fine-tuning epochs. Pre-trained accuracy ("PT Acc.") is also reported. Models and datasets with lower pre-trained performance show a stronger correlation between faithfulness and accuracy gains.

variations and complicates the evaluation process. This performance issue likely led to conclusions in other papers that faithfulness is highly model- and task-dependent. Thus, our study argues that the traditional BERT-based faithfulness evaluation methods are inadequate for current LLMs. Fine-tuning is one approach to addressing this evaluation issue, nonetheless, further research is needed to develop more accurate methods for evaluating the faithfulness of LLM rationales.

4.4.2 Faithfulness after Fine-Tuning

Table 4 also displays the faithfulness flip rate of the fine-tuned models. A comparison of results after fine-tuning (where the near-zero flip rate issue is addressed) reveals that attribution methods outperform prompting methods in faithfulness. This difference can stem from attribution methods basing their explanations on the model's internal processes, whereas prompting may provide plausible answers without direct access to this information, potentially diverging from the truth of the model's inner workings. Additionally, prompting is affected by the model's ability to follow instructions, which may result in the generation of an inaccurate number of words or the inclusion of words not present in the input sentence, leading to less faithful results.

Another finding from Table 4 and Figure 4 is that for MedicalBios, where the model was already performing well, further fine-tuning leads to a decrease in faithfulness. This can occur because additional fine-tuning allows the model to learn more cues from the training set, making it less sensitive to word masking. As a result, masking the same number of words may no longer flip its predictions, since the model can rely more effectively on signals from the unmasked portions.

Finally, we present the flip rate after masking human rationales in Table 4. Despite expectations that the model would better recognize the importance of words for its own decisions, these methods consistently underperform human rationales. This result emphasizes that while the current methods demonstrate a degree of faithfulness, there remains room for further refinement and enhancement.

5 Conclusions

In this study, we investigated extracting rationales from LLMs, focusing on human alignment and model faithfulness. We experimented with prompting and attribution methods across different LLM architectures and datasets. Our study shows that attribution methods generally provide rationales that align more closely with human reasoning compared to prompting methods, especially after fine-tuning. Of greater importance, attribution methods also outperform prompting approaches in faithfulness, as they more accurately reflect the model's internal decision-making process. Moreover, we showed that fine-tuning reduces label bias and correlates with faithfulness in low-accuracy datasets, explaining the model and task-dependent results of previous works. Despite these improvements, a gap persists between the models' rationales and human rationales in alignment and faithfulness, underscoring the need for the development of more advanced explanation methods to bridge this gap.

Limitations

LLM instruction-following abilities. In our implementation of prompting strategies, we heavily rely on the LLM's capability to follow instructions accurately. For example, when requesting the top-k words separated by a specific delimiter character, we expect the model to output a list of words in our desired format and quantity with no extra explanations. However, LLMs are still not fully adept at adhering to prompts precisely (Sun et al., 2023), which can lead to outputs in various formats different from our expectations. Since our primary focus in this paper is not to evaluate the format-following ability of LLMs, we have taken measures to address discrepancies in the outputs as much as possible.

To mitigate these discrepancies, we adopt tailored parsing approaches to handle unexpected output formats. For instance, if a model separates words in the output with a "," character instead of the instructed character "|", we adjust our parsing method accordingly. Fortunately, each model tends to adhere to a relatively consistent output format across the dataset, which enables us to adapt our parsing approach accordingly. Nonetheless, it's worth noting that an LLM with enhanced instruction-following abilities could potentially yield even better parsing results and consequently achieve higher performance levels.

Attribution-based methods In selecting the explanation methods based on the inner workings of the models we opted for the ones that were already implemented for LLMs and were relatively efficient to execute given the large size of the models. Nonetheless, we acknowledge that recent vector-based methods have shown promising faithfulness results by decomposing the representations (Kobayashi et al., 2020, 2021; Modarressi et al., 2022; Ferrando et al., 2022; Modarressi et al., 2023) on smaller models such as BERT (Devlin et al., 2019) compared with the gradient-based methods. Our study highlights the gap that could be filled by implementing these methods for LLMs.

Prompt Engineering Although we reported various versions of prompts for extracting rationales in this paper and conducted preliminary prompt engineering, we acknowledge that better prompts could potentially achieve higher performance. However, this approach diverges from realistic use cases where users may ask questions in various wordings. This limitation is inherent to prompting methods, whereas attribution-based methods are not susceptible to this issue. Therefore, addressing this limitation calls for continued exploration and refinement of both prompting and attribution-based methods in rationale extraction.

Larger Models In our experiments, we evaluated open models with less than 8B parameters due to resource limitations. However, we acknowledge that larger models could potentially perform better in following instructions, leading to improved human alignment and model faithfulness in their self-explanations.

Perturbation-based faithfulness evaluation In this paper, we conduct faithfulness evaluation of LLM rationales using perturbation-based metrics. Those metrics assume that removing critical features based on rationales would largely affect model performance. However, Whether perturbation-based metrics truly reflect rationale faithfulness is a widely discussed but unsolved question, as they would produce out-of-distribution counterfactuals. For example, Yin et al. (2022) show that with different kinds of perturbations such as removal or noise in hidden representations, the faithful sets vary significantly. For consistency, we follow previous work (DeYoung et al., 2019; Huang et al., 2023a). We leave deeper study into faithfulness measurements of LLM rationales to future work.

References

Samira Abnar and Willem Zuidema. 2020. Quantifying attention flow in transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4190–4197, Online. Association for Computational Linguistics.

Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, and Isabelle Augenstein. 2020. A diagnostic study of explainability techniques for text classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3256–3274, Online. Association for Computational Linguistics.

Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. 2018a. e-snli: Natural language inference with natural language explanations. In *NeurIPS*.

Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. 2018b. e-snli: Natural language inference with natural language explanations. In *Advances in Neural Information Processing Systems*, pages 9539–9549.

Samuel Carton, Surya Kanoria, and Chenhao Tan. 2021. What to learn, and how: Toward effective learning from rationales. *ArXiv*, abs/2112.00071.

Yanda Chen, Ruiqi Zhong, Narutatsu Ri, Chen Zhao, He He, Jacob Steinhardt, Zhou Yu, and Kathleen McKeown. 2023. Do models explain themselves? counterfactual simulatability of natural language explanations. *arXiv preprint arXiv:2307.08678*.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. 2019. What does BERT look at? an analysis of BERT's attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 276–286, Florence, Italy. Association for Computational Linguistics.

Maria De-Arteaga, Alexey Romanov, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, and Adam Tauman Kalai. 2019. Bias in bios: A case study of semantic representation bias in a high-stakes setting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, page 120–128, New York, NY, USA. Association for Computing Machinery.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. 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 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. 2019. Eraser: A benchmark to evaluate rationalized nlp models. *Preprint*, arXiv:1911.03429.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. 2020. ERASER: A benchmark to evaluate rationalized NLP models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4443–4458, Online. Association for Computational Linguistics.

Oliver Eberle, Ilias Chalkidis, Laura Cabello, and Stephanie Brandl. 2023. Rather a nurse than a physician - contrastive explanations under investigation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6907–6920, Singapore. Association for Computational Linguistics.

Javier Ferrando, Gerard I. Gállego, and Marta R. Costa-jussà. 2022. Measuring the mixing of contextual information in the transformer. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8698–8714, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Peter Hase and Mohit Bansal. 2022. When can models learn from explanations? a formal framework for understanding the roles of explanation data. In *Proceedings of the First Workshop on Learning with Natural Language Supervision*, pages 29–39, Dublin, Ireland. Association for Computational Linguistics.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Quzhe Huang, Shengqi Zhu, Yansong Feng, and Dongyan Zhao. 2021. Exploring distantly-labeled rationales in neural network models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5571–5582, Online. Association for Computational Linguistics.

Shiyuan Huang, Siddarth Mamidanna, Shreedhar Jangam, Yilun Zhou, and Leilani H Gilpin. 2023a. Can large language models explain themselves? a study of llm-generated self-explanations. *arXiv preprint arXiv:2310.11207*.

Shiyuan Huang, Siddarth Mamidanna, Shreedhar Jangam, Yilun Zhou, and Leilani H. Gilpin. 2023b. Can large language models explain themselves? a study of llm-generated self-explanations. *Preprint*, arXiv:2310.11207.

Alon Jacovi and Yoav Goldberg. 2020. Towards faithfully interpretable nlp systems: How should we define and evaluate faithfulness? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4198–4205.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Goro Kobayashi, Tatsuki Kuribayashi, Sho Yokoi, and Kentaro Inui. 2020. Attention is not only a weight: Analyzing transformers with vector norms. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7057–7075, Online. Association for Computational Linguistics.

Goro Kobayashi, Tatsuki Kuribayashi, Sho Yokoi, and Kentaro Inui. 2021. Incorporating Residual and Normalization Layers into Analysis of Masked Language Models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4547–4568, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2016. Rationalizing neural predictions. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 107–117, Austin, Texas. Association for Computational Linguistics.

Jiwei Li, Xinlei Chen, Eduard Hovy, and Dan Jurafsky. 2016. Visualizing and understanding neural models in nlp. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 681–691.

Andreas Madsen, Sarath Chandar, and Siva Reddy. 2024. Are self-explanations from large language models faithful? In *Proceedings of The 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*.

Ali Modarressi, Mohsen Fayyaz, Ehsan Aghazadeh, Yadollah Yaghoobzadeh, and Mohammad Taher Pilehvar. 2023. DecompX: Explaining transformers decisions by propagating token decomposition. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2649–2664, Toronto, Canada. Association for Computational Linguistics.

Ali Modarressi, Mohsen Fayyaz, Yadollah Yaghoobzadeh, and Mohammad Taher Pilehvar. 2022. GlobEnc: Quantifying global token attribution by incorporating the whole encoder layer in transformers. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 258–271, Seattle, United States. Association for Computational Linguistics.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong,

Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. 2023. The refinedweb dataset for falcon llm: Outperforming curated corpora with web data, and web data only. *Preprint*, arXiv:2306.01116.

Marco Tulio Ribeiro, UW EDU, Sameer Singh, and Carlos Guestrin. 2016. Model-Agnostic Interpretability of Machine Learning. In *ICML Workshop on Human Interpretability in Machine Learning*.

Gabriele Sarti, Nils Feldhus, Ludwig Sickert, Oskar van der Wal, Malvina Nissim, and Arianna Bisazza. 2023. Inseq: An interpretability toolkit for sequence generation models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 421–435, Toronto, Canada. Association for Computational Linguistics.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2014a. Deep inside convolutional networks: Visualising image classification models and saliency maps. *Preprint*, arXiv:1312.6034.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2014b. Deep inside convolutional networks: Visualising image classification models and saliency maps. *CoRR*, abs/1312.6034.

Julia Strout, Ye Zhang, and Raymond Mooney. 2019. Do human rationales improve machine explanations? In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 56–62, Florence, Italy. Association for Computational Linguistics.

Jiao Sun, Anjali Narayan-Chen, Shereen Oraby, Alessandra Cervone, Tagyoung Chung, Jing Huang, Yang Liu, and Nanyun Peng. 2022a. ExPUNations: Augmenting puns with keywords and explanations. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4590–4605, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jiao Sun, Swabha Swayamdipta, Jonathan May, and Xuezhe Ma. 2022b. Investigating the benefits of free-form rationales. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5867–5882, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jiao Sun, Yufei Tian, Wangchunshu Zhou, Nan Xu, Qian Hu, Rahul Gupta, John Wieting, Nanyun Peng, and Xuezhe Ma. 2023. Evaluating large language models on controlled generation tasks. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3155–3168, Singapore. Association for Computational Linguistics.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. *Preprint*, arXiv:1703.01365.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018a. FEVER: a large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819, New Orleans, Louisiana. Association for Computational Linguistics.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018b. FEVER: a Large-scale Dataset for Fact Extraction and VERification. In *Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 809–819.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2023. Attention is all you need. *Preprint*, arXiv:1706.03762.

Yuxia Wang, Minghan Wang, and Preslav Nakov. 2024. Rethinking STS and NLI in large language models. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 965–982, St. Julian's, Malta. 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.

Fan Yin, Zhouxing Shi, Cho-Jui Hsieh, and Kai-Wei Chang. 2022. On the sensitivity and stability of model interpretations in nlp. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2631–2647.

Qihuang Zhong, Liang Ding, Juhua Liu, Bo Du, and Dacheng Tao. 2023. Can chatgpt understand too? a comparative study on chatgpt and fine-tuned bert. *Preprint*, arXiv:2302.10198.

A Appendix

Model	Access
meta-llama/Meta-Llama-3-8B-Instruct	Open Source
meta-llama/Llama-2-7b-chat-hf	Open Source
mistralai/Mistral-7B-Instruct-v0.2	Open Source
gpt-3.5-turbo-1106	Proprietary
gpt-4-turbo-2024-04-09	Proprietary

Table A.1: The details of the models we used in this work.

Hyperparameter	Value
Total Batch Size	64
Learning Rate E-SNLI	7e-06
Learning Rate FEVER	7e-06
Learning Rate MedicalBios	3e-06
Num Epochs	10
Learning Rate Scheduler Warmup Steps	10
Training Dataset Size	5000
LoRA r	32
LoRA alpha	16
LoRA drop out	0.05

Table A.2: The hyperparameters used for fine-tuning the models using LoRA.

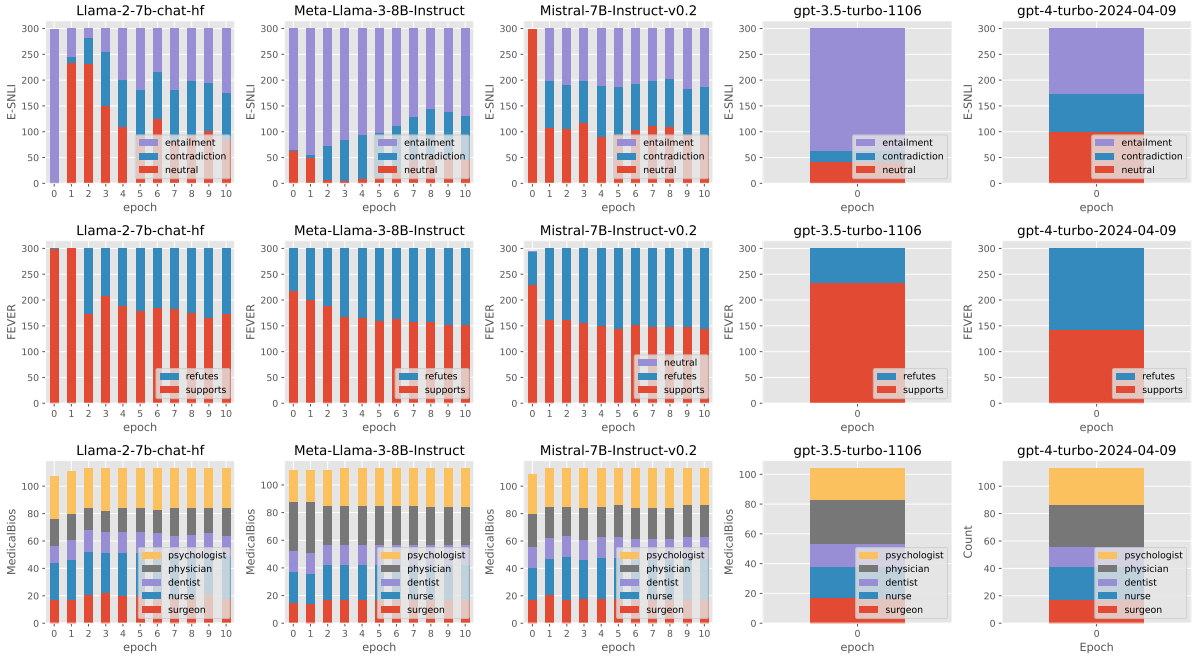

Figure A.1: The distribution of predicted labels across epochs of fine-tuning. Pre-trained off-the-shelf models tend to be heavily biased toward a label in poorly performing datasets.

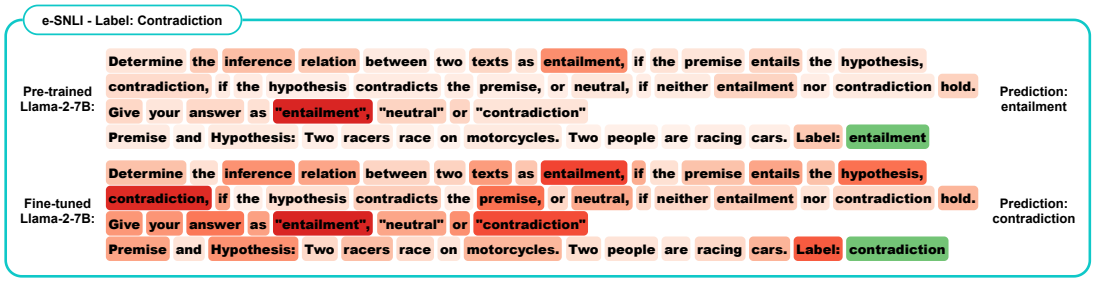

Figure A.2: Token Importance before and after fine-tuning Llama-2-7B based on the InputXGradient explainability method. The predicted word by the model is shown in green and its attributions to previous words are shown in red. The attributions before fine-tuning are more skewed (Fisher-Pearson coefficient of skewness over all the dataset: 3.37±0.31), and become less skewed after fine-tuning (1.42±0.36).

Figure A.3: Token Importance before and after fine-tuning Llama-2-7B based on the Input×Gradient explainability method. The predicted/true label is shown in green, with its attributions to input words in red. The human rationale for the examples are ["dog", "cat], ["twins"], and ["grazes", "touching"]. The fine-tuned model identified these words solely through training on classification data, without any rationale data.

Method	Selection	Prompt
PROMPT	UNBOUND	Premise: “‘ {premise} “‘ Hypothesis: “‘ {hypothesis} “‘ Label: {label} Identify the most important words from the text that are most relevant to understanding the relationship between the premise and the hypothesis. Write the words as a pipe-separated (|) list of words with spaces. Do not output any other text, symbols, or explanations.
PROMPT	TOP-VAR	Premise: “‘ {premise} “‘ Hypothesis: “‘ {hypothesis} “‘ Label: {label} Identify the top {k} most important words from the text that are most relevant to understanding the relationship between the premise and the hypothesis. Write the top {k} words as a pipe-separated (|) list of words with spaces. Do not output any other text, symbols, or explanations.
ATTRIBUTION-BASED	TOP-VAR (Selected later from tokens with the highest attribution scores)	Premise and Hypothesis: “‘ {premise} {hypothesis} “‘ Determine the inference relation between two (short, ordered) texts as entailment, if the premise entails the hypothesis, contradiction, if the hypothesis contradicts the premise, or neutral, if neither entailment nor contradiction hold. Respond with exactly one of the following: "entailment", "contradiction", or "neutral." Do not output any other text, symbols, or explanations—just the label. Label: {label}
CLASSIFICATION		Premise and Hypothesis: “‘ {premise} {hypothesis} “‘ Determine the inference relation between two (short, ordered) texts as entailment, if the premise entails the hypothesis, contradiction, if the hypothesis contradicts the premise, or neutral, if neither entailment nor contradiction hold. Respond with exactly one of the following: "entailment", "contradiction", or "neutral." Do not output any other text, symbols, or explanations—just the label. Label:

Table A.3: The prompts utilized for the e-SNLI dataset.

Method	Selection	Prompt
PROMPT	UNBOUND	Claim: {claim} Evidence: “‘ {evidence} “‘ Label: {label} Identify the most important words from the given evidence that are most essential for verifying the factual relationship between the claim and the evidence. Provide these words as a pipe-separated (|) list. Do not output any other text, symbols, or explanations.
PROMPT	TOP-VAR	Claim: {claim} Evidence: “‘ {evidence} “‘ Label: {label} Identify the top {k} most important words from the given evidence that are most essential for verifying the factual relationship between the claim and the evidence. Provide these top {k} words as a pipe-separated (|) list. Do not output any other text, symbols, or explanations.
ATTRIBUTION-BASED	TOP-VAR (Selected later from tokens with the highest attribution scores)	Claim: {claim} Evidence: “‘ {evidence} “‘ Determine the factual relationship between the claim and the supporting evidence as "supports" if the evidence supports the claim or "refutes" if the evidence refutes the claim. Respond with exactly one of the following: "supports" or "refutes". Do not output any other text, symbols, or explanations—just the label. Label: {label}
CLASSIFICATION		Claim: {claim} Evidence: “‘ {evidence} “‘ Determine the factual relationship between the claim and the supporting evidence as "supports" if the evidence supports the claim or "refutes" if the evidence refutes the claim. Respond with exactly one of the following: "supports" or "refutes". Do not output any other text, symbols, or explanations—just the label. Label:

Table A.4: The prompts utilized for the FEVER dataset.

Method	Selection	Prompt
PROMPT	UNBOUND	Bio: “‘ {bio} “‘ Identify the most important words from the text that indicate why is the person in the following short bio described as a {label}. Write the key words as a pipe-separated (|) list of single words. Do not output any other text, symbols, or explanations.
PROMPT	TOP-VAR	Bio: “‘ {bio} “‘ Identify the top {k} most important words from the text that indicate why is the person in the following short bio described as a {label}. Write the key words as a pipe-separated (|) list of single words. Do not output any other text, symbols, or explanations.
ATTRIBUTION-BASED	TOP-VAR (Selected later from tokens with the highest attribution scores)	Bio: “‘ {bio} “‘ Determine the occupation of the person in the short bio. Respond with exactly one of the following: (psychologist, surgeon, nurse, dentist, physician). Do not output any other text, symbols, or explanations—just the label. Occupation: {label}
CLASSIFICATION		Bio: “‘ {bio} “‘ Determine the occupation of the person in the short bio. Respond with exactly one of the following: (psychologist, surgeon, nurse, dentist, physician). Do not output any other text, symbols, or explanations—just the label. Occupation:

Table A.5: The prompts utilized for the MedicalBios dataset.