# Evaluating Human Alignment and Model Faithfulness of LLM Rationale

Anonymous ACL submission

## Abstract

We study how well large language models (LLMs) explain their generations with rationales – a set of tokens extracted from the input texts that reflect the decision process of LLMs. We examine LLM rationales extracted with two methods: 1) attribution-based methods that use attention or gradients to locate important tokens, and 2) prompting-based methods that guide LLMs to extract rationales using prompts. Through extensive experiments, we show that prompting-based rationales align better with human-annotated rationales than attribution-based rationales, and demonstrate reasonable alignment with humans even when model performance is poor. We additionally find that the faithfulness limitations of prompting-based methods, which are identified in previous work, may be linked to their collapsed predictions. By fine-tuning these models on the corresponding datasets, both prompting and attribution methods demonstrate improved faithfulness. Our study sheds light on more rigorous and fair evaluations of LLM rationales, especially for prompting-based ones.[1]

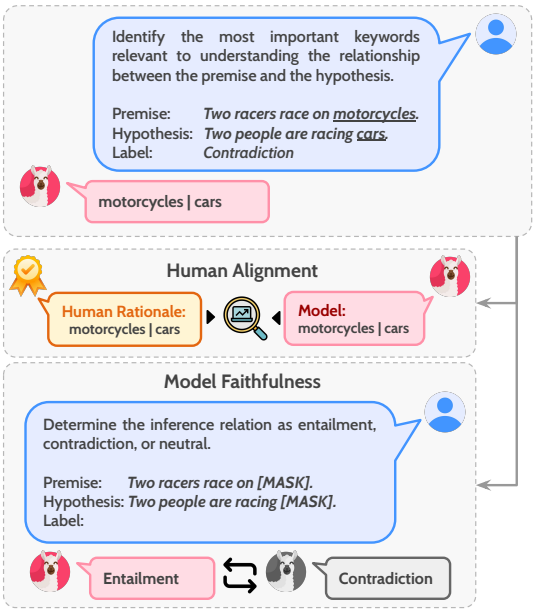


Figure 1: An example of our analysis methodology on the e-SNLI dataset. *Human alignment* compares model rationales with human-annotated rationale; *Model faithfulness* measures the rates when model prediction changes (e.g. from `Contradiction` to `Entailment`) after masking the rationales identified by the model.[3]

## 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, they are not always reliable and usually cannot clearly explain their outputs (Ji et al., 2023), which limits the deployment of these models in high-stakes scenarios.

*Rationales*[2], 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 smaller and open-source models like BERT (Devlin et al., 2019), rationales are extracted with attribution-based methods like attention weights (Wiegreffe and Pinter, 2019) or gradients (Li et al., 2016). For decoder-only LLMs, besides attribution-based methods, 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

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

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

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

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. A longstanding debate exists regarding the relationship between these two aspects (Agarwal et al., 2024). 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. Specifically, recent works (Huang et al., 2023a; Madsen et al., 2024) study prompting-based methods and show that they might not be faithful to the reasoning process of LLMs. Moreover, they only consider LLMs as out-of-box models, without fine-tuning on specific tasks. How fine-tuning of LLMs on downstream tasks influences the human alignment and faithfulness of LLM rationales is under-explored.

In this paper, we conduct extensive experiments to evaluate LLM rationales more comprehensively 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 two annotated natural language classification datasets, e-SNLI (Camburu et al., 2018a) and MedicalBios (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 while prompting-based rationales are generally less faithful than attribution-based methods, they tend to align better with human-annotated rationales, both before and after fine-tuning. Surprisingly, even when prompting-based rationales exhibit poor performance, they can still produce explanations that reasonably align with human reasoning.

We also observe that low classification performance and collapsing predictions might be related to the faithfulness limitation of LLM rationales. Fine-tuning LLMs on specific datasets improves the quality of rationales for both prompting and, particularly, attribution techniques in terms of faithfulness and human alignment. This finding complements the observations in Madsen et al. (2024), where faithful evaluation was conducted using only out-of-the-box LLMs.

In summary, our work contributes to the ongoing efforts to enhance the interpretability and trustworthiness of LLMs by providing empirical evidence and practical recommendations for extracting and evaluating rationales from these models.

## 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) attribution-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). More recently, researchers have been using rationales, written in natural language, to serve interpretability where they can reveal the "reasoning" behind model decisions.

**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

| | Input | Human | Mistral-7B | LLama-3-8B | GPT-4 Turbo |
|---|---|---|---|---|---|
| e-SNLI | Premise: **Five** children playing soccer chase after a ball.<br>Hypothesis: There are **ten** children playing.<br>Label: Contradiction | **[Five, ten]** | [**Five** children, **ten** children] | [children, playing] | [children, **ten**] |
| | Premise: Asian individuals are **taking pictures** of some unidentified thing.<br>Hypothesis: Asians **play soccer** in the park.<br>Label: Contradiction | **[taking, pictures, play, soccer]** | [Asian, individuals, **taking**, **pictures**, unidentified, thing, **soccer**, park] | [Asian, individuals, unidentified, **soccer**] | [Asian, **pictures**, **play**, **soccer**] |
| 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<br>Label: Surgeon | **[surgical, training, MRCS]** | [His, **MRCS**, **surgical**] | [**surgical**, **training**, **MRCS**] | [**surgical**, **MRCS**, SpR] |
| | She works at both the Park Nicollet Women's Center and Regions Hospital – Cancer Care Center. Kandice dedicates herself to providing **nurturing**, **sensitive care** along with good educational support to her patients.<br>Label: Nurse | **[nurturing, sensitive, care]** | [Kandice, works, nurse, Cancer Care Center, Women's Center, provides, **sensitive care**, educational support, patients] | [**care**, patients, support] | [**care**, patients] |

Table 1: Examples from e-SNLI and MedicalBios datasets are presented, showcasing rationales from human annotations alongside those generated by three models (Using "Short Prompt Top-Var" method). Human rationales and the corresponding generated words are highlighted in **bold font**.

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), 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 utilize two natural language classification datasets that have been annotated with human rationales indicating which input words were pivotal for the ground truth 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) is a natural language inference task with 3 classes including Entailment, Contradiction, and Neutral, showing the relation between the premise and hypothesis sentences. This dataset is annotated for rationales supporting the classification label by DeYoung et al. (2019). We utilize 5,000 examples from the training set and 300 examples from the test set.

**MedicalBios** 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 large language models, encompassing both open-source and proprietary ones. 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. You can see the exact model information in Table 6 in the appendix.

### 3.3 Methods

#### 3.3.1 Prompting-Based Method

We employ various prompting strategies to explore the effects of prompt wording and model alignment in generating text of similar length to human-annotated rationale (Tables 12&13). The following prompts are used to evaluate these aspects.

We test two versions of prompts to examine how the clarity and length of the prompt influence the model's output:

**Normal Prompt** This version provides a detailed explanation, including all the points the model should consider. It is longer and aims to ensure the model fully understands the task.

| Method | Model | Mistral-7B Instruct-v0.2 | | Llama-2-7b chat | | Llama-3-8B Instruct | | GPT-3.5 Turbo 1106 | | GPT-4 Turbo 2024-04-09 | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | Dataset | E-SNLI | MedBios | E-SNLI | MedBios | E-SNLI | MedBios | E-SNLI | MedBios | E-SNLI | MedBios |
| **Attribution-based** | | | | | | | | | | | |
| Attention | Top-Ratio | 28.59 | 29.34 | 40.92 | 22.11 | 23.80 | 15.01 | - | - | - | - |
| Saliency | Top-Ratio | 37.16 | 28.97 | 28.64 | 31.35 | 32.19 | 27.80 | - | - | - | - |
| Input×Gradient | Top-Ratio | 31.41 | 27.51 | 26.85 | 30.96 | 36.32 | 29.83 | - | - | - | - |
| Attention | Top-Var | 36.03 | 41.01 | 48.69 | 35.08 | 30.56 | 24.30 | - | - | - | - |
| Saliency | Top-Var | 46.07 | 38.30 | 37.25 | 41.49 | 39.78 | 36.36 | - | - | - | - |
| Input×Gradient | Top-Var | 38.86 | 37.08 | 36.66 | 40.45 | **44.76** | 39.48 | - | - | - | - |
| **Prompting-based** | | | | | | | | | | | |
| Normal Prompt | Unbound | 38.95 | 42.50 | 45.45 | 49.16 | 39.58 | 49.05 | 43.96 | 45.14 | 47.24 | 51.89 |
| Short Prompt | Unbound | 41.84 | 42.26 | 43.54 | 43.74 | 43.60 | 51.28 | 43.28 | 45.28 | 46.58 | 53.50 |
| Normal Prompt | Top-Ratio | 41.17 | 41.17 | 45.90 | 43.98 | 35.98 | 47.58 | 36.58 | 43.68 | 45.27 | 47.81 |
| Short Prompt | Top-Ratio | 41.83 | 42.05 | 40.91 | 45.65 | 37.46 | 46.91 | 37.92 | 41.00 | 47.31 | 45.64 |
| Normal Prompt | Top-Var | 45.93 | **48.43** | **50.93** | **52.23** | 41.67 | **58.85** | 43.85 | **57.16** | 55.33 | **59.28** |
| Short Prompt | Top-Var | **46.97** | 48.08 | 47.34 | 50.54 | 44.65 | 58.17 | **44.31** | 53.60 | **55.78** | 59.10 |

Table 2: Human alignment F1↑ score in e-SNLI and MedicalBios datasets. Random baseline (Selecting Top-Var random words) is **27±4** and **22±1** for e-SNLI and MedicalBios respectively over 100 seeds. The top two alignments in each column are indicated by **bold** and underlined formatting.

**Short Prompt** Given that long prompts may confuse LLMs, we also use a shorter version that conveys the necessary information in a few sentences.

We also experiment with three versions of the introduced prompts to manage the number of words the model generates.

**Unbound Prompt** In this method, the model is not restricted in the number of words it can generate. It needs to establish the appropriate length autonomously.

**Top-Var Prompt** This prompt requires the model to generate exactly the same number of words as in the human rationale annotations for each sentence. This method controls for word count in our experiments, enabling us to assess model alignment independently of its word importance threshold.

**Top-Ratio Prompt** In this approach, the model is guided to identify the top $k$ most important words within a given sentence. The value of $k$ is derived from a predetermined percentage of the total word count in the input sentence, a ratio established based on the training set. For example, this ratio is set at 20% for sentences in the e-SNLI dataset and 13% for those in the MedicalBios dataset.

#### 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.

## 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 this process influences alignment and faithfulness.

### 4.1 Human Alignment

The annotated rationales provide explanations for the ground truth label. Therefore, for the evaluation of human alignment, we first request the model to provide a rationale for the provided label. With this, we create an array of binary values where for each word in the input sentence, we indicate a *1* if it is present among the generated words (*0* otherwise). By comparing these vectors with the equivalent binary representations of human annotations, we

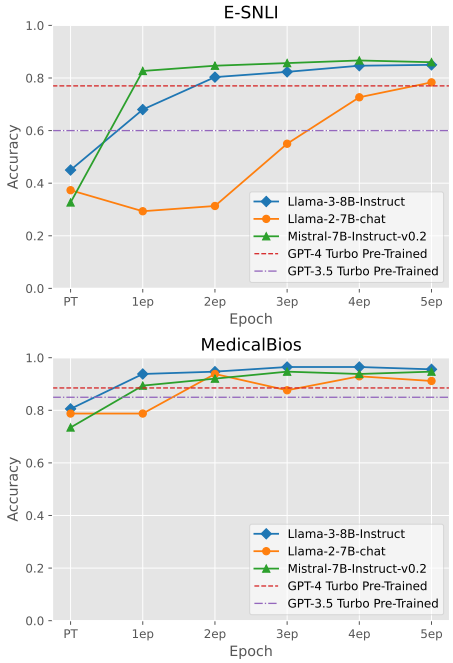


Figure 2: Accuracy changes throughout 5 epochs of fine-tuning. (PT denotes the pre-trained model's accuracy)

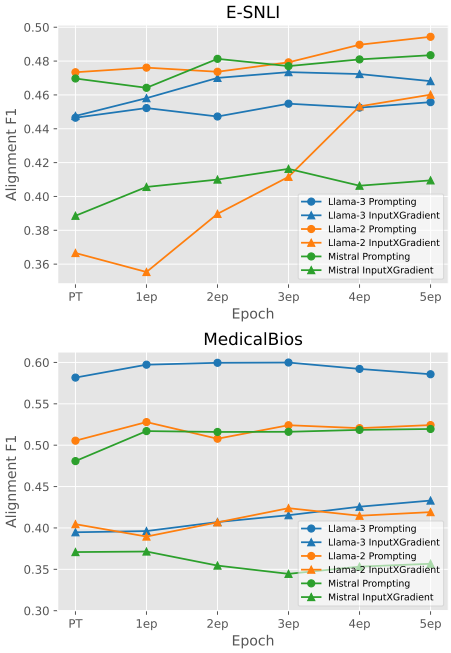


Figure 3: Human alignment F1↑ changes throughout 5 epochs of fine-tuning. (PT denotes the pre-trained model)

calculate the F1 score. This F1 score for human alignment is reported in Table 2 for the e-SNLI and MedicalBios datasets.

Firstly, prompting-based methods outperform attribution-based methods on human alignment. Short or normal prompting demonstrate superior performance compared to attritbution-based methods in nearly all datasets and models except for Llama-3-8B on e-SNLI, which is also comparable. This gap can be attributed to the reliance of attribution-based methods on the classification capability of LLMs (which might be subpar).

Secondly, 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. Furthermore, the random baseline, which involves selecting Top-Var random words in each sentence, yields F1 scores of 0.27±0.04 and 0.22±0.01 for e-SNLI and MedicalBios respectively across 100 seeds. Contrasted with the *Top-Var* rows in Table 2, this indicates that both attribution-based and prompting-based approaches exhibit superior alignment compared to a random baseline.

Thirdly, across models, the performance comparison between normal and short prompts is varying and inconclusive. For example, short prompts perform better for Mistral, LLama-3, and GPT-4-Turbo, while they perform worse for LLama-2 and GPT-3.5 models.

Finally, comparing the evaluated models reveals GPT-4-Turbo to be the most aligned with humans. However, other models demonstrate task-dependent alignment, with some excelling in e-SNLI and others in MedicalBios.

### 4.2 Effect of Fine-tuning on Alignment

As described by Wang et al. (2024) and Zhong et al. (2023), zero-shot LLMs may underperform small fine-tuned models such as BERT. And smaller LLMs like LLaMA-2-7B might even collapse entirely. We observe similar failure patterns, where models generate a single label from the possible options regardless of the input sentence. This phenomenon is illustrated in Figure 2, where the pre-trained (PT) open LLMs achieve near-random accuracy (33%) on the e-SNLI dataset. This issue raises the question of whether fine-tuning these models to improve their classification performance also aids in aligning their explanations more closely with human expectations.

To address this issue, we fine-tune the LLMs using LoRA (Hu et al., 2022), a parameter-efficient fine-tuning technique. The hyperparameters for fine-tuning are provided in Table 7. Figure 2 demonstrates that the classification performance

| Method | Model | Mistral-7B FT Instruct-v0.2 | | Llama-2-7b FT chat | | Llama-3-8B FT Instruct | |
|---|---|---|---|---|---|---|---|
| | Dataset Selection | E-SNLI | MedicalBios | E-SNLI | MedicalBios | E-SNLI | MedicalBios |
| **Attribution-based** | | | | | | | |
| Saliency | Top-Ratio | 37.36 (+0.20) | 27.29 (**–1.69**) | 36.34 (**+7.71**) | 31.73 (+0.38) | 34.76 (**+2.57**) | 30.18 (**+2.38**) |
| Input×Gradient | Top-Ratio | 33.71 (**+2.30**) | 27.47 (–0.04) | 36.98 (**+10.13**) | 32.64 (**+1.68**) | 38.59 (**+2.27**) | 31.91 (**+2.08**) |
| Saliency | Top-Var | 44.87 (–1.19) | 37.12 (–1.18) | 45.63 (**+8.37**) | 42.80 (+1.31) | 42.90 (**+3.13**) | 39.57 (**+3.21**) |
| Input×Gradient | Top-Var | 40.95 (**+2.09**) | 35.67 (–1.41) | 46.01 (**+9.36**) | 41.91 (+1.46) | **46.81** (**+2.06**) | 43.31 (**+3.83**) |
| **Prompting-based** | | | | | | | |
| Normal Prompt | Unbound | 40.48 (**+1.53**) | 41.66 (–0.84) | 45.28 (–0.17) | 47.26 (**–1.89**) | 39.04 (–0.54) | 48.41 (–0.64) |
| Short Prompt | Unbound | 39.49 (**–2.35**) | 42.74 (+0.48) | 43.11 (–0.43) | 43.45 (–0.29) | 43.86 (+0.26) | 50.45 (–0.83) |
| Normal Prompt | Top-Ratio | 40.99 (–0.17) | 41.59 (+0.43) | 44.40 (–1.50) | 43.29 (–0.70) | 35.42 (–0.56) | 47.00 (–0.58) |
| Short Prompt | Top-Ratio | 42.01 (+0.18) | 43.91 (**+1.85**) | 43.23 (**+2.33**) | 45.29 (–0.36) | 38.15 (+0.69) | 47.68 (+0.77) |
| Normal Prompt | Top-Var | 45.70 (–0.24) | 48.51 (+0.08) | **50.58** (–0.35) | 51.35 (–0.88) | 41.69 (+0.02) | **59.70** (+0.85) |
| Short Prompt | Top-Var | **48.35** (+1.38) | **51.96** (**+3.88**) | 49.43 (**+2.09**) | **52.43** (**+1.89**) | 45.57 (+0.92) | 58.57 (+0.41) |

Table 3: Human alignment F1↑ score in E-SNLI and MedicalBios datasets after fine-tuning for 5 epochs on the respective dataset. The difference in alignment between the fine-tuned and pre-trained model (Table 2) is reported in the parentheses. (Changes greater than 1.5% are denoted using a bold font.)

of LLaMA-2, LLaMA-3, and Mistral improves significantly after fine-tuning on the e-SNLI dataset and shows slight improvements on the MedicalBios dataset, outperforming GPT-4-Turbo in both cases.

We rerun the human alignment experiments for "Short Prompt Top-Var" and "Input×Gradient Top-Var" across all epochs, as shown in Figure 3. The results suggest a general trend of improved alignment of both methods with increasing epochs.

To analyze the generalization of this trend more comprehensively, we conduct alignment experiments on all prompting and attribution methods in Table 3 on the final epoch (5th epoch) of fine-tuning. The difference in human alignment between the fine-tuned and pre-trained models (Table 2) is reported in parentheses.

Overall, the results show more positive change than negative in alignment. Among the prompting-based methods, "Short Prompt Top-Var" demonstrates the highest gains from fine-tuning. Attribution-based methods, particularly in LLaMA models, exhibit significant alignment improvements after fine-tuning. 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. We have demonstrated qualitative examples of such cases in Figure 5. 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.

### 4.3 Faithfulness to the Model

While human alignment provides a useful measure of the plausibility of LLM rationales, it is also 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.

#### 4.3.1 Limitations of Faithfulness Evaluation before Fine-Tuning

Table 4 presents the faithfulness flip rate of the pre-trained LLMs. A noteworthy finding is that in the e-SNLI dataset, where classification accuracy was notably low (Figure 2), both attribution-based and prompting-based methods resulted in a very small flip rate. Even more concerning, masking all the words in the input sentence led to less than a

| Masking Part | Method | Dataset Selection | Mistral-7B Instruct-v0.2 E-SNLI | Mistral-7B Instruct-v0.2 MedBios | Llama-2-7b chat E-SNLI | Llama-2-7b chat MedBios | Llama-3-8B Instruct E-SNLI | Llama-3-8B Instruct MedBios | GPT-3.5 Turbo 1106 E-SNLI | GPT-3.5 Turbo 1106 MedBios |
|---|---|---|---|---|---|---|---|---|---|---|
| **Attribution-based** | | | | | | | | | | |
| Input | Saliency | Top-Ratio | 3.00 | 50.00 | 2.36 | 31.82 | 29.33 | 36.11 | - | - |
| | Input×Gradient | Top-Ratio | 2.33 | 44.32 | 2.02 | 29.09 | 26.67 | 40.74 | - | - |
| **Prompting-based** | | | | | | | | | | |
| Input | Normal Prompt | Unbound | 3.33 | 59.09 | 1.01 | 30.91 | 38.67 | 41.67 | 77.00 | 41.82 |
| | Short Prompt | Unbound | 3.67 | 59.09 | 1.68 | 35.45 | 44.00 | 47.22 | 73.67 | 37.27 |
| | Normal Prompt | Top-Ratio | 1.67 | 28.41 | 1.68 | 33.64 | 24.00 | 45.37 | 41.33 | 24.55 |
| | Short Prompt | Top-Ratio | 2.33 | 30.68 | 1.01 | 26.36 | 22.67 | 46.30 | 44.33 | 20.91 |
| **Baselines** | | | | | | | | | | |
| Input | Human | Human | 4.00 | 60.23 | 2.02 | 38.18 | 31.67 | 43.12 | 55.67 | 36.36 |
| | Random | Random | 1.33 | 9.09 | 1.01 | 10.91 | 24.67 | 14.68 | 41.33 | 5.45 |
| | Everything | Everything | 2.33 | 97.73 | 0.34 | 60.00 | 69.67 | 70.64 | 89.00 | 74.55 |

Table 4: Faithfulness Flip Rate↑ percentage in E-SNLI and MedicalBios datasets. The number of words to mask is enforced in Top-Ratio, and no method could mask more than the specified number for each sentence.

3% flip rate for the Mistral and LLaMA-2 models (Mask Everything).

In further exploration, Figure 4 illustrates the Input×Gradient attributions of the predicted label, shown in green, to all instruction and input words, shown in shades of red. We notice that in the pre-trained Llama-2 model, the prediction for the label "entailment" is incorrect, with the model placing excessive emphasis on the word "entailment" in the instruction while largely ignoring the input sentence including the premise and hypothesis sentences. However, after fine-tuning, the attribution distribution becomes less skewed, leading to a correct prediction by the model.

Therefore, we hypothesize that the pre-trained model focuses more on the instruction rather than the input sentence. Moreover, in Table 10, we investigate two masking scenarios. The first scenario, denoted as Input, involves masking only the words from the input sentence (e.g., the premise and hypothesis in e-SNLI) while leaving the instruction intact, similar to Table 4. The second scenario, denoted as Input&Instruction, extends masking to the entire instruction and input, constituting the entire prompt. When we extend the masking to include the instruction, the flip rate can increase up to 100%. This indicates that the model relies heavily on the instruction for its decisions, regardless of the input sentence. This phenomenon aligns with findings by Yin et al. (2023) and Kung and Peng (2023), who both found that, among all segments of a prompt, label information or output space is essential for the model's performance. This raises concerns about the reliability of this experiment for measuring model faithfulness in LLMs.

Consequently, we argue that to conduct a more robust faithfulness experiment on LLMs, it is not advisable to solely rely on pre-trained models, as their classification accuracy can vary depending on the model and dataset (Madsen et al., 2024). Instead, we suggest aligning the experiment more closely with the scenario of fine-tuned encoder-based models (Ferrando et al., 2022; Modarressi et al., 2023) by training the LLMs and assessing faithfulness on the fine-tuned model.

### 4.3.2 Faithfulness after Fine-Tuning

Table 5 displays the faithfulness flip rate of the fine-tuned open models on e-SNLI and MedicalBios. The number of masked words can directly influence the prediction flip rate in this experiment. To ensure a fair comparison, we limit the number of words considered in the Top-Ratio and Top-Var selections to the top $k$ words for each sentence across different methods. This approach is particularly important for prompting-based methods, as LLMs struggle to follow instructions involving fine-grained hard constraints (Sun et al., 2023).

First, we see that fine-tuning has effectively addressed the near-zero flip rate (Table 4) in e-SNLI, indicating that the model is no longer completely disregarding the input sentence.

Second, a comparison of results in each selection group of "Top-Ratio" and "Top-Var" (with a similar number of masked words) reveals that attribution-

| Selection | Model / Dataset / Method | Mistral-7B FT Instruct-v0.2 E-SNLI | Mistral-7B FT Instruct-v0.2 MedicalBios | Llama-2-7B FT chat E-SNLI | Llama-2-7B FT chat MedicalBios | Llama-3-8B FT Instruct E-SNLI | Llama-3-8B FT Instruct MedicalBios |
|---|---|---|---|---|---|---|---|
| Unbound | Normal Prompt | **64.00** | 34.51 | 67.00 | 17.70 | 64.33 | 18.75 |
| | Short Prompt | 53.00 | **37.17** | **67.33** | **21.24** | **72.00** | **25.00** |
| Top-Ratio | Attention | 50.33 | **23.89** | **69.67** | 14.16 | 49.33 | 5.36 |
| | Saliency | **51.00** | 23.89 | 59.67 | 21.24 | **60.00** | 17.86 |
| | Input×Gradient | 45.33 | 21.24 | 61.33 | 22.12 | 58.00 | 18.75 |
| | Normal Prompt | 35.33 | 17.70 | 52.33 | 19.47 | 45.00 | **20.54** |
| | Short Prompt | 40.33 | 17.70 | 60.33 | **23.89** | 51.33 | **20.54** |
| Top-Var | Attention | 48.00 | **20.35** | **61.00** | 11.50 | 48.67 | 5.36 |
| | Saliency | **52.67** | 14.16 | 54.67 | 17.70 | 55.00 | 14.29 |
| | Input×Gradient | 45.33 | 19.47 | 53.33 | **19.47** | **56.00** | 15.18 |
| | Normal Prompt | 37.00 | 15.04 | 50.33 | 16.81 | 47.00 | **16.07** |
| | Short Prompt | 40.33 | 19.47 | 55.00 | 17.70 | 50.67 | **16.07** |
| **Baselines** | | | | | | | |
| Top-Var | Human | 49.67 | 21.24 | 60.00 | 24.78 | 62.00 | 24.11 |
| Top-Var | Random | 32.67 | 9.73 | 45.00 | 7.96 | 36.67 | 6.25 |
| Everything | Everything | 66.33 | 84.07 | 63.00 | 75.22 | 77.33 | 74.11 |

Table 5: Faithfulness Flip Rate↑ percentage for fine-tuned models on the respective dataset after 5 epochs. The number of words to mask is enforced in Top-Ratio and Top-Var by limiting masked words per sentence. (Results are comparable in each similar selection technique, ensuring a similar number of masked words.)

based methods generally outperform prompting. Consistent results can be seen in our top-$k$ experiments in the appendix A.1 and Table 8. This difference can be attributed to the fact that attribution methods base 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.

Third, we also present the flip rate after masking human rationales in Table 5. These rates are comparable to the "Top-Var" selection of other methods, as they involve the same number of masked words. Despite expectations that the model would better recognize the importance of words for its own decisions, prompting methods consistently underperformed human rationales, and attribution methods did so in half of the cases. 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 the extraction of rationales in Large Language Models (LLMs) with a focus on human alignment and model faithfulness. Our experiments encompassed both prompting-based and attribution-based methods across various LLM architectures and datasets. Before fine-tuning, we observed that prompting generally yielded better human alignment, even when classification performance was poor. However, the reliability of faithfulness evaluations was compromised by low classification performance and collapsing predictions in pre-trained models highlighting the need for refining faithfulness evaluation setup.

To address this, we fine-tuned the models to enhance their accuracy on classification tasks, which led to improvements in aligning their explanations more closely with human expectations. In this scenario, although prompting showed superior alignment before, its faithfulness in reflecting model decision-making was not as strong as that of attribution-based methods.

Despite these improvements, a gap remained between the models' rationales and human rationales in both alignment and faithfulness. This highlights 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.

Chirag Agarwal, Sree Harsha Tanneru, and Himabindu Lakkaraju. 2024. Faithfulness vs. plausibility: On the (un) reliability of explanations from large language models. *arXiv preprint arXiv:2402.04614*.

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.

Po-Nien Kung and Nanyun Peng. 2023. Do models really learn to follow instructions? an empirical study of instruction tuning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1317–1328, Toronto, Canada. 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*.

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.

Fan Yin, Jesse Vig, Philippe Laban, Shafiq Joty, Caiming Xiong, and Chien-Sheng Wu. 2023. Did you read the instructions? rethinking the effectiveness of task definitions in instruction learning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3063–3079, Toronto, Canada. Association for Computational Linguistics.

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

## A.1 Top-k Faithfulness

As previously noted, the number of masked words significantly impacts the flip rate. To explore this further, we conducted a Top-$k$ experiment, masking $k = 1, 2, 3, 4, 5, 10$ words and calculating the flip rate (see Table 8). The results consistent with Table 5 indicate that on average, attribution-based methods surpass prompting in terms of faithfulness.

| 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 6: The details of the models we used in this work.

| Hyperparameter | Value |
|---|---|
| Total Batch Size | 64 |
| Learning Rate E-SNLI | 1e-05 |
| Learning Rate MedicalBios | 5e-06 |
| Num Epochs | 5 |
| Learning Rate Scheduler Warmup Steps | 10 |
| Training Dataset Size | 5000 |
| LoRA r | 32 |
| LoRA alpha | 16 |
| LoRA drop out | 0.05 |

Table 7: The hyperparameters used for fine-tuning the models using LoRA.

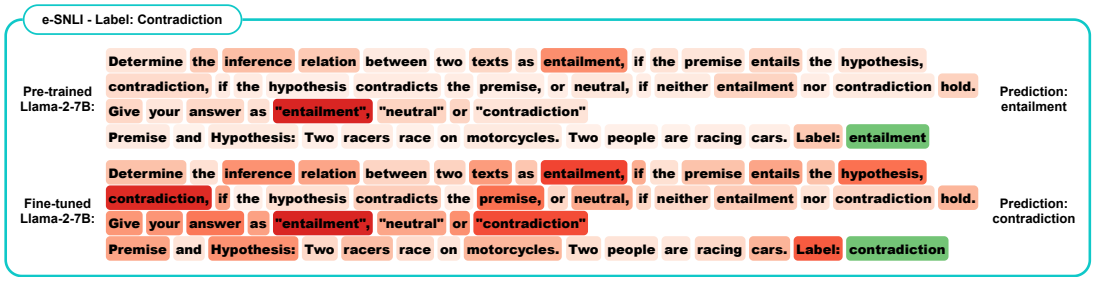


Figure 4: Token Importance before and after fine-tuning Llama-2-7B based on the InputXGradient explainability method. The predicted word by the model is shown in green and its attributions to previous words are shown in red. The attributions before fine-tuning are more skewed (Fisher-Pearson coefficient of skewness over all the dataset: 3.37±0.31), and become less skewed after fine-tuning (1.42±0.36).

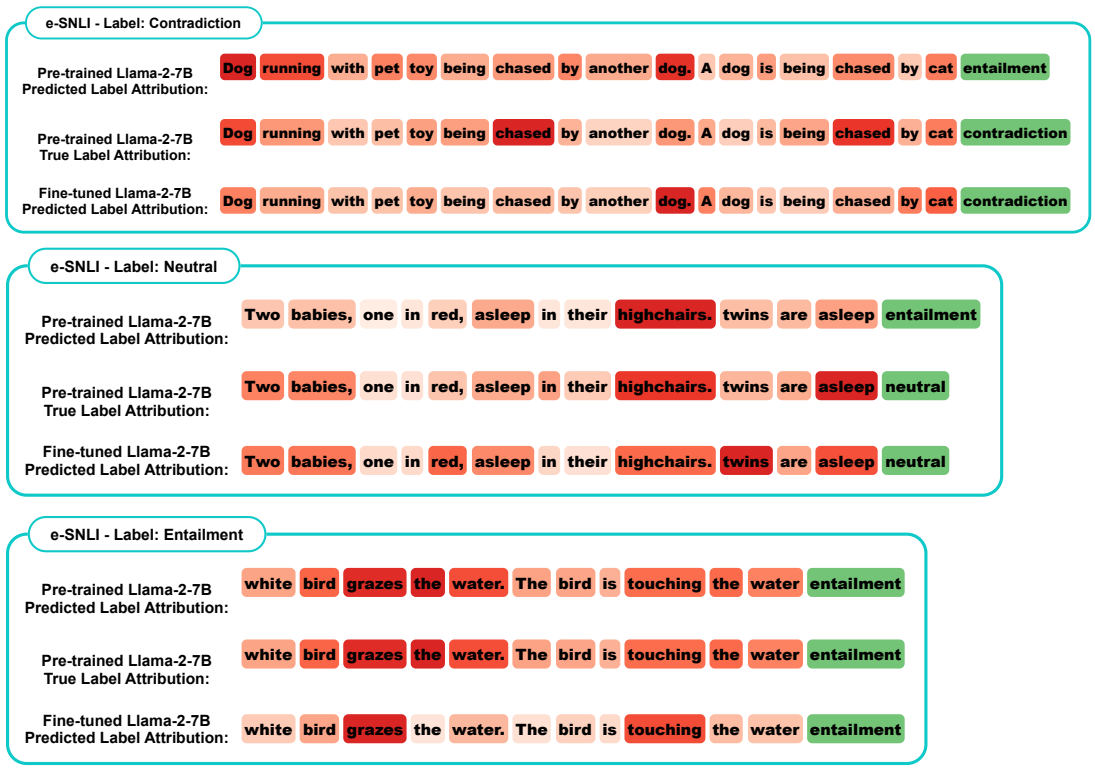


Figure 5: Token Importance before and after fine-tuning Llama-2-7B based on the Input×Gradient explainability method. The predicted/true label is shown in green, with its attributions to input words in red. The human rationale for the examples are ["dog", "cat], ["twins"], and ["grazes", "touching"]. The fine-tuned model identified these words solely through training on classification data, without any rationale data.

| Dataset | Model / Top k / Method | Mistral-7B FT Instruct-v0.2 1 | 2 | 3 | 4 | 5 | 10 | Avg | Llama-2-7b FT chat 1 | 2 | 3 | 4 | 5 | 10 | Avg | Llama-3-8B FT Instruct 1 | 2 | 3 | 4 | 5 | 10 | Avg |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **E-SNLI** | Attention | **41.0** | 36.7 | 48.0 | 47.7 | 54.0 | 62.3 | 48.3 | **40.0** | **50.0** | **63.7** | **66.3** | **69.3** | **71.0** | **60.1** | 31.0 | 35.3 | 41.7 | 49.7 | 56.0 | 69.7 | 47.2 |
| | Saliency | 40.0 | **46.0** | **49.3** | **52.3** | **54.7** | **63.0** | **50.9** | 36.7 | 45.0 | 54.3 | 58.0 | 59.7 | 68.3 | 53.7 | 30.3 | 37.7 | 50.0 | 56.3 | **63.7** | 77.0 | 52.5 |
| | Input×Gradient | 34.0 | 39.0 | 44.0 | 46.7 | 50.0 | 57.7 | 45.2 | 37.0 | 42.3 | 50.7 | 60.3 | 60.7 | 70.0 | 53.5 | **33.0** | **38.0** | **54.3** | **59.0** | 61.0 | **77.3** | **53.8** |
| | Prompting | 31.3 | 30.0 | 41.3 | 40.7 | 47.0 | 55.3 | 40.9 | 35.3 | 44.3 | 54.7 | 60.7 | 64.3 | 64.0 | 53.9 | 29.3 | 28.7 | 47.0 | 48.7 | 57.7 | 76.7 | 48.0 |
| **MedicalBios** | Attention | 7.1 | 10.6 | 15.0 | 15.0 | 19.5 | 31.0 | 16.4 | 5.3 | 8.0 | 7.1 | 8.8 | 9.7 | 13.3 | 8.7 | 0.9 | 3.6 | 3.6 | 5.4 | 6.2 | 15.2 | 5.8 |
| | Saliency | 5.3 | 12.4 | 14.2 | **17.7** | **20.4** | **31.0** | **16.8** | 14.2 | **18.6** | 18.6 | 15.9 | 16.8 | 18.6 | 17.1 | 7.1 | 11.6 | 13.4 | 17.0 | **18.8** | 18.8 | 14.4 |
| | Input×Gradient | **8.8** | **16.8** | 15.0 | 14.2 | 17.7 | 25.7 | 16.4 | **15.9** | 17.7 | **19.5** | 15.0 | **18.6** | **19.5** | **17.7** | **8.0** | **12.5** | **14.3** | 16.1 | 17.0 | 22.3 | **15.0** |
| | Prompting | 7.1 | 10.6 | **16.8** | 16.8 | 17.7 | 28.3 | 16.2 | 0.9 | 9.7 | 14.2 | **19.5** | 17.7 | **19.5** | 13.6 | 6.2 | 7.1 | 12.5 | **19.6** | 16.1 | **27.7** | 14.9 |

Table 8: Faithfulness Flip Rate↑ percentage in E-SNLI and MedicalBios datasets. The Top-$k$ is enforced in masking and no method could mask more than the specified $k$. The highest faithfulness in each dataset and top-$k$ is in **bold** font.

| Dataset | Model / Top k / Method | Mistral-7B FT Instruct-v0.2 1 | 2 | 3 | 4 | 5 | 10 | Llama-2-7b FT chat 1 | 2 | 3 | 4 | 5 | 10 | Llama-3-8B FT Instruct 1 | 2 | 3 | 4 | 5 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **E-SNLI** | Attention | 1.0 | 2.0 | 3.0 | 4.0 | 5.0 | 10.0 | 1.0 | 2.0 | 3.0 | 4.0 | 5.0 | 10.0 | 1.0 | 2.0 | 3.0 | 4.0 | 5.0 | 10.0 |
| | Saliency | 1.0 | 2.0 | 3.0 | 4.0 | 5.0 | 10.0 | 1.0 | 2.0 | 3.0 | 4.0 | 5.0 | 10.0 | 1.0 | 2.0 | 3.0 | 4.0 | 5.0 | 10.0 |
| | Input×Gradient | 1.0 | 2.0 | 3.0 | 4.0 | 5.0 | 10.0 | 1.0 | 2.0 | 3.0 | 4.0 | 5.0 | 10.0 | 1.0 | 2.0 | 3.0 | 4.0 | 5.0 | 10.0 |
| | Prompting | 1.0 | 1.8 | 2.9 | 3.8 | 4.7 | 8.2 | 0.9 | 1.9 | 2.9 | 3.8 | 4.8 | 7.4 | 0.9 | 1.8 | 2.8 | 3.8 | 4.8 | 9.1 |
| **MedicalBios** | Attention | 1.0 | 2.0 | 3.0 | 4.0 | 5.0 | 10.0 | 1.0 | 2.0 | 3.0 | 4.0 | 5.0 | 10.0 | 1.0 | 2.0 | 3.0 | 4.0 | 5.0 | 10.0 |
| | Saliency | 1.0 | 2.0 | 3.0 | 4.0 | 5.0 | 10.0 | 1.0 | 2.0 | 3.0 | 4.0 | 5.0 | 10.0 | 1.0 | 2.0 | 3.0 | 4.0 | 5.0 | 10.0 |
| | Input×Gradient | 1.0 | 2.0 | 3.0 | 4.0 | 5.0 | 10.0 | 1.0 | 2.0 | 3.0 | 4.0 | 5.0 | 10.0 | 1.0 | 2.0 | 3.0 | 4.0 | 5.0 | 10.0 |
| | Prompting | 1.0 | 1.8 | 2.9 | 3.9 | 4.9 | 9.9 | 0.2 | 1.6 | 2.7 | 3.7 | 4.7 | 8.5 | 0.8 | 1.7 | 2.7 | 3.7 | 4.5 | 7.7 |

Table 9: The average number of masked words in each experiment of Table 8. Prompting can have fewer masked words as a result of generating words out of the input sentence or not following the instruction on how many words it should generate (We limit the masks if the model generates more than $k$ words).

| Masking Part | Method | Model / Dataset Selection | Mistral-7B Instruct-v0.2 E-SNLI | MedicalBios | Llama-2-7b chat E-SNLI | MedicalBios | Llama-3-8B Instruct E-SNLI | MedicalBios | GPT-3.5 Turbo 1106 E-SNLI | MedicalBios |
|---|---|---|---|---|---|---|---|---|---|---|
| **Attribution-based** | | | | | | | | | | |
| Input | Saliency | Top-Ratio | 3.00 | 50.00 | 2.36 | 31.82 | 29.33 | 36.11 | - | - |
| | Input×Gradient | Top-Ratio | 2.33 | 44.32 | 2.02 | 29.09 | 26.67 | 40.74 | - | - |
| Input&Instr. | Saliency | Top-Ratio | 97.00 | 71.59 | 100.00 | 90.91 | 86.00 | 72.22 | - | - |
| | Input×Gradient | Top-Ratio | 97.33 | 63.64 | 100.00 | 93.64 | 86.33 | 74.07 | - | - |
| **Prompting-based** | | | | | | | | | | |
| Input | Normal Prompt | Unbound | 3.33 | 59.09 | 1.01 | 30.91 | 38.67 | 41.67 | 77.00 | 41.82 |
| | Short Prompt | Unbound | 3.67 | 59.09 | 1.68 | 35.45 | 44.00 | 47.22 | 73.67 | 37.27 |
| | Normal Prompt | Top-Ratio | 1.67 | 28.41 | 1.68 | 33.64 | 24.00 | 45.37 | 41.33 | 24.55 |
| | Short Prompt | Top-Ratio | 2.33 | 30.68 | 1.01 | 26.36 | 22.67 | 46.30 | 44.33 | 20.91 |
| Input&Instr. | Extended Prompt | Unbound | 68.67 | 61.36 | 8.75 | 28.18 | 58.33 | 45.37 | 99.67 | 50.00 |
| **Baselines** | | | | | | | | | | |
| Input | Human | Human | 4.00 | 60.23 | 2.02 | 38.18 | 31.67 | 43.12 | 55.67 | 36.36 |
| | Everything | Everything | 2.33 | 97.73 | 0.34 | 60.00 | 69.67 | 70.64 | 89.00 | 74.55 |

Table 10: Faithfulness Flip Rate↑ percentage in E-SNLI and MedicalBios datasets. The number of words to mask is enforced in Top-Ratio, and no method could mask more than the specified number for each sentence.

| Selection | Model / Dataset / Method | Mistral-7B FT Instruct-v0.2 E-SNLI | Mistral-7B FT Instruct-v0.2 MedicalBios | Llama-2-7B FT chat E-SNLI | Llama-2-7B FT chat MedicalBios | Llama-3-8B FT Instruct E-SNLI | Llama-3-8B FT Instruct MedicalBios |
|---|---|---|---|---|---|---|---|
| Unbound | Normal Prompt | 64.00 | 34.51 | 67.00 | 17.70 | 64.33 | 18.75 |
| | Short Prompt | 53.00 | 37.17 | 67.33 | 21.24 | 72.00 | 25.00 |
| Top-Ratio | Attention | 50.67 | 28.32 | 70.33 | 15.04 | 53.67 | 9.82 |
| | Saliency | 52.33 | 30.09 | 62.67 | 24.78 | 66.00 | 23.21 |
| | Input×Gradient | 47.67 | 25.66 | 62.33 | 26.55 | 62.67 | 25.89 |
| | Normal Prompt | 54.67 | 32.74 | 63.00 | 23.01 | 61.67 | 24.11 |
| | Short Prompt | 52.67 | 35.40 | 64.00 | 25.66 | 59.00 | 25.00 |
| Top-Var | Attention | 47.67 | 23.01 | 67.33 | 15.93 | 53.67 | 6.25 |
| | Saliency | 50.00 | 19.47 | 58.67 | 26.55 | 60.00 | 17.86 |
| | Input×Gradient | 44.33 | 20.35 | 58.00 | 24.78 | 61.67 | 19.64 |
| | Normal Prompt | 51.33 | 30.97 | 64.67 | 19.47 | 61.33 | 21.43 |
| | Short Prompt | 50.67 | 30.09 | 60.67 | 20.35 | 59.00 | 20.54 |
| **Baselines** | | | | | | | |
| Top-Var | Human | 49.67 | 21.24 | 60.00 | 24.78 | 62.00 | 24.11 |
| Everything | Everything | 66.33 | 84.07 | 63.00 | 75.22 | 77.33 | 74.11 |

Table 11: Faithfulness Flip Rate↑ percentage in E-SNLI and MedicalBios datasets for fine-tuned models on the respective dataset for 5 epochs. No limitations were applied to the number of masked words which means that LLMs could mask more words in prompting techniques by generating more words than instructed.

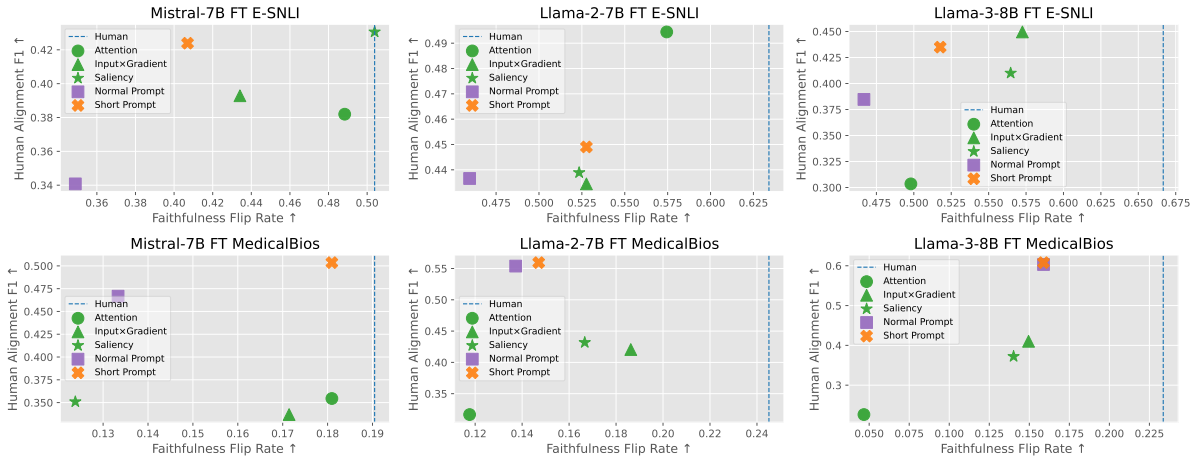


Figure 6: The human alignment (F1) and model faithfulness flip rate, evaluated at the intersection of correctly predicted examples (important for alignment) and the samples that were predicted within the output space of classes prior to fine-tuning (important for faithfulness). Additionally, the number of masks is constrained to the Top-Var selection.

| Method | Selection | Prompt |
|---|---|---|
| NORMAL PROMPT | UNBOUND | You will be provided a short premise and a hypothesis, each delimited with triple backticks.<br>Your task is to identify the most important key words in the premise and hypothesis that are most relevant to understanding the relationship between the premise and the hypothesis given a label (entailment/neutral/contradiction).<br>• These key words should be selected to capture the main idea or reasoning behind why the given label indicates the relationship between the premise and the hypothesis.<br>• The key words should be copied verbatim from the text.<br>• Write the key words as a pipe-separated (\|) list of words with spaces.<br>• Example output: <word1> \| <word2> \| <word3><br>• Make sure to only output keyword tokens and do not explain the answer.<br><br>Premise: “‘ {premise} “‘<br>Hypothesis: “‘ {hypothesis} “‘<br>Label: {label} |
| SHORT PROMPT | UNBOUND | Identify the most important key words from the text that are most relevant to understanding the relationship between the premise and the hypothesis.<br>Write the words as a pipe-separated (\|) list of words with spaces and do not explain the answer.<br><br>Premise: “‘ {premise} “‘<br>Hypothesis: “‘ {hypothesis} “‘<br>Label: {label} |
| EXTENDED PROMPT | UNBOUND | Consider the following prompt: " 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.<br>Give your answer as "entailment", "neutral" or "contradiction" and nothing else.<br>Premise and Hypothesis: “‘ {premise} {hypothesis} “‘<br>Label: "<br><br>Identify the most important single key words from the whole prompt (Including instruction and sentence) that are most relevant to understanding the relationship between the premise and the hypothesis as {label}.<br>Just write the words as a pipe-separated (\|) list of words with spaces and do not explain the answer. |
| NORMAL PROMPT | TOP-VAR / TOP-RATIO | You will be provided a short premise and a hypothesis, each delimited with triple backticks.<br>Your task is to identify the top {k} most important key words in the premise and hypothesis that are most relevant to understanding the relationship between the premise and the hypothesis given a label (entailment/neutral/contradiction).<br>• These key words should be selected to capture the main idea or reasoning behind why the given label indicates the relationship between the premise and the hypothesis.<br>• The key words should be copied verbatim from the text.<br>• Write the key words as a pipe-separated (\|) list of words with spaces.<br>• Example output: <word1> \| <word2> \| <word3><br>• Make sure to only output top {k} keyword tokens and do not explain the answer.<br><br>Premise: “‘ {premise} “‘<br>Hypothesis: “‘ {hypothesis} “‘<br>Label: {label} |
| SHORT PROMPT | TOP-VAR / TOP-RATIO | Identify the top {k} most important key words from the text that are most relevant to understanding the relationship between the premise and the hypothesis.<br>Write the top {k} words as a pipe-separated (\|) list of words with spaces and do not explain the answer.<br>Premise: “‘ {premise} “‘<br>Hypothesis: “‘ {hypothesis} “‘<br>Label: {label} |
| ATTRIBUTION-BASED | TOP-VAR / TOP-RATIO (Selected later from tokens with the highest attribution scores) | 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.<br>Give your answer as "entailment", "neutral" or "contradiction" and nothing else.<br><br>Premise and Hypothesis: “‘ {premise} {hypothesis} “‘<br>Label: {predicted_class} |
| CLASSIFICATION | | 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.<br>Give your answer as "entailment", "neutral" or "contradiction" and nothing else.<br><br>Premise and Hypothesis: “‘ {premise} {hypothesis} “‘<br>Label: |

Table 12: The prompts utilized for the e-SNLI dataset.

| Method | Selection | Prompt |
|---|---|---|
| NORMAL PROMPT | UNBOUND | You will be provided a biography which is delimited with triple backticks.<br>Your task is to identify the most important key words from the text that indicate why is the person in the following short bio described as a {label}.<br>• The key words should be copied verbatim from the text.<br>• Write the key words as a pipe-separated (\|) list of single words.<br>• Make sure to only output keyword tokens and do not explain the answer.<br>• Example output: <word1> \| <word2> \| <word3><br><br>Bio: ‘‘‘ {bio} ‘‘‘ |
| SHORT PROMPT | UNBOUND | Identify the most important key words from the text that indicate why is the person in the following short bio described as a {label}.<br>• Write the key words as a pipe-separated (\|) list of single words.<br><br>Bio: ‘‘‘ {bio} ‘‘‘ |
| EXTENDED PROMPT | UNBOUND | Consider the following prompt:<br>" Determine the occupation of the person in the following short bio.<br>Give your answer as a single word from (psychologist, surgeon, nurse, dentist, physician) and nothing else.<br><br>Bio: ‘‘‘ {bio} ‘‘‘<br>Occupation: "<br><br>Identify the most important single key words from the whole prompt (Including instruction and bio) that indicate why is the person in the short bio described as a {label}.<br>Just write the words as a pipe-separated (\|) list of words with spaces and do not explain the answer. |
| NORMAL PROMPT | TOP-VAR / TOP-RATIO | You will be provided a biography which is delimited with triple backticks.<br>Your task is to identify the top {k} most important key words from the text that indicate why is the person in the following short bio described as a {label}.<br>• The key words should be copied verbatim from the text.<br>• Write the key words as a pipe-separated (\|) list of single words.<br>• Make sure to only output top {k} keyword tokens and do not explain the answer.<br>• Example output: <word1> \| <word2> \| <word3><br><br>Bio: ‘‘‘ {bio} ‘‘‘ |
| SHORT PROMPT | TOP-VAR / TOP-RATIO | Identify the top {k} most important key words from the text that indicate why is the person in the following short bio described as a {label}.<br>• Write the key words as a pipe-separated (\|) list of single words and do not explain.<br><br>Bio: ‘‘‘ {bio} ‘‘‘ |
| ATTRIBUTION-BASED | TOP-VAR / TOP-RATIO (Selected later from tokens with the highest attribution scores) | Determine the occupation of the person in the following short bio.<br>Give your answer as a single word from (psychologist, surgeon, nurse, dentist, physician) and nothing else.<br><br>Bio: ‘‘‘ {bio} ‘‘‘<br>Occupation: {predicted_label} |
| CLASSIFICATION | | Determine the occupation of the person in the following short bio.<br>Give your answer as a single word from (psychologist, surgeon, nurse, dentist, physician) and nothing else.<br><br>Bio: ‘‘‘ {bio} ‘‘‘<br>Occupation: |

Table 13: The prompts utilized for the MedicalBios dataset.