Self-Rationalization in the Wild: A Large Scale Out-of-Distribution Evaluation on NLI-related tasks

Anonymous TACL submission

Abstract

Free-text explanations are expressive and easy to understand, but many datasets lack annotated explanation data, making it challenging to train models for explainable predictions. To address this, we investigate how to use existing explanation datasets for self-rationalization and evaluate models' out-of-distribution (OOD) performance. We fine-tune T5-Large and OLMo-7B models and assess the impact of fine-tuning data quality, the number of fine-tuning samples, and few-shot selection methods. The models are evaluated on 19 diverse OOD datasets across three tasks: natural language inference, fact-checking, and hallucination detection in abstractive summarization. For the generated explanation evaluation, we conduct a human study on 13 selected models and study its correlation with the Acceptability score (T5-11B) and three other LLM-based reference-free metrics. Human evaluation shows that the Acceptability score correlates most strongly with human judgments, demonstrating its effectiveness in evaluating free-text explanations. Our findings reveal: 1) few annotated examples effectively adapt models for OOD explanation generation; 2) compared to sample selection strategies, fine-tuning data source has a larger impact on OOD performance; and 3) models with higher label prediction accuracy tend to produce better explanations, as reflected by higher Acceptability scores.[1]

1 Introduction

Generating textual explanations has been a major focus in machine learning and NLP (Wei et al., 2022; Kunz and Kuhlmann, 2024; Calderon and Reichart, 2024), as the explanations are expressive and do not require readers to have model-level knowledge to understand. One popular line of work is self-rationalization (Wiegreffe et al., 2021; Marasović et al., 2022), in which a model jointly generates the task label and a free-text explanation for the predicted label. Compared with highlighting words and phrases (DeYoung et al., 2020), free-text explanations can express unstated knowledge and common-sense in easily understandable forms. However, datasets containing annotated free-text explanations are rare due to expensive annotations.

A few datasets for free-text explanation generation (Camburu et al., 2018; Wang et al., 2019b; Sap et al., 2020; Aggarwal et al., 2021; Chen et al., 2022) exist, with e-SNLI (Camburu et al., 2018) being one of the seminal datasets in the NLI area. Based on SNLI (Bowman et al., 2015), the dataset focuses on reasoning over fine-grained nuances of common-sense knowledge. However, datasets containing longer or more domain-specific text, such as fact-checking on real-world claims, lack annotated explanations (Hanselowski et al., 2019; Saakyan et al., 2021). This poses severe challenges for (*i*) training and (*ii*) evaluating self-rationalizing models on these tasks. No large scale analysis exists to understand how well self-rationalization models can transfer from existing data to unknown datasets.

We fill the gap by learning self-rationalization from established sources with annotated explanations and evaluating its generalization performance on 19 out-of-distribution (OOD) datasets over three related tasks (see evaluation setup in Figure 1): NLI, fact-checking (FC) and hallucination detection of abstractive summarization (HDAS). NLI focuses on textual entailment within a controlled context, FC extends to reason over real-world claims with retrieved evidence, and HDAS centers around machine-generated text. Our OOD datasets vary in *domains* (e.g., news, Wikipedia, social media,

[1] We will make all our code available upon acceptance under the MIT license.

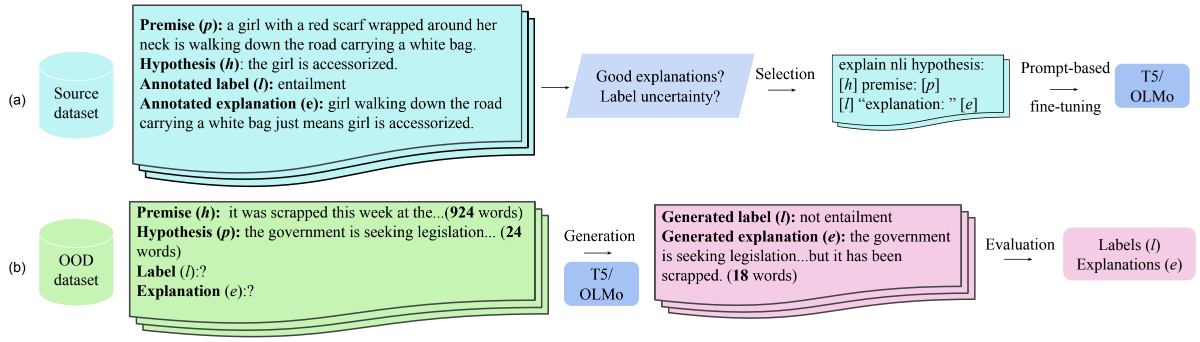


Figure 1: OOD evaluation pipeline of self-rationalization. The pipeline comprises two main parts. The first part (a) relates to **learning to self-rationalize** with a source dataset (Section 3); it involves sample selection and fine-tuning a generative model. The second part (b) relates to **OOD generation and evaluation** (Section 4); we evaluate the model on three categories of OOD tasks: NLI, fact-checking, and hallucination detection.

science), and *textual structures* (e.g., synthetic template-based, multiple premises, sentence compositions, long documents), presenting a diverse and challenging OOD setting (see details of each dataset in Table 1).

Despite the popularity of LLMs, using them in a large experimental design is prohibitive, as they are computationally expensive to perform inference and evaluation, especially when the input text is long. Further, data contamination is a concern when performing evaluations on OOD datasets (Sainz et al., 2023), as the training data of most LLMs are not transparent, such as Llama 2 (Touvron et al., 2023) and GPT-4 (Achiam et al., 2023). To address this, we selected two open-source models—T5-Large (Raffel et al., 2020) and OLMo-7B (Groeneveld et al., 2024)—to study self-rationalization, both of which have fully transparent pretraining datasets. They also require fewer computational resources than many LLMs, allowing us to perform a large scale study.

We study the impact of data size and quality on OOD performance, focusing on these three factors: the source dataset for fine-tuning, the number of selected samples, and sample selection strategies for few-shot fine-tuning. To enhance the quality of generated explanations in OOD datasets, we introduce a new approach with an acceptability filtering model (Wiegreffe et al., 2022) to select better training samples. We address the lack of gold reference explanations by studying the effectiveness of the Acceptability score with a human evaluation and comparing it against three LLM-based reference-free metrics. Out of the automatic metrics, the Acceptability score correlates highest with humans in all three tasks. Our evaluation results show that: 1) OOD performances are comparable between models fine-tuned with few-shot selected samples and a full training set; 2) fine-tuning data source has a high impact on OOD performance, while sample selection has a lower impact; 3) higher Acceptability scores are associated with better label prediction performances, providing a new perspective on the task performance vs explainability trade-off.

2 Related Work

Free-text explanation generation and evaluation Self-rationalization has been a popular approach for generating free-text explanations (Wiegreffe et al., 2021; Marasovic et al., 2022; Ross et al., 2022; Veerubhotla et al., 2023; Ramnath et al., 2024). Wiegreffe et al. (2021) shows that joint learning of label prediction and explanation generation results in explanations more aligned with predicted labels. Marasovic et al. (2022) addressed the scarcity of annotated explanation data by using prompt-based fine-tuning on a few examples, though their evaluation was limited to in-distribution datasets. Few works have studied how such models can generalize to OOD. Zhou and Tan (2021) studied how learning with few-shot instances with template-based explanations influences OOD generalization. Their OOD dataset (e-HANS) is limited with constructed templates based on the HANS dataset (McCoy et al., 2019). Ross et al. (2022) studied the effect of self-rationalization on reducing models' reliance on spurious cues in out-of-domain datasets, and they showed that self-rationalization improves models robustness when fine-tuning data size is small.

Yordanov et al. (2022) studied the setup where the target dataset has few annotated free-text explanations but abundant labels. Their approach is limited to target datasets in which free-text explanations exist. In contrast to the above OOD evaluations, we focus on the OOD evaluation of self-rationalization for 19 diverse datasets, and our evaluation does not rely on reference explanations.

Reliable evaluation is crucial for explanation generation. Traditional metrics that measure text overlap with references have shown low correlation with human judgments (Sulem et al., 2018), and reference explanations are not always available. Recent works, like TigerScore (Jiang et al., 2023), Auto-J (Li et al., 2024a), and Themis (Hu et al., 2024), use LLMs as evaluators. These metrics rely on detailed instructions specifying evaluation aspects (e.g., relevance, accuracy, coherence) and formatted inputs for the task. The trained metric then generates a rating along with a textual analysis. To test their suitability for the explanation generated with self-rationalization, in this work, we study their correlations with human judgments.

Few-shot sample selection Recent studies show that fine-tuning with smaller, high-quality datasets can outperform larger datasets (Li et al., 2024b; Xia et al., 2024). Li et al. (2024b) proposed to use a relatively small language model to evaluate and select a few instances for instruction-tuning on larger models. To select data to perform well in transfer learning, Xia et al. (2024) proposed data selection for instruction-tuning on a target-specific domain. They show that training with 5% of the data outperforms training with the full dataset. The main constraint is that the validation set needs to be from the target domains. Chen and Mueller (2024) proposed to improve data quality by estimating their model's confidence, and for the low-quality data, they either filter or correct them. Most methods for sample selection are designed to perform well on in-distribution or known target domains, and the goal is for better classification performance. In contrast, our work focuses on selecting data that should help OOD performance on both label prediction and explanation generation.

3 Learning to Self-rationalize

Figure 1 shows our out-of-distribution (OOD) evaluation pipeline. We first (a) fine-tune a language model on a source dataset to learn self-rationalization. Specifically, we require a fully annotated source dataset S, in which each instance contains input $x_s = (h_i, p_i)$ and output $y_s = (l_i, e_i)$, where h_i, p_i represent a hypothesis and premise pair, l_i and e_i represent the annotated label and explanation. We select m representative instances per class from S for fine-tuning by following a sample selection process. Our sample selection method deliberately restrains from using data from the OOD datasets, preserving them untouched. Finally, we fine-tune a language model to generate a label and explanation. In (b), we evaluate the fine-tuned model performance on OOD datasets (Section 4). Given an OOD dataset O, with instances $x_o = (h_j, p_j)$, where h_j, p_j represents a new hypothesis and premise pair, the fine-tuned model generates the label ($\hat{l_j}$) and explanation ($\hat{e_j}$).

3.1 Source dataset

To learn self-rationalization for NLI-related tasks, we select two large source datasets that contain explanations: (a) **e-SNLI** (Camburu et al., 2018), derived from the NLI dataset SNLI (Bowman et al., 2015) by adding human annotated explanations. (b) **e-FEVER** (Stammbach and Ash, 2020), originated from the fact-checking dataset FEVER (Thorne et al., 2018) with GPT-3 generated synthetic explanations. To improve data quality, we heuristically filter out incorrect explanations from the dataset (see details in Appendix A.1). We selected these two datasets as they are representative for our OOD datasets and have abundant explanations.

3.2 Acceptability-based sample selection

Inspired by Schiller et al. (2022), we examine how varying the size and quality of fine-tuning data (source dataset) affects OOD performance. Since self-rationalization includes joint label prediction and explanation generation, we propose our method considering both the label and explanation quality:

Data filtering with acceptability score To improve explanation quality, we filter the fine-tuning data using the acceptability model from Wiegreffe et al. (2022). This model, trained on SNLI data, predicts whether a generated explanation is acceptable based on human judgment. We remove samples with acceptability scores (the predicted probability for the label "acceptable") below a 0.3 threshold.

Data selection For data quality estimation in label prediction, we adapt two methods from the literature: (1) **ambiguous**: Following Swayamdipta et al. (2020), we select samples with high ambiguity, which has been shown to improve OOD generalization. Ambiguity is measured as the distance between an instance's predicted label probability and the mean of all predicted label probabilities using the pre-fine-tuning model (details in Appendix A.2). (2) **FastVote-*k*** (Su et al., 2022): A graph-based method to select diverse and representative samples. We use the recommended k = 150.

With the combined two steps (data filtering + selection), we denote the sample methods as **accept-ambiguous** and **accept-FastVote-*k***.

3.3 Fine-tuning on source datasets

For fine-tuning T5-Large, we use the standard NLI template from (Marasovic et al., 2022), which has been shown to give the best results for e-SNLI dataset with T5. The encoder and decoder prompts are (also shown in Figure 1) :

> **Input**: *explain nli hypothesis:* [hypothesis] *premise:* [premise]
> **Output**: [label] *"explanation: "* [explanation]

For fine-tuning OLMo-7B, as the model is relative large, we choose parameter-efficient tuning with LoRA (Hu et al., 2022) using the following instruction (Zarharan et al., 2024). The response is in a JSON format to facilitate extraction of labels and explanations:

```
### Premise: [premise] Hypothesis: [hypothesis]
### Response: {"relationship": [label], "explanation": [explanation]}
```

For the number of shots, we compare 1, 2, 4, 8, 16, 32, 64, and 128 shots. To ensure robustness, we create five subsets from each source dataset, with 5,000 randomly selected samples per subset (with no overlap between subsets). We apply the sample selection methods from Section 3.2 to each subset and report the average results (see Appendix A.2 for additional fine-tuning details). In total, we fine-tuned 402 T5 models and 302 OLMo models[2].

Baselines We compare the few-shot fine-tuned models with two full-set fine-tuned models on e-SNLI and e-FEVER, respectively. In addition, we include the random sample selection baseline to compare few-shot sample selection methods.

4 OOD Generation and Evaluation

In this section, we introduce part (b) of the pipeline in Figure 1. For all fine-tuned models, we perform inference on all OOD datasets.

4.1 Out-of-Distribution datasets

For a comprehensive evaluation, we collect datasets that resemble the NLI task and divide them into three categories: **NLI**, Fact-checking (**FC**), and Hallucination Detection of Abstractive Summarization (**HDAS**). Table 1 lists the OOD datasets used (see Appendix A.1 for dataset details and pre-processing). To ensure no data contamination in our OOD evaluation, we specifically excluded datasets used for supervised fine-tuning of T5 (Raffel et al., 2020). OLMo model was pre-trained on Dolma (Soldaini et al., 2024) corpus, which contains data from diverse sources but is not fine-tuned with curated NLI datasets.

NLI NLI datasets access models' ability to infer relationships between sentences, with challenges ranging from compositional meaning (Marelli et al., 2014), adjective-noun composition (Pavlick and Callison-Burch, 2016), common-sense inference (Zhang et al., 2017), to multiple premise entailment (Lai et al., 2017). DNC (Poliak et al., 2018a) expands the challenge by incorporating diverse semantic phenomena into the NLI format. HANS (McCoy et al., 2019) and WNLI (Wang et al., 2019a) are two adversarial datasets designed to reveal models' underlying heuristic biases. Glue Diagnostics (Wang et al., 2019a) and ConjNLI (Saha et al., 2020) further diversify the NLI task, testing models against a wide array of linguistic challenges and over conjunctive sentences.

FC FC datasets aim to evaluate the veracity of claims against evidence from various sources, including fact-checking platforms (Hanselowski et al., 2019), scientific articles (Wadden et al., 2020), Wikipedia (Schuster et al., 2021; Eisenschlos et al., 2021), and information related to climate change and COVID-19 (Diggelmann et al., 2020; Saakyan et al., 2021). The domain-specific nature of some datasets, such as SciFact's focus on biomedicine and Climate FEVER's on climate change, requires models to be domain-aware and handle evidence with varying granularity. FC

[2] For T5: 2 source datasets ×5 subsets ×8#shots ×5 sampling methods +2 full-shot models. For OLMo, we discard 1 and 2 shots as our primary results show that models fail to learn with too few examples.

	OOD dataset	Size	#L.	Domain	#words (Hyp.)	#words (Pre.)	IAA
NLI	SICK (Marelli et al., 2014)	4,906	3	news, image captions	10	10	0.84^{O}
	AddOneRTE (Pavlick and Callison-Burch, 2016)	387	2	news, image captions, forums, literature	13	12	0.77^{O}
	JOCI (Zhang et al., 2017)	39,092	3	image captions, commonsense stories	6	14	0.54^{C}
	MPE (Lai et al., 2017)	1,000	3	image captions	4	48	0.70^{O}
	DNC (Poliak et al., 2018a)	60,036	2	events, named entities, puns, sentiments	5	19	-
	HANS (McCoy et al., 2019)	30,000	2	template-based (synthetic)	6	9	-
	WNLI (Wang et al., 2019a)	71	2	fiction books	7	21	-
	Glue Diagnostics (Wang et al., 2019a)	1,104	3	news, Reddit, Wikipedia, academic papers	16	16	0.73^{F}
	ConjNLI (Saha et al., 2020)	623	3	Wikipedia	13	13	0.83^{C}
FC	Snopes Stance (Hanselowski et al., 2019)	1651	3	Snopes (fact-checking platform)	16	126	0.70^{C}
	SciFact (Wadden et al., 2020)	300	3	biomedicine, scientific articles	13	247	0.75^{C}
	Climate-FEVER (Diggelmann et al., 2020)	1,381	3	climate change, Google searches	20	136	0.33^{K}
	VitaminC (Schuster et al., 2021)	55,197	3	Wikipedia, COVID-19	13	28	0.71^{F}
	COVID-FACT (Saakyan et al., 2021)	4,086	2	Reddit, COVID-19	12	73	0.50^{C}
	FM2 (Eisenschlos et al., 2021)	1,380	2	Wikipedia	14	32	-
HDAS	FactCC (Kryscinski et al., 2020)	503	2	news (CNN/DailyMail), rule-based	14	644	0.75^{C}
	QAGs CNNDM (Wang et al., 2020)	714	2	news (CNN/DailyMail), BART-based	16	318	0.51^{K}
	QAGs XSUM (Wang et al., 2020)	239	2	news (XSUM), BART-based	18	351	0.34^{K}
	XSUM Hallucination (Maynez et al., 2020)	1,869	2	news (XSUM), 7 different models	19	361	0.92^{O}

Table 1: OOD datasets categories and details. NLI: yellow, FC: pink, and HDAS: blue. Hyp.: hypothesis, Pre.: premise, #words: number of words in average, IAA: inter-annotator agreement (numbers are from the original papers). L.: labels, C: Cohen's kappa, F: Fleiss's kappa, K: Krippendorff's alpha, O: other metrics, -: unspecified. The sizes are reported on test/dev split; if the split is not provided, we report and evaluate on the entire dataset.

datasets challenge models to evaluate the truthfulness of claims in real-world scenarios with applied NLI techniques. For all FC datasets, we use gold evidence, considering that retrieved evidence may change the gold label of the claim).

HDAS HDAS datasets encompass a variety of model-generated summaries, reflecting the evolving landscape of automatic text generation and its implications for information integrity. FactCC (Kryscinski et al., 2020) challenges models to identify inaccuracies in summaries generated through five rule-based transformations. QAGS CNN and QAGS XSUM (Wang et al., 2020), derived from CNN/DailyMail and XSUM datasets, consist of summaries generated by the BART model (Lewis et al., 2020). XSUM Hallucination (Maynez et al., 2020) contains factuality annotated summaries generated by seven models.

In comparison, the three tasks vary in objective, domain, and text length. NLI targets logical relationships between sentences, requiring models to handle linguistic subtleties and logic-based reasoning in a controlled textual context. FC focuses on real-world applicability, requiring external information and complex reasoning between sentences and documents. HDAS addresses the problems of automatic document summarization. Regarding text length, FC datasets typically have longer premises than NLI, with HDAS having the longest. Together, these datasets present a challenging NLI-related OOD scenario.

4.2 Inference on OOD datasets

During OOD inference, fine-tuned models may not generate a label and explanation following the output template. To address this, for T5 models, we take the first token to represent the predicted label. For datasets that only include two classes ("entailment" and "non-entailment"), we merge the "contradiction" and "neutral" labels into the "non-entailment" label (see more details on label extraction in Appendix A.3). We detect explanations by searching for the pattern "explanation: " and, if absent, treat all text after the first word as the explanation. For OLMo models, as we instruction-tuned the model to generate a JSON-formatted output, we extract the labels and explanations by finding their keys and if not found, we set both to be none.

5 Results and Analysis

In this section, we first present label prediction performance results. Next, we evaluate explanations through human judgments and analyze their correlation with reference-free metrics. We then report explanation evaluation results across all datasets using the most correlated reference-free metric. Finally, we present the overall OOD performance on all 19 datasets on the best-performing models.

5.1 OOD Performance on Label Prediction

We compare the OOD label prediction performance of fine-tuned T5-Large and OLMo-7B models on two source datasets, considering various sample selection methods and number of shots, as shown in Figure 2. Label prediction performance is measured using the Macro F1 score.

T5 vs. OLMo: As shown in Figure 2, T5 and OLMo models exhibit distinct trends in label prediction performance as the number of shots increases. OLMo starts with low performance, improving almost monotonically with more shots. T5, however, shows less variation, starting with slightly higher performance and then reaching levels similar to full-shot models. This difference may be because of T5's pre-training on NLI datasets (MNLI, QNLI, RTE, CB), allowing it to handle NLI tasks effectively without much benefit from additional fine-tuning (see detailed discussion in Section 6.1). This is further indicted by the results: T5 full-shot fine-tuning with both source datasets have similar F1 scores, and neither yields better results than their best few-shot counterparts.

e-SNLI vs. e-FEVER: Overall, e-FEVER models achieve better average OOD F1 than e-SNLI, and the OLMo model fine-tuned on e-FEVER full-shot has the highest OOD F1 score. For e-SNLI, T5 and OLMo models reach similar performances at 128 shots, but the trends are the opposite. For e-FEVER, T5 models' performance tends to stabilize after just 2-shots, while OLMo models' performance continues to increase and eventually outperform T5 models.

Sample Selection As depicted in Figure 2, no sample selection method consistently outperforms others in label prediction. For T5, selection methods perform similarly, especially with e-SNLI, though "accept-ambiguous" is slightly better with e-FEVER. For OLMo, "FastVote-k" excels with e-SNLI, while "random" selection outperforms others with e-FEVER (after 32 shots), nearly matching full-shot performance. Surprisingly, "FastVote-k" and "ambiguous" do not surpass the random baseline, possibly due to outliers and training instability when using small numbers of samples (Karamcheti et al., 2021; Su et al., 2022).

Acronym	Source	Model	#Shots	Selection
$\mathrm{T}^{Fev}_{64,AFk}$	e-FEVER	T5	64	accept-FastVote-k
$\mathrm{T}^{Fev}_{128,R}$	e-FEVER	T5	128	random
$\mathrm{T}^{Fev}_{128,Fk}$	e-FEVER	T5	128	FastVote-k
$\mathrm{T}^{Fev}_{128,AFk}$	e-FEVER	T5	128	accept-FastVote-k
T^{Fev}_{Full}	e-FEVER	T5	Full	-
$\mathrm{T}^{Sn}_{64,Fk}$	e-SNLI	T5	64	FastVote-k
$\mathrm{T}^{Sn}_{64,AFk}$	e-SNLI	T5	64	accept-FastVote-k
T^{Sn}_{Full}	e-SNLI	T5	Full	-
$\mathrm{O}^{Fev}_{16,AFk}$	e-FEVER	OLMo	16	accept-FastVote-k
$\mathrm{O}^{Fev}_{128,AFk}$	e-FEVER	OLMo	128	accept-FastVote-k
O^{Fev}_{Full}	e-FEVER	OLMo	Full	-
$\mathrm{O}^{Sn}_{128,AFk}$	e-SNLI	OLMo	128	accept-FastVote-k
O^{Sn}_{Full}	e-SNLI	OLMo	Full	-

Table 2: Selected models for human evaluation for the models **T**5 and **O**LMo. The left most column shows the acronym of the models, which will be used throughout the rest of the paper.

5.2 OOD Explanation Quality Evaluation

We evaluate the generated explanations using both human evaluation and reference-free automatic metrics, and analyze the correlation between them.

5.2.1 Human evaluation setup

Conducting a human study is challenging due to the extensive number of models and OOD datasets. Thus, we select three OOD datasets (SICK, VitaminC, XSUM Hallucination) representing NLI, FC, and HDAS, respectively. To study the impact of fine-tuning factors on OOD explanations, we select models that demonstrated high and comparable F1 scores averaged across the three OOD datasets (see Figure 6 in Appendix B with the selected models highlighted). Table 2 lists the 13 selected mode details, with first column provides models' acronyms for across reference later (examples of generated explanations by the selected models can be found in Table 7, 8 and 9 in Appendix A.6).

For instance selection, following Marasovic et al. (2022), we shuffle each dataset and select the first 15 correctly predicted instances per class and model. This results in 1560 instances, including those with identical hypothesis-premise pairs but different model-generated explanations. Each instance is evaluated by three different workers, and each worker evaluate 10 instances, requiring in total 468 crowd-workers. Evaluators are shown the hypothesis-premise pair, its relationship (gold label), and the generated explanation and then asked to answer two questions (see the evaluation page in

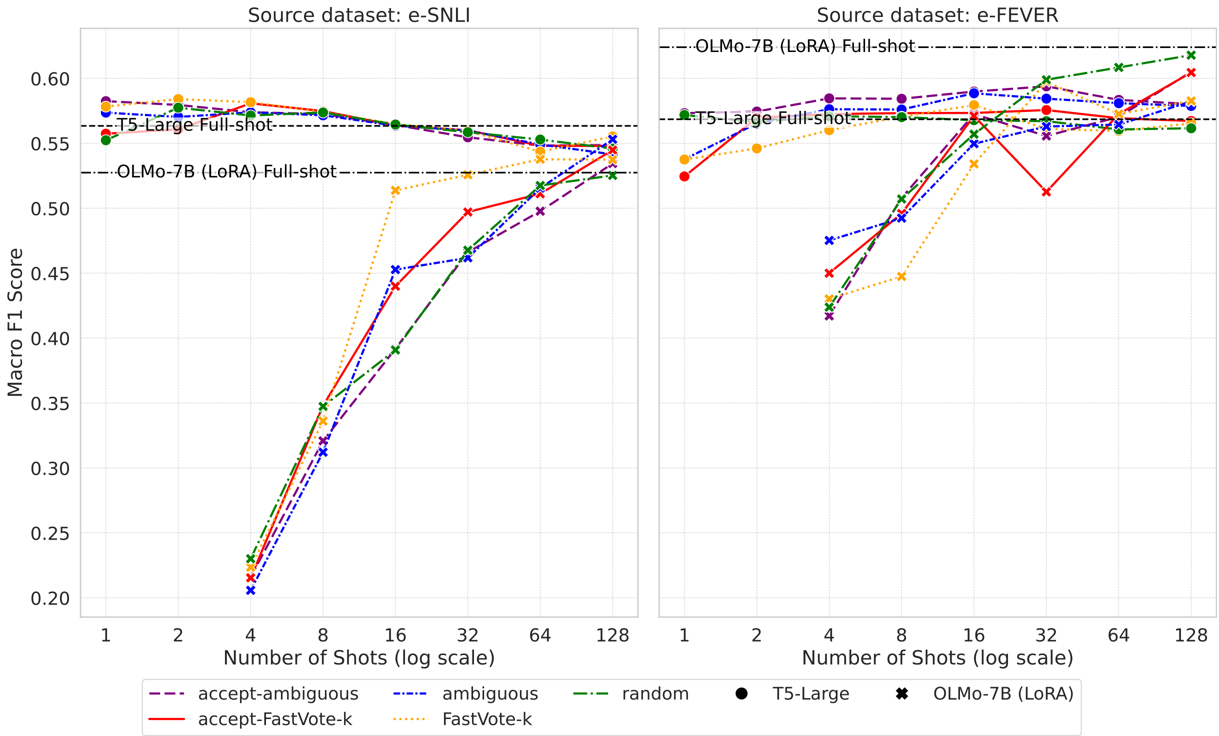


Figure 2: Average Macro F1 score across different number of shots and sample selection methods. Each point is the average of all 19 OOD datasets, and 5 models from the 5 subsets.

Figure 5 of Appendix A.4).

- Given the Hypothesis and Premise, does the Explanation justify the given Relationship (Single-selection)? Options: *Yes*, *Weakly Yes*, *Weakly No* and *No*.
- What are the shortcomings of the Explanation (Multi-selection)? Options: *Does not make sense*, *Insufficient justification*, *Irrelevant to the task*, *Too trivial (only repeating one of the sentences)*, *Contains hallucinated content (not present the premise)* and *None (only if the previous answer is Yes)*.

We calculate the average score of each instance from 3 evaluators by assigning the weight to the selected answers as follows (Marasovic et al., 2022; Yordanov et al., 2022): Yes: 1, Weakly Yes: 2/3, Weakly No: 1/3 and No: 0.

We use the Prolific platform for recruiting workers, and the open-source POTATO annotation tool (Pei et al., 2022) for the evaluation interface.

5.2.2 Evaluation with reference-free metrics

We propose to use the **Acceptability score**[3] (Wiegreffe et al., 2022) as a reference-free metric, considering it is designed for accessing NLI explanations. We choose the largest size of the model variance: T5-11B. The model assigns a score between 0 and 1. We compare this metric against the state-of-the-art NLG reference-free evaluation metrics (see Appendix A.5 for the instructions of the evaluation models):

- **Auto-J** (Li et al., 2024a): trained with LLaMA-2-13B-chat model to evaluate LLM-generated responses. The metric generates an explanation for its judgment and a final integer rating from 1 to 10.
- **TigerScore** (Jiang et al., 2023): trained with LLaMA-2 on MetricInstruct dataset. We choose the larger size of the metric: TIGERScore-13B. It generates a breakdown error analysis and a final error score from 0 to infinity (the smaller, the better).
- **Themis** (Hu et al., 2024): trained with Llama-3-8B based on their constructed dataset NLG-Eval. It offers flexible aspect-based evaluations across different tasks. We tested three aspects—relevance, coherence, and consistency—and selected relevance due to its highest correlation with human judgments. The metric outputs an evaluation analysis and provides a scale rating from 1 to 5.

[3]In this paper, when mentioning the acceptability filter (T5-Large), we start with lowercase "a", and the Acceptability metric (T5-11B) capital "A".

Dataset	Auto-J	TigerScore	Themis	Accept.
SICK	-0.011	-0.220	0.400	**0.466**
VitaminC	0.163	-0.263	0.394	**0.469**
XSUM H.	0.223	-0.216	0.326	**0.475**
All	0.123	-0.219	0.387	**0.484**

Table 3: Spearman's correlation between human scores and automatic scores in different OOD datasets. All correlation coefficients are significant with $\rho < 0.001$, except for Auto-J on SICK.

5.2.3 Correlation between human evaluation and automatic evaluation metrics

Table 3 shows the Spearman's correlation[4] between human and reference-free metrics for the three OOD datasets. The Acceptability score (T5-11B) has the highest correlation with human evaluation for all datasets, followed by Themis, and Auto-J has the lowest. The highest correlations on all three datasets demonstrate the usability of the Acceptability score as a reference-free metric for the explanation evaluation of NLI-related tasks.

5.2.4 Evaluation results on selected models and instances

The average scores of human evaluations in the three OOD datasets are shown in Table 10 in Appendix B. The scores show that SICK has the highest explanation scores, with VitaminC slightly lower than SICK's, and XSUM Hallucination the lowest, agreed by humans and two automatic metrics. This may be due to the extremely long premise/document in the XSUM dataset, making it difficult for the model to generate good explanations. For shortcomings of explanations, see the detailed results in Figure 7 in Appendix B).

Table 4 shows the evaluation results on the 13 selected models. We include Acceptability and Themis scores as they have moderate correlations with humans. In addition, we show the average Acceptability score on all 19 datasets for overall results. We discuss the evaluation results regarding each factor in the following.

T5 vs OLMo As shown in Table 4, the difference between the two base models is most pronounced with e-SNLI full-shot. T5 fine-tuned on full shot e-SNLI (T^{Sn}_{Full}) provides the best explanations (besides $T^{Sn}_{64,AFk}$), whereas OLMo on full-shot e-SNLI (O^{Sn}_{Full}) generates the worse explanations. This may be due to catastrophic forgetting in the OLMo model when fine-tuned on too many e-SNLI samples, as its few-shot version produces explanations comparable to those of the T5 model.

[4]We choose Spearman over Pearson correlation as Pearson correlation assumes variables to be continuous and from a normal distribution.

Model	Human	Themis	Accept. (3)	Accept. (19)
$T^{Fev}_{64,AFk}$	0.631	2.058	0.317	0.250
$T^{Fev}_{128,R}$	0.623	1.983	0.276	0.206
$T^{Fev}_{128,Fk}$	0.589	1.867	0.216	0.201
$T^{Fev}_{128,AFk}$	0.611	**2.092**	**0.328**	**0.256**
T^{Fev}_{Full}	**0.653**	1.958	0.309	0.191
$T^{Sn}_{64,Fk}$	0.621	2.133	0.369	0.259
$T^{Sn}_{64,AFk}$	**0.679**	**2.367**	0.418	0.281
T^{Sn}_{Full}	0.678	2.050	**0.519**	**0.343**
$O^{Fev}_{16,AFk}$	0.631	**2.417**	**0.423**	0.305
$O^{Fev}_{128,AFk}$	0.639	2.250	0.384	**0.307**
O^{Fev}_{Full}	**0.656**	1.917	0.311	0.219
$O^{Sn}_{128,AFk}$	**0.643**	**2.300**	**0.491**	**0.303**
O^{Sn}_{Full}	0.408	1.208	0.194	0.111

Table 4: Evaluation results on OOD datasets of the 13 selected models. 3 means on the three selected datasets, 19 means all datasets. Models are grouped by base models and source datasets.

e-SNLI vs e-FEVER Most e-SNLI models outperform e-FEVER in explanation quality (under the same model type and number of shots), except for OLMO full-shot. This could be attributed to the higher quality of explanations in the e-SNLI source dataset, while e-FEVER explanations are generated by GPT-3 (see more detailed comparison in Section 6.2).

Few vs Full Overall, few-shot models achieved similar human scores to their full-shot counterparts, except for the OLMo full-shot e-SNLI model. Although full-shot models showed slightly higher human scores, reference-free metrics favored the explanations generated by few-shot models, particularly for e-FEVER models.

Sample Selection As shown in Table 4, using the acceptability filter ("accept-FastVote-k") improves explanation quality compared with the same sample selection without the filter ("FastVote-k"); however, $T^{Fev}_{128,AFk}$ is not better than random selection ($T^{Fev}_{128,R}$) according to humans. Nevertheless, based on the scores from the two reference-free metrics, using the acceptability filter improves generated explanation quality (see more detailed discussion in Section 6.2).

5.3 Self-Rationalization in the Wild: Overall OOD Performance

A good self-rationalization model should perform well both on label prediction and explanation generation. Thus, we first evaluate the generated explanations from a large number of models using the Acceptability score (for all instances, we use the gold labels for calculating the Acceptability score). Due to computational constraints, we limit the number of shots to 4, 16, 64, 128, and full, with data selected from the first subset (the Acceptability scores across different number of shots and sample selections can be found in Figure 8 of Appendix B). We then show models' overall performance considering both the F1 and Acceptability score. Finally, we select the best-performing models to demonstrate overall performance on the 19 OOD datasets.

5.3.1 Relationship between label prediction performance and explanation quality

Figure 3 shows the distribution of models under different fine-tuning factors, with the x-axis showing the Acceptability score and the y-axis the macro F1 score (scores are averaged over all datasets). We select the best models based on the Pareto fronts[5].

As depicted in Figure 3, higher Acceptability scores are usually associated with better F1 scores. Regarding each factor, we see that 1) OLMo models' OOD performances are less stable than T5 models' but achieve better results with higher numbers of shots; 2) Sample selection methods with the acceptability filter have higher Acceptability scores; 3) Comparing the source datasets, fine-tuning on e-SNLI in general achieve higher Acceptability scores while on e-FEVER yield better F1 scores (see more discussions on the impact of each factor in Section 6).

Regarding the best-performing models that consider both labels and explanations, two models are selected based on the Pareto front: $\mathrm{O}^{Fev}_{128,AFk}$ (OLMo, 128 shots, accept-Fastvote-*k*, e-FEVER) and T^{Sn}_{Full} (T5, full-shot, e-SNLI). The first achieves the highest F1 score, while the second has the best Acceptability score, with both models performing competitively on the other metric.

[5]For each point if no other point is strictly higher in both scores, the point is part of the Pareto front. See definition in https://en.wikipedia.org/wiki/Pareto_front.

5.3.2 Performance on the 19 OOD Datasets

Table 5 shows the F1 score and Acceptability score on the best models across each OOD dataset (state-of-the-art results on each dataset can be found in Table 11 of Appendix B). As a comparison, we also include two other models with the same configurations as the best models but trained on a different source dataset: T^{Fev}_{Full} and $\mathrm{O}^{Sn}_{128,AFk}$.

As shown in Table 5, the $\mathrm{O}^{Fev}_{128,AFk}$ model achieves the highest F1 score on most OOD datasets, though its Acceptability score is slightly lower than that of the T^{Sn}_{Full} model. When comparing e-SNLI and e-FEVER fine-tuned models, e-FEVER models generally outperform in F1 scores on FC and HDAS datasets, with $\mathrm{O}^{Fev}_{128,AFk}$ scoring about 10 percentile higher on average for FC (slightly less) and HDAS (slightly more). In terms of explanation generation, OLMo-based models exhibit better performance. Even on e-FEVER, OLMo achieves competitive scores across most OOD datasets, whereas the T5 model fine-tuned on e-FEVER (T^{Fev}_{Full}) produces the worst explanations, except for the HDAS task (this might also be due to the number of shots difference, as fine-tuned on more number of shots with e-FEVER do not always lead to better explanations). Finally, the Acceptability scores show a decreasing trend from NLI to HDAS tasks, consistent with previous human evaluation results (see Table 10 in Appendix B), where datasets with longer premises generally resulted in lower Acceptability scores.

6 Discussions

This section explains the reasons for our earlier findings. First, we discuss how fine-tuning data and the model affect label prediction and explanation generation. Then, we analyze the relationship between label prediction performance and Acceptability score across the three OOD tasks.

6.1 Impact of fine-tuning dataset and base model on OOD label prediction

Source dataset Generally, OOD label prediction performance is better with models fine-tuned on the e-FEVER dataset. To explore the reasons, we show the F1 score per class for both ID and OOD test datasets (including cross-source and 9 OOD three-label datasets) in Table 12 in Appendix B, based on $\mathrm{O}^{Sn}_{128,AFk}$ and $\mathrm{O}^{Fev}_{128,AFk}$ models. $\mathrm{O}^{Sn}_{128,AFk}$ (e-SNLI) model has a better ID performance (0.86) but generalizes poorly to OOD (0.54), whereas

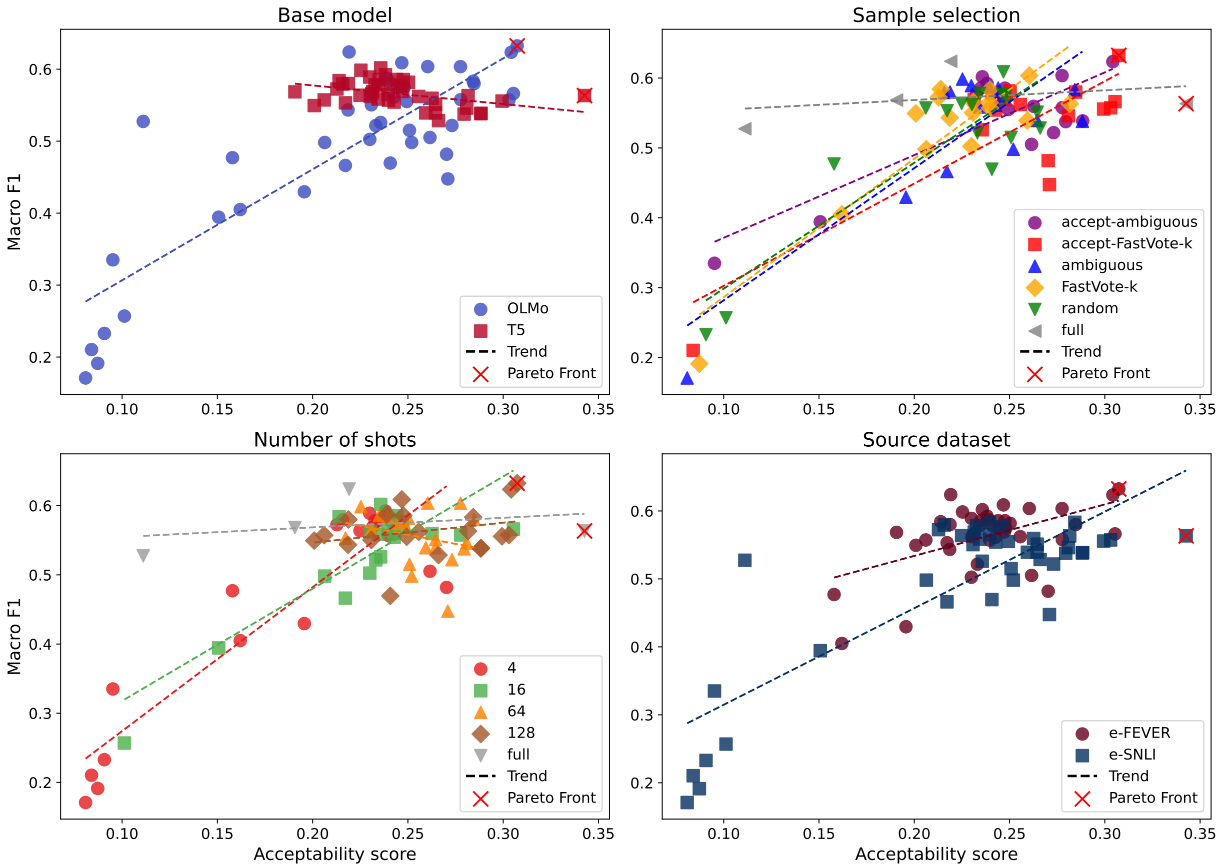


Figure 3: Distribution of models under different fine-tuning factors, with the x-axis showing the Acceptability score, and the y-axis the macro F1 score (scores are averaged over all datasets). The dashed lines are the estimated linear trends of the Acceptability score and macro F1 score.

$O^{Fev}_{128,AFk}$ (e-FEVER) model has a worse ID (0.69) but better OOD performance (0.59). For both source datasets, models perform better on e-SNLI test set than e-FEVER test set, indicating that e-FEVER is a harder dataset to learn. In addition, fine-tuning on e-FEVER helped improving performance on harder classes ("Neural (NEI)") and "Entailment (Supports)".

Base model We observed that T5 models' OOD label prediction performances are much more stable than OLMo. We believe it is due to two reasons: (1) T5 was fine-tuned for the supervised text-to-text language modeling objective (Raffel et al., 2020) including NLI datasets, and FC and HDAS are relatively similar tasks. Since we formatted the claims/summaries and evidence/documents as hypothesis/premise pairs, T5 can perform relatively well with very few shots. On the downside, the model did not improve with more fine-tuning data (especially with e-SNLI). In contrast, although OLMo models started with low performance, they eventually outperformed T5 with increased number fine-tuning samples. (2) The prompt for fine-tuning T5 matches the one used during its original supervised fine-tuning on NLI datasets, so T5 models do not need to adapt to the format for predicting NLI labels. In contrast, OLMo models perform poorly with few samples due to output formatting issues (expected in JSON format with specific keys for labels and explanations).

6.2 Impact of fine-tuning data on OOD explanation quality

Source Dataset We observed that models fine-tuned on e-SNLI generally have higher OOD Acceptability scores (when having similar F1 scores). To understand the effect of fine-tuning data on OOD explanations, Table 6 compares the two source datasets based on input length (hypothesis, premise, and explanations), average Acceptability scores of the original data (128 shots), and Acceptability and F1 scores for ID and OOD test sets. The results, based on $O^{Sn}_{128,AFk}$ and $O^{Fev}_{128,AFk}$, show that the input length has a large impact on the ID Acceptability score, but the impact on OOD is

	Macro F1 score				Acceptability score			
Dataset	$\mathbf{T}^{Sn}_{Full}$	T^{Fev}_{Full}	$\mathrm{O}^{Sn}_{128,AFk}$	$\mathbf{O}^{Fev}_{128,AFk}$	$\mathbf{T}^{Sn}_{Full}$	T^{Fev}_{Full}	$\mathrm{O}^{Sn}_{128,AFk}$	$\mathbf{O}^{Fev}_{128,AFk}$
SICK	58.5	**78.8**	55.4	<u>65.1</u>	**53.0**	18.5	<u>47.5</u>	40.2
AddOneRTE	<u>72.3</u>	**75.6**	65.0	72.0	<u>44.5</u>	9.3	**44.9**	39.4
JOCI	<u>52.5</u>	41.8	49.2	**53.7**	**51.9**	12.4	<u>43.6</u>	41.6
MPE	**68.7**	37.7	<u>62.4</u>	60.7	**49.8**	6.4	<u>45.8</u>	39.2
DNC	<u>60.1</u>	**66.9**	53.4	58.5	**35.1**	10.0	25.8	<u>32.8</u>
HANS	<u>58.2</u>	43.3	51.7	**65.9**	**38.6**	27.6	24.0	<u>27.8</u>
WNLI	35.0	32.4	<u>42.1</u>	**55.1**	<u>29.9</u>	22.7	**31.7**	28.0
Glue Diagnostics	57.9	<u>59.3</u>	57.7	**61.3**	**47.9**	29.0	<u>42.7</u>	41.9
Conj	<u>62.6</u>	**65.4**	58.1	56.9	**48.7**	30.4	<u>41.4</u>	38.7
Snopes Stance	36.8	44.1	<u>45.7</u>	**58.4**	**20.1**	9.9	18.1	<u>20.1</u>
SciFACT	60.7	<u>62.5</u>	56.2	**70.0**	<u>25.7</u>	17.6	22.5	**25.8**
Climate FEVER	46.9	<u>47.5</u>	42.4	**51.3**	**20.9**	12.8	18.4	<u>20.8</u>
VitaminC	55.8	**58.8**	55.3	<u>56.5</u>	**40.3**	29.8	<u>39.2</u>	37.2
COVID-Fact	63.3	<u>65.9</u>	55.3	**69.8**	**28.1**	12.2	19.8	<u>23.5</u>
FM2	70.2	71.7	<u>76.0</u>	**79.3**	<u>38.4</u>	24.1	**39.0**	38.1
FactCC	56.4	<u>59.6</u>	56.0	**65.2**	16.8	**27.6**	19.1	<u>24.6</u>
QAGS CNN	51.8	59.3	<u>60.0</u>	**72.5**	20.2	**26.4**	19.0	<u>25.8</u>
QAGS XSUM	55.0	59.3	<u>61.4</u>	**72.6**	**24.0**	15.9	19.0	<u>23.0</u>
XSUM H.	47.9	50.4	<u>55.8</u>	**56.9**	<u>17.3</u>	11.6	**17.6**	15.1
Avg NLI	<u>58.4</u>	55.7	55.0	**61.0**	**44.4**	18.5	<u>38.6</u>	36.6
Avg FC	55.6	<u>58.4</u>	55.2	**64.2**	**28.9**	17.7	26.2	<u>27.6</u>
Avg HDAS	52.8	57.1	<u>58.3</u>	**66.8**	19.6	**22.4**	17.9	<u>22.1</u>
Avg All	56.3	<u>56.9</u>	55.7	**63.2**	**34.3**	19.1	30.3	<u>30.7</u>

Table 5: Macro F1 and Acceptability Scores on each OOD Dataset on the best models ($\mathrm{O}^{Fev}_{128,AFk}$ and T^{Sn}_{Full}) and the different source dataset counterpart (T^{Fev}_{Full} and $\mathrm{O}^{Sn}_{128,AFk}$). The best score is bold, and second-best is underlined.

Source	Input Length	Source Accept.	ID Accept.	OOD Accept.	ID F1	OOD F1
e-SNLI	38	**0.671**	**0.565**	0.262	**82.8**	54.3
e-FEVER	118	0.394	0.367	**0.263**	58.9	**59.9**

Table 6: Performance comparison across the two source datasets.

minor (as it should depend on OOD input length). Despite lower OOD F1 scores, $\mathrm{O}^{Sn}_{128,AFk}$ (e-SNLI) model has similar OOD Acceptability scores to $\mathrm{O}^{Fev}_{128,AFk}$ (e-FEVER) model. This could be because part of the SNLI dataset was used to train the Acceptability model. Nevertheless, Acceptability score is more impacted by models' label prediction performance, as reflected by the F1 Scores.

Data Filtering Our acceptability-based (T5-Large) filtering model had only slight impacts on label prediction but improved explanation quality, according to the Acceptability score. One hypothesis is that since the Acceptability score metric (T5-11b) is a larger version of the filter model (only differing in size), the metric may favor explanations generated from models fine-tuned on acceptability-filtered samples. To investigate this, we conducted an experiment using the Themis metric as the filter for selecting samples (called "Themis-FastVote-k"), filtering out samples with ratings below 3 (on a 1-5 scale). The experiment is based on the OLMo best model ($\mathrm{O}^{Fev}_{128,AFk}$), and the results are shown in Table 13 in Appendix B. The Acceptability score with "Themis-FastVote-k"(0.303) is similar to "accept-FastVote-k"(0.307), despite having a lower F1 score. This suggests that using the acceptability filter does not cause the Acceptability metric to overestimate explanations generated from the filtered data.

6.3 Relationship between label prediction performance and Acceptability score

In Figure 3, we observed a positive correlation between F1 and Acceptability scores across models. We analyze on the best e-SNLI and e-FEVER models to further explore the relationship between label prediction performance and the Acceptability score within a model. We calculated the average balanced accuracy (used instead of F1 to account for varying class counts across datasets) for each task within different Acceptability score ranges, shown in Fig-

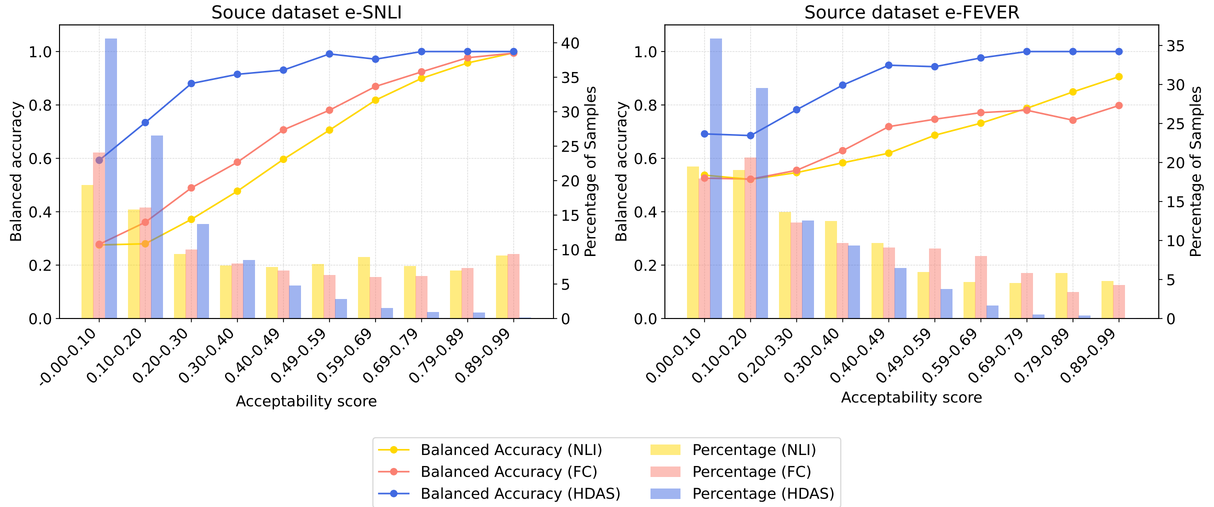


Figure 4: Distribution of label prediction accuracy (balanced) across different Acceptability score ranges. The left y-axis shows the balanced accuracy of samples from that Acceptability score range, and the right y-axis shows the percentage of samples in that range.

ure 4. Among the three tasks, most HDAS samples have Acceptability scores below 0.3, while FC and NLI samples are distributed more evenly, indicating lower explanation quality in HDAS. When comparing source datasets, the e-SNLI model shows a steeper accuracy curve, suggesting that lower Acceptability scores often correspond to incorrect predictions of the model. In both models, the Acceptability score is positively linked to label prediction performance, especially in the lower score ranges (below 0.6).

7 Conclusion

This work investigated self-rationalization models' ability to generalize to NLI-related OOD tasks through the evaluation on 19 diverse datasets. We achieve this by fine-tuning T5-large and OLMo-7B under different configurations (varying fine-tuning dataset source, size, and instance selection strategies) to study the impact of data size and quality on OOD task performance and explanation quality. We also examined the Acceptability score as a reference-free metric for the generated explanation evaluation through a human evaluation. Through the study, we gained some important insights: i) fine-tuning a model on few-shot examples can perform surprisingly well in OOD datasets compared to fine-tuning on a large full-size dataset; ii) fine-tuning data source, compared to sample selection, has a larger impact on OOD performance; iii) Acceptability score is positively related to models label prediction performance.

Future work could explore ensemble learning with multiple few-shot models, as our findings suggest that few-shot models are comparable to full-shot ones. Additionally, e-FEVER appeared to be a more challenging dataset than e-SNLI, as its model demonstrated worse ID but better OOD performance, thus future work may explore fine-tuning harder tasks for better OOD generalization.

Limitations

We did not compare with other LLMs, as the opacity of the training data for LLMs means we cannot confirm whether our OOD datasets are genuinely OOD for them. Our fine-tuned models were selected based on in-distribution (ID) validation sets (for T5-Large), which may limit their OOD performance, as ID and OOD performance are not always correlated. Since our OOD datasets are sourced from English-only data, this study is limited to English. We found that different sample selection methods had a minor impact on OOD label prediction performance, though this conclusion may not generalize to other selection methods. With up to 128 shots, we observed performance similar to or better than full-shot models, though increasing the number of shots could yield further improvements, which we leave for future exploration.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni

Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. GPT-4 Technical Report. *arXiv preprint arXiv:2303.08774*.

Shourya Aggarwal, Divyanshu Mandowara, Vishwajeet Agrawal, Dinesh Khandelwal, Parag Singla, and Dinesh Garg. 2021. Explanations for CommonsenseQA: New Dataset and 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 3050–3065, Online. Association for Computational Linguistics.

Payal Bajaj, Chenyan Xiong, Guolin Ke, Xiaodong Liu, Di He, Saurabh Tiwary, Tie-Yan Liu, Paul Bennett, Xia Song, and Jianfeng Gao. 2022. METRO: Efficient Denoising Pretraining of Large Scale Autoencoding Language Models with Model Generated Signals. *arXiv preprint arXiv:2204.06644*.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Nitay Calderon and Roi Reichart. 2024. On Behalf of the Stakeholders: Trends in NLP Model Interpretability in the Era of LLMs. *arXiv preprint arXiv:2407.19200*.

Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. 2018. e-SNLI: Natural Language Inference with Natural Language Explanations. *Advances in Neural Information Processing Systems*, 31.

Jiuhai Chen and Jonas Mueller. 2024. Automated Data Curation for Robust Language Model Fine-Tuning. *arXiv preprint arXiv:2403.12776*.

Wei-Lin Chen, An-Zi Yen, Hen-Hsen Huang, and Hsin-Hsi Chen. 2022. Learning to generate explanation from e-hospital services for medical suggestion. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2946–2951, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Zeming Chen, Qiyue Gao, and Lawrence S. Moss. 2021. NeuralLog: Natural language inference with joint neural and logical reasoning. In *Proceedings of *SEM 2021: The Tenth Joint Conference on Lexical and Computational Semantics*, pages 78–88, Online. Association for Computational Linguistics.

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.

Thomas Diggelmann, Jordan Boyd-Graber, Jannis Bulian, Massimiliano Ciaramita, and Markus Leippold. 2020. Climate-FEVER: A Dataset for Verification of Real-World Climate Claims. In *Tackling Climate Change with Machine Learning workshop at NeurIPS 2020*.

Julian Eisenschlos, Bhuwan Dhingra, Jannis Bulian, Benjamin Börschinger, and Jordan Boyd-Graber. 2021. Fool Me Twice: Entailment from Wikipedia Gamification. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 352–365, Online. Association for Computational Linguistics.

Dirk Groeneveld, Iz Beltagy, Evan Walsh, Akshita Bhagia, Rodney Kinney, Oyvind Tafjord, Ananya Jha, Hamish Ivison, Ian Magnusson, Yizhong Wang, Shane Arora, David Atkinson, Russell Authur, Khyathi Chandu, Arman Cohan, Jennifer Dumas, Yanai Elazar, Yuling Gu, Jack Hessel, Tushar Khot, William Merrill, Jacob Morrison, Niklas Muennighoff, Aakanksha Naik, Crystal Nam, Matthew Peters, Valentina Pyatkin, Abhilasha Ravichander, Dustin Schwenk, Saurabh Shah, William Smith, Emma Strubell, Nishant Subramani, Mitchell Wortsman, Pradeep Dasigi, Nathan Lambert, Kyle Richardson, Luke Zettlemoyer, Jesse Dodge, Kyle Lo, Luca Soldaini, Noah

Smith, and Hannaneh Hajishirzi. 2024. OLMo: Accelerating the science of language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15789–15809, Bangkok, Thailand. Association for Computational Linguistics.

Jian Guan, Jesse Dodge, David Wadden, Minlie Huang, and Hao Peng. 2024. Language models hallucinate, but may excel at fact verification. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1090–1111, Mexico City, Mexico. Association for Computational Linguistics.

Andreas Hanselowski, Christian Stab, Claudia Schulz, Zile Li, and Iryna Gurevych. 2019. A richly annotated corpus for different tasks in automated fact-checking. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 493–503, Hong Kong, China. Association for Computational Linguistics.

Or Honovich, Roee Aharoni, Jonathan Herzig, Hagai Taitelbaum, Doron Kukliansy, Vered Cohen, Thomas Scialom, Idan Szpektor, Avinatan Hassidim, and Yossi Matias. 2022. TRUE: Re-evaluating Factual Consistency Evaluation. In *Proceedings of the Second DialDoc Workshop on Document-grounded Dialogue and Conversational Question Answering*, pages 161–175, 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*.

Xinyu Hu, Li Lin, Mingqi Gao, Xunjian Yin, and Xiaojun Wan. 2024. Themis: Towards Flexible and Interpretable NLG Evaluation. *arXiv preprint arXiv:2406.18365*.

Dongfu Jiang, Yishan Li, Ge Zhang, Wenhao Huang, Bill Yuchen Lin, and Wenhu Chen. 2023. TIGERScore: Towards Building Explainable Metric for All Text Generation Tasks. *arXiv preprint arXiv:2310.00752*.

Siddharth Karamcheti, Ranjay Krishna, Li Fei-Fei, and Christopher Manning. 2021. Mind your outliers! investigating the negative impact of outliers on active learning for visual question answering. 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 7265–7281, Online. Association for Computational Linguistics.

Rabeeh Karimi Mahabadi, Yonatan Belinkov, and James Henderson. 2020a. End-to-End Bias Mitigation by Modelling Biases in Corpora. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8706–8716, Online. Association for Computational Linguistics.

Rabeeh Karimi Mahabadi, Yonatan Belinkov, and James Henderson. 2020b. End-to-end bias mitigation by modelling biases in corpora. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8706–8716, Online. Association for Computational Linguistics.

Najoung Kim, Roma Patel, Adam Poliak, Patrick Xia, Alex Wang, Tom McCoy, Ian Tenney, Alexis Ross, Tal Linzen, Benjamin Van Durme, Samuel R. Bowman, and Ellie Pavlick. 2019. Probing what different NLP tasks teach machines about function word comprehension. In *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (*SEM 2019)*, pages 235–249, Minneapolis, Minnesota. Association for Computational Linguistics.

Wojciech Kryscinski, Bryan McCann, Caiming Xiong, and Richard Socher. 2020. Evaluating the Factual Consistency of Abstractive Text Summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9332–9346, Online. Association for Computational Linguistics.

Jenny Kunz and Marco Kuhlmann. 2024. Properties and challenges of LLM-generated explanations. In *Proceedings of the Third Workshop on Bridging Human–Computer Interaction and Natural Language Processing*, pages 13–27, Mexico City, Mexico. Association for Computational Linguistics.

Alice Lai, Yonatan Bisk, and Julia Hockenmaier. 2017. Natural Language Inference from Multiple Premises. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 100–109, Taipei, Taiwan. Asian Federation of Natural Language Processing.

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

Junlong Li, Shichao Sun, Weizhe Yuan, Run-Ze Fan, hai zhao, and Pengfei Liu. 2024a. Generative judge for evaluating alignment. In *The Twelfth International Conference on Learning Representations*.

Ming Li, Yong Zhang, Shwai He, Zhitao Li, Hongyu Zhao, Jianzong Wang, Ning Cheng, and Tianyi Zhou. 2024b. Superfiltering: Weak-to-Strong Data Filtering for Fast Instruction-Tuning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14255–14273, Bangkok, Thailand. Association for Computational Linguistics.

Hanmeng Liu, Ruoxi Ning, Zhiyang Teng, Jian Liu, Qiji Zhou, and Yue Zhang. 2023. Evaluating the Logical Reasoning Ability of ChatGPT and GPT-4. *arXiv preprint arXiv:2304.03439*.

Ana Marasovic, Iz Beltagy, Doug Downey, and Matthew Peters. 2022. Few-Shot Self-Rationalization with Natural Language Prompts. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 410–424, Seattle, United States. Association for Computational Linguistics.

Marco Marelli, Stefano Menini, Marco Baroni, Luisa Bentivogli, Raffaella Bernardi, and Roberto Zamparelli. 2014. A SICK cure for the evaluation of compositional distributional semantic models. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 216–223, Reykjavik, Iceland. European Language Resources Association (ELRA).

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On Faithfulness and Factuality in Abstractive Summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online. Association for Computational Linguistics.

Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the Wrong Reasons: Diagnosing Syntactic Heuristics in Natural Language Inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy. Association for Computational Linguistics.

Ellie Pavlick and Chris Callison-Burch. 2016. Most "babies" are "little" and most "problems" are "huge": Compositional Entailment in Adjective-Nouns. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2164–2173, Berlin, Germany. Association for Computational Linguistics.

Jiaxin Pei, Aparna Ananthasubramaniam, Xingyao Wang, Naitian Zhou, Apostolos Dedeloudis, Jackson Sargent, and David Jurgens. 2022. POTATO: The portable text annotation tool. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 327–337, Abu Dhabi, UAE. Association for Computational Linguistics.

Adam Poliak, Aparajita Haldar, Rachel Rudinger, J. Edward Hu, Ellie Pavlick, Aaron Steven White, and Benjamin Van Durme. 2018a. Collecting Diverse Natural Language Inference Problems for Sentence Representation Evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 67–81, Brussels, Belgium. Association for Computational Linguistics.

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. 2018b. Hypothesis only baselines in natural language inference. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics, pages 180–191, New Orleans, Louisiana. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Sahana Ramnath, Brihi Joshi, Skyler Hallinan, Ximing Lu, Liunian Harold Li, Aaron Chan, Jack Hessel, Yejin Choi, and Xiang Ren. 2024. Tailoring Self-Rationalizers with Multi-Reward Distillation. In *International Conference on Learning Representations (ICLR*.

Alexis Ross, Matthew Peters, and Ana Marasovic. 2022. Does self-rationalization improve robustness to spurious correlations? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7403–7416, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Arkadiy Saakyan, Tuhin Chakrabarty, and Smaranda Muresan. 2021. COVID-Fact: Fact Extraction and Verification of Real-World Claims on COVID-19 Pandemic. 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 2116–2129, Online. Association for Computational Linguistics.

Swarnadeep Saha, Yixin Nie, and Mohit Bansal. 2020. ConjNLI: Natural Language Inference Over Conjunctive Sentences. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8240–8252, Online. Association for Computational Linguistics.

Oscar Sainz, Jon Campos, Iker García-Ferrero, Julen Etxaniz, Oier Lopez de Lacalle, and Eneko Agirre. 2023. NLP evaluation in trouble: On the need to measure LLM data contamination for each benchmark. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10776–10787, Singapore. Association for Computational Linguistics.

Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A. Smith, and Yejin Choi. 2020. Social Bias Frames: Reasoning about Social and Power Implications of Language. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5477–5490, Online. Association for Computational Linguistics.

Benjamin Schiller, Johannes Daxenberger, and Iryna Gurevych. 2022. Diversity Over Size: On the Effect of Sample and Topic Sizes for Argument Mining Datasets. *arXiv preprint arXiv:2205.11472*.

Tal Schuster, Adam Fisch, and Regina Barzilay. 2021. Get your vitamin C! robust fact verification with contrastive evidence. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 624–643, Online. Association for Computational Linguistics.

Luca Soldaini, Rodney Kinney, Akshita Bhagia, Dustin Schwenk, David Atkinson, Russell Authur, Ben Bogin, Khyathi Chandu, Jennifer Dumas, Yanai Elazar, Valentin Hofmann, Ananya Jha, Sachin Kumar, Li Lucy, Xinxi Lyu, Nathan Lambert, Ian Magnusson, Jacob Morrison, Niklas Muennighoff, Aakanksha Naik, Crystal Nam, Matthew Peters, Abhilasha Ravichander, Kyle Richardson, Zejiang Shen, Emma Strubell, Nishant Subramani, Oyvind Tafjord, Evan Walsh, Luke Zettlemoyer, Noah Smith, Hannaneh Hajishirzi, Iz Beltagy, Dirk Groeneveld, Jesse Dodge, and Kyle Lo. 2024. Dolma: an Open Corpus of Three Trillion Tokens for Language Model Pretraining Research. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15725–15788, Bangkok, Thailand. Association for Computational Linguistics.

Dominik Stammbach and Elliott Ash. 2020. e-FEVER: Explanations and summaries for automated fact checking. In *Truth and Trust Online (TTO)*.

Hongjin Su, Jungo Kasai, Chen Henry Wu, Weijia Shi, Tianlu Wang, Jiayi Xin, Rui Zhang, Mari Ostendorf, Luke Zettlemoyer, Noah A Smith, et al. 2022. Selective Annotation Makes Language Models Better Few-Shot Learners. In *The Eleventh International Conference on Learning Representations*.

Elior Sulem, Omri Abend, and Ari Rappoport. 2018. BLEU is not suitable for the evaluation of text simplification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 738–744, Brussels, Belgium. Association for Computational Linguistics.

Swabha Swayamdipta, Roy Schwartz, Nicholas Lourie, Yizhong Wang, Hannaneh Hajishirzi, Noah A. Smith, and Yejin Choi. 2020. Dataset Cartography: Mapping and Diagnosing Datasets with Training Dynamics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9275–9293, Online. Association for Computational Linguistics.

Yi Tay, Mostafa Dehghani, Vinh Q Tran, Xavier Garcia, Jason Wei, Xuezhi Wang, Hyung Won Chung, Dara Bahri, Tal Schuster, Steven Zheng, et al. 2022. UL2: Unifying Language Learning Paradigms. In *International Conference on Learning Representations*.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. 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.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. *arXiv preprint arXiv:2307.09288*.

Aditya Srikanth Veerubhotla, Lahari Poddar, Jun Yin, György Szarvas, and Sharanya Eswaran. 2023. Few shot rationale generation using self-training with dual teachers. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4825–4838, Toronto, Canada. Association for Computational Linguistics.

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. 2020. Fact or Fiction: Verifying Scientific Claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7534–7550, Online. Association for Computational Linguistics.

Alex Wang, Kyunghyun Cho, and Mike Lewis. 2020. Asking and Answering Questions to Evaluate the Factual Consistency of Summaries. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5008–5020, Online. Association for Computational Linguistics.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2019a. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. In *7th International Conference on Learning Representations, ICLR 2019*.

Cunxiang Wang, Shuailong Liang, Yue Zhang, Xiaonan Li, and Tian Gao. 2019b. Does it Make Sense? And Why? A Pilot Study for Sense Making and Explanation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4020–4026, Florence, Italy. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Sarah Wiegreffe, Jack Hessel, Swabha Swayamdipta, Mark Riedl, and Yejin Choi. 2022. Reframing Human-AI Collaboration for Generating Free-Text Explanations. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 632–658, Seattle, United States. Association for Computational Linguistics.

Sarah Wiegreffe, Ana Marasović, and Noah A. Smith. 2021. Measuring Association Between

Labels and Free-Text Rationales. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10266–10284, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Robert Wolfe, Isaac Slaughter, Bin Han, Bingbing Wen, Yiwei Yang, Lucas Rosenblatt, Bernease Herman, Eva Brown, Zening Qu, Nic Weber, et al. 2024. Laboratory-scale ai: Open-weight models are competitive with chatgpt even in low-resource settings. In *ACM Conference on Fairness, Accountability, and Transparency*, pages 1199–1210.

Yuxiang Wu, Matt Gardner, Pontus Stenetorp, and Pradeep Dasigi. 2022. Generating data to mitigate spurious correlations in natural language inference datasets. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2660–2676, Dublin, Ireland. Association for Computational Linguistics.

Mengzhou Xia, Sadhika Malladi, Suchin Gururangan, Sanjeev Arora, and Danqi Chen. 2024. LESS: Selecting influential data for targeted instruction tuning. In *International Conference on Machine Learning (ICML)*.

Jiuding Yang, Hui Liu, Weidong Guo, Zhuwei Rao, Yu Xu, and Di Niu. 2024. Reassess summary factual inconsistency detection with large language model. In *Proceedings of the 1st Workshop on Towards Knowledgeable Language Models (KnowLLM 2024)*, pages 27–31, Bangkok, Thailand. Association for Computational Linguistics.

Yordan Yordanov, Vid Kocijan, Thomas Lukasiewicz, and Oana-Maria Camburu. 2022. Few-Shot Out-of-Domain Transfer Learning of Natural Language Explanations in a Label-Abundant Setup. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3486–3501.

Majid Zarharan, Pascal Wullschleger, Babak Behkam Kia, Mohammad Taher Pilehvar, and Jennifer Foster. 2024. Tell me why: Explainable public health fact-checking with large language models. In *Proceedings of the 4th Workshop on Trustworthy Natural Language Processing (TrustNLP 2024)*, pages 252–278, Mexico City, Mexico. Association for Computational Linguistics.

Sheng Zhang, Rachel Rudinger, Kevin Duh, and Benjamin Van Durme. 2017. Ordinal Common-sense Inference. *Transactions of the Association for Computational Linguistics*, 5:379–395.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Yangqiaoyu Zhou and Chenhao Tan. 2021. Investigating the Effect of Natural Language Explanations on Out-of-Distribution Generalization in Few-shot NLI. In *Proceedings of the Second Workshop on Insights from Negative Results in NLP*, pages 117–124, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

A Category 1: Additional details

A.1 Data pre-processing

For the following datasets, we applied pre-processing as defined below:

e-FEVER We filter out incorrect explanations from e-FEVER based on the following rules (around 14% of samples are removed from the training set):

- The explanation is: “The relevant information about the claim is lacking in the context.” but the label is not NEI (NOT ENOUGH INFO).
- The explanation repeats the claim, and the label is not SUPPORTS.

AddOneRTE (Pavlick and Callison-Burch, 2016) We convert the mean human scores into two classes *entailed* (when the score is no less than 4) and *not_entailment* (when the score is no greater than 3, anything between 3 and 4 are removed), following the literature convention (Karimi Mahabadi et al., 2020a).

Ordinal Common-sense Inference (JOCI) (Zhang et al., 2017) We follow Karimi Mahabadi et al. (2020a) by mapping the labels *very likely* to *entailment*; *likely*, *plausible* and *technically possible* to *neutral*; and *impossible* to *contradiction*.

Multiple Premise Entailment (MPE) (Lai et al., 2017) We concatenate the premise sentences together to form one premise paragraph.

SciFact (Wadden et al., 2020) The dataset does not have public available labels for test set, thus we use the dev set. We do not perform evidence retrieval and use the cited document abstracts as evidence.

Climate FEVER (Diggelmann et al., 2020) We use the paragraph-level evidence labels.

FactCC (Kryscinski et al., 2020) We map label *factual* as *entailment* and *non-factual* to *not_entailment*.

QAGS CNN (Wang et al., 2020) We aggregate with majority voting from the provided human annotations.

QAGS XSUM (Wang et al., 2020) We aggregate with majority voting from the provided human annotations.

XSUM Hallucination (Maynez et al., 2020) We aggregate with majority voting from the provided human annotations.

A.2 Ambiguous sample selection method

We input the (h_i, p_i) to the T5-large model, and take the probability of the first most likely output token, since the first token represent the classification label. We denote the probability as p_i. To select ambiguous samples, we calculate a mean probability score p_{mean} as follows:

$$p_{mean} = (p_{max} + p_{min})/2 \quad (1)$$

where p_{max} and p_{min} represents the highest and lowest probability score among all sample scores respectively. Then we re-calculate the score based on its absolute distance with p_{mean}:

$$p'_i = |(p_i - p_{mean})| \quad (2)$$

with the absolute distance, we re-rank the samples from low to high to select the most ambiguous ones. The lowest value represents the most ambiguous sample and the highest the least ambiguous.

A.3 Additional implementation details

For T5-Large model fine-tuning, we perform a hyper-parameter search over the learning rate for each number of shots for each source dataset separately, with random sample selection from the first subset. We select the learning rate based on the highest performance on the in-distribution validation set within 50 epochs. The performance is based on the summation of label accuracy and explanation BERTscore (Zhang et al., 2020). The same hyper-parameters are used for all sample selection methods, which share the same m and source dataset for fine-tuning. To calculate the labels’ accuracy and explanations’ BERTscore, we divide the output sequence into the label and explanation. With the template format, T5 learns to generate a text label, followed by a separation pattern, “explanation:”, and then the explanation tokens. Thus, we take the token before the separation pattern as the text label and after as the explanation. During hyper-parameter search, we test these learning rates: 3e-7, 3e-6, 3e-5, and 3e-4. For the validation set in fine-tuning, we randomly select 300 samples in the original validation set as the in-distribution set, as the original one is too large; thus, validation takes much longer. We follow the same settings as FEB (Marasovic et al., 2022) for the validation instances; for the ones with more than one explanation annotated, we merge them into one sequence separated by [SEP] token.

For OLMo-7B fine-tuning with LoRA, we follow recommended hyperparameters studied in Zarharan et al. (2024): LoRA r and alpha values are both 16, the learning rate is 2e-4, and the optimizer is "paged_adamw_32bit". We fine-tune all few-shot models with 50 epochs and use the models from the last epoch. For full-shot fine-tuning, the number of epochs is ten instead of 50.

The sentence-transformer model used in embedding the input for the Fast-Vote-k method is *paraphrase-mpnet-base-v2*.

In inference, for label mapping of T5 models, we focus on probabilities of tokens corresponding to our target labels: "entailment", "contradiction", "neutral", disregarding others (except for "entailment", as this word contains three-word tokens: "en", "tail" and "ment", we take the token number of "en"). The label is then determined based on the highest probability among these three tokens.

A.4 Human evaluation interface

The evaluation interface is shown in Figure 5, including the task instruction, some examples, and the evaluation page. To select eligible participants, our screening requires participants to have at least an undergraduate degree, and primary language as English, with an approval rate above 99%. For high-quality evaluation, we inserted 2 attentions questions to filter out low-quality evaluations (an evaluation is rejected if the worker failed on both attention checks, or failed on one and contains invalid answers through our manual checking).

A.5 Input template for explanation evaluation with the reference-free metrics

- **Acceptability score**

> *premise:* [premise] *hypothesis:* [hypothesis] *answer:* [gold label] *explanation:* [explanation]

- **TigerScore** and **Auto-J**

> *Given a hypothesis and its premise, please explain why the hypothesis is entailment, neutral, or contradiction.*
> *Hypothesis:* [hypothesis], *Premise:* [premise].
> *Please explain why the hypothesis is* [gold label].

- **Themis (relevance aspect, input in JSON format)**

> {"task": "Controllable Generation", "aspect": "Coherence: Given the explanation for the relationship between the hypothesis and premise pair, how much does the generated explanation make sense?", "source_des": "Hypothesis and Premise Pair", "source": "Hypothesis: [hypothesis], Premise: [premise], please explain why the Hypothesis is [gold label].", "target_des'': "Explanation", "target": [explanation]}

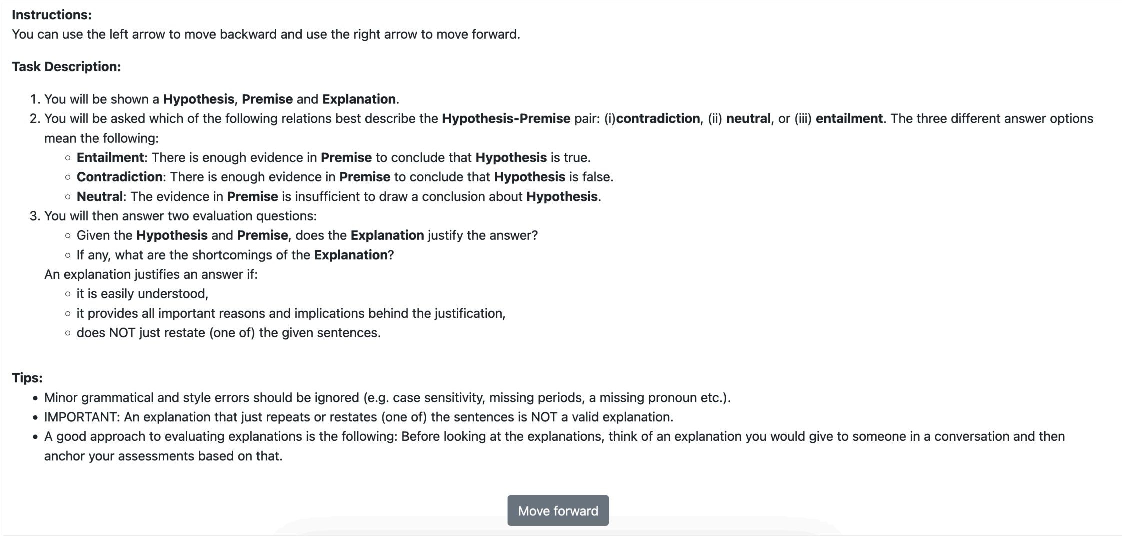
A.6 Generated explanations by different models and their evaluation scores

Hypothesis (contradiction)	A person in a blue jacket is jumping onto a tall cement wall
Premise	The man is performing a large jump
$\mathbf{T}^{Fev}_{64,AFk}$ Human: 0.444, Accept.: 0.027	not all men wear a blue jacket.
$\mathbf{T}^{Fev}_{128,R}$ Human: 0.556, Accept.: 0.450	A man performing a large jump is not necessarily in a blue jacket and jumping onto a tall cement wall.
$\mathbf{T}^{Fev}_{128,Fk}$ Human: 0.889, Accept.: 0.036	"the relevant information about the claim is lacking in the context."
$\mathbf{T}^{Fev}_{128,AFk}$ Human: 1.0, Accept.: 0.494	The man is not necessarily wearing a blue jacket and the jump is not necessarily onto a tall wall.
$\mathbf{T}^{Fev}_{Full}$ Human: 0.222, Accept.: 0.036	"the relevant information about the claim is lacking in the context."
$\mathbf{T}^{Sn}_{64,Fk}$ Human: 0.556, Accept.: 0.520	the man is performing a large jump onto a tall cement wall.
$\mathbf{T}^{Sn}_{64,AFk}$ Human: 1.0, Accept.: 0.394	the man is performing a large jump onto a tall cement wall does not indicate that he is in a blue jacket.
$\mathbf{T}^{Sn}_{Full}$ Human: 1.0, Accept.: 0.842	a large jump does not imply that the man is jumping onto a tall cement wall.
$\mathbf{O}^{Fev}_{16,AFk}$ Human: 0.778, Accept.: 0.036	"the relevant information about the claim is lacking in the context."
$\mathbf{O}^{Fev}_{128,AFk}$ Human: 0.889, Accept.: 0.473	A person in a blue jacket is performing a large jump onto a tall cement wall
$\mathbf{O}^{Fev}_{Full}$ Human: 0.667, Accept.: 0.285	The man is performing a large jump, however it does not say what kind of jacket he is wearing.
$\mathbf{O}^{Sn}_{128,AFk}$ Human: 0.778, Accept.: 0.536	a man in a blue jacket is performing a large jump onto a tall cement wall.
$\mathbf{O}^{Sn}_{Full}$ Human: 0.667, Accept.: 0.805	The man is performing a large jump, but it does not say whether he is jumping onto a tall cement wall.

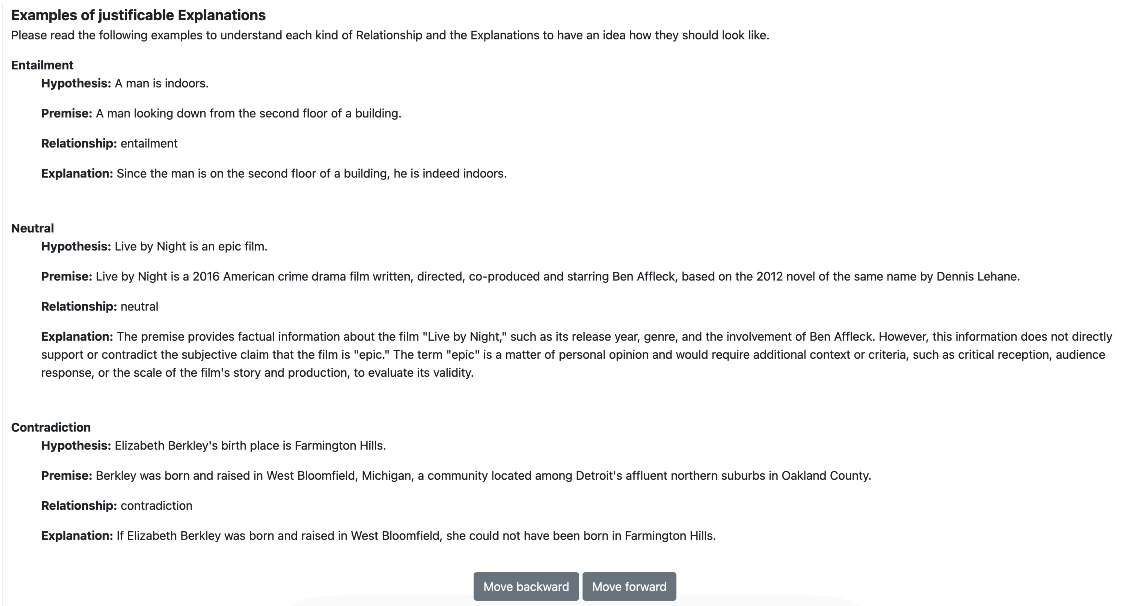
Table 7: An example of generated explanations by the 13 selected models for human evaluation, with the instance randomly selected from the SICK dataset (all models have the correct prediction).

Hypothesis (entailment)	Guns N ' Roses is an American hard rock band.
Premise	Guns N ' Roses, often abbreviated as GnR, is an American hard rock band from Los Angeles formed in 1985.
$\mathbf{T}^{Fev}_{64,AFk}$ Human: 0.889, Accept.: 0.588	guns n roses is an american hard rock band.
$\mathbf{T}^{Fev}_{128,R}$ Human: 1.0, Accept.: 0.510	guns n'roses is an american hard rock band from los angeles formed in 1985.
$\mathbf{T}^{Fev}_{128,Fk}$ Human: 0.333, Accept.: 0.174	guns n'roses is an american hard rock band, however it does not say whether it is a rock band.
$\mathbf{T}^{Fev}_{128,AFk}$ Human: 0.778, Accept.: 0.527	guns n roses is an american hard rock band from los angeles formed in 1985.
$\mathbf{T}^{Fev}_{Full}$ Human: 1.0, Accept.: 0.556	guns n'roses is an american hard rock band from los angeles formed in 1985.
$\mathbf{T}^{Sn}_{64,Fk}$ Human: 0.778, Accept.: 0.498	the band was formed in 1985.
$\mathbf{T}^{Sn}_{64,AFk}$ Human: 0.889, Accept.: 0.573	the band is from Los Angeles and formed in 1985.
$\mathbf{T}^{Sn}_{Full}$ Human: 1.0, Accept.: 0.302	guns n'roses is an american hard rock band is a rephrasing of guns n'roses, often abbreviated as gnr, is an american hard rock band
$\mathbf{O}^{Fev}_{16,AFk}$ Human: 1.0, Accept.: 0.603	Guns N'Roses is an American hard rock band, formed in Los Angeles in 1985.
$\mathbf{O}^{Fev}_{128,AFk}$ Human: 0.889, Accept.: 0.556	Guns N'Roses is an American hard rock band from Los Angeles formed in 1985.
$\mathbf{O}^{Fev}_{Full}$ Human: 1.0, Accept.: 0.556	Guns N'Roses is an American hard rock band from Los Angeles formed in 1985.
$\mathbf{O}^{Sn}_{128,AFk}$ Human: 0.889, Accept.: 0.618	Guns N'Roses is a hard rock band.
$\mathbf{O}^{Sn}_{Full}$ Human: 0.111, Accept.: 0.088	Guns is hard to form a hard hard hard hard.

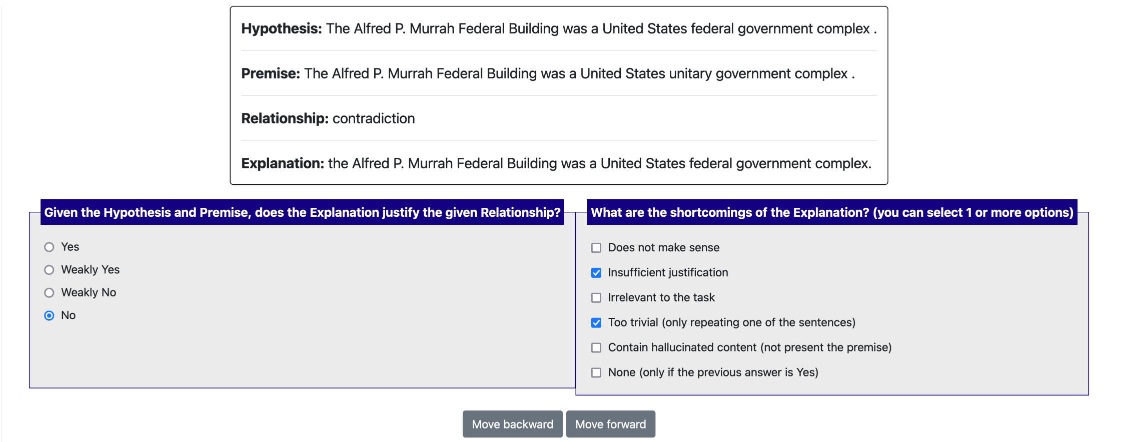
Table 8: An example of generated explanations by the 13 selected models for human evaluation, with the instance randomly selected from the VitaminC dataset (all models have the correct prediction).



(a) Task instructions



(b) Examples



(c) The evaluation page

Figure 5: Screenshots of human evaluation interface

Hypothesis (entailment)
a hospital trust is being investigated by the health watchdog over its finances.

Premise
Monitor is looking into the financial sustainability of Southend University Hospital Foundation Trust “on behalf of patients”, the NHS regulator said. Finances became a concern when a planned £7.8m deficit grew, for which Monitor could see no recovery plan. The hospital blamed the larger-than-anticipated deficit on growth in demand and increased staff recruitment. Hospital chairman Alan Tobias OBE, said: “The overspend is a result of vital investment in services and more staff to ensure high standards of patient care as well as the rise in attendances. "We welcome this review to clearly demonstrate to Monitor - as well as patients and stakeholders - our future financial plans are both robust and sustainable.” The hospital has pledged to cut the deficit while maintaining “high quality care for local people” and said it was committed to returning to a surplus within three years. Monitor said its investigation will look into the state of the hospital trustś finances, assess the strength of its financial management and explore ways of improving its sustainability, on behalf of patients across the region. The investigation will also examine how the trust works with other local health and care organisations to respond to the financial challenges it faces. Katherine Cawley, regional director at Monitor, said: “People are relying on Southend to provide them with quality healthcare now and in the future, and we are concerned that this may be at risk because of the trustś poor financial performance. "We have launched this investigation to find out more about the financial situation at the trust and to establish what can be done to improve things.”

$\mathbf{T}^{Fev}_{64,AFk}$ Human: 0.667, Accept.: 0.267
monitor is looking into the financial sustainability of southend university hospital foundation trust “on behalf of patients”.

$\mathbf{T}^{Fev}_{128,R}$ Human: 0.889, Accept.: 0.276
Monitor is looking into the financial sustainability of Southend University Hospital Foundation Trust ‘on behalf of patients’.

$\mathbf{T}^{Fev}_{128,Fk}$ Human: 1.0, Accept.: 0.310
southend university hospital foundation trust is being investigated by monitor over its finances.

$\mathbf{T}^{Fev}_{128,AFk}$ Human: 0.889, Accept.: 0.319
a hospital trust is being investigated by monitor over its finances.

$\mathbf{T}^{Fev}_{Full}$ Human: 0.889, Accept.: 0.266
monitor is investigating the financial sustainability of southend university hospital foundation trust “on behalf of patients”.

$\mathbf{T}^{Sn}_{64,Fk}$ Human: 0.778, Accept.: 0.240
the health watchdog is looking into the financial sustainability of Southend University Hospital Foundation Trust “on behalf of patients”

$\mathbf{T}^{Sn}_{64,AFk}$ Human: 1.0, Accept.: 0.263
monitor is looking into the financial sustainability of southend university hospital foundation trust “on behalf of patients”

$\mathbf{T}^{Sn}_{Full}$ Human: 0.778, Accept.: 0.219
monitor is the health watchdog.

$\mathbf{O}^{Fev}_{16,AFk}$ Human: 1.0, Accept.: 0.262
a hospital trust is being investigated by the health watchdog over its finances.

$\mathbf{O}^{Fev}_{128,AFk}$ Human: 1.0, Accept.: 0.302
The hospital trust is being investigated by the health watchdog over its finances.

$\mathbf{O}^{Fev}_{Full}$ Human: 0.889, Accept.: 0.310
southend university hospital foundation trust is being investigated by monitor over its finances.

$\mathbf{O}^{Sn}_{128,AFk}$ Human: 1.0, Accept.: 0.358
Monitor is looking into the financial sustainability of Southend University Hospital Foundation Trust “on behalf of patients”,“explanation”:“The hospital trust’s poor financial performance is being investigated by the health watchdog over its finances.

$\mathbf{O}^{Sn}_{Full}$ Human: 0.444, Accept.: 0.151
The financial services watch the financial policy of the financial and financial management to the financial services to the financial services.

Table 9: An example of generated explanations by the 13 selected models for human evaluation, with the instance randomly selected from the XSUM Hallucination dataset (all models have the correct prediction).

B Category 2: Complementary results

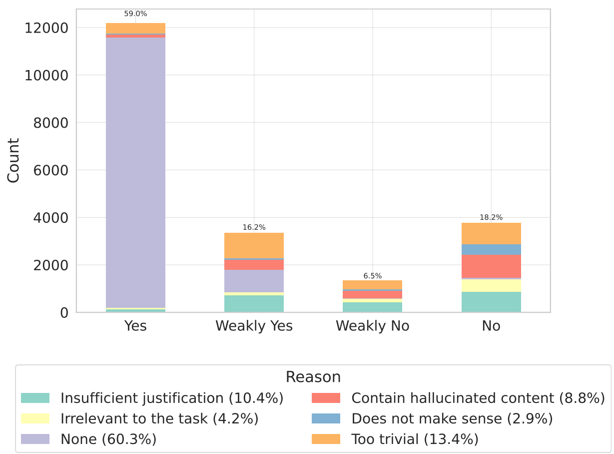


Figure 7: Distribution of reasons of shortcomings from by four answers for the question "Does the explanation justify the answer?". The overall explanation quality is high according to the crowd workers, around 59% instances have "Yes" for the question "Does the explanation justify the answer?". The most common shortcoming across all answers is "Too trivial", followed by "Insufficient justification" and "Contain hallucinated content".

Dataset	Human	Themis	Accept.
SICK	**0.655**	**2.185**	**0.437**
VitaminC	0.621	2.183	0.363
XSUM H.	0.567	1.633	0.202
All	0.620	2.046	0.350

Table 10: Human scores and automatic scores in different OOD datasets.

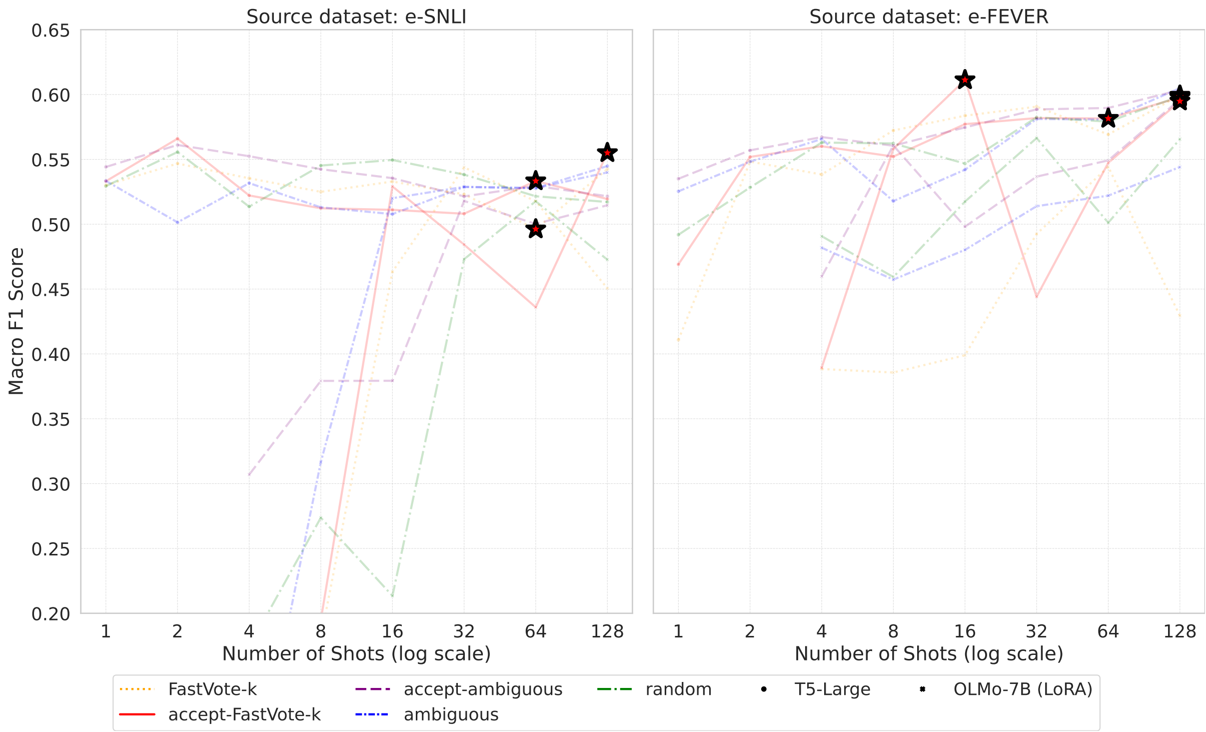


Figure 6: F1 scores of the 3 selected OOD datasets (SICK, VitaminC, XSUM Hallucination) on models fine-tuned with data from the first subset. Models marked with the asterisks are the selected ones for human evaluation (besides the full-shot models which we all include). We did not consider 1- and 2-shots fine-tuned T5 models on e-SNLI, as we observed very low quality explanations in those models.

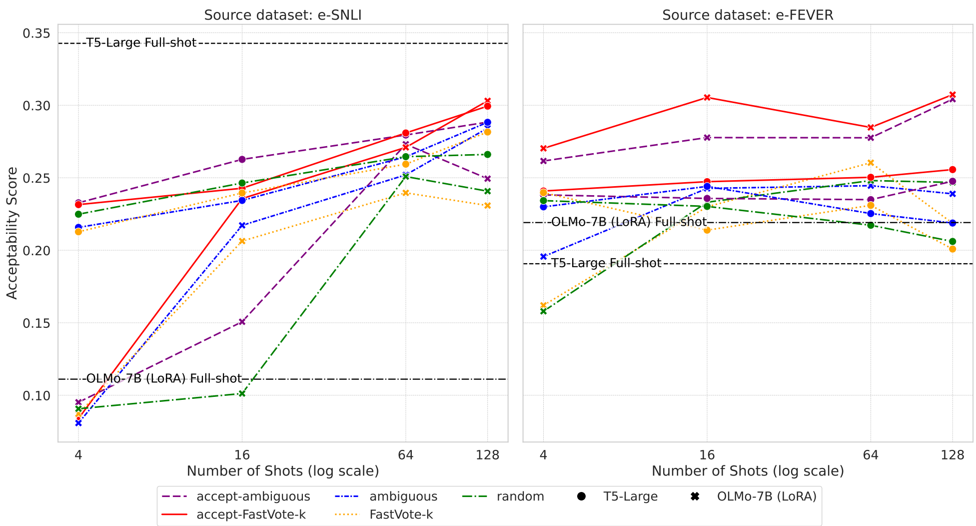


Figure 8: Acceptability score across different number of shots and sample selection methods. Selection methods with "accept-" has highest Acceptability scores for all models on both source datasets.

Dataset	T^{Sn}_{Full}	T^{Fev}_{Full}	$O^{Sn}_{128,AFk}$	$O^{Fev}_{128,AFk}$	MAJ	SOTA
SICK	57.1	82.4	53.7	64.2	56.9	90.3 (Chen et al., 2021)
AddOneRTE	88.6	88.4	81.9	85.5	85.3	92.2 (Pavlick and Callison-Burch, 2016)
JOCI	53.6	61.5	47.1	57.9	57.9	62.6 (Poliak et al., 2018b)
MPE	71.0	41.6	65.6	60.2	42.4	70.2 (Karimi Mahabadi et al., 2020b)
DNC	60.8	68.3	55.2	62.1	50.3	69.0 (Kim et al., 2019)
HANS	63.7	54.9	59.3	68.6	50.0	79.1 (Wu et al., 2022)
WNLI	45.1	43.7	49.3	56.3	56.3	85.6 (Raffel et al., 2020)
Glue Diagnostics	60.1	61.9	58.2	62.7	41.7	57.0^{M} (Bajaj et al., 2022)
Conj	62.6	66.9	58.3	57.3	45.1	72.7 (Liu et al., 2023)
Snopes Stance	36.6	60.3	45.4	61.1	45.9	59.6^{F1} (Hanselowski et al., 2019)
SciFACT	65.3	67.7	54.3	70.0	41.3	91.4^{F1} (Wadden et al., 2020)
Climate FEVER	47.9	49.5	43.5	51.3	47.4	75.0 (Wolfe et al., 2024)
VitaminC	59.8	63.0	58.4	61.0	50.1	91.1 (Tay et al., 2022)
COVID-Fact	66.5	74.3	65.1	76.3	68.3	83.5 (Saakyan et al., 2021)
FM2	71.7	73.2	76.6	79.7	50.7	88.5 (Guan et al., 2024)
FactCC	88.3	89.3	68.6	79.1	87.7	91.3^{BA} (Yang et al., 2024)
QAGS CNN	75.6	78.2	62.9	76.8	74.4	81.3 (Honovich et al., 2022)
QAGS XSUM	60.3	62.8	61.5	72.8	51.5	77.4 (Honovich et al., 2022)
XSUM H.	58.9	62.4	82.9	80.0	90.1	66.4^{BA} (Yang et al., 2024)

Table 11: Comparison of accuracy on the 19 OOD datasets with different models. MAJ: majority voting baseline, SOTA: state-of-the-art, M: Matthews coefficient, F1: F1 score, BA: balanced accuracy.

Source	Test Set	E.	N.	C.	A.
e-SNLI	ID (Sn)	**86.56**	**79.62**	**91.76**	**85.98**
	OOD (Fev)	78.17	38.65	68.82	61.88
	OOD (9)	59.26	49.56	51.97	53.60
e-FEVER	ID (Fev)	83.22	48.07	76.39	69.23
	OOD (Sn)	**89.04**	**78.18**	**86.63**	**84.61**
	OOD (9)	69.17	56.64	52.12	59.31

Table 12: F1 score performance on different test sets, contrasting the two source datasets. E.: entailment, N.: neutral, C.: contradiction, A.: average F1 score. Fev: e-FEVER, Sn: e-SNLI.

Selection	Accept.	Themis	F1
Themis-FastVote-*k*	0.303	3.027	58.24
accept-FastVote-*k*	0.307	2.774	63.24

Table 13: Evaluation results using Themis as a filter and as Acceptability a metric (T5-11B), compared to using acceptability as a filter (T5-Large) and Themis as a metric.