Cross-Refine: Improving Natural Language Explanation Generation by Learning in Tandem

Qianli Wang[1,2]   **Tatiana Anikina**[1,3]   **Nils Feldhus**[1]
Simon Ostermann[1,3,4]   **Sebastian Möller**[1,2]   **Vera Schmitt**[1,2]

[1]German Research Center for Artificial Intelligence (DFKI)
[2]Technische Universität Berlin   [3]Saarland Informatics Campus
[4] Centre for European Research in Trusted AI (CERTAIN)
{firstname.lastname}@dfki.de

Abstract

Natural language explanations (NLEs) are vital for elucidating the reasoning behind large language model (LLM) decisions. Many techniques have been developed to generate NLEs using LLMs. However, like humans, LLMs might not always produce optimal NLEs on first attempt. Inspired by human learning processes, we introduce CROSS-REFINE[1], which employs role modeling by deploying two LLMs as *generator* and *critic*, respectively. The generator outputs a first NLE and then refines this initial explanation using feedback and suggestions provided by the critic. CROSS-REFINE does not require any supervised training data or additional training. We validate CROSS-REFINE across three NLP tasks using three state-of-the-art open-source LLMs through automatic and human evaluation. We select SELF-REFINE ([Madaan et al., 2023](#)) as the baseline, which only utilizes self-feedback to refine the explanations. Our findings from automatic evaluation and a user study indicate that CROSS-REFINE outperforms SELF-REFINE. Meanwhile, CROSS-REFINE can perform effectively with less powerful LLMs, whereas SELF-REFINE only yields strong results with ChatGPT. Additionally, we conduct an ablation study to assess the importance of feedback and suggestions. Both of them play an important role in refining explanations. We further evaluate CROSS-REFINE on a bilingual dataset in English and German.

1 Introduction

As the complexity of LLMs continues to increase, NLEs are pivotal in explainable artificial intelligence (XAI) ([Madsen et al., 2022](#); [Lyu et al., 2024](#); [Zhao et al., 2024](#)). NLEs can serve as a bridge between XAI and humans, providing justifications in a format that humans can easily understand ([Camburu et al., 2018](#); [Wiegreffe et al., 2021](#)). LLMs are widely employed to generate NLEs across diverse
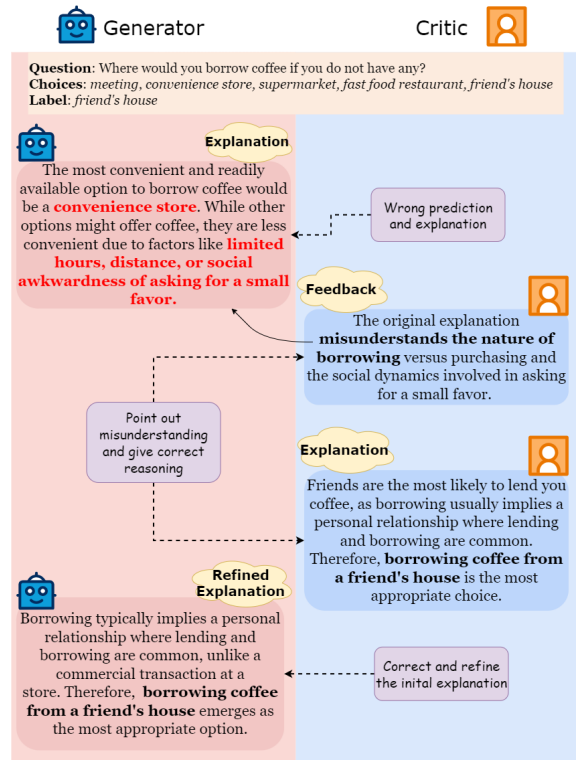

Figure 1: CROSS-REFINE example of the question *"Where would you borrow coffee if you do not have any?"* from ECQA. The initial explanation by the **generator** has been accurately corrected and refined based on the feedback and explanations provided by the **critic**.

domains ([Singh et al., 2023](#); [Wang et al., 2024b](#); [Kwon et al., 2024](#); [Stern et al., 2024](#); [Wang et al., 2024c](#)). However, similar to humans, LLMs may not consistently generate optimal explanations in their initial attempt ([Madaan et al., 2023](#)), e.g., due to lack of faithfulness ([Chuang et al., 2024](#)). LLMs have the potential to enhance their reasoning abilities through self-improvement without relying on external inputs ([Huang et al., 2023](#)). Based on this observation, [Madaan et al. (2023)](#) proposed SELF-REFINE, where LLMs use their own feedback to refine and improve their performance iteratively. This is shown to work only with large and powerful

---

[1]https://github.com/qiaw99/Cross-Refine

models; smaller models tend to hallucinate or generate repeated outputs. Moreover, Tyen et al. (2024) highlighted that LLMs generally struggle to identify reasoning errors and, therefore, cannot always self-correct their reasoning (Huang et al., 2024b).

In this paper, we first propose CROSS-REFINE, which draws inspiration from how humans benefit from learning from others (Foster and Rosenzweig, 1995; De Felice et al., 2023) and additional feedback or suggestions. CROSS-REFINE involves deploying a base LLM as the *generator* to generate an NLE and a second LLM as the *critic* (Figure 1). While the generator outputs initial explanations, the critic provides the generator with feedback and suggestions based on initial explanations. Feedback and suggestions are then cross-referenced by the generator to refine the initial explanations. The cross-referencing process involves the refinement by the *critic*, helping to mitigate the limitation of not being able to self-correct to some extent compared to SELF-REFINE (Madaan et al., 2023).

Secondly, we validate CROSS-REFINE on three NLP tasks - commonsense question answering, natural language inference, and fact-checking. We perform an automatic evaluation using three model-based metrics, as well as a user study to assess explanations based on perceived faithfulness, insightfulness, and coherence. Both results suggest that CROSS-REFINE can outperform SELF-REFINE when LLMs have substantial knowledge relevant to the given task. However, when LLMs are required to reason about topics beyond their domain of expertise, e.g., in the medical domain, CROSS-REFINE and SELF-REFINE both perform poorly. We find that CROSS-REFINE works effectively with less powerful LLMs, while SELF-REFINE delivers strong results only with ChatGPT (OpenAI, 2023).

Thirdly, compared to SELF-REFINE, we incorporate the critic's feedback and suggestions instead of self-feedback. We conduct an ablation study to assess the importance of each deployed component. The ablation study reveals that both components contribute significantly and equally to the refinement of the explanations.

Lastly, we evaluate CROSS-REFINE on a bilingual dataset HealthFC (Vladika et al., 2024) in English and German. The evaluation shows that CROSS-REFINE can outperform SELF-REFINE and consistently performs better in generating NLEs in German compared to SELF-REFINE.

2 Background

2.1 In-Context Learning for NLE

Several Chain-of-Thought (CoT) (Wei et al., 2022) prompting techniques have been introduced that yield remarkable performance improvements in NLE generation, e.g., Zero-Shot CoT (Kojima et al., 2022), Plan-and-Solve (Wang et al., 2023a), and optimization by prompting (Yang et al., 2024b). Self-consistency further demonstrates that self-evaluation can help LLMs improve reasoning (Wang et al., 2023b). CROSS-REFINE also considers in-context learning to generate NLEs, and the critic deployed in CROSS-REFINE plays a similar role to that of Wang et al.'s (2023b) self-evaluation.

2.2 NLE Evaluation

Regarding automated metrics for evaluating NLEs, BLEURT (Sellam et al., 2020) calculates the semantic similarity between human annotated explanations and generated explanations. BARTScore (Yuan et al., 2021) treats the evaluation process as a text generation task and measures the likelihood of generating the reference text given the generated text. RORA (Jiang et al., 2024d) measures the new information provided by a NLE to justify a label by evaluating the conditional ν-information (Hewitt et al., 2021). Huang et al. (2024a) asked ChatGPT to evaluate the output of the generation on multiple scales. TIGERScore (Jiang et al., 2024b) uses natural language instructions to provide error analysis, pinpointing errors in the outputs.

For human evaluation of NLEs, prevalent metrics such as plausibility, faithfulness, simulatability, and insightfulness are used to evaluate factual correctness and logical coherence (Chan et al., 2022; Atanasova et al., 2023); consistency with the model's decision process (Lakkaraju et al., 2019; Jacovi and Goldberg, 2020; Agarwal et al., 2024); how well a human can imitate model's behaviour based on explanations (Doshi-Velez and Kim, 2017; Arora et al., 2022); and how relevant is the information of an explanation (Clinciu et al., 2021), respectively. To validate CROSS-REFINE, BLEURT, BARTScore, and TIGERScore are included for automatic evaluation (§5.1), while perceived faithfulness, coherence, and insightfulness are included for the human evaluation (§5.2).

3 Methodology

CROSS-REFINE is inspired by how humans learn from others and employs two LLMs separately for
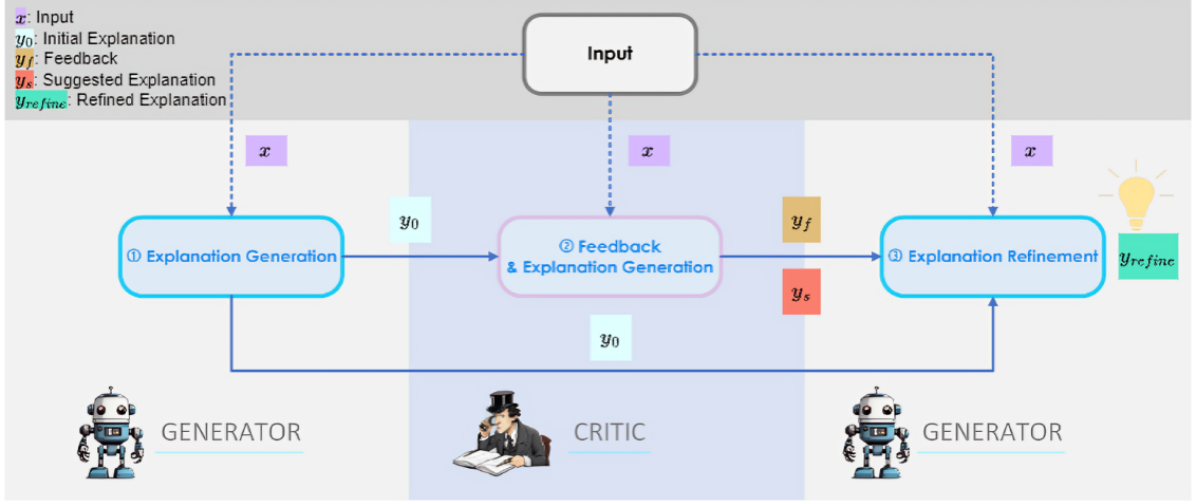
Figure 2: Pipeline of CROSS-REFINE. (1) Generator: produces an initial explanation. (2): Critic: provides feedback and an suggested explanation based on the generator's initial output. (3) Generator: utilizes the feedback and suggested explanation from the critic to refine and improve the initial explanation.

role modeling: one as the critic and the other as the generator (Figure 2). The generator outputs the *initial explanation*, while the critic offers *feedback* and *suggestions* on it, which can be used by the generator to refine the initial explanation.

3.1 CROSS-REFINE Example

Figure 1 provides an example for how generator and critic collaboratively improve NLEs. In the example, the generator initially chooses an incorrect choice (*"convenience store"*), resulting in the explanation that is untruthful for the given question. In the feedback and suggested explanation provided by the critic, the errors made by the generator are identified (*"misunderstanding the nature of borrowing"*), with the help of which the generator can recognize its mistakes and subsequently refine and correct both the prediction and the explanation[2].

3.2 Pipeline

We describe the pipeline of CROSS-REFINE as shown in Figure 2 and denote the generator by $\mathcal{G}$ and the critic by $\mathcal{C}$.

Initial Generation   The generator outputs the initial explanation independently using CoT prompting (Wei et al., 2022) with 3 to 20 shots depending on the input length (i.d. fewer shots with longer inputs) following the FEB template (Marasovic et al., 2022). Given an input x and prompt p_{gen}, CROSS-

REFINE generates the initial explanation y_0 :

$$y_0 = \mathcal{G}(x | p_{gen}) \qquad (1)$$

Quality Assessment   Given an input x, the initial explanation y_0 and a prompt p_{imp}, the critic determines whether the initial explanation needs improvement y_{imp}:

$$y_{imp} = \mathcal{C}(y_0, x | p_{imp}) \qquad (2)$$

Feedback and Suggestion   Afterwards, the critic offers feedback y_f on the initial explanation y_0 from the generator based on the provided input x with the prompt p_f:

$$y_f = \mathcal{C}(y_0, x | p_f) \qquad (3)$$

Meanwhile, the critic generates a suggested explanation y_s by considering the input x, the initial explanation y_0, and feedback y_f generated by the critic with the prompt p_s:

$$y_s = \mathcal{C}(y_f, y_0, x | p_s) \qquad (4)$$

Refinement   Lastly, the feedback y_f and the suggested explanation y_s generated by the critic are forwarded to the generator, which the generator uses to obtain the refined explanation y_{refine} with the prompt p_{refine}:

$$y_{refine} = \mathcal{G}(y_s, y_f, y_0, x | p_{refine}) \qquad (5)$$

In such a way, the generator can take into account the critic's *feedback* and *suggested explanation*. The feedback and suggested explanation are

---

[2]More CROSS-REFINE examples are in Appendix A.

cross-referenced by the generator, which serves as a guide, ultimately enhancing the quality of the generator's initial explanations.

4 Experimental Setup

4.1 Baseline

We employ SELF-REFINE (Madaan et al., 2023) as the baseline, which can enhance the initial outputs of the LLM only through iterative self-feedback. Unlike CROSS-REFINE, it does not involve multiple reasoning steps and the model does not distinguish between the roles of critic and generator.

4.2 Datasets

Following Atanasova et al. (2023), we demonstrate the validity of our approach, CROSS-REFINE, by applying it to three typical NLP tasks: natural language inference, commonsense question answering, and fact-checking. We select the subsequent three datasets[3] because of their sufficient size and the high quality of human-annotated NLEs.

e-SNLI Natural Language Inference (Dagan et al., 2006) involves determining whether a given relationship between a *premise* and a *hypothesis* can be classified as *entailment*, *contradiction*, or *neutrality*. The e-SNLI dataset (Camburu et al., 2018) is an extension of the Stanford Natural Language Inference (SNLI) corpus (Bowman et al., 2015), enriched with human-authored NLEs.

ECQA Compared to question answering, commonsense question answering requires the application of implicit background knowledge that extends beyond the information explicitly presented in the given context (Talmor et al., 2019). Each instance in the ECQA dataset comprises a *question*, several answer *options*, and human annotated explanations (Aggarwal et al., 2021).

HealthFC The significance of fact-checking has greatly increased due to the swift spread of mis- and disinformation and accurate information (Guo et al., 2022). HealthFC (Vladika et al., 2024) is a bilingual fact-checking dataset (English and German) and consists of *questions*, *documents* as well as *veracity annotations* (whether the answer is *true*, *false* or *unknown* based on the provided document) and the corresponding *explanations*.

There are several reasons why we chose HealthFC[4] for our experiments. Firstly, this dataset is new and it is unlikely that it was seen during training by the employed LLMs. Secondly, it involves claims and documents from the medical domain and includes some specific terminology and domain knowledge that differs from more general-purpose data which LLMs are typically trained on. Thirdly, it is a bilingual dataset which means that we can check the performance of CROSS-REFINE also with German.

4.3 Models

We select three state-of-the-art open-source general-purpose LLMs with increasing sizes from different model families: Qwen2-7B (Yang et al., 2024a), Mixtral-8x7B (Jiang et al., 2024a), and Llama3-70B (AI@Meta, 2024)[5].

4.4 Demonstrations for In-Context Learning

To refine the initial explanation, we employ in-context learning to prompt the critic for feedback and suggestions and prompt the generator for refined explanations. For this purpose, we create a collection of demonstrations FiXer[6], which comprises the initial explanations of the generator, the feedback and suggested explanations of the critic, and the refined explanations of the generator.

4.5 Prompts

Conforming to the FEB template (Marasovic et al., 2022), the prompt instructions used for explanation refinement include the task description, a list of information provided, and a few demonstrations for in-context learning (§4.4), as depicted in Appendix F.

5 Evaluation

5.1 Automatic Evaluation

The refined explanations are evaluated using the following three automated reference-based or reference-free metrics[7].

---

[3]Examples from each dataset and label distributions of three employed dataset can be found in Appendix B.

[4]Note that due to the input length constraints, we extract only those sentences from the documents that were annotated as relevant in the original dataset. This results in shorter, more claim-focused documents that are then included in the prompt.

[5]More details about models and inference time can be found in Appendix C.

[6]Abbreviation of "Feedback of initial eXplanation and explanation refinement" (FiXer). More details about data collection are provided in Appendix D.

[7]The models used for automatic evaluation metrics are detailed in Appendix E.

BLEURT BLEURT (Sellam et al., 2020) utilizes BERT (Devlin et al., 2019), which is fine-tuned on a collection of human ratings, to deliver ratings of generated outputs, ranging from -1 to 1.

BARTScore BARTScore (Yuan et al., 2021) leverages BART (Lewis et al., 2020) to score the generated text based on how well the generated text matches the reference text. Additionally, BARTScore evaluates both "from generated to reference" and "from reference to generated" directions, providing a more robust assessment.

TIGERScore TIGERScore (Jiang et al., 2024b) utilizes natural language instructions to perform error analysis, identifying mistakes in the generated text using fine-tuned Llama2 (Touvron et al., 2023) and delivering corresponding explanations for each mistake. TIGERScore assigns a penalty score between $[-5, -0.5]$ for each mistake.

5.2 Human Evaluation

To further validate CROSS-REFINE, we conduct a user study in which participants subjectively evaluate the refined explanations according to three dimensions.

5.2.1 Subjective Ratings

Based on how Feldhus et al. (2023) and Chiang and Lee (2023) design Likert scales for explanation evaluation, we ask human annotators to assess reasoning outputs generated by CROSS-REFINE based on the following dimensions used in the user study conducted by Tsai et al. (2024):

- **Perceived Faithfulness (Binary)**: Investigate whether the generated reasoning exhibits hallucination and if it includes any misinformation;
- **Coherence (5-point Likert)**: Assess whether the generated reasoning is sensible, clear and coherent and reflects the reasons behind the user's preference;
- **Insightfulness (5-point Likert)**: Evaluate the extent to which the generated reasoning provides informative insights into the user's preferences.

Coherence and insightfulness are rated on a 5-point Likert scale ranging from "strongly disagree" to "strongly agree", corresponding to points from 1 to 5. Perceived faithfulness is assessed using a binary scale, with a score of 0 assigned for unfaithful explanations and 1 for faithful explanations.

5.2.2 User Study Setup

Given the large number of combinations shown in Table 1, we limit the user study to the easiest and most difficult datasets, ECQA and HealthFC, respectively. Additionally, we focus on Qwen2 and Llama3 as the generators, since Mixtral does not perform well with SELF-REFINE and CROSS-REFINE (Table 1). In this way, we maintain a feasible scope of our user study.

We sample subsets ($n = 10$) of ECQA and HealthFC randomly among the instances that fulfill the selection criteria described in Appendix H, which makes the task more manageable for the annotators, reducing the risk of performance decline over time (Mangin et al., 2022), and ensuring the quality of the annotations. Based on the inputs, explanations are generated using different combinations of three deployed LLMs and the baseline, as illustrated in Table 2. Each explanation is rated by two annotators based on three subjective evaluation dimensions (§5.2.1). The inputs and their corresponding explanations are provided to the annotators in the form of questionnaires[8]. We use the Crowdee crowdsourcing platform[9] to recruit annotators, distribute questionnaires, and store the annotators' responses. We recruit a total of 32 annotators who are all English native speakers and do not necessarily have expertise in XAI.

5.3 Ablation Study

As illustrated in Figure 2, the generator receives feedback and a suggested explanation from the critic in the final step to refine its initial explanation. To analyze the impact of individual components, namely *feedback* and *suggested explanation*, on the quality of the refined explanation, we conduct an comprehensive ablation experiment (§6.3).

Influence of Suggestions Compared to SELF-REFINE, CROSS-REFINE additionally introduces suggestions from the critic to guide the generator, we explore the extent to which the suggestions can influence the refined explanations.

5.4 CROSS-REFINE on German Data

While the data we have described thus far is only in English, we also investigate the effectiveness of CROSS-REFINE on the German data provided in HealthFC dataset (§6.4).

---

[8]The annotation instructions can be found in Appendix G.
[9]https://www.crowdee.com/

Dataset		ECQA			eSNLI			HealthFC		
Critic	*Generator*	BLEURT ↑	BART Score ↑	TIGER Score ↑	BLEURT ↑	BART Score ↑	TIGER Score ↑	BLEURT ↑	BART Score ↑	TIGER Score ↑
Self-Refine	Qwen2	-0.68	-3.91	-4.38	-0.88	-4.19	-4.63	-0.25	-3.09	-1.09
Qwen2	Qwen2	-0.33	**-3.64**	-2.20	-0.97	-3.33	-4.33	**-0.24**	**-3.02**	**-0.79**
Qwen2	Mixtral	-0.67	-4.13	-2.88	**-0.71**	-3.44	-3.65	-0.33	-3.15	-1.11
Qwen2	Llama3	**-0.30**	-3.65	**-1.71**	-0.99	**-3.21**	**-2.55**	-0.83	-3.60	-2.87
Self-Refine	Mixtral	-0.75	-4.03	-4.72	-0.83	-3.72	-4.50	-0.60	-3.37	-2.28
Mixtral	Qwen2	-0.50	-4.08	**-1.68**	-0.71	**-3.44**	**-3.66**	-0.76	-3.60	-2.67
Mixtral	Mixtral	-0.66	-3.98	-2.25	**-0.64**	-3.49	-3.87	**-0.38**	**-3.21**	**-1.41**
Mixtral	Llama3	**-0.36**	**-3.46**	-4.48	-0.69	-3.52	-4.46	-0.81	-3.61	-2.87
Self-Refine	Llama3	-0.59	-3.79	-5.64	-0.99	-4.20	-4.19	-0.33	-3.14	-1.85
Llama3	Qwen2	**-0.37**	-3.72	-2.72	**-0.51**	**-3.25**	-3.74	-0.76	-3.55	-2.65
Llama3	Mixtral	-0.45	-3.64	-3.78	-0.70	-3.47	**-3.43**	-0.30	-3.13	**-0.63**
Llama3	Llama3	-0.68	**-3.62**	**-2.16**	-0.66	-3.26	-3.90	**-0.29**	**-3.07**	-0.97

Table 1: Automatic evaluation results of refined explanations generated by SELF-REFINE, and CROSS-REFINE with `Qwen2-7B`, `Mixtral-8*7B`, and `Llama3-70B` using BLEURT, BARTScore, and TIGERScore on the ECQA, eSNLI, and HealthFC datasets.

6 Results

6.1 Automatic Evaluation

Table 1 demonstrates that CROSS-REFINE can easily outperform SELF-REFINE on ECQA and eSNLI, although the scores for each automated metric are lower compared to the results of HealthFC. This discrepancy can be attributed to the shorter length of the gold rationales in ECQA and eSNLI relative to those in HealthFC. The longer context inherent in CROSS-REFINE, which includes feedback and suggestions from the critic, tends to generate relatively longer explanations, contributing to this variation in scores.

Interestingly, Table 1 reveals that for HealthFC, CROSS-REFINE with the same LLM as both generator and critic ("self CROSS-REFINE") outperforms SELF-REFINE[10], indicating that suggestions play a crucial role in refining explanations (a further proof is shown in §6.3). However, CROSS-REFINE underperforms compared to SELF-REFINE on HealthFC when using different combinations of LLMs instead of "self CROSS-REFINE". The poorer performance might be caused by the lack of domain-specific knowledge, particularly in the medical domain, as the three LLMs that we deploy are general purpose models (Yang et al., 2024c). Furthermore, since HealthFC was released very recently (Vladika et al., 2024), it is highly unlikely that three LLMs were trained on HealthFC, unlike the other two datasets. This result aligns with our intuition that models which lack knowledge

in a particular domain are less likely to provide constructive and helpful feedback and suggestions to others (Valero Haro et al., 2019). Moreover, it suggests that cross-referencing could potentially lead to worse performance if feedback and suggestions are incorrect or hallucinated (Tan, 2022; Augenstein et al., 2023).

6.2 User Study

Table 2 shows that, for ECQA, CROSS-REFINE overall outperforms SELF-REFINE, particularly in terms of coherence, where the margin is relatively large. Similarly, for HealthFC, the findings align with those mentioned in §6.1: "self CROSS-REFINE" can outperform SELF-REFINE, but other combinations other than "self CROSS-REFINE" perform worse than SELF-REFINE. Furthermore, we discover a correlation between TIGERScore and the results of the user study.

Since each combination from Table 2 is evaluated by two annotators, we report that our inter-annotator agreements (IAA) are at Krippendorff's α of 0.45 for ECQA and 0.39 for HealthFC. The low IAA scores can be attributed to the factor that we evaluate perceived faithfulness and insightfulness using a 5-point Likert scale, which is more fine-grained compared to a binary choice. The IAA on HealthFC is lower compared to ECQA due to its intrinsic difficulty. Additionally, we calculate the exact match between the two annotators, but in many cases, their scores are very close, such as 4 (agree) and 5 (strongly agree) or 1 (strongly disagree) and 2 (disagree).

From Table 2, we observe that the scores for per-

---

[10]"Self CROSS-REFINE" differs from SELF-REFINE in that it additionally incorporates explicit suggestions from the critic. SELF-REFINE instead just improves itself in few shots.

| Dataset | | ECQA | | | HealthFC | | |
Generator	Critic	Faith.	Coh.	Insight.	Faith.	Coh.	Insight.
Self-Refine (Qwen)		0.75	3.15	4.10	0.75	3.75	3.90
Qwen2	Qwen2	**0.75**	**4.40**	4.05	**1.00**	3.20	**4.15**
Qwen2	Mixtral	0.50	3.80	**4.15**	0.50	3.85	3.40
Qwen2	Llama3	0.50	3.65	3.20	0.25	**4.20**	3.80
Self-Refine (Llama3)		0.50	2.80	2.35	1.00	4.35	4.10
Llama3	Qwen2	1.00	4.19	4.05	1.00	3.45	3.60
Llama3	Mixtral	0.75	**4.50**	4.00	0.50	3.15	2.75
Llama3	Llama3	**1.00**	4.05	**4.15**	1.00	**4.55**	**4.35**

Table 2: The results (in average scores from two annotators for each combination) of the user study on the quality of the refined explanations generated by SELF-REFINE and CROSS-REFINE using Qwen2 and Llama3 as the generator. The refined explanations are evaluated based on *Perceived Faithfulness* (Faith.), *Coherence* (Coh.), and *Insightfulness* (Insight.), conducted on the ECQA and HealthFC datasets.

ceived faithfulness are sightly higher for HealthFC compared to ECQA. In some cases, where medical domain knowledge is required, annotators might not fully grasp the context of instances from HealthFC, especially when the explanations seem to be plausible. Meanwhile, recruiting annotators with specific expertise, especially in the medical field, is very challenging through crowdsourcing platforms. Moreover, we lack the expertise in the medical domain to filter qualified recruited annotators. These findings can partially highlight the risks of over-trusting LLM outputs when individuals are not well-versed in the given topic (Li et al., 2024).

Figure 10 presents examples of SELF-REFINE and CROSS-REFINE. Like the examples shown in Figure 10, we observe several cases where the explanations generated by SELF-REFINE are untruthful, while the critic in CROSS-REFINE can correct errors, making the explanations more trustworthy.

Therefore, based on automatic evaluation and the user study results, we can draw the conclusion that CROSS-REFINE can outperform SELF-REFINE, when LLMs possess substantial knowledge relevant to the given task. However, when LLMs are required to provide reasoning on topics outside of their domain of expertise, CROSS-REFINE outperforms SELF-REFINE only in the "self CROSS-REFINE" setting, i.e. utilizing the same model for both critic and generator.

6.3 Ablation Study

For the evaluation, we randomly select samples from eSNLI and deploy Qwen2-7B as both the generator and the critic, maintaining an analogous experimental setting to SELF-REFINE (Madaan et al., 2023), as SELF-REFINE shares the most similarity

Feedback	Suggestion	BLEURT	BARTScore	TIGERScore
✓	✓	-0.97	-3.33	-4.33
✓	✗	-0.72 (↓ 0.25)	-3.11 (↑ 0.22)	-4.86 (↓ 0.53)
✗	✓	-0.84 (↓ 0.13)	-3.35 (↓ 0.02)	-4.67 (↓ 0.34)

Table 3: Ablation Study of CROSS-REFINE: Impact of different components on the refinement of explanations.

| Model | | ECQA | | eSNLI | | HealthFC | |
Generator	Critic	Init.	Sug.	Init.	Sug.	Init.	Sug.
Qwen2	Qwen2	0.76	**0.87**	0.45	**0.85**	0.90	**0.96**
Qwen2	Mixtral	**0.54**	0.49	0.19	**0.66**	0.47	**0.50**
Qwen2	Llama3	0.72	**0.72**	0.62	**0.73**	0.49	**0.51**
Mixtral	Qwen2	0.46	**0.50**	0.37	**0.84**	0.56	**0.78**
Mixtral	Mixtral	0.51	**0.60**	0.56	**0.56**	0.53	**0.91**
Mixtral	Llama3	0.76	**0.81**	0.51	**0.73**	0.52	**0.93**
Llama3	Qwen2	0.67	**0.74**	0.31	**0.73**	0.39	**0.45**
Llama3	Mixtral	**0.69**	0.65	0.51	**0.70**	0.40	**0.50**
Llama3	Llama3	**0.63**	0.61	0.46	**0.64**	0.73	**0.92**

Table 4: The semantic similarities between the refined explanations and the initial explanations (**Init.**) and between the refined explanations and suggestions (**Sug.**).

to our approach. We then generate explanations with and without a certain component and the automated metrics (§5.1) are applied to each set of generated explanations to assess their quality.

Table 3 shows that while BARTScore slightly increases when using CROSS-REFINE without suggestions to refine the explanations, BLEURT and TIGERScore experience a sharp reduction. In contrast, when using CROSS-REFINE without feedback, all scores decline to some extent, but not as significantly as in the case of CROSS-REFINE without suggestions. Meanwhile, since we use the same LLM for both the generator and the critic in the ablation study ("self CROSS-REFINE"), and SELF-REFINE relies solely on self-feedback, making the feedback comparable between two approaches, we deduce that suggestions play an equally important role in the refinement of explanations.

Influence of Suggestions To measure the influence of suggestions, we evaluate the semantic similarity using SBERT[11] between the refined explanation and the initial explanation, as well as between the refined explanation and the suggestions individually. Table 4 indicates that, in general, the refined explanations align more closely with the suggestions than with the initial explanations, which implies that the "cross-refinement" steps effectively prompt changes to the initial explanation. This pro-

---
[11]https://huggingface.co/sentence-transformers/all-mpnet-base-v2

Dataset		HealthFC (German)		
Generator	*Critic*	BERTScore ↑	BARTScore ↑	MoverScore ↑
Self-Refine (Qwen)		0.6935	-5.6894	0.5246
Qwen2	Qwen2	**0.7068**	-4.4023	**0.5271**
Qwen2	Mixtral	0.6240	-6.5103	0.5068
Qwen2	Llama3	0.7036	**-4.3785**	0.5258
Self-Refine (Mixtral)		0.6519	-6.4200	0.5009
Mixtral	Qwen2	**0.6789**	-5.3713	0.5132
Mixtral	Mixtral	0.6776	**-5.1785**	**0.5173**
Mixtral	Llama3	0.6782	-5.2327	0.5161
Self-Refine (Llama3)		0.6626	-6.1267	0.5083
Llama3	Qwen2	0.6574	-5.6861	0.5078
Llama3	Mixtral	0.6220	-6.5031	0.5059
Llama3	Llama3	**0.6656**	**-5.4474**	**0.5088**

Table 5: Automatic evaluation results on HealthFC (German) dataset using BERTScore, BARTScore, and MoverScore.

6.4 CROSS-REFINE on German Data

For automatic evaluation, we discard BLEURT and TIGERSCORE, as they only support English. For BARTScore, we use a different model that is compatible with German. In addition, we deploy MoverScore (Zhao et al., 2019) and BERTScore (Zhang et al., 2020). MoverScore measures the semantic distance by contextualized representations and distance metrics, while BERTScore evaluates the token-level similarity between the reference texts and the LLM outputs by leveraging contextual embeddings. Table 5 demonstrates that, overall, CROSS-REFINE produces better NLEs than SELF-REFINE on HealthFC (German).

For the German portion of the HealthFC dataset, we compare different configurations based on the number of explanations generated in German, English, or another language. Language identification is performed using FASTTEXT-LANGDETECT[12]. Table 6 presents the percentage of explanations generated in German. The results, summarized in Table 6 show that Qwen2 and Llama3 consistently outperform Mixtral in generating NLEs in German. Additionally, SELF-REFINE outputs explanations in English notably more often compared to CROSS-REFINE, e.g., Mixtral-8x7B generates a higher percentage of self-refined explanations in English (57.5%) compared to German (39%), despite explicit prompts instructing that *"Your response should be in German"* and several German-language demonstrations.

Dataset		HealthFC (German)		
Generator	*Critic*	German ↑	English ↓	Other ↓
Self-Refine (Qwen)		88.00	11.00	1.00
Qwen2	Qwen2	96.86	2.29	0.86
Qwen2	Mixtral	93.43	3.71	2.86
Qwen2	Llama3	**97.43**	**2.29**	**0.29**
Self-Refine (Mixtral)		39.00	57.50	3.50
Mixtral	Qwen2	74.29	23.43	2.29
Mixtral	Mixtral	**86.00**	**13.14**	**0.86**
Mixtral	Llama3	76.00	22.29	1.71
Self-Refine (Llama3)		57.50	41.00	**1.50**
Llama3	Qwen2	80.06	18.21	1.73
Llama3	Mixtral	**92.20**	**4.62**	3.18
Llama3	Llama3	82.95	14.74	2.31

Table 6: Percentage of the generated explanations in different languages (English, German and other) for HealthFC (German).

Interestingly, in some cases, outputs are mixed, containing both English and German, e.g.: *"Refined explanation: Die Antwort ist unbekannt, weil das Dokument aufzeigt ..."*, while English is typically used at the beginning to indicate the type of output, e.g., to indicate the type of the generated output such as *"Refined explanation:"* in the example above.

Overall, CROSS-REFINE proves beneficial for generating explanations in a language different from English, even when the underlying model is predominantly trained on English data.

7 Related Work

Refined Explanations Krishna et al. (2023) proposed to take advantage of post-hoc explanations in in-context learning. Tong et al. (2024) found that LLMs can benefit from correct examples and learn from mistakes, while An et al. (2024) fine-tuned LLMs using pairs consisting of errors and their respective corrections. The mixture of agents (MoA) (Wang et al., 2024a) approach collects the strengths of multiple LLMs by constructing a layered MoA architecture and improves the reasoning by providing criticism. Moreover, LLMs can use self-generated feedback, refinement, or introspection as means to enhance reasoning abilities (Huang et al., 2023; Madaan et al., 2023; Zhang et al., 2023; Xu et al., 2024). Welleck et al. (2023) suggested to use a base generator that proposes an initial hypothesis and a trained corrector that iteratively improves its quality. Compared to Welleck et al.'s (2023) approach, CROSS-REFINE does not necessarily require the critic can completely correct the hypothesis, as

it can be very challenging ([Huang et al., 2024b](); [Tyen et al., 2024]()). Instead, CROSS-REFINE focuses on providing feedback and suggested explanations generated by the critic, which the generator can use to refine its initial explanations. Furthermore, CROSS-REFINE does not require supervised training data collection that is used for corrector fine-tuning ([Welleck et al., 2023]()). Meanwhile, SELF-REFINE leverages self generated feedback to refine the explanation iteratively ([Madaan et al., 2023]()). Madaan et al. (2023) showed that with SELF-REFINE, less powerful LLMs struggle with explanation refinement, because they have difficulties in generating suitable feedback and thus tend to repeat the same output or generate hallucinated output. In comparison, since CROSS-REFINE deploys the critic, the generator has an external source (i.e., feedback and suggestion) except for itself, which can be cross-referenced. Because of cross-reference, CROSS-REFINE can be highly effective for tasks where LLMs have substantial knowledge. Moreover, CROSS-REFINE performs well with less powerful LLMs, compared to SELF-REFINE.

8 Conclusion

We introduced CROSS-REFINE, an approach that improves NLEs through cross-refinement based on automated and human evaluation across various tasks. CROSS-REFINE uses two LLMs for role modeling: One as the generator and the other as the critic. The generator refines its initial explanations by cross-referencing feedback and suggestions provided by the critic. Overall, CROSS-REFINE can outperform similar state-of-the-art approaches such as SELF-REFINE and can refine the explanations well with less powerful LLMs compared to SELF-REFINE. For tasks that fall outside of the LLMs' domain expertise, e.g., in the medical domain, and require more structured domain knowledge, CROSS-REFINE using the same LLM both as the generator and the critic can surpass SELF-REFINE. Furthermore, since CROSS-REFINE introduces feedback along with suggestions from the critic to refine the generator's initial explanation, through the ablation study, we observe that suggestions are as crucial as feedback in refining explanations. Additionally, we find that CROSS-REFINE outperforms SELF-REFINE when data is in German (HealthFC), and with CROSS-REFINE, NLEs are more likely to be generated in German compared to SELF-REFINE.

9 Future Work

Future work includes exploring whether human-crafted feedback and suggestions can align with LLM generated ones. We plan to conduct a more fine-grained error analysis to inspect to what extent CROSS-REFINE can address the errors contained in the explanations. We will explore how the interpretation of terminology of quality metrics, e.g., faithfulness or insightfulness, can impact the quality of the user study. Furthermore, we will investigate whether CROSS-REFINE using LLMs trained on medical data can perform better on the HealthFC dataset. In addition, we plan to incorporate human interactions into the CROSS-REFINE workflow.

Limitations

CROSS-REFINE, while not inherently iterative like SELF-REFINE, already demonstrates superior performance compared to the latter. Moreover, its structure allows for straightforward adaptation into an adaptive framework, potentially enhancing its refinement capabilities further.

Despite we created a collection of demonstrations, FiXer, which includes various instances consisting of initial explanations (generator), feedback and suggested explanations (critic) and refined explanations (generator), we are limited to using a small number of demonstrations ($n \in [3, 10]$) depending on input length for few-shot prompting to refine NLEs with CROSS-REFINE. This limitation is primarily due to constraints on context length, e.g. Mixtral 7*8B has the context window with only 8k tokens[13]. We will consider using LONGLLMLIN-GUA ([Jiang et al., 2024c]()) to compress the prompt while the model performance can be enhanced.

We only performed experiments using datasets in English and German (only for HealthFC). In other languages, current models might not offer the same advantages.

We have to use different automatic evaluation metrics or models for the German data in HealthFC, as BLEURT and TIGERScore do not support languages other than English.

Ethical statement

The conducted user study was ethically approved by the Ethics Committee of Faculty IV of Technische Universität Berlin. The 32 annotators in our

---

user study were paid at least the minimum wage according to the standards of our host institutions' regions. The annotation took each annotator 30 minutes on average.

Acknowledgment

References

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

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.

AI@Meta. 2024. Llama 3 model card.

Shengnan An, Zexiong Ma, Zeqi Lin, Nanning Zheng, Jian-Guang Lou, and Weizhu Chen. 2024. Learning from mistakes makes llm better reasoner. *Preprint*, arXiv:2310.20689.

Siddhant Arora, Danish Pruthi, Norman Sadeh, William W. Cohen, Zachary C. Lipton, and Graham Neubig. 2022. Explain, edit, and understand: Rethinking user study design for evaluating model explanations. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(5):5277–5285.

Pepa Atanasova, Oana-Maria Camburu, Christina Lioma, Thomas Lukasiewicz, Jakob Grue Simonsen, and Isabelle Augenstein. 2023. Faithfulness tests for natural language explanations. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 283–294, Toronto, Canada. Association for Computational Linguistics.

Isabelle Augenstein, Timothy Baldwin, Meeyoung Cha, Tanmoy Chakraborty, Giovanni Luca Ciampaglia, David Corney, Renee DiResta, Emilio Ferrara, Scott Hale, Alon Halevy, Eduard Hovy, Heng Ji, Filippo Menczer, Ruben Miguez, Preslav Nakov, Dietram Scheufele, Shivam Sharma, and Giovanni Zagni. 2023. Factuality challenges in the era of large language models. *Preprint*, arXiv:2310.05189.

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.

Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. 2018. e-snli: natural language inference with natural language explanations. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, page 9560–9572, Red Hook, NY, USA. Curran Associates Inc.

Chun Sik Chan, Huanqi Kong, and Liang Guanqing. 2022. A comparative study of faithfulness metrics for model interpretability methods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5029–5038, Dublin, Ireland. Association for Computational Linguistics.

Cheng-Han Chiang and Hung-yi Lee. 2023. Can large language models be an alternative to human evaluations? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15607–15631, Toronto, Canada. Association for Computational Linguistics.

Yu-Neng Chuang, Guanchu Wang, Chia-Yuan Chang, Ruixiang Tang, Shaochen Zhong, Fan Yang, Mengnan Du, Xuanting Cai, and Xia Hu. 2024. Faithlm: Towards faithful explanations for large language models. *Preprint*, arXiv:2402.04678.

Miruna-Adriana Clinciu, Arash Eshghi, and Helen Hastie. 2021. A study of automatic metrics for the evaluation of natural language explanations. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2376–2387, Online. Association for Computational Linguistics.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2006. The pascal recognising textual entailment challenge. In *Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Tectual Entailment*, pages 177–190, Berlin, Heidelberg. Springer Berlin Heidelberg.

Sara De Felice, Antonia F de C Hamilton, Marta Ponari, and Gabriella Vigliocco. 2023. Learning from others is good, with others is better: the role of social interaction in human acquisition of new knowledge. *Philosophical Transactions of the Royal Society B*, 378(1870):20210357.

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.

Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. *Preprint*, arXiv:1702.08608.

Nils Feldhus, Qianli Wang, Tatiana Anikina, Sahil Chopra, Cennet Oguz, and Sebastian Möller. 2023. InterroLang: Exploring NLP models and datasets through dialogue-based explanations. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5399–5421, Singapore. Association for Computational Linguistics.

Andrew D Foster and Mark R Rosenzweig. 1995. Learning by doing and learning from others: Human capital and technical change in agriculture. *Journal of political Economy*, 103(6):1176–1209.

Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. 2023. OPTQ: Accurate quantization for generative pre-trained transformers. In *The Eleventh International Conference on Learning Representations*.

Zhijiang Guo, Michael Schlichtkrull, and Andreas Vlachos. 2022. A survey on automated fact-checking. *Transactions of the Association for Computational Linguistics*, 10:178–206.

John Hewitt, Kawin Ethayarajh, Percy Liang, and Christopher Manning. 2021. Conditional probing: measuring usable information beyond a baseline. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1626–1639, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Fan Huang, Haewoon Kwak, Kunwoo Park, and Jisun An. 2024a. ChatGPT rates natural language explanation quality like humans: But on which scales? In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3111–3132, Torino, Italia. ELRA and ICCL.

Jiaxin Huang, Shixiang Gu, Le Hou, Yuexin Wu, Xuezhi Wang, Hongkun Yu, and Jiawei Han. 2023. Large language models can self-improve. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1051–1068, Singapore. Association for Computational Linguistics.

Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, and Denny Zhou. 2024b. Large language models cannot self-correct reasoning yet. In *The Twelfth International Conference on Learning Representations*.

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, Online. Association for Computational Linguistics.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2024a. Mixtral of experts. *Preprint*, arXiv:2401.04088.

Dongfu Jiang, Yishan Li, Ge Zhang, Wenhao Huang, Bill Yuchen Lin, and Wenhu Chen. 2024b. TIGER-Score: Towards building explainable metric for all text generation tasks. *Transactions on Machine Learning Research*.

Huiqiang Jiang, Qianhui Wu, Xufang Luo, Dongsheng Li, Chin-Yew Lin, Yuqing Yang, and Lili Qiu. 2024c. LongLLMLingua: Accelerating and enhancing LLMs in long context scenarios via prompt compression. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1658–1677, Bangkok, Thailand. Association for Computational Linguistics.

Zhengping Jiang, Yining Lu, Hanjie Chen, Daniel Khashabi, Benjamin Van Durme, and Anqi Liu. 2024d. RORA: Robust free-text rationale evaluation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1070–1087, Bangkok, Thailand. Association for Computational Linguistics.

Takeshi Kojima, Shixiang (Shane) Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems*, volume 35, pages 22199–22213. Curran Associates, Inc.

Satyapriya Krishna, Jiaqi Ma, Dylan Slack, Asma Ghandeharioun, Sameer Singh, and Himabindu Lakkaraju. 2023. Post hoc explanations of language models can improve language models. In *Advances in Neural Information Processing Systems*, volume 36, pages 65468–65483. Curran Associates, Inc.

Taeyoon Kwon, Kai Tzu-iunn Ong, Dongjin Kang, Seungjun Moon, Jeong Ryong Lee, Dosik Hwang, Beomseok Sohn, Yongsik Sim, Dongha Lee, and Jinyoung Yeo. 2024. Large language models are clinical reasoners: Reasoning-aware diagnosis framework with prompt-generated rationales. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(16):18417–18425.

Himabindu Lakkaraju, Ece Kamar, Rich Caruana, and Jure Leskovec. 2019. Faithful and customizable explanations of black box models. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '19, page 131–138, New York, NY, USA. Association for Computing Machinery.

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.

Moxin Li, Wenjie Wang, Fuli Feng, Fengbin Zhu, Qifan Wang, and Tat-Seng Chua. 2024. Think twice before trusting: Self-detection for large language models through comprehensive answer reflection. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 11858–11875, Miami, Florida, USA. Association for Computational Linguistics.

Qing Lyu, Marianna Apidianaki, and Chris Callison-Burch. 2024. Towards Faithful Model Explanation in NLP: A Survey. *Computational Linguistics*, 50(2):657–723.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-Refine: Iterative refinement with self-feedback. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Andreas Madsen, Siva Reddy, and Sarath Chandar. 2022. Post-hoc interpretability for neural nlp: A survey. *ACM Comput. Surv.*, 55(8).

Thomas Mangin, Michel Audiffren, Alison Lorcery, Francesco Mirabelli, Abdelrhani Benraiss, and Nathalie André. 2022. A plausible link between the time-on-task effect and the sequential task effect. *Frontiers in Psychology*, 13:998393.

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.

OpenAI. 2023. ChatGPT (july 2023 version). Large language model.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Chandan Singh, John X. Morris, Jyoti Aneja, Alexander Rush, and Jianfeng Gao. 2023. Explaining data patterns in natural language with language models. In *Proceedings of the 6th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 31–55, Singapore. Association for Computational Linguistics.

William Stern, Seng Jhing Goh, Nasheen Nur, Patrick J Aragon, Thomas Mercer, Siddhartha Bhattacharyya, Chiradeep Sen, and Van Minh Nguyen. 2024. Natural language explanations for suicide risk classification using large language models. In *ML4CMH@ AAAI*, pages 74–83.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A question answering challenge targeting commonsense knowledge. 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 4149–4158, Minneapolis, Minnesota. Association for Computational Linguistics.

Chenhao Tan. 2022. On the diversity and limits of human explanations. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2173–2188, Seattle, United States. Association for Computational Linguistics.

Yongqi Tong, Dawei Li, Sizhe Wang, Yujia Wang, Fei Teng, and Jingbo Shang. 2024. Can LLMs learn from previous mistakes? investigating LLMs' errors to boost for reasoning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3065–3080, Bangkok, Thailand. Association for Computational Linguistics.

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.

Alicia Tsai, Adam Kraft, Long Jin, Chenwei Cai, Anahita Hosseini, Taibai Xu, Zemin Zhang, Lichan Hong, Ed Chi, and Xinyang Yi. 2024. Leveraging LLM reasoning enhances personalized recommender systems. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 13176–13188,

Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Gladys Tyen, Hassan Mansoor, Victor Carbune, Peter Chen, and Tony Mak. 2024. LLMs cannot find reasoning errors, but can correct them given the error location. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 13894–13908, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Anahuac Valero Haro, Omid Noroozi, Harm JA Biemans, and Martin Mulder. 2019. The effects of an online learning environment with worked examples and peer feedback on students' argumentative essay writing and domain-specific knowledge acquisition in the field of biotechnology. *Journal of Biological Education*, 53(4):390–398.

Juraj Vladika, Phillip Schneider, and Florian Matthes. 2024. HealthFC: Verifying health claims with evidence-based medical fact-checking. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 8095–8107, Torino, Italia. ELRA and ICCL.

Junlin Wang, Jue Wang, Ben Athiwaratkun, Ce Zhang, and James Zou. 2024a. Mixture-of-agents enhances large language model capabilities. *Preprint*, arXiv:2406.04692.

Lei Wang, Wanyu Xu, Yihuai Lan, Zhiqiang Hu, Yunshi Lan, Roy Ka-Wei Lee, and Ee-Peng Lim. 2023a. Plan-and-solve prompting: Improving zero-shot chain-of-thought reasoning by large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2609–2634, Toronto, Canada. Association for Computational Linguistics.

Qianli Wang, Tatiana Anikina, Nils Feldhus, Josef Genabith, Leonhard Hennig, and Sebastian Möller. 2024b. LLMCheckup: Conversational examination of large language models via interpretability tools and self-explanations. In *Proceedings of the Third Workshop on Bridging Human–Computer Interaction and Natural Language Processing*, pages 89–104, Mexico City, Mexico. Association for Computational Linguistics.

Qianli Wang, Tatiana Anikina, Nils Feldhus, Simon Ostermann, and Sebastian Möller. 2024c. CoXQL: A dataset for parsing explanation requests in conversational XAI systems. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 1410–1422, Miami, Florida, USA. Association for Computational Linguistics.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023b. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc.

Sean Welleck, Ximing Lu, Peter West, Faeze Brahman, Tianxiao Shen, Daniel Khashabi, and Yejin Choi. 2023. Generating sequences by learning to self-correct. In *The Eleventh International Conference on Learning Representations*.

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.

Tianyang Xu, Shujin Wu, Shizhe Diao, Xiaoze Liu, Xingyao Wang, Yangyi Chen, and Jing Gao. 2024. SaySelf: Teaching LLMs to express confidence with self-reflective rationales. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5985–5998, Miami, Florida, USA. Association for Computational Linguistics.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jianxin Yang, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Xuejing Liu, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, Zhifang Guo, and Zhihao Fan. 2024a. Qwen2 technical report. *Preprint*, arXiv:2407.10671.

Chengrun Yang, Xuezhi Wang, Yifeng Lu, Hanxiao Liu, Quoc V Le, Denny Zhou, and Xinyun Chen. 2024b. Large language models as optimizers. In *The Twelfth International Conference on Learning Representations*.

Yuncheng Yang, Yulei Qin, Tong Wu, Zihan Xu, Gang Li, Pengcheng Guo, Hang Shao, Yucheng Shi, Ke Li, Xing Sun, Jie Yang, and Yun Gu. 2024c. Leveraging open knowledge for advancing task expertise in large language models. *Preprint*, arXiv:2408.15915.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. Bartscore: Evaluating generated text as text generation. In *Advances in Neural Information Processing Systems*, volume 34, pages 27263–27277. Curran Associates, Inc.

Haodi Zhang, Min Cai, Xinhe Zhang, Chen Jason Zhang, Rui Mao, and Kaishun Wu. 2023. Self-convinced prompting: Few-shot question answering with repeated introspection. *Preprint*, arXiv:2310.05035.

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*.

Haiyan Zhao, Hanjie Chen, Fan Yang, Ninghao Liu, Huiqi Deng, Hengyi Cai, Shuaiqiang Wang, Dawei Yin, and Mengnan Du. 2024. Explainability for large language models: A survey. *ACM Trans. Intell. Syst. Technol.*, 15(2).

Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M. Meyer, and Steffen Eger. 2019. MoverScore: Text generation evaluating with contextualized embeddings and earth mover distance. 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 563–578, Hong Kong, China. Association for Computational Linguistics.

A CROSS-REFINE Examples

Figure 3, Figure 4, and Figure 5 show CROSS-REFINE examples from ECQA, eSNLI, and HealthFC.

Question: Why would a person put flowers in a room with dirty gym socks?

Choices: continue to grow, plant themselves, many colors, smell good, make pretty

Label: smell good

Initial Explanation: The most likely reason for putting flowers in a room with dirty gym socks is to make the room prettier with their vibrant colors and pleasant fragrance, even amidst the unpleasant odor of the socks.

Feedback: The explanation provided suggests that flowers are placed in the room to make it prettier despite the presence of dirty gym socks. While this is a plausible reason, it doesn't directly address the interaction between the flowers and the socks. It's important to consider that the primary function of the flowers in this scenario is likely related to their fragrance rather than just visual aesthetics.

Suggestion: Flowers are likely placed in a room with dirty gym socks primarily to improve the room's scent by counteracting the unpleasant odor with their own natural fragrance, thereby making the environment more pleasant despite the socks' smell.

Refined Explanation: The choice is "smell good". The most likely reason for putting flowers in a room with dirty gym socks is to mask the unpleasant odor of the socks with their pleasant fragrance.

Figure 3: CROSS-REFINE example on ECQA dataset.

B Dataset

B.1 Dataset Example

Figure 6 shows data points from ECQA, eSNLI, and HealthFC.

Premise: Three firefighter come out of subway station.

Hypothesis: Three firefighters putting out a fire inside of a subway station.

Label: neutral

Initial Explanation: The premise that three firefighters come out of a subway station logically implies that three firefighters are putting out a fire inside of the subway station.

Feedback: The explanation is incorrect because it makes an unsupported assumption. The premise only mentions the firefighters coming out of the subway station and does not specify their actions inside the station. Therefore, the hypothesis is not logically implied by the premise.

Suggestion: The premise states that three firefighters come out of a subway station but does not provide details about their activities inside the station. The hypothesis claims that the firefighters were putting out a fire inside the subway station. Since the premise does not specify what the firefighters were doing, it neither supports nor contradicts the hypothesis, making the relationship neutral.

Refined Explanation: The premise states that three firefighters come out of a subway station, but it does not mention their activities inside. The hypothesis suggests they were putting out a fire inside the station, which is not explicitly stated. Therefore, the relationship is neutral as the premise does not provide enough evidence to support or contradict the hypothesis.

Figure 4: CROSS-REFINE example on eSNLI dataset.

Question: Can homeopathic preparations with Anamirta cocculus help against dizziness?

Document: But without success: We could not find any meaningful scientific evidence for effectiveness in dizziness. It is now well proven by studies that homeopathic preparations usually do not work better than a sham medication (placebo) [8]. However, it is not well established that betahistine is effective against dizziness [11,13]. The fact that betahist helps with dizziness is not well documented [11, 13].

Label: no

Initial Explanation: There is no meaningful scientific evidence for effectiveness in dizziness. Source: [8, 11, 13] 1. A randomized, double-blind, placebo-controlled, multicenter study (the "Betahistine Study") was conducted to evaluate the efficacy of betahistine in patients with vertigo and dizziness. The study included 124 patients with vertigo and dizziness, and it was found that betahistine was not more effective than a placebo in reducing symptoms of vertigo and dizziness. [8]
2. Another study was conducted to evaluate the effectiveness of betahistine in patients with benign paroxysmal positional vertigo (BPPV).

Feedback: The initial explanation correctly identifies the lack of meaningful scientific evidence for the effectiveness of homeopathic preparations, including Anamirta cocculus, in treating dizziness. It also notes that betahistine, a different medication, is not well-documented as effective against dizziness. However, the explanation includes details about studies on betahistine, which are not directly related to the question about Anamirta cocculus and might introduce unnecessary complexity.

Suggestion: No, homeopathic preparations with Anamirta cocculus have not been shown to be effective against dizziness. The document indicates that there is no meaningful scientific evidence supporting the effectiveness of homeopathic treatments for dizziness. Additionally, studies have demonstrated that homeopathic remedies generally do not perform better than a placebo. Therefore, there is no reliable evidence to suggest that Anamirta cocculus or similar homeopathic preparations can help with dizziness.

Refined Explanation: There is no meaningful scientific evidence to support the effectiveness of Anamirta cocculus, a homeopathic preparation, in treating dizziness. Studies have consistently shown that homeopathic treatments, including Anamirta cocculus, do not work better than a placebo. Therefore, it is not possible to conclude that Anamirta cocculus or similar homeopathic preparations can help alleviate dizziness.

Figure 5: CROSS-REFINE example on HealthFC dataset.

B.2 Label Distribution

Figure 7 displays label distributions of eSNLI and HealthFC, as ECQA does not have fixed labels.

C Experiment

C.1 Models

Table 7 demonstrates LLMs that are used for CROSS-REFINE. To reduce memory consump-
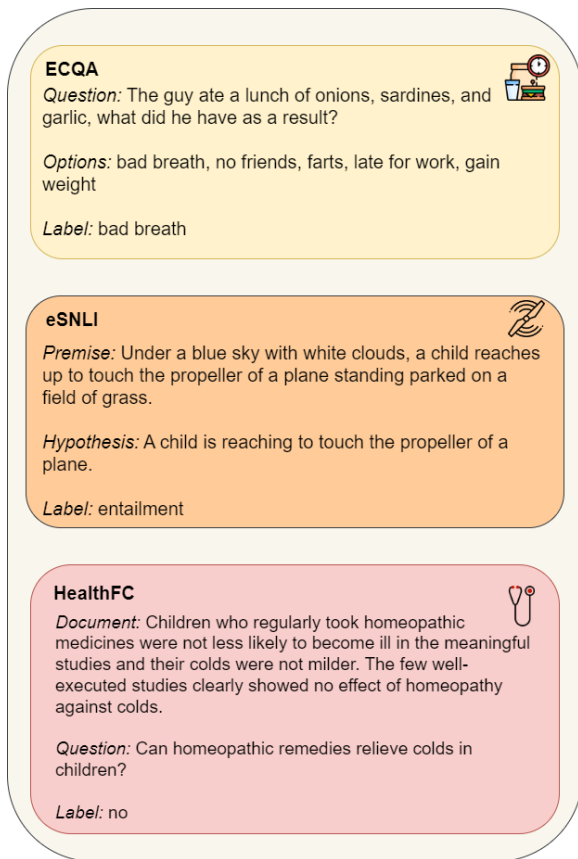
Figure 6: Data points from ECQA, eSNLI, and HealthFC.

(a) eSNLI dataset

(b) HealthFC dataset

Figure 7: Label distributions of eSNLI and HealthFC.

tion, we use a GPTQ-quantized version (Frantar et al., 2023). All LLMs are directly downloaded from Huggingface and run on a single NVIDIA RTXA6000, A100 or H100 GPU.
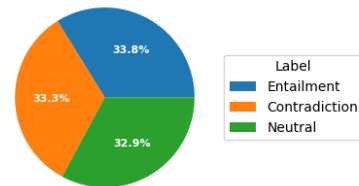
C.2 Inference Time

Table 8 shows inference time for feedback & suggestions generation and refinement of explanations using Qwen2-7B, Mixtral 8*7B and Llama3-70B on ECQA, eSNLI and HealthFC.

D Demonstrations for In-Context Learning

Firstly, we prompt Llama3-8B (AI@Meta, 2024) to generate the initial explanations, which potentially has more room for improvement compared to larger LLMs[14]. Afterwards, we ask ChatGPT to provide corresponding feedback and suggestions. Then we manually create a small subset of data points that can be used as demonstrations for refining explanations, which are reviewed by two au-
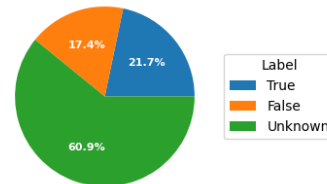
thors of this paper. Lastly, Llama3-8B is prompted with created demonstrations to refine the initial explanations based on feedback and suggestions. The generated outputs then undergo a review process and are post-processed if necessary. For instance, if the initial explanation is of good quality and does not require improvement, or if the refined explanation is of lower quality than the initial explanation, we annotate whether examples need further refinement. Finally, we gather a total of 60 data points for FiXer.

E Models Used for Automatic Evaluation Metrics

Table 9 displays the models used for automatic evaluation metrics.

F Prompt Instruction

The prompts used by CROSS-REFINE for explanation refinement are given in Figure 8.

G User Study

Figure 9 displays the descriptions and instructions that we give the annotators for the user study.

H Sample Selection for User Study

For the HealthFC dataset we observe different quality of generated explanations and to make sure that the explanations involved in the user

---

[14]Note that for HealthFC (German), we use ChatGPT instead of Llama3-8B to ensure that the generated outputs are consistently in German.

Name	Citation	Size	Link
Qwen2	Yang et al. (2024a)	7B	https://huggingface.co/Qwen/Qwen2-7B
Mixtral	Jiang et al. (2024a)	8*7B	https://huggingface.co/mistralai/Mixtral-8x7B-v0.11
Llama3	AI@Meta (2024)	70B	https://huggingface.co/meta-llama/Meta-Llama-3-70B

Table 7: Three open sourced LLMs used in CROSS-REFINE.

	ECQA		eSNLI		HealthFC	
Model	Feedback & Suggestions	Refinement	Feedback & Suggestions	Refinement	Feedback & Suggestions	Refinement
Qwen2-7B	2h	5h	2h	4h	6h	14h
Mixtral 8*7B	7h	15h	7h	12h	15h	21h
Llama3-70B	9h	15h	8h	16h	21h	48h

Table 8: Inference time for feedback & suggestions generation and refinement of explanations using Qwen2-7B, Mixtral 8*7B and Llama3-70B on ECQA, eSNLI and HealthFC.

Metric	Model	Link
BLEURT	BERT	https://huggingface.co/prajjwal1/bert-tiny
BARTScore	BART	https://huggingface.co/facebook/bart-large-cnn
TIGERScore	Llama2	https://huggingface.co/TIGER-Lab/TIGERScore-7B
BARTScore (DE)	mBART	https://huggingface.co/facebook/mbart-large-50
MoverScore	BERT	https://huggingface.co/google-bert/bert-base-german-cased
BERTScore	BERT	https://huggingface.co/google-bert/bert-base-german-cased

Table 9: Models used for automatic evaluation metrics.

study are meaningful we apply some selection criteria to filter out suboptimal generations (tokenization was performed with NLTK[15] and cosine similarity was computed with SENTENCE-TRANSFORMER[16] using the pre-trained model multi-qa-mpnet-base-cos-v1[17]):

1. Explanation length within 20 to 50 tokens.

2. Bigram ratio: $\frac{num\_bigram\_types}{total\_num\_bigrams} >= 0.8$ to ensure the diversity of generated samples without too many repetitions of the same token(s).

3. Digit ratio: $\frac{num\_digit\_tokens}{total\_num\_tokens} <= 0.3$ to ensure that the explanation does not contain too many digits.

4. Cosine similarity between the embeddings of the original question and generated explanation is at least 0.6 to avoid including such cases where e.g. the model generates an explanation for one of the demonstrations instead of the input question-document pair.

From those samples that fulfill all the requirements, we randomly sample 10 explanations per setting. The same procedure is applied to all combinations of models in both SELF-REFINE and CROSS-REFINE settings.

I Examples of SELF-REFINE and CROSS-REFINE

Figure 10 shows examples of SELF-REFINE and CROSS-REFINE.

---

[15] https://www.nltk.org/
[16] https://sbert.net/
[17] https://huggingface.co/sentence-transformers/multi-qa-mpnet-base-cos-v1

ECQA:

You are an excellent assistant to improve explanations through feedback and suggestion. Your task is to firstly understand the question and choices. Then please refine the initial explanation based on the provided feedback and suggestion. Belows are some examples. Please follow the structure of given examples to generate refine explanation.
Question: [*question*]
Choices: [*choices*]
Answer: [*answer*]
Feedback: [*feedback*]
Suggestion: [*suggestion*]
Refined explanation:

eSNLI:

You are an excellent assistant to improve explanations through feedback and suggestion. Your task is to firstly understand the premise and hypothesis. Then please refine the initial explanation based on the provided feedback and suggestion. Belows are some examples. Please follow the structure of given examples to generate refine explanation.
Premise: [*premise*]
Hypothesis: [*hypothesis*]
Answer: [*answer*]
Feedback: [*feedback*]
Suggestion: [*suggestion*]
Refined explanation:

HealthFC:

You are an excellent assistant to improve explanations through feedback and suggestion. Your task is to firstly understand the question and document. Then please refine the initial explanation based on the provided feedback and suggestion. Belows are some examples. Please follow the structure of given examples to generate refine explanation.
Question: [*question*]
Document: [*document*]
Answer: [*answer*]
Feedback: [*feedback*]
Suggestion: [*suggestion*]
Refined explanation:

Figure 8: Prompt instructions for ECQA, eSNLI, and HealthFC.

User Study Description

Dear participants,

Our user study focuses on the evaluation of natural language explanations. We select two datasets: **ECQA** (commonsense question answering) and **HealthFC** (fact-checking). The explanations are generated by different model combinations. You will receive explanations generated by one specific model combination and from one single dataset.

Dataset Structure:

1. ECQA consists of **"question"** and **"choices"**. Based on the given question, the most suitable choice should be chosen.
2. HealthFC consists of **"question"** and **"document"**. Based on the provided document, answer the question with *unknown, yes, no*.

User Study Instruction

You will be provided with 10 instances to evaluate. Each instance contains two fields depending on the dataset: the model prediction and the model explanation. Please assess the explanation based on the following dimensions:

1. **Faithfulness (Binary)**: Investigate whether the generated reasoning exhibits hallucination and if it includes any misinformation. (unfaithful or faithful)
2. **Coherence (5-point Likert)**: Assess whether the generated reasoning is sensible, clear and coherent and reflects the reasons behind the user's preference. (strong disagree, disagree, neutral, agree, strong agree)
3. **Insightfulness (5-point Likert)**: Evaluate the extent to which the generated reasoning provides informative insights into the user's preferences. (strong disagree, disagree, neutral, agree, strong agree)

Please note that if you find the explanation is not relevant to the given context, assign **"strong disagree"** for **insightfulness**.

Figure 9: Descriptions and instructions of the user study.

Question: What grows well in the sun?
Choices: west-beach-beans-universe-plants
Prediction: universe
Explanation: A sun is a star and a star is an object in the universe. Therefore, a sun grows in the universe.

(a) SELF-REFINE

Question: What grows well in the sun?
Choices: west-beach-beans-universe-plants
Prediction: plants
Explanation: 1. Plants need sunlight to grow. 2. The sun is a source of light and warmth. 3. Plants absorb sunlight through their leaves. 4. Sunlight provides energy for photosynthesis, the process by which plants produce their own food. 5. Sunlight also helps plants regulate their temperature

(b) CROSS-REFINE

Figure 10: Examples of SELF-REFINE and CROSS-REFINE. The gold label is *plants*.