# Enhancing Relation Extraction via Supervised Rationale Verification and Feedback

(/pdf?id=yRFQDiVxdA)

*Yongqi Li (/profile?id=~Yongqi_Li3), Xin Miao (/profile?id=~Xin_Miao4), Shen Zhou (/profile?id=~Shen_Zhou2), Mayi Xu (/profile?id=~Mayi_Xu1), Yuyang Ren (/profile?id=~Yuyang_Ren1), Tieyun Qian (/profile?id=~Tieyun_Qian1)*

Published: 10 Dec 2024, Last Modified: 18 Jan 2025 AAAI 2025 Oral Conference, Area Chairs, Senior Program Committee, Program Committee, Authors Revisions (/revisions?id=yRFQDiVxdA) BibTeX 

**Primary Keyword:** Speech & Natural Language Processing (SNLP) -> SNLP: Information Extraction
**Secondary Keywords:** Speech & Natural Language Processing (SNLP) -> SNLP: Language Models
**TL;DR:** A novel automated feedback framework for LLM based relation extraction.

**Abstract:**
Despite the rapid progress that existing automated feedback methods have made in correcting the output of large language models (LLMs), these methods cannot be well applied to the relation extraction (RE) task due to their designated feedback objectives and correction manner. To address this problem, we propose a novel automated feedback framework for RE, which presents a rationale supervisor to verify the rationale and provide re-selected demonstrations as feedback to correct the initial prediction. Specifically, we first design a causal intervention and observation method for to collect biased/unbiased rationales for contrastive training the rationale supervisor. Then, we present a verification-feedback-correction procedure to iteratively enhance LLMs' capability of handling the RE task. Extensive experiments prove that our proposed framework significantly outperforms existing methods.

**Supplementary Material:** zip (/attachment?id=yRFQDiVxdA&name=supplementary_material)
**iThenticate Agreement:** Yes, I agree to iThenticate's EULA agreement version: v1beta
**Reproducibility Checklist:** I certify all co-authors of this work have read and completed the Reproducibility Checklist.
**Submission Number:** 12736



## Paper Decision

Decision by Program Chairs 10 Dec 2024, 06:13 (modified: 18 Jan 2025, 03:05)
Program Chairs, Area Chairs, Senior Program Committee, Authors Revisions (/revisions?id=AT3oZa0hEb)

**Decision:** Accept (Oral)
**Comment:**
The paper aims to improve relation extraction metrics for LLMs. The central idea is to optimize the selection of in-context examples through a feedback loop that helps resample the examples used in LLM input. There was consensus in the PC that this paper would be a good addition to the AAAI program.

## Mitigating LLM Bias in Relation Extraction

Official Review by Program Committee WooT 02 Nov 2024, 08:23 (modified: 10 Dec 2024, 06:47)
Program Chairs, Area Chairs, Senior Program Committee, Program Committee, Authors
Revisions (/revisions?id=9BU7DIiFh9)

**Review:**
The authors introduces a novel framework designed to improve RE by addressing challenges in biased predictions from large language models LLMs. The proposed SRVF framework incorporates a rationale supervisor that identifies biases in LLM predictions and adjusts by supplying refined demonstrations as feedback. This iterative correction mechanism is shown to

enhance RE performance across several datasets, with the framework outperforming existing methods.

- The authors propose a causal intervention method to collect biased and unbiased rationales, addressing a challenging aspect of the RE task.
- They conduct comprehensive experiments using several datasets and multiple LLMs, with results consistently showing performance gains.
- The authors acknowledge concerns regarding inference speed arising from the use of feedback and have addressed this issue.

**Rating:** 7: Good paper, accept
**Confidence:** 1: The reviewer's evaluation is an educated guess

### Rebuttal by Authors

Rebuttal
by Authors (Xin Miao (/profile?id=~Xin_Miao4), Shen Zhou (/profile?id=~Shen_Zhou2), Tieyun Qian (/profile?id=~Tieyun_Qian1), Mayi Xu (/profile?id=~Mayi_Xu1), +2 more (/group/info?id=AAAI.org/2025/Conference/Submission12736/Authors))
08 Nov 2024, 16:57 (modified: 12 Nov 2024, 00:47)
Program Chairs, Area Chairs, Senior Program Committee, Program Committee, Authors
Revisions (/revisions?id=qbvmkTNUa1)

**Rebuttal:**
Dear Reviewer,

We are very pleased that our exploration on automated feedback for LLM-based RE has received your recognition, and we sincerely appreciate the time and efforts you dedicated to reviewing our paper. Once again, thank you very much!

Best wishes!

Paper 12736 Authors.

### Review

Official Review by Program Committee HzHE 15 Oct 2024, 06:33 (modified: 10 Dec 2024, 06:47)
Program Chairs, Area Chairs, Senior Program Committee, Program Committee, Authors
Revisions (/revisions?id=3sowylY9pj)

**Review:**

# Summary

The paper proposes a framework to improve relation extraction (RE) by a rationale supervisor that verifies the rationale and reselects demonstrations as feedback.

# Strengths

1. The use of rationale is an innovative way to improve on the RE task. It is based on an interesting observation of the stereotypes arising from pre-training. Also, the rationale is only a short piece of text in the output, which makes the inference cost low.
2. This method of reselecting demonstrations can potentially extend to other ICL tasks as well.
3. The paper comprehensively tests the proposed framework on a number of LLMs. Even on one of the SoTA LLMs Llama-3-70B, the proposed framework has a consistent and large improvement in performance.

# Weaknesses

1. The writing is a bit confusing. It is not clear until Figure 3 that rationale is the explanation for RE. While an example ("stereotype") has been given in the introduction, the rationale is not explicitly defined so it is not clear which part of the example is the rationale.
2. The rational supervisor is a BERT model. The fact that the rationale supervisor is a small model is quite important, but it is only mentioned in the appendix and not in the main paper.
3. It would be great, if budget permits, that the experiments in Table 1 are conducted with a stronger model than Llama-2-7b-chat. The results from stronger models are presented in Table 3, but Table 3 is limited to much fewer combinations of methods and hyperparameters.

**Rating:** 8: Top 50% of accepted papers, clear accept
**Confidence:** 4: The reviewer is confident but not absolutely certain that the evaluation is correct

### Rebuttal by Authors

Rebuttal
by Authors (Xin Miao (/profile?id=~Xin_Miao4), Shen Zhou (/profile?id=~Shen_Zhou2), Tieyun Qian (/profile?id=~Tieyun_Qian1), Mayi Xu (/profile?id=~Mayi_Xu1), +2 more (/group/info?id=AAAI.org/2025/Conference/Submission12736/Authors))
08 Nov 2024, 16:58 (modified: 12 Nov 2024, 00:47)
Program Chairs, Area Chairs, Senior Program Committee, Program Committee, Authors
Revisions (/revisions?id=ajFpn8cbbk)

**Rebuttal:**
Dear Reviewer,

We are delighted that our work has received your recognition, and we deeply appreciate the constructive feedback you've provided. Below are our responses to your suggestions and questions.

> Q1: No explicit definition of "rationale"

We apologize for not clearly defining "rationale" in our initial writing. In the revised version, we will explicitly define "rationale" in the Introduction section as "the generated explanation when LLMs perform RE" to prevent any confusion for readers.

> Q2: The fact that the rationale supervisor is a small model is only mentioned in the appendix.

Due to space constraints, we moved all experimental implementation details to the appendix in the submitted version, which led to the omission of the statement "The rationale supervisor is a BERT model" in the main paper. We will include this important information in the Method section of the revised version.

> Q3: It would be great, if budget permits, that the experiments in Table 1 are conducted with a stronger model than Llama-2-7b-chat.

To address your concern, we adopt the stronger **Llama-3-70B-Instruct** LLM for all methods in Table 1. However, due to time, resource, and budget constraints during the rebuttal period, we only completed the SemEval 5-shot setting experiments. The results are as follows:

| Backbone | Method | SemEval |
|---|---|---|
| Random | ICL | 65.36 |
| | w/ Self-Refine | 65.32 |
| | w/ Self-Consistency | 66.26 |
| | w/ GRACE | 66.74 |
| | w/ our SRVF | **72.01** |
| SimCSE | ICL | 69.40 |
| | w/ Self-Refine | 69.59 |
| | w/ Self-Consistency | 69.88 |
| | w/ GRACE | 70.34 |
| | w/ our SRVF | **73.87** |
| Task-specific | ICL | 71.21 |
| | w/ Self-Refine | 71.34 |
| | w/ Self-Consistency | 71.67 |
| | w/ GRACE | 72.35 |
| | w/ our SRVF | **74.68** |

Table: Micro-F1 scores using three backbones on the SemEval 5-shot setting. Here we adopt the Llama-3-70B-Instruct as the LLM.

These results show that with sronger LLM, our method can yield even better performance, further validating our method's effectiveness. We will strive to complete experiments for other settings after the rebuttal period.

Thank you once again for your efforts during the review process!

Best wishes!

Paper 12736 Authors.

## Improving Relation Extraction via Re-Sampling In-Context Examples

Official Review by Program Committee BNff 20 Sept 2024, 14:37 (modified: 10 Dec 2024, 06:47)

Program Chairs, Area Chairs, Senior Program Committee, Program Committee, Authors

Revisions (/revisions?id=azqhU26WFN)

**Review:**

This paper aims to improve relation extraction metrics for large language models (LLMs). The central idea is to optimize the selection of in-context examples through a feedback loop that helps resample the examples used in LLM input. The proposed feedback approach enhances the target metrics across three datasets, outperforming baselines for various models.

Pros:

- The extensive experiments support the proposed feedback methodology, demonstrating substantial improvements.
- The methodology has the potential to improve a variety of similar tasks in a similar way.

Cons:

- An iterative approach to improvement increases inference costs. Key questions to consider: Is the increase in computation justified by the performance gains for relation extraction? Is the proposed method effective for smaller models (e.g., those with fewer than 7 billion parameters)?
- There is little discussion of how the baselines are implemented and set up. For example, how is the Self-Consistency method applied? What is considered a majority vote?

Relation extraction (RE) is often regarded as a foundational component for more complex applications. Should we continue focusing on improving RE with LLMs, or should we shift our attention to solving more complex use cases directly?

Required Fixes:

- Table 2, column 7 is misleading: the best result is achieved by "w/o RCT," whereas "Our SRVF" is incorrectly bolded.

**Rating:** 7: Good paper, accept
**Confidence:** 1: The reviewer's evaluation is an educated guess

## Rebuttal by Authors

Rebuttal
by Authors (Xin Miao (/profile?id=~Xin_Miao4), Shen Zhou (/profile?id=~Shen_Zhou2), Tieyun Qian (/profile?id=~Tieyun_Qian1), Mayi Xu (/profile?id=~Mayi_Xu1), +2 more (/group/info?id=AAAI.org/2025/Conference/Submission12736/Authors))
08 Nov 2024, 16:59 (modified: 12 Nov 2024, 00:47)
Program Chairs, Area Chairs, Senior Program Committee, Program Committee, Authors
Revisions (/revisions?id=i8e9MWR92B)

**Rebuttal:**
Dear Reviewer,

We greatly appreciate your recognition and constructive comments on our work. Below are clarifications to address your concerns.

> Q1: Q1.1 Computation cost. Q1.2 Effectiveness for smaller LLMs.

A1.1 We believe that achieving significant performance gains with minimal inference costs is worthwhile. For example, in our experimental environment, on the SemEval dataset, our SRVF based on a 7B-size LLM increases the average inference time per sample by only 0.09 seconds (from 0.31s to 0.40s), while improving performance by 2.13%-21.96%.

A1.2 We conduct experiments with the recently released lightweight LLMs from the Qwen2.5 and Llama3.2 series. The results are as follows:

| | **Qwen2.5-0.5B-Instruct** | | | **Qwen2.5-1.5B-Instruct** | | | **Qwen2.5-3B-Instruct** | | | **Llama-3.2-1B-Instruct** | | | **Llama3.2-3B-Instruct** | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | SemEval | TACRED | Re-TACRED | SemEval | TACRED | Re-TACRED | SemEval | TACRED | Re-TACRED | SemEval | TACRED | Re-TACRED | SemEval | TA |
| ICL | 41.42 | 11.01 | 15.33 | 41.50 | 17.88 | 27.29 | 59.78 | 24.30 | 27.13 | 41.13 | 13.23 | 18.85 | 59.55 | 2 |
| w/ SRVF | 42.69 | 14.76 | 29.03 | 42.02 | 22.63 | 36.99 | 64.70 | 31.59 | 38.61 | 42.56 | 19.53 | 28.37 | 65.14 | 2 |

Table: Micro-F1 scores averaged over three backbones on the most challenging 5-shot setting with lightweight LLMs.

We can observe that our SRVF maintains a substantial improvement over basic ICL, which validates the effectiveness of our SRVF for lightweight LLMs.

> Q2 Details of baselines

Due to space constraints, we move the baseline details to the supplementary material. In the next version, we will re-organize the core details of the baselines into the main paper.

> Q3 Continue improving RE with LLMs?

Here are our thoughts for this excellent question:

1. RE is crucial for building trustworthy knowledge bases like knowledge graphs (KGs), which are still essential nowadays, such as ensuring the trustworthiness of LLMs based on KGs. Thus, improving LLM-based RE remains important for these applications.
2. For future work, we think both directions are OK. On one hand, we can continue exploring LLMs for more complex document-level, dynamic, or open RE to better serve downstream applications. On the other hand, we can explore complex tasks like reasoning or dialogue systems directly.

> Q4 Required Fixes

Thank you for your careful review and for spotting the bolding error. We apologize and will fix it in a future version.

Best wishes!

Paper 12736 Authors.

## Enhancing Relation Extraction via Supervised Rationale Verification and Feedback

Official Review by Program Committee 98Zn 11 Sept 2024, 09:48 (modified: 10 Dec 2024, 06:47)

Program Chairs, Area Chairs, Senior Program Committee, Program Committee, Authors

Revisions (/revisions?id=P0OG1biMpY)

**Review:**

This paper propose an automated feedback framework for RE. In this framework, a rationale supervisor is adpplied to provide re-selected demonstrations as feedback to correct the predication. This this paper is novel and show clearly improvement for LLMs-based relation extraction. This paper is well written and easy to follows. The presentation of this paper , including pictures, tables are clear and nice. In my opinion, this paper can be accepted.

**Rating:** 6: Marginally above acceptance threshold

**Confidence:** 4: The reviewer is confident but not absolutely certain that the evaluation is correct

### Rebuttal by Authors

Rebuttal

by Authors (Xin Miao (/profile?id=~Xin_Miao4), Shen Zhou (/profile?id=~Shen_Zhou2), Tieyun Qian (/profile?id=~Tieyun_Qian1), Mayi Xu (/profile?id=~Mayi_Xu1), +2 more (/group/info?id=AAAI.org/2025/Conference/Submission12736/Authors))

08 Nov 2024, 16:59 (modified: 12 Nov 2024, 00:47)

Program Chairs, Area Chairs, Senior Program Committee, Program Committee, Authors

Revisions (/revisions?id=cayZQ7LUVQ)

**Rebuttal:**

Dear Reviewer,

We are truly grateful for your generous praise of our work, and we are thrilled that our efforts and explorations have been recognized. Thank you once again for the time and efforts you dedicated during the review process.

Best wishes!

Paper 12736 Authors.