Reasoning vs. Critic-Based Verification for Biomedical Relation Extraction with Large Language Models

Ali Assi[1], Nour Elislem Karabadji[2,4], Mohamed Elati[3] and Wajdi Dhifli[3,*]

[1]Rochester Institute of Technology, Dubai, United Arab Emirates
[2]National Higher School of Technology and Engineering, Laboratoire De Technologies Des Systemes Energetiques (LTSE) E3360100, 23005, Annaba, Algeria
[4]Electronic Document Management Laboratory (LabGED), Badji Mokhtar-Annaba University, P.O. Box 12 Annaba, Algeria
[3]Univ. Lille, CNRS, Inserm, CHU Lille, UMR9020-U1277 – CANTHER – Cancer Heterogeneity Plasticity and Resistance to Therapies, Lille, F-59000, France

Abstract

Large Language Models (LLMs) have shown promising results in biomedical relation extraction tasks, yet they frequently produce false positives due to hallucinations and alignment issues with specific schemas. To address this, we investigate a critic-based verification mechanism where a second "critic" prompt reviews and verifies extracted relations. Through comprehensive benchmarking on the BioRED dataset, we demonstrate that this approach is highly effective, reducing false positives by 58.5% and achieving 74.1% precision with top models, competitive with supervised methods. Furthermore, we reveal critical trade-offs in extraction strategies: while extended reasoning dramatically improves initial single-pass precision (Claude: from 33.1% to 55.7%), the critic layer serves as a "great equalizer" for mid-tier models, boosting Llama-3.2 performance with a 40.8% reduction in errors, whereas frontier models show diminishing returns.

Keywords

Prompt Engineering, Large Language Models, Biomedical Relation Extraction, Self-Correction, BioRED

1. Introduction

Biomedical relation extraction (RE) aims to identify semantic relationships between biological entities (e.g., genes, diseases, chemicals) from scientific literature. While recent Large Language Models (LLMs) have demonstrated impressive capabilities in various natural language processing tasks, they often suffer from *hallucinations*—generating plausible-sounding but factually incorrect outputs [1, 2].

In biomedical domains, false positives are particularly problematic as they can propagate into knowledge bases [3] and mislead downstream research. A common challenge in biomedical RE is handling complex linguistic phenomena such as negation ("Drug A does *not* treat Disease B") and conditional statements, which LLMs frequently misinterpret when using standard single-pass extraction prompts.

We investigate whether LLMs can *self-correct* their extraction errors through a multi-step prompting strategy. Specifically, we propose a critic-based verification approach where:

1. **Phase 1 (Extraction)**. The LLM extracts candidate relations from biomedical text using entity-aware prompts.
2. **Phase 2 (Verification)**. A "critic" prompt reviews each extracted relation against the original text, filtering out unsupported or incorrectly interpreted relationships.

We evaluate our approach on the BioRED benchmark [4], a curated dataset of 600 PubMed abstracts annotated with biomedical entities and relations. Our empirical results on 75 test documents demonstrate a 58.5% reduction in false positives compared to single-pass extraction (1089 → 452), with a

[PromptEng '26] 3rd International Workshop on Prompt Engineering, April 13–14, 2026, Dubai, United Arab Emirates
*Corresponding author.
ali.assi@rit.edu (A. Assi); n.karabadji@ensti-annaba.dz (N. E. Karabadji); mohamed.elati@univ-lille.fr (M. Elati); wajdi.dhifli@univ-lille.fr (W. Dhifli)
0000-0003-3709-1519 (A. Assi); 0000-0002-5307-7090 (N. E. Karabadji); 0000-0002-4440-2904 (M. Elati); 0000-0002-6580-8778 (W. Dhifli)

corresponding improvement in precision and F1-score. This suggests that structured critic prompts effectively serve as a quality control mechanism for LLM-based extraction.

2. Related Work

Biomedical Relation Extraction. Traditional approaches to biomedical RE have relied on supervised learning with neural architectures [5]. Recent work has explored LLM-based extraction using few-shot prompting [6], but these methods often lack explicit error correction mechanisms.

Self-Correction in LLMs. The concept of LLMs verifying their own outputs has gained attention with techniques like self-consistency [7], Reflexion [8], and Self-Refine [9]. However, these methods typically focus on reasoning tasks rather than structured information extraction.

Prompt Engineering for Extraction. Entity-aware prompting, where models are provided with pre-annotated entities to constrain relation extraction, has been shown to improve precision in structured prediction tasks [10]. Our work builds on this by adding a verification layer.

Reasoning-Enhanced Generation. The recent emergence of "reasoning" or "system 2" models (e.g., DeepSeek-R1 [11], OpenAI o1) represents a shift from pure pattern matching to inference-time compute. While these models excel at complex logical puzzles, their application to information extraction remains underexplored. Our study provides one of the first empirical comparisons between internal "thought processes" (extended thinking) and external verification (critic loops) in the biomedical domain.

3. Methodology

3.1. Task Formulation

Given a biomedical text T and a set of pre-annotated entities $E = \{e_1, e_2, \ldots, e_n\}$, our goal is to extract a set of valid relations $R = \{(e_i, r, e_j) \mid e_i, e_j \in E, r \in \mathcal{R}\}$, where $\mathcal{R}$ is a predefined set of relation types (e.g., `association`, `positive_correlation`, `negative_correlation`).

Following standard practice in RE evaluation [4], we use pre-annotated entities from BioRED and focus specifically on the relation classification subtask.

3.2. Single-Pass Baseline

Our baseline uses an entity-aware extraction prompt that provides the LLM with the list of entities and asks it to identify relationships:

```
You are given a text and a list of
biomedical entities. Identify relationships
ONLY between the provided entities.

Text: "{text}"
Entities:
- EntityA (Type)
- EntityB (Type)
...

Output (JSON): [
  {"subject": "EntityA",
   "predicate": "relation_type",
   "object": "EntityB"}
```

]

3.3. Few-Shot Enhancement

To address the alignment issue, we enhance the prompt with a single curated 1-shot example (PMID 15623763) that demonstrates:

- **Exact Entity Usage**: Copying entity text verbatim.
- **Ontology Compliance**: Mapping natural language predicates to BioRED types (e.g., "causing" $\rightarrow$ `positive_correlation`).

The inclusion of this single example serves to calibrate the model's output distribution, shifting it from a generic "open information extraction" mode to a specific "BioRED-compliant" mode. This alignment is crucial, as many zero-shot errors stem not from a lack of understanding, but from a mismatch in expected output format. Our few-shot approach follows established prompt engineering practices [12], incorporating a curated example to demonstrate the desired output format and ontology compliance.

```
**Few-Shot Example:**
Text: "TGFBI gene mutations causing lattice..."
Entities:
- TGFBI (GeneOrGeneProduct)
...
Output:
[{"subject": "TGFBI", "predicate": "association", ...}]
```

3.4. Critic-Based Two-Phase Approach

Our proposed approach adds a verification phase:

Phase 1 (Extraction): Same as the single-pass baseline.

Phase 2 (Critic Verification): For each extracted relation (s, p, o), we construct a verification prompt:

```
Given the original text and this extracted
triple:
- Subject: {subject}
- Predicate: {predicate}
- Object: {object}

Does the text EXPLICITLY support this
relationship? Consider:
1. Negation (e.g., "does not")
2. Speculation (e.g., "might", "possibly")
3. Correct directionality

Answer: CORRECT/NEGATED/SPECULATIVE/INCORRECT
with brief justification.
```

The design of this critic prompt is intentional. By explicitly asking for "Negation" and "Speculation", we force the model to perform specific checks that are often skipped during the primary generation phase. The constraint to find explicit support discourages the hallucination of relations that might be true in general medical knowledge but are not present in the specific text segment.

We retain only relations marked as CORRECT by the critic.

4. Results

4.1. Experimental Setup

Dataset: We use BioRED [4], which contains 600 PubMed abstracts with entity and relation annotations. We evaluate on the standard test split (75 documents) to ensure comparability.

Models: We evaluate three tiers of LLMs:

- **Local Models (Ollama):** Llama-3.2 (temp=0.1), MedLlama2 (7B, temp=0.1). Llama-3.2 was selected as the efficiency baseline due to its widespread adoption and low latency.
- **Reasoning Models:** DeepSeek V3 (temp=0.1), Claude 4.5 Sonnet with extended thinking, Claude Opus 4.5 with extended thinking.
- **Frontier Models:** Gemini 3 Pro (temp=0.1), Claude 4.5 Sonnet (standard mode, temp=0.0).

Metrics: We report micro-averaged precision, recall, and F1-score. Additionally, we analyze false positive (FP) counts and categorize error types.

4.2. Role of Prompting Strategies

Table 1 presents our main findings comparing single-pass extraction against the critic-based approach.

Table 1
Performance Comparison: Zero-Shot vs. Few-Shot Prompting strategies (Llama-3.2) on 75 BioRED test documents.

Strategy	Method	Precision	Recall	F1	False Positives
Zero-Shot	Single-Pass	0.050	0.090	0.064	1089
	Critic-Based	0.079	0.062	0.070	452 (**-58.5%**)
Few-Shot (1-shot)	Single-Pass	0.130	0.094	0.109	395
	Critic-Based	**0.182**	0.083	**0.114**	**234** (**-40.8%**)

Zero-Shot Baseline: Initial zero-shot experiments revealed a high noise rate (1089 FPs). The critic successfully filtered 58% of these, but absolute precision remained constrained by output formatting issues.

Few-Shot Improvement: Incorporating a single curated example (PMID 15623763) significantly aligned the extraction with BioRED's schema. As shown in the bottom half of Table 1, precision improved dramatically, confirming that the "Precision Floor" was largely an alignment issue rather than a reasoning failure. The Critic continued to provide value even in this high-precision regime.

False Positive Reduction: The critic successfully filtered false positives in both settings. In the zero-shot baseline, it removed 58.5% of FPs. In the few-shot setting, even with a higher quality initial extraction, the critic further reduced FPs by 40.8% (395 $\rightarrow$ 234), demonstrating its utility across different levels of base model performance.

Precision Gains: The combination of few-shot prompting and critic verification achieved the highest precision (18.2%), more than doubling the zero-shot baseline performance. This confirms that guiding the LLM with both positive examples (few-shot) and negative constraints (critic) is the optimal strategy for BioRED extraction.

4.3. Model Comparison

We benchmarked six models across three tiers (Table 2): local models (Llama-3.2, MedLlama2), reasoning-focused models (DeepSeek V3, Claude 4.5 (Thinking), Claude Opus 4.5 (Thinking)), and frontier SOTA models (Gemini 3 Pro, Claude 4.5 Sonnet).

Table 2
Model Comparison: Local vs SOTA Models (Few-Shot Strategy with Critic enabled).

Model	Type	Precision	Recall	F1	Inference
Local Models (Ollama)					
Llama-3.2	General	0.182	0.083	0.114	∼10s/doc
MedLlama2 (7B)	Specialized	0.154	0.050	0.075	∼3s/doc
SOTA Models (API/Manual Benchmark)					
Gemini 3 Pro	Frontier	**0.383**	**0.257**	**0.307**	N/A
DeepSeek V3	Reasoning	0.363	0.182	0.243	N/A
Claude 4.5 (Thinking)	Reasoning	0.557	0.062	0.111	N/A
Claude Opus 4.5 (Thinking)	Reasoning	0.433	0.082	0.138	N/A
Claude 4.5 Sonnet	General	0.331	0.074	0.122	N/A

Local Models. The general-purpose Llama-3.2 outperformed the specialized MedLlama2 (18.2% vs 15.4% precision), demonstrating that general reasoning capabilities often outweigh domain specialization for relation extraction tasks.

SOTA Models. Manual benchmarking on the full 75-document test set revealed significant performance gaps. Gemini 3 Pro achieved the highest overall F1 (30.7%) with 38.3% precision and 25.7% recall, establishing a strong upper bound. DeepSeek V3 followed closely with 36.3% precision, demonstrating that reasoning-optimized models can approach frontier performance.

Extended Thinking Impact. We compared Claude 4.5 Sonnet in standard and extended thinking modes. Extended thinking dramatically improved Phase 1 precision from 33.1% to 55.7% (+22.6 percentage points), representing the highest single-pass precision among all tested models. This demonstrates that extended reasoning enables more selective extraction with 67% fewer false positives (31 vs 95). However, standard mode achieved higher final precision when combined with critic verification (74.1% vs 66.7%).

Key Finding. The 2.2x performance gap between the best SOTA model (Gemini 3 Pro) and best local model (Llama-3.2) highlights the trade-off between deployment flexibility and extraction quality. However, all models remain far below supervised SOTA (62.8% F1 [4]), indicating substantial room for prompt engineering improvements.

4.4. Critic Phase Effectiveness

Critic Phase Analysis: The effectiveness of the Critic phase varied dramatically across models and reasoning modes (Table 3). Standard Claude 4.5 Sonnet demonstrated the most dramatic improvement, achieving 74.1% precision (Phase 2) from 33.1% (Phase 1) with a remarkable 92.6% false positive reduction. Extended thinking mode showed more balanced behavior: 55.7% → 66.7% precision with moderate 64.5% FP reduction. Llama-3.2 showed moderate improvement (40.8% FP reduction), while Gemini 3 Pro exhibited minor improvement (10.9% FP reduction). This pattern reveals: (1) highly capable models may already produce near-optimal outputs in single-pass extraction, (2) mid-tier models benefit substantially from self-correction, (3) extended thinking improves initial extraction quality, reducing the need for aggressive critic filtering, and (4) aggressive critic filtering can achieve state-of-the-art precision (74.1%)

Table 3
Critic Phase Effectiveness Across Models

Model	Phase 1 P	Phase 2 P	Δ P	FP Reduction
Claude 4.5 Sonnet	0.331	**0.741**	**+0.410**	**92.6%**
Claude 4.5 (Thinking)	0.557	0.667	+0.110	64.5%
Claude Opus 4.5 (Thinking)	0.433	0.486	+0.053	45.6%
Llama-3.2	0.182	0.307	+0.125	40.8%
Gemini 3 Pro	0.383	0.399	+0.016	10.9%

competitive with supervised methods (60-70%). This precision-first approach mirrors certainty-based sampling strategies in systematic review screening [13], where high precision is prioritized to minimize manual review burden.

Table 4
Examples of relations filtered by the Critic.

Text: "...pemt gene polymorphisms were associated with NAFLD..." **Single-Pass Extraction:** (pemt gene polymorphisms, associated with, NAFLD) **Critic Verdict:** CORRECT $\rightarrow$ **Retained** (Matched to ground truth via relaxed evaluation)
Text: "Metformin is not recommended for patients with..." **Single-Pass Extraction:** (Metformin, treats, patients) **Critic Verdict:** NEGATED (Explicit negation found) $\rightarrow$ **Filtered**

Error Analysis. Manual inspection of results reveals two main challenges:

1. **Ontological Mismatch** (Major). The LLM extracts correct relations but uses natural language predicates (e.g., "results in") that do not map 1:1 to BioRED's specific types (e.g., "Positive_Correlation"), reducing precision scores.
2. **Implicit Relations**. Relations implied by context (e.g., A causes B, B causes C $\rightarrow$ A causes C) are often extracted by the LLM but absent in sentence-level ground truth.

4.5. Qualitative Analysis

To better understand the critic's decision boundary, we analyzed specific instances of correction (see Table 4 for examples). A recurrent pattern involved conditional statements. For example, in text stating "Drug A *may* potentially interact with Drug B," single-pass models frequently extracted a definitive 'Association'. The critic, prompted to verify explicit support, correctly flagged this as 'SPECULATIVE' and filtered it out.

Conversely, we observed correctable "over-rejection" errors. In sentences with complex appositives (e.g., "The drug, an inhibitor of Protein X, showed..."), the critic occasionally failed to resolve the syntactic link between the subject and the distant predicate, labeling the relation as 'unsupported'. This suggests that while the critic excels at semantic verification (truthfulness), it can struggle with complex syntactic parsing, a known limitation of current LLMs.

Another interesting phenomenon was the critic's handling of "generic" versus "specific" relations. The prompt instructed the model to verify relations between specific entities. In cases where the text discussed a class of drugs (e.g., "statins") but the extracted relation named a specific instance not mentioned in the immediate context, the critic successfully identified this as a hallucination of specificity, rejecting the relation. This behavior is particularly valuable for biocuration, where precise grounding is mandatory.

5. Discussion

5.1. Comparison with Supervised Baselines

The original BioRED benchmarks established by Luo et al. [4] achieve significantly higher performance (F1 $\approx$ 70%) using PubMedBERT models fine-tuned on the full training set. In contrast, our approach relies on *in-context learning* with a single example and no parameter updates.

- **Data Efficiency**. Our method requires 1 labelled example, whereas BioRED baselines require 500+ fully annotated documents.
- **Trade-off**. The gap in F1 score (30.7% vs 70%) highlights that off-the-shelf LLMs are not yet direct replacements for supervised extractors in high-precision domains, even with critic-based refinement.

5.2. Why Does the Critic Work?

Our results suggest that the critic-based approach is effective because:

- **Dual perspective**. The extraction phase is optimized for recall ("find all potential relations"), while the critic focuses on precision ("verify each claim").
- **Explicit negation detection**. The verification prompt specifically instructs the model to check for negation and speculation, addressing common biomedical RE errors.
- **Lower temperature for verification**. Using temperature=0.0 for critic promotes conservative, deterministic judgments.

5.3. Limitations

Computational cost. The two-phase approach requires approximately 2× inference time compared to single-pass extraction.

Precision Improvements. Our results show that specific prompting strategies can significantly lift the "Precision Floor". By providing a single few-shot example, we improved alignment with the BioRED schema, nearly doubling precision even before critic verification. This suggests that much of the error budget in zero-shot LLM extraction comes from definition mismatch rather than hallucination.

Dataset scale. Our evaluation on 75 documents provides statistically significant evidence of the false positive reduction capability.

6. Conclusion and Future Work

We introduced a critic-based prompting strategy for biomedical relation extraction that effectively mitigates the hallucination problem inherent in LLMs. By adding a verification layer, we achieved a 58.5% reduction in false positives and state-of-the-art precision of 74.1% with frontier models. Our analysis highlights that while extended reasoning capabilities can significantly improve initial extraction quality, the critic mechanism provides crucial robustness, particularly for mid-tier models where it serves as a powerful performance equalizer. This work establishes multi-step prompting as an essential quality control mechanism for reliable knowledge extraction.

Future directions include: (1) incorporating entity normalization/linking [14] into the prompts, (2) testing on larger benchmark datasets, (3) exploring multi-critic ensembles where different prompts vote on extraction validity, and (4) investigating active learning strategies to iteratively refine critic prompts based on error patterns.

Declaration on Generative AI

During the preparation of this work, the author(s) used Gemini 3 in order to: grammar and spelling check.

References

[1] Y. Zhang, Y. Li, L. Cui, D. Cai, L. Liu, T. Fu, X. Huang, E. Zhao, Y. Zhang, Y. Chen, L. Wang, A. T. Luu, W. Bi, F. Shi, S. Shi, Siren's song in the ai ocean: A survey on hallucination in large language models, ArXiv abs/2309.01219 (2023).

[2] L. Huang, W. Yu, W. Ma, W. Zhong, Z. Feng, H. Wang, Q. Chen, W. Peng, X. Feng, B. Qin, T. Liu, A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions, ACM Trans. Inf. Syst. 43 (2025).

[3] W. Dhifli, A. B. Diallo, PGR: a novel graph repository of protein 3d-structures, Journal of Data Mining in Genomics & Proteomics 6 (2015) 1.

[4] L. Luo, P.-T. Lai, C.-H. Wei, C. N. Arighi, Z. Lu, Biored: a rich biomedical relation extraction dataset, Briefings in Bioinformatics 23 (2022) bbac282.

[5] J. Chen, B. Hu, W. Peng, Q. Chen, B. Tang, Biomedical relation extraction via knowledge-enhanced reading comprehension, BMC bioinformatics 23 (2022) 20.

[6] M. Agrawal, S. Hegselmann, H. Lang, Y. Kim, D. Sontag, Large language models are few-shot clinical information extractors, in: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 2022, pp. 1998–2022.

[7] X. Wang, J. Wei, D. Schuurmans, Q. V. Le, E. H. Chi, S. Narang, A. Chowdhery, D. Zhou, Self-consistency improves chain of thought reasoning in language models, in: The Eleventh International Conference on Learning Representations, Kigali, Rwanda, 2023, pp. 1–20.

[8] N. Shinn, F. Cassano, A. Gopinath, K. Narasimhan, S. Yao, Reflexion: language agents with verbal reinforcement learning, in: Proceedings of the 37th International Conference on Neural Information Processing Systems, NeuRIPS '23, Curran Associates Inc., Red Hook, NY, USA, 2023.

[9] A. Madaan, N. Tandon, P. Gupta, S. Hallinan, L. Gao, S. Wiegreffe, U. Alon, N. Dziri, S. Prabhumoye, Y. Yang, S. Gupta, B. P. Majumder, K. Hermann, S. Welleck, A. Yazdanbakhsh, P. Clark, Self-refine: iterative refinement with self-feedback, in: Proceedings of the 37th International Conference on Neural Information Processing Systems, NeuRIPS '23, Curran Associates Inc., Red Hook, NY, USA, 2023.

[10] S. Wadhwa, S. Amir, B. C. Wallace, Revisiting relation extraction in the era of large language models, Proceedings of the conference. Association for Computational Linguistics. Meeting 2023 (2023) 15566–15589.

[11] DeepSeek-AI, D. Guo, D. Yang, H. Zhang, J. Song, R. Zhang, R. Xu, Q. Zhu, S. Ma, P. Wang, X. Bi, X. Zhang, X. Yu, Y. Wu, Z. Wu, Z. Gou, Z. Shao, Z. Li, Z. Gao, Z. Zhang, Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, arXiv preprint arXiv:2501.12948, 2025.

[12] U. Halil, J. Huang, D. Graux, J. Z. Pan, Llm shots: Best fired at system or user prompts?, in: Companion Proceedings of the ACM on Web Conference 2025, 2025, pp. 1605–1613.

[13] T. K. Saha, M. Ouzzani, H. M. Hammady, A. K. Elmagarmid, W. Dhifli, M. A. Hasan, A large scale study of svm based methods for abstract screening in systematic reviews, 2016. arXiv:1610.00192.

[14] A. Assi, W. Dhifli, Instance matching in knowledge graphs through random walks and semantics, Future Generation Computer Systems 123 (2021) 73–84.