AUT Journal of Modeling and Simulation

AUT J. Model. Simul., 58(2) (2026) 175-188
DOI: 10.22060/miscj.2025.24382.5421

Prompt Engineering for Biomedical NLI: An Exploratory Study

Hasan Fadhil Qasim Alkhawaf[1]*, Heshaam Faili[2]

[1] Alborz Campus, University of Tehran, Tehran, Iran
[2] College of Engineering, School of Electrical and Computer Engineering, University of Tehran, Tehran, Iran.

ABSTRACT: Biomedical Natural Language Inference (BioNLI) is a core task in biomedical NLP that looks to identify whether the biomedical premise entails the hypothesis. Prompt-based methods are gaining traction as one of the simplest and fastest approaches for effectively using large language models (LLMs) for these sorts of inference tasks without the need for complex, time-consuming fine-tuning of the model. However, the inherent difficulty of biomedical NLI problems represents a significant challenge for prompt engineering given the heavy use of terminology specific to the biomedical domain. During in-context prompting or using predefined examples for zero-shot or few-shot prompting, there is often a lack of contextual information or generalizability to the heterogeneity found in biomedical texts for determining entailment decisions. In this work, we present a comprehensive evaluation of a variety of prompting methods (zero-shot, few-shot static, few-shot dynamic, Chain-of-Thought, self-consistency, and Tree-of-Thought) with two LLMs, DeepSeek-R1-Distill-Qwen-14B and LLaMA-3.1-8B-Instruct, from the prompt-engineering perspective. We applied these methods to the BioNLI dataset and reported on key evaluation metrics across all methods. Our results show that dynamic contextual in-context prompting, together with structured reasoning, produces high-quality inference in our context. Between all of the models and configurations, few-shot ToT prompting using the DeepSeek model produced the best results, scoring a macro-F1 score of 71.05, even outperforming retrieval-augmented models reported in prior studies. These findings show that prompt engineering alone can handle complex biomedical reasoning effectively, without needing retrieval or full fine-tuning.

Review History:

Received: Jul.11, 2025
Revised: Nov. 09, 2025
Accepted: Dec. 07, 2025
Available Online: Sep. 15, 2026

Keywords:

Biomedical NLI
Natural Language Inference
Prompt Engineering
Chain-of-Thought
Tree-of-Thought

1- Introduction

Prompt-based learning has rapidly become a flexible and powerful approach to tapping pre-trained Large Language Models (LLMs) in a variety of Natural Language Processing (NLP) tasks with little or no fine-tuning. [1]. Given task-specific prompts and a few labeled examples, LLMs can complete a wide range of tasks in a few-shot or zero-shot mode, employing only their pre-trained knowledge and natural language instructions. This has been particularly successful across Natural Language Inference (NLI), question answering, and classification [2, 3].

Biomedical Natural Language Inference (BioNLI) differs significantly from general NLI, given its specialized and safety-critical nature. Biomedical texts include dense clinical vocabulary, domain-specific abbreviations, implicit modalities of causality, and complicated negation, all of which require advanced reasoning skills beyond those needed for commonsense language understanding [4].

To demonstrate the BioNLI task, an example is given in Figure 1. Each instance has a premise, a biomedical statement taken from the literature, and a hypothesis, which denotes a possible interpretation or conclusion. The model must determine whether the premise adequately supports the hypothesis or if it is contradicted by it. For example, the premise "Increased serum ALT levels indicate liver injury" and the hypothesis "Elevated ALT is a marker of hepatic damage" have the label Entailment, since they indicate the same underlying biological relationship.

Unlike general NLI datasets such as SNLI or MNLI, biomedical benchmarks demand careful interpretation of clinical evidence, as the effects of misinterpretation can lead to dangerous conclusions in real-world clinical environments [5]. Prompt-based approaches have improved recently, and adapting them for BioNLI is a critical research question. Biomedical literature is inherently complex, including domain-specific vocabulary, implicit modalities of causality, and complex reasoning, which may require multi-step inferences [6]. Without access to external context or structured reasoning, generic prompting may not yield accurate entailment decisions. For example, zero-shot prompting does not always have a contextual signal for confident inference, and static few-shot prompting does not always generalize to

*Corresponding author's email: hasan.fadhil@uomisan.edu.iq

Copyrights for this article are retained by the author(s) with publishing rights granted to Amirkabir University Press. The content of this article is subject to the terms and conditions of the Creative Commons Attribution 4.0 International (CC-BY-NC 4.0) License. For more information, please visit https://www.creativecommons.org/licenses/by-nc/4.0/legalcode.

UNCORRECTED PROOF

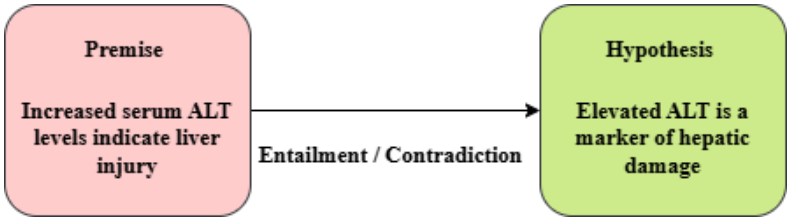


Fig. 1. An illustrative example of the BioNLI task showing the relationship between premise, hypothesis, and inference label.

UNCORRECTED PROOF

a wide range of biomedical claims [7]. This will mean that BioNLI is more difficult than general NLI, and demands models that have domain-aware reasoning capabilities and can robustly adapt to clinical semantics.

Recent approaches, such as Chain-of-Thought (CoT) [8] and Tree-of-Thought (ToT) prompting [9]have provided structured reasoning into the prompting process. These methods have encouraged the model to logically decompose step-by-step or to try out different reasoning paths before giving the final answer and hence can achieve more robust performance on tasks where complex inference is involved.

In this paper, we perform an extensive comparison of prompt engineering approaches for biomedical NLI with two fine-tuned LLMs: DeepSeek-R1-Distill-Qwen-14B and LLaMA-3. 1-8B-Instruct. We investigate a wide range of prompting approaches (including zero-shot, static few-shot, dynamic few-shot, CoT, self-consistency prompting [10] and ToT) to examine the impact of diverse prompting methods on the BioNLI dataset.

The remaining sections in this paper are organized as follows: Section 2 reviews related work on prompt engineering and biomedical NLI. Section 3 describes the dataset and models, and the prompting schemes utilized. The experimental results are presented in Section 4. Section 5 discusses comparisons with retrieval-based benchmarks. Section 6 concludes the paper and proposes future work.

2- Related works

This section reviews prior studies that have employed prompt engineering as a core strategy in their investigation of NLI tasks, particularly within the biomedical domain. Jiang et al. [11] devised a benchmarking framework, denoted Bio-benchmark, to address existing benchmarks for creating LLMs related to tasks in bioinformatics. The tasks included 30 tasks, such as protein structure prediction, drug interactions, etc. The model consists of the six most common LLMs that are used in experiments in zero-shot and few-shot scenarios without fine-tuning. They created BioFinder for the enhancement of answer extraction and validation through the usage of NLP and NLI techniques. Results showed satisfactory LLM performance in many tasks, particularly in few-shot settings, while highlighting areas needing improvement, such as RNA structure and drug-drug interaction predictions.

Rohanian et al. [12] study the inadequacies of general-purpose LLMs and how they relate to the field of bio-NLP with applications to Named Entity Recognition (NER), Relation Extraction (RE), and NLI, for example. They find that while models like BioBERT perform well, generative LLMs struggle with structured outputs. To improve performance, they suggest instruction tuning models and created a new dataset, named Llama2-MedTuned-Instructions, consisting of ~200,000 samples from different biomedical sources available presently. When the Llama2 models were tuned on this dataset, Llama2-MedTuned was obtained, which showed much improved behaviour in zero-shot performance, particularly in the case of NLI, and also competitive results in NER and RE benchmarks. This indicates that instruction-tuned LLMs could be a promising approach for research in bio-NLP.

Heston and Khun [13] have investigated some of the risks and benefits of Global Language Models (GLM) in medical education. The authors emphasized the importance of prompt engineering to obtain accurate and high-quality responses. They also raised their concerns about misinformation, biases, ethical issues, and over-reliance on technology. The authors categorized the prompts into three classes: zero-shot, few-shot, and structured prompts, and discussed their importance in medical education and the need for repetitive prompting to enhance the quality and minimize the imprecision of their output.

Câmpeanu et al. [1] presented a methodology for addressing the NLI4CT task using the SOLAR 10B instruction-tuned model, without any fine-tuning of the model. They focused on some prompting and summarizing specific parts within the CTR. They used zero-shot and few-shot prompting adaptations to tailor the prompts for the parts of the report in concern, such as eligibility, intervention, results, and adverse events, etc. They used the NLI4CT, with 1,000 breast cancer trial reports and over 24,000 entailment relation annotations. Their overall approach received a consistency score of 0.72, and they ranked 14th in the SemEval-2024 Task 2 leaderboard.

Mairesse and Verlingue [14] also participated in the SemEval-2024 Task 2, which dealt with NLI using pairs of statements and CTR sections, from the clinicaltrials.gov dataset. They looked at prompt optimization strategies with

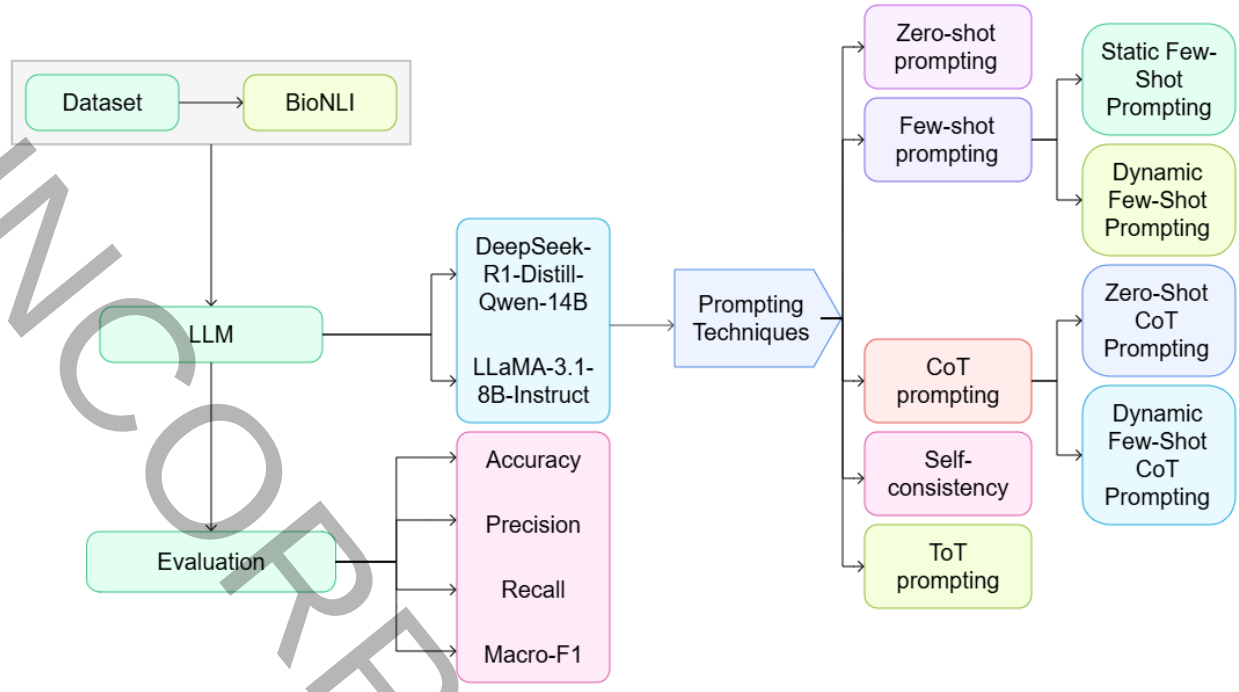


Fig. 2. Five prompting strategies used for BioNLI.

LLMs, including zero-shot CoT, dynamic one-shot CoT, and OPRO (Optimization by Prompting). Of their approaches, the zero-shot CoT produced the highest performance with an F1 score of 0.70 for case number 1 in the test set.

3- Methodology

We leveraged prompt engineering methods to investigate the reasoning capabilities of LLMs employed on a biomedical NLI task, motivated by the task of determining whether the premise entails (Entailment) or contradicts (Contradiction) the hypothesis. We employed the BioNLI dataset and two pre-trained large language models, DeepSeek-R1-Distill-Qwen-14B [15] and LLaMA-3.1-8B-Instruct [16]. We designed five prompting methods on the BioNLI task: zero-shot, static few-shot, dynamic few-shot, CoT, self-consistency, and ToT (Figure 2). Each method incorporated examples of the task and examples of reasoning, hoping to lead to better performance from the LLMs. We evaluate each of these methods using standard metrics, accuracy, precision, recall, and macro-F1, and compare our results with existing benchmarks in the literature and state-of-the-art methods.

3- 1- Dataset

For this study, we used the BioNLI dataset [17], a benchmark for biomedical NLI, which is the first in this area to present sentence pairs from biomedical literature that include mechanistic premises from nine different methods of generation. BioNLI is publicly available [17] and was split into training and test sets; descriptive statistics can be seen in Table 1. This dataset contains two classes: ‘Entailment’ and ‘Contradiction’. There are 2,790 and 2,754 samples for each class in the train set. In the test set, 5,084 and 1,224 samples are included for each class. The dataset has a total of 11,852 samples, 7,874 of which are in the class ‘Entailment’ and 3,978 are in ‘Contradiction’. BioNLI provides a publicly available means of assessing model performance in sophisticated biomedical scenarios. Each instance in the dataset has a premise (a scientific claim or observation) and a hypothesis (a conclusion or interpretation) for which the task is to determine whether the hypothesis is logically supported by the premise. In this way, each pair is labeled as either:

Entailment: The hypothesis logically follows from the premise.

Contradiction: The hypothesis is contradictory or not supported by the premise.

3- 2- Zero-Shot Prompting

Zero-shot prompting will assess how well an LLM model can perform a task without seeing any examples at all [18]. Thus, we created a basic prompt with only task instructions and a test pair (premise and hypothesis) to assess the inherent NLI capability of LLMs in the biomedical domain as a baseline. The prompt template used is shown in Figure 2. Our goal was to see the extent to which a pre-trained model was able to generalize to the NLI task without any previous exposure to biomedical sentence structure.

UNCORRECTED PROOF

Table 1. Data distribution in the BioNLI dataset.

Class	Train	Test	Total
Entailment	2,790	5,084	7,874
Contradiction	2,754	1,224	3,978
Total	5,544	6,308	11,852

UNCORRECTED PROOF

```
This is a Natural Language Inference task in the biomedical domain. Identify whether a given premise and hypothesis hold an entailment or contradiction relationship. It is an `Entailment` if the premise claim supports the hypothesis; otherwise, it is a `Contradiction`.

Instruction: Return only one word: 'Entailment' or 'Contradiction'.

Premise: `{premise}`

Hypothesis: `{hypothesis}`
```

Fig. 3. The structured template for the zero-shot prompting.

3- 3- Few-Shot Prompting

After the zero-shot evaluation, we were interested in seeing if performance could be improved by adding exemplars, which is called few-shot prompting. This research aimed to examine how LLMs respond to demonstrations and whether guided prompting would improve accuracy. At this stage, we evaluated two changing few-shot prompting techniques: static few-shot prompting and dynamic few-shot prompting.

Static Few-Shot Prompting: In this approach, we manually selected four labeled examples (two entailments and two contradictions) from the training set, as shown in Appendix A. These instances were integrated into the prompts and remained invariant across all test instances. This approach provides the model with representative patterns, allowing for improvement with minimal effort.

Dynamic Few-Shot Prompting: To improve the predictions, we dynamically retrieve the most relevant examples from the training set for each test instance based on semantic similarity that has been calculated using the Sentence-BERT model [19]. This will ensure that the model sees contextually relevant demonstrations for every case of inference. This prompt and its relevant examples go beyond what is done with static prompting to see if the relevant aspects of the prompting affect the prompt itself. This prompting goes beyond traditional static prompts to investigate whether the relevance of information in the prompt can affect its usefulness. The dynamic few-shot prompting template used is shown in Appendix B.

3- 4- CoT Prompting

Motivated by the fact that model performance significantly benefits from examples, notably contextually relevant ones, we enrich our analysis with reasoning traces via Chain-of-Thought (CoT) prompts. CoT prompting enables advanced reasoning abilities through intermediate reasoning steps [8, 20]. CoT can be combined with zero-shot and few-shot prompting to get better results on more complex tasks that require reasoning before responding. Also, the hypothesis can affirm or refute the premises. In this study, as it was conducted in the medical space, not only must there be reasoning clearly stated for the task, but the model can parse the clinical statements and understand key clinical entities to ensure there is a logical relationship between the premise (entail) or not (contradict).

3- 4- 1- Zero-Shot CoT Prompting

We started with zero-shot CoT Prompting, in which the model was provided with reasoning instructions (without any examples) and was asked to think through the relationship between premise and hypothesis before answering. The

UNCORRECTED PROOF

```
This is a Natural Language Inference task in the biomedical domain. Identify whether a given premise and hypothesis hold an entailment or contradiction relationship. It is an `Entailment` if the premise claim supports the hypothesis; otherwise, it is a `Contradiction`.

Instruction: Think step by step. Return only one word: 'Entailment' or 'Contradiction'.

Reasoning: Let's carefully analyze this step by step before answering.

1. Identify the main clinical statement or claim in the premise.

2. Identify the main clinical statement or claim in the hypothesis.

3. Highlight key clinical entities (e.g., diseases, treatments, lab values) mentioned in both.

4. Compare whether the clinical facts in the hypothesis are present, implied, or contradicted by the premise.

5. Consider the context, such as temporality, causality, or negation, and whether it affects consistency between statements.

6. Based on this analysis, determine the logical relationship: does the premise entail or contradict the hypothesis?

Premise: `{premise}`

Hypothesis: `{hypothesis}`

Answer:
```

Fig. 4. The structured template for the zero-shot CoT prompting.

structured template for this stage is illustrated in Figure 3. This condition was used to assess LLMs' ability to perform capable reasoning in a biomedical context, with no demonstrations.

3- 4- 2- Dynamic Few-Shot CoT Prompting

Acknowledging the limitations of zero-shot CoT, we further conducted an experiment based on the combination of examples and reasoning paths. This was followed by the few-shot CoT prompting phase. We followed the previous findings that dynamic selection performs better and employed a dynamic few-shot CoT strategy, as depicted in Appendix C. In this configuration, the prompt consisted of semantically related instances, along with detailed reasoning. The idea was to provide the model with explicit reasoning demonstrations tailored to the input instance.

3- 5- Self-Consistency

Sometimes, LLMs produce different outputs for the same input due to the randomness inherent in the model architecture involved in generating language. To mitigate the variability and obtain better predictions, we used a self-consistency prompting strategy. In this method, the model is prompted with the same input multiple times, generating a different reasoning path each time, and the final answer is selected by majority vote. To enhance the probability of diverse reasoning paths, we set the temperature to 0.8. The temperature parameter regulates how stochastic the next prediction will be in the model output: the closer to one the temperature, the more diverse the next sampling will be; the lower the temperature, the more deterministic the output will be. The use of a moderately high temperature along with sampling allowed the model to create next-generation reasoning chains before responding. The use of self-consistency prompting allows us to leverage the previous Dynamic Few-Short Chain-of-Thought prompting setup, in which we think about each reasoning sequence of example-driven logical steps. Thus, self-consistency is preserved and improves both the robustness and accuracy of the model, particularly on more demanding tasks requiring complex multi-step inference.

3- 6- Tree of Thought

Tree-of-Thought (ToT) prompting is a sophisticated reasoning method typically utilized to simulate higher forms of decision-making through multiple parallel reasoning branches. Unlike simpler linear prompting methods, ToT provides additional structure to the exploration of parallel thinking across different logical perspectives, such as semantic, relational, and implicit knowledge in clinical entities [21]. This is a better simulation of how clinicians look at evidence, how they take different perspectives, and perform an evaluation before making a decision. We used Few-Shot ToT prompting in this study, combining few-shot learning with CoT reasoning, as the tree branching structure of ToT provides an environment for the model to perform reasoning in multiple distinct paths (e.g., A, B, C for each case) using different reasoning approaches to see how each pathway evaluated the relationship between the premise and hypothesis. At the end of the entire reasoning process across the branches, the model produced a final output through majority voting of the conclusions of each branch. We were primarily motivated to explore the feasibility of ToT in evidence synthesis after a

few limitations in conventional CoT prompting, particularly in edge cases when single-path reasoning fails. Our objective was to examine whether ToT could improve the consistency, interpretability, and accuracy of the model's predictions of complex biomedical inferences. ToT provided opportunities for deeper-level analysis and improved robustness in cases that required reasoning beyond simple, surface-level pattern matching. The structured template for this stage is shown in Appendix D. ToT expands upon reasoning by chains to allow the model to explore multiple reasoning paths in parallel rather than conforming to a single inferential progression. This format is especially beneficial in biomedical NLI, where the clinical logic often demands simultaneous consideration of alternative diagnostic hypotheses, verification of semantic and causal coherence, and integration of implicit biomedical knowledge before concluding. By requiring the model to branch and compare reasoning paths, ToT is more in accord with human-inspired medical decision-making and decreases the likelihood of premature and/or biased conclusions in complex inference problems [9, 22].

4- Results

The outcomes of the various prompt engineering methods that were tested using the LLMs on the BioNLI dataset are shown in Table 2, using four traditional metrics (accuracy, precision, recall, and Macro-F1). Concerning the prompt-only approaches, the best Macro-F1 score (0.5936) was observed in Dynamic few-shot Prompting with DeepSeek, which is better than both Static few-shot (0.5255) and Zero-Shot prompting (0.5200). These results indicate that the use of contextual examples, which are dynamically retrieved using semantic matching, promotes model performance in biomedical NLI tasks. The Static examples afforded some level of benefit, but there is little ability to consider the heterogeneous nature of biomedical text. With the addition of reasoning via few-shot CoT prompting, considerable improvements were seen, with DeepSeek returning a Macro-F1 of 0.6124. The use of Self-Consistency CoT, where multiple reasoning paths were created and the majority voting technique was used to create a final output, improved this to 0.6361. In addition, and combined with the higher performance scores, these results also support the view that performance on inference biomedical tasks may benefit from multi-step reasoning and 'aggregation' of model outputs that help reduce uncertainty. The Zero-shot CoT prompting produced the highest error (Macro-F1 of 0.3835 for DeepSeek), which suggested to us that an attempt to perform reasoning with no contextual or demonstrational examples created disconnected and incoherent reasoning chains. The important part to stress here is that CoT prompting alone is ineffective in complex domains such as biomedicine, unless it is associated with examples or domain adaptation circumstances.

The strongest results across the board were with few-shot ToT prompting, where DeepSeek achieved a Macro-F1 of 0.7105, and LLaMA achieved 0.5796. This structured multi-path reasoning approach was very effective when making decisions about complex entailments and suggests a reasonable likelihood of usefulness as a clinical reasoning simulation. Across all prompting configurations, DeepSeek-R1-Distill-Qwen-14B outperformed LLaMA-3.1-8B-Instruct. We attribute DeepSeek's better performance to its domain-oriented pretraining and distillation processes, which we expect to support better biomedical language grounding. There was also a very large performance difference between the two models in identical prompting conditions, which we expect was partly attributable to the larger random seed-derived differences concerning their architecture and training.

5- Discussion

The experimental results reported in this study provide evidence supporting the efficacy of prompt engineering strategies in improving the performance of LLMs on the BioNLI task. The method producing the best macro-F1 score of 0.7105 using the DeepSeek-R1-Distill-Qwen-14B model was few-shot ToT prompting, surpassing all other prompting configurations, including zero-shot and standard few-shot techniques.

We contrasted our results against those from Li et al.'s study [23], which introduced BioRAB, the most robust biomedical benchmark that enables the analysis of Retrieval-Augmented Language models (RALs) on a variety of biomedical NLP tasks, including BioNLI. Li et al.'s best-performing setup, the LLaMA2-13B with the MedCPT retriever, achieved a Macro-F1 score of 69.21, outperforming the base LLM (62.62). Our best-performing prompt-only setup (DeepSeek + few-shot ToT) surpassed the retrieval-augmented model from Li et al.'s study in Macro-F1 score by +1.34 (71.05 vs 69.21). This was a significant finding, particularly given that the prompt engineering leveraged no external knowledge retrieval, only structured design of prompts and reasoning modalities. This suggests that, given the proper prompting circumstances, state-of-the-art LLMs can match or exceed the retrieval-augmented mode performance, at least for BioNLI. Taken together, our results demonstrate that prompt engineering, especially for reasoning-based types like CoT and ToT, can allow for similar or complementary approaches to retrieval-based approaches, particularly when it can be established that the domain-specific knowledge is already well-encoded into the LLM. Table 3 provides a direct comparison between our best-performing prompt-based approach and the retrieval-augmented configurations evaluated by Li et al. on the BioNLI dataset.

The performance gains reported may appear modest in absolute numerical terms, but our method can achieve this gain without any fine-tuning whatsoever and by using prompt-based reasoning. By contrast, the comparison retrieval-augmented baselines require supervised task-specific training as well as domain-adapted retrievers. The reported Macro-F1 of 71.05% for our method exceeds the BioRAB baseline performance of 69.21% obtained with LLaMA2-13B + MedCPT retriever, and significantly exceeds that of 62.62% observed with the base LLaMA2-13B model. This indicates that structured prompting, in particular, ToT reasoning, is a viable and effective alternative to such parameter-intensive

UNCORRECTED PROOF

Table 2. Evaluation results of different prompt engineering strategies.

Prompting technique / Model	Accuracy	Precision	Recall	Macro-F1
Zero-shot				
DeepSeek-R1-Distill-Qwen-14B	0.5588	0.5722	0.6152	0.5200
Llama-3.1-8B-Instruct	0.3707	0.5103	0.5132	0.3666
Static Few-shot				
DeepSeek-R1-Distill-Qwen-14B	0.5713	0.5680	0.6093	0.5255
Llama-3.1-8B-Instruct	0.4692	0.5543	0.5807	0.4528
Dynamic Few-shot				
DeepSeek-R1-Distill-Qwen-14B	0.7355	0.5936	0.6027	0.5936
Llama-3.1-8B-Instruct	0.6150	0.5385	0.5574	0.5250
Zero-shot CoT				
DeepSeek-R1-Distill-Qwen-14B	0.3910	0.5126	0.5171	0.3835
Llama-3.1-8B-Instruct	0.3450	0.5161	0.5182	0.3450
Few-shot CoT				
DeepSeek-R1-Distill-Qwen-14B	0.6574	0.6368	0.7176	0.6124
Llama-3.1-8B-Instruct	0.6480	0.5150	0.5192	0.5118
Self-Consistency CoT				
DeepSeek-R1-Distill-Qwen-14B	0.6794	0.6565	0.7487	0.6361
Llama-3.1-8B-Instruct	0.6746	0.5231	0.5271	0.5231
Few-Shot ToT				
DeepSeek-R1-Distill-Qwen-14B	**0.7829**	**0.6932**	**0.7623**	**0.7105**
Llama-3.1-8B-Instruct	0.6304	0.6056	0.6677	0.5796

supervised pipelines, with competitive performance, but clearly requiring much less compute and data.

6- Conclusion

In this paper, we examined several prompt engineering approaches to improve LLM performance on the BioNLI task. Our experiments showed the importance of a variety of different prompting techniques in obtaining an effective model, particularly in complex reasoning tasks. Among all methods explored in this work, the Few-Shot ToT prompting type provided the best performance, with a DeepSeek-R1-Distill-Qwen-14B model reaching a performance of 71.05 Macro-F1. This level of performance was not only superior to basic prompting techniques, but also the best

UNCORRECTED PROOF

Table 3. Comparison of BioNLI performance between our best prompt-based approach and the RALs model [26].

Study	Method Type	Model Configuration	Macro-F1
This Study (Ours)	Prompt Engineering (ToT)	DeepSeek-R1-Distill-Qwen-14B + Few-Shot ToT	**71.05**
Li et al. [23]	Retrieval-Augmented (RAL)	LLaMA2-13B + MedCPT Retriever	69.21
Li et al. [23]	Base LLM (No Retrieval)	LLaMA2-13B	62.62

retrieval-augmented configuration of all previous work from newly released literature [23], relying purely upon prompt engineering and internal reasoning only. Our analyses demonstrate the efficacy of structured prompting and reasoning-based methodologies as useful and efficient alternatives to retrieval-augmented means when access to external knowledge is tenuous or non-existent. Additionally, DeepSeek consistently outperformed LLaMA-3.1 across all comparisons, suggesting that the distillation method and pretraining on domain specificity, specifically for biomedical-related inference tasks, are beneficial and important. For future work, we propose parameter-efficient adaptation by testing the limits of prompt tuning, LoRA, or a combination of both.

References

[1] M. Micluta-Campeanu, C. Creanga, A.-M. Bucur, A.S. Uban, L.P. Dinu, UniBuc at SemEval-2024 Task 2: Tailored Prompting with Solar for Clinical NLI, in: Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024), 2024, pp. 586-595.

[2] T. Kojima, S.S. Gu, M. Reid, Y. Matsuo, Y. Iwasawa, Large language models are zero-shot reasoners, Advances in neural information processing systems, 35 (2022) 22199-22213.

[3] Z. Al Nazi, M.R. Hossain, F. Al Mamun, Evaluation of open and closed-source LLMs for low-resource language with zero-shot, few-shot, and chain-of-thought prompting, Natural Language Processing Journal, 10 (2025) 100124.

[4] A. Seerat, S. Nasir, M. Wasim, N.M. Garcia, Biomedical Natural Language Inference on clinical trials using the BERT-based Models, Procedia Computer Science, 241 (2024) 576-581.

[5] M. Jullien, M. Valentino, A. Freitas, SemEval-2024 Task 2: Safe Biomedical Natural Language Inference for Clinical Trials, in: Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024), 2024, pp. 1947-1962.

[6] H. Wang, C. Liu, N. Xi, S. Zhao, M. Ju, S. Zhang, Z. Zhang, Y. Zheng, B. Qin, T. Liu, Prompt Combines Paraphrase: Teaching Pre-trained Models to Understand Rare Biomedical Words, in: Proceedings of the 29th International Conference on Computational Linguistics, 2022, pp. 1422-1431.

[7] T. Korbak, K. Shi, A. Chen, R.V. Bhalerao, C. Buckley, J. Phang, S.R. Bowman, E. Perez, Pretraining language models with human preferences, in: International Conference on Machine Learning, PMLR, 2023, pp. 17506-17533.

[8] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q.V. Le, D. Zhou, Chain-of-thought prompting elicits reasoning in large language models, Advances in neural information processing systems, 35 (2022) 24824-24837.

[9] S. Yao, D. Yu, J. Zhao, I. Shafran, T. Griffiths, Y. Cao, K. Narasimhan, Tree of thoughts: Deliberate problem solving with large language models, Advances in neural information processing systems, 36 (2023) 11809-11822.

[10] 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.

[11] J. Jiang, P. Chen, J. Wang, D. He, Z. Wei, L. Hong, L. Zong, S. Wang, Q. Yu, Z. Ma, Benchmarking Large Language Models on Multiple Tasks in Bioinformatics NLP with Prompting, CoRR, (2025).

[12] O. Rohanian, M. Nouriborji, S. Kouchaki, F. Nooralahzadeh, L. Clifton, D.A. Clifton, Exploring the effectiveness of instruction tuning in biomedical language processing, Artificial Intelligence in Medicine, 158 (2024) 103007.

[13] T.F. Heston, C. Khun, Prompt Engineering in Medical

UNCORRECTED PROOF

Education, International Medical Education, 2(3) (2023) 198-205.

[14] C. Brutti-mairesse, L. Verlingue, CRCL at SemEval-2024 Task 2: Simple prompt optimizations, in: Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024), 2024, pp. 437-442.

[15] DeepSeek-AI, DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning, (2025).

[16] Meta, LLaMA-3.1-8B-Instruct, (2024).

[17] M. Bastan, M. Surdeanu, N. Balasubramanian, BioNLI: Generating a Biomedical NLI Dataset Using Lexico-semantic Constraints for Adversarial Examples, in: Findings of the Association for Computational Linguistics: EMNLP 2022, 2022, pp. 5093-5104.

[18] Y. Li, A Practical Survey on Zero-Shot Prompt Design for In-Context Learning, in: Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing, 2023, pp. 641-647.

[19] A.R. (FAIR), all-roberta-large-v1, (2023).

[20] J. Miao, C. Thongprayoon, S. Suppadungsuk, P. Krisanapan, Y. Radhakrishnan, W. Cheungpasitporn, Chain-of-thought utilization in large language models and application in nephrology, Medicina, 60(1) (2024) 148.

[21] K. Dadgar, Generating collaborative teaching case studies using GenAI: Tree of Thoughts prompt engineering, (2025).

[22] S. Chen, B. Li, D. Niu, Boosting of Thoughts: Trial-and-Error Problem Solving with Large Language Models, in: The Twelfth International Conference on Learning Representations.

[23] M. Li, Z. Zhan, H. Yang, Y. Xiao, J. Huang, R. Zhang, Benchmarking retrieval-augmented large language models in biomedical NLP: Application, robustness, and self-awareness, arXiv preprint arXiv:2405.08151, (2024).

HOW TO CITE THIS ARTICLE
H. F. Q. Alkhawaf, H. Faili, Prompt Engineering for Biomedical NLI: An Exploratory Study, AUT J. Model. Simul., 58(2) (2026) 149-162.

DOI: 10.22060/miscj.2025.24382.5421

UNCORRECTED PROOF

Appendix A: The structured template for the static few-shot prompting.

This is a Natural Language Inference task in the biomedical domain. Identify whether a given premise and hypothesis hold an entailment or contradiction relationship. It is an `Entailment` if the premise claim supports the hypothesis; otherwise, it is a `Contradiction`.

Instruction: Return only one word: 'Entailment' or 'Contradiction'.

Premise: `Class IA phosphoinositide 3-kinase (PI3K) is a heterodimer composed of a p85 regulatory and a p110 catalytic subunit that regulates a variety of cell responses, including cell division and survival. PI3K is activated following Tyr kinase stimulation and by Ras. We found that the C-terminal region of p85, including the C-Src homology 2 (C-SH2) domain and part of the inter-SH2 region, protects the p110 catalytic subunit from Ras-induced activation. Although the p110 activity associated with a C-terminal p85 deletion mutant increased significantly in the presence of an active form of Ras, purified wild type p85-p110 was only slightly stimulated by active Ras. Nonetheless, incubation of purified p85-p110 with Tyr-phosphorylated peptides, which mimic the activated platelet-derived growth factor receptor, restored Ras-induced p85-p110 activation. `

Hypothesis: `In conclusion, p85 inhibits p110 activation by Ras ; this blockage is released by Tyr kinase stimulation, showing that the classical mechanism of class IA PI3K stimulation mediated by Tyr kinases also regulates Ras -induced PI3K activation.`

Answer: Entailment

Premise: `To determine the relation between left ventricular filling properties and plasma atrial natriuretic peptide in systemic sclerosis, we evaluated 30 consecutive patients and 48 age- and sex-matched controls. The venous plasma atrial natriuretic peptide was measured by radio-immunoassay. Left ventricular involvement was evaluated by echocardiography and mitral regurgitation was evaluated by Doppler. The patient group had markedly elevated plasma atrial natriuretic peptide as compared to the matched controls, (239.4 +/- 59 vs. 178.2 +/- 36 pmol/l, P < 0.0005). We found signs of impaired left ventricular filling properties among the patients with an increase of the Doppler A-wave velocity and A/E ratio. A relative reduction of early filling was found in spite of some degree of mitral regurgitation in two-thirds of the patients. The plasma atrial natriuretic peptide concentration was related to the A-wave velocity (r = 0.44, P < 0.0005), the A/E ratio (r = 0.40, P < 0.005), and also to the degree of mitral regurgitation (r = 0.43, P < 0.005). The relationship to the A-wave velocity remained when considering possible confounding factors.`

Hypothesis: `We conclude that systemic sclerosis leads to elevated plasma atrial natriuretic peptide .`

Answer: Contradiction

Premise: `A total of 151 galactose tolerance tests (GTT) and liver biopsies were performed in a consecutive study of 45 psoriatic patients with methotrexate-induced liver fibrosis; of these, 23 had cirrhosis. Most patients with an abnormal liver histology had a normal GTT.`

Hypothesis: `We conclude that an oral GTT is not sensitive enough to reveal methotrexate -induced liver fibrosis or cirrhosis.`

Answer: Entailment

Premise: `Blood glucose levels were measured over a 24-h period in eight insulin-dependent diabetic subjects who were difficult to control and who presented with morning fasting hyperglycemia. At least seven exhibited clinical characteristics suggestive of the Somogyi phenomenon. A continuous glucose monitoring apparatus was used to relate the concentrations of glucose during the day to concomitant levels of free insulin and cortisol. In all patients a significant (P < 0.01) rise in fasting morning glucose started at about 0600 h, while they were still asleep. In six patients the morning elevation of blood glucose was preceded by stable, almost normal glucose levels during the night (117 ± 2.5 mg/dl); one of the two remaining patients (no. 7) exhibited high overnight glucose levels (268 ± 7 . 2 mg/dl), whereas the other (no. 8) had a mild hypoglycemic episode (45 mg/dl) 6 h before the hyperglycemic period. In all patients the fasting glucose rise was associated with the usual morning cortisol surge (P < 0.05) and with a significant decrease in the concentration of serum free insulin (P < 0.01). The free insulin levels in patient no. 8 were higher, while those of patient no. 7 were lower than in the other six patients.`

Hypothesis: `We conclude that morning fasting hyperglycemia in insulin-dependent diabetic patients may be triggered by the usual morning cortisol surge.`

Answer: Contradiction

Premise: `{premise}`

Hypothesis: `{hypothesis}`

Answer:

UNCORRECTED PROOF

Appendix B: The structured template for the dynamic few-shot prompting.

This is a Natural Language Inference task in the biomedical domain. Identify whether a given premise and hypothesis hold an entailment or contradiction relationship. It is an `Entailment` if the premise claim supports the hypothesis; otherwise, it is a `Contradiction`.

Instruction: Return only one word: 'Entailment' or 'Contradiction'.

Premise: {entailment_example_1_premise}

Hypothesis: {entailment_example_1_hypothesis}

Answer: Entailment

Premise: {contradiction_example_1_premise}

Hypothesis: {contradiction_example_1_hypothesis}

Answer: Contradiction

Premise: {entailment_example_2_premise}

Hypothesis: {entailment_example_2_hypothesis}

Answer: Entailment

Premise: {contradiction_example_2_premise}

Hypothesis: {contradiction_example_2_hypothesis}

Answer: Contradiction

Premise: `{premise}`

Hypothesis: `{hypothesis}`

Answer:

UNCORRECTED PROOF

Appendix C: The structured template for the dynamic few-shot CoT prompting.

This is a Natural Language Inference task in the biomedical domain. Identify whether a given premise and hypothesis hold an entailment or contradiction relationship. It is an `Entailment` if the premise claim supports the hypothesis; otherwise, it is a `Contradiction`.

Instruction: Think step by step. Return only one word: 'Entailment' or 'Contradiction'.

Reasoning: Let's carefully analyze this step by step before answering.

1. Identify the main clinical statement or claim in the premise.

2. Identify the main clinical statement or claim in the hypothesis.

3. Highlight key clinical entities (e.g., diseases, treatments, lab values) mentioned in both.

4. Compare whether the clinical facts in the hypothesis are present, implied, or contradicted by the premise.

5. Consider the context, such as temporality, causality, or negation, and whether it affects consistency between statements.

6. Based on this analysis, determine the logical relationship: does the hypothesis entail or contradict the premise?

Premise: {entailment_example_1_premise}

Hypothesis: {entailment_example_1_hypothesis}

Answer: Entailment

Premise: {contradiction_example_1_premise}

Hypothesis: {contradiction_example_1_hypothesis}

Answer: Contradiction

Premise: {entailment_example_2_premise}

Hypothesis: {entailment_example_2_hypothesis}

Answer: Entailment

Premise: {contradiction_example_2_premise}

Hypothesis: {contradiction_example_2_hypothesis}

Answer: Contradiction

Premise: `{premise}`

Hypothesis: `{hypothesis}`

Answer:

UNCORRECTED PROOF

Appendix D: The structured template for the few-shot ToT prompting.

This is a Natural Language Inference task in the biomedical domain. Identify whether a given premise and hypothesis hold an entailment or contradiction relationship. It is an `Entailment` if the premise claim supports the hypothesis; otherwise, it is a `Contradiction`.

Instruction: Think through different possible reasoning paths. Explore multiple interpretations or logical branches and then make a decision based on the most consistent conclusion. Return only one word: 'Entailment' or 'Contradiction'.
Reasoning Tree: Let's explore multiple reasoning paths step by step.

Thought Path A:

1. Identify the main clinical statement or claim in the premise.

2. Identify the main clinical statement or claim in the hypothesis.

3. Do the facts in the hypothesis logically follow from or align with those in the premise?

4. If yes, does that indicate an entailment? If not, move to contradiction.

Conclusion A: [Entailment/Contradiction]

Thought Path B:

1. Highlight clinical entities (e.g., diseases, symptoms, labs) in both statements.

2. Analyze relationships among these entities (e.g., is a disease being treated or ruled out?).

3. Check for mismatches in negation, temporality, or causality.

4. Do any conflicts or absences lead to contradiction?

Conclusion B: [Entailment/Contradiction]

Thought Path C:

1. Consider alternative interpretations or edge cases (e.g., implicit knowledge, clinical reasoning).

2. Does the hypothesis stretch beyond what the premise justifies?

3. Could it be partially implied or completely opposed?

Conclusion C: [Entailment/Contradiction]

Final Decision:

Based on the majority agreement among Thought Paths A, B, and C, what is the most supported relation? If two or more paths agree, return that as the final answer.

Premise: {entailment_example_1_premise}
Hypothesis: {entailment_example_1_hypothesis}
Answer: Entailment

Premise: {contradiction_example_1_premise}
Hypothesis: {contradiction_example_1_hypothesis}
Answer: Contradiction

Premise: {entailment_example_2_premise}
Hypothesis: {entailment_example_2_hypothesis}
Answer: Entailment

Premise: {contradiction_example_2_premise}
Hypothesis: {contradiction_example_2_hypothesis}
Answer: Contradiction

Premise: `{premise}`
Hypothesis: `{hypothesis}`
Answer:

UNCORRECTED PROOF

UNCORRECTED PROOF