Enhancing Clinical Trial Analysis through Large Language Models for Multi-Evidence Natural Language Inference

Shobanapriyan Chandrasegaran, Amal Htait
Aston University, Birmingham, UK
shobanapriyan.chand@gmail.com, a.htait@aston.ac.uk

Abstract

The exponential growth of clinical trial reports (CTRs) presents a critical challenge for evidence-based medicine, with manual systematic reviews requiring months to synthesise findings. This paper evaluates Large Language Models (LLMs) and retrieval methods for automated Natural Language Inference (NLI) and evidence extraction from CTRs, and seeks to improve upon previously reported results in this domain. Using the NLI4CT dataset containing 2,400 annotated statement-evidence pairs from breast cancer trials, we conducted a comparative evaluation of general-purpose LLMs, domain-specific LLMs, and transformer-based baselines across entailment classification and evidence retrieval tasks. Reasoning-capable, general-purpose LLMs (such as Qwen-32B) demonstrated superior performance in the entailment classification task, exceeding both the performance of other models evaluated in this study and the previously reported state-of-the-art results. Although domain-specific adaptations showed improvements at comparable scale, larger general-purpose language models maintained superior absolute performance. For evidence retrieval, Large Language embedding models (such as bge-large-en-v1.5) surpassed classic transformer-based ranking approaches. These findings demonstrate that modern LLMs with reasoning capabilities can effectively support real-time clinical evidence synthesis without task-specific fine-tuning, offering a pathway toward scalable automated systems for clinical trial interpretation that could substantially reduce the evidence-to-practice gap in medical decision-making.

Keywords: Large Language Models, Natural Language Inference, Clinical Trail Analysis

1. Introduction

Clinical trial reports (CTRs) constitute the foundation of evidence-based medicine, yet their effective synthesis remains a persistent challenge. These documents, often spanning hundreds of pages with dense technical terminology and heterogeneous reporting structures, demand sophisticated interpretation to support timely clinical decision-making (Ioannidis et al., 2014; Higgins et al., 2019). An example of how CTR evidence is used for reasoning over patient statements is presented in Figure 1. The traditional manual review process, while thorough, struggles to keep pace with the exponential growth of published trials, creating a widening gap between available evidence and clinical practice (Bastian et al., 2010; Garcia-Alamino and Lopez-Cano, 2020). This bottleneck in evidence synthesis delays the translation of research findings into patient care, potentially impacting treatment outcomes and healthcare quality. Natural Language Inference (NLI) presents a compelling computational approach to address this challenge. By determining whether hypotheses are entailed, contradicted, or neutral with respect to given premises, NLI systems can automate the verification of clinical claims against trial evidence (Dagan et al., 2005). However, clinical NLI poses unique challenges that extend beyond general-domain applications. Medical texts require understanding of specialised terminology, complex numerical relationships, and implicit domain knowledge. Moreover, clinical conclusions rarely derive from single statements; they emerge from multiple pieces of evidence distributed across trial protocols, results, and statistical analyses. This multi-evidence nature of clinical reasoning necessitates models capable of aggregating information from disparate sources while maintaining interpretability and accuracy (Jullien et al., 2023).

Recent advances in Large Language Models (LLMs) offer new opportunities for clinical text understanding. These models demonstrate remarkable zero-shot capabilities across various medical reasoning tasks, suggesting their potential for CTR analysis without extensive task-specific training (Nori et al., 2023). When combined with sophisticated retrieval mechanisms, LLMs can identify relevant evidence spans and perform compositional reasoning over them (Fan et al., 2024). Domain-adapted variants of LLMs, such as BioMistral and OpenBioLLM, further promise enhanced performance through specialised pretraining on biomedical corpora (Lee et al., 2020; Phan et al., 2021; Gu et al., 2021). Nevertheless, critical questions remain regarding their reliability compared to traditional fine-tuned transformers models and their ability to provide verifiable, source-grounded justifications essential for clinical applications.

This study presents a comprehensive evaluation of LLM-based approaches for multi-evidence NLI in clinical trial analysis. Our primary aim is to improve upon the accuracy of previously reported

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 515–523
11–16 May 2026.

Example 1 — Entailment

Premise (CTR Evidence — Eligibility Criteria):

"Patients aged 18 years or older with confirmed HER2-positive breast cancer were eligible for the trial."

Hypothesis (Patient Record Statement):

"A 45-year-old patient with HER2-positive breast cancer is eligible for the trial."

Label: Entailment — the patient meets both eligibility conditions (age $\geq$ 18 and HER2-positive).

Example 2 — Contradiction

Premise (CTR Evidence — Eligibility Criteria):

"Patients aged 18 years or older with confirmed HER2-positive breast cancer were eligible for the trial."

Hypothesis (Patient Record Statement):

"A 45-year-old patient with HER2-negative breast cancer is eligible for the trial."

Label: Contradiction — the patient does not meet the HER2-positive requirement, directly conflicting with the eligibility criteria.

Figure 1: Illustrative examples of clinical trial eligibility criteria, showing entailment and contradiction between trial eligibility statements and patient records.

methods while systematically comparing performance across two key dimensions: (1) the performance of general-purpose LLMs, domain-adapted LLMs (e.g., Bio-Mistral7B , LLaMA3-OpenBioLLM-8B), and domain-specific transformer models (e.g., BioBERT, BioMegatron) in identifying entailment and contradiction relations between clinical hypotheses and CTR content, (2) the performance of dense embedding-based retrieval large models, transformer-based ranking models, and BM25 in clinical evidence retrieval. In this work, we use the NLI4CT benchmark, a widely adopted dataset, to enable rigorous and comparable assessment on clinically relevant premise-hypothesis pairs (Jullien et al., 2023). This work contributes to bridging the evidence-practice gap by providing a scalable approach for faster evidence synthesis, supporting informed medical decision-making while maintaining the transparency required for clinical adoption (Garcia-Alamino and Lopez-Cano, 2020).

2. Related Work

The NLI4CT dataset (Jullien et al., 2023) was introduced to advance NLI methods specifically for clinical trial data. This resource was motivated by the need for systems capable of integrating and reasoning over multiple pieces of evidence, often scattered across complex and structured CTRs, to determine entailment relationships. Unlike traditional NLI tasks that rely on a single premise-hypothesis pair, NLI4CT focus on multi-evidence reasoning, reflecting real-world scenarios in evidence-based medicine where accurate interpretation of aggregated information is crucial. The resource includes two main tasks, **Task 1**: Textual Entailment, which involves determining the inference relation (entailment or contradiction) between a natural language statement and a CTR; and **Task 2**: Evidence Retrieval, which requires extracting supporting facts from the CTR to justify the label predicted in the entailment task. Several approaches have been proposed to address this challenge. Kanakarajan and Sankarasubbu (2023) implemented Flan-T5, an instruction-tuned T5 model, for both zero-shot evaluation and task-specific fine-tuning, where best results were achieved by the finetuned Flan-T5-xxl model with F1 score of 0.834, for Task 1. Pahwa and Pahwa (2023) explored three different strategies: a sentence transformer-based bi-encoder model, cross-encoder models fine-tuned on the task training dataset, and GPT-3.5 using zero-shot and few-shot methods. In their work, GPT-3.5 Davinci model outperformed with F1 score of 0.694, for Task 1. In comparison, Vladika and Matthes (2023) adopted a contrasting methodology by developing a unified joint architecture that employed multi-task learning (MTL) to simultaneously address both tasks. Their system integrated ensemble techniques with multiple domain-adapted BERT variants, including a DeBERTa-based model, achieving F1 scores of 0.798 for Task 1 and 0.818 for Task 2. As for the group with the best results, Zhou et al. (2023) proposed the Multigranularity Inference Network (MGNet), which uses a joint inference approach and integrates four components: (1) a transformer-based Joint Semantics Encoder to capture contextual representations of the hypothesis and premise while embedding medical domain knowledge; (2) a sentence-level encoder to model higher-level contextual semantics; (3) a token-level encoder to extract reasoning-related semantics from individual sentences within the premise; and (4) task-specific classifiers for final predictions. Their system achieved F1 scores of 0.856 for Task 1 and 0.853 for Task 2.

3. Methodology

3.1. Dataset

The dataset employed in this study is the NLI4CT corpus, which was specifically curated to advance research in clinical natural language inference (NLI). It comprises 1,000 breast cancer CTRs sourced from ClinicalTrials.gov, a registry maintained by the U.S. National Library of Medicine (Zarin et al., 2011). The dataset is HIPAA-compliant, ensuring all personally identifiable information is removed before public release. Each CTR is structured into four key sections: Eligibility Criteria, Intervention, Results, and Adverse Events through annotation process, capturing patient selection, treatment details, trial outcomes, and side effects, respectively. The final dataset consists of 2,400 annotated statements, each paired with: (1) a medical statement, (2) a primary and secondary CTR, (3) an entailment or contradiction label, and (4) the supporting evidence. The dataset is split into 1,700 training, 500 test, and 200 development instances, balanced across labels and CTR sections. It includes both single-trial statements (60%) and comparative statements (40%), enabling evaluation of intra-trial and inter-trial reasoning. The comparative focus is especially valuable, reflecting real-world scenarios where outcomes must be compared across studies.

3.2. System Overview

Recent evaluations suggest that general-purpose LLMs can exhibit strong zero-shot performance on complex medical reasoning tasks when guided by carefully designed prompts (Nori et al., 2023; Wei et al., 2022). Accordingly, for Task 1 (Textual Entailment), the experiments focused on evaluating prompt engineering approaches across a range of general-purpose and biomedical domain-adapted LLMs (Lee et al., 2020; Gu et al., 2021). Several prompt variations were tested as part of the prompt engineering process, and the best-performing prompt was used in the final experiments. The system input comprises (1) a patient medical statement, (2) primary evidence extracted from CTRs, and (3) secondary evidence drawn from related CTRs. These components are combined into a structured prompt and passed to an LLM, which outputs an entailment label either (Entailment or Contradiction) based on the provided evidence. Multiple LLMs were evaluated to enable comparative analysis.

For Task 2 (Evidence Retrieval), semantic evidence retrieval is performed using embedding models. Such methods have proven effective for semantic similarity tasks, including information retrieval and evidence ranking (Reimers and Gurevych, 2019). Similar to Task 1, the input compromises a patient medical statement, primary evidence, and secondary evidence. Each input is encoded into an embedding vector, and cosine similarity is computed between the statement and candidate evidence spans. A reranking logic then orders the evidence by similarity, returning the most relevant supporting facts. The overall system design for Task 1 and Task 2 is illustrated in Figure 2.

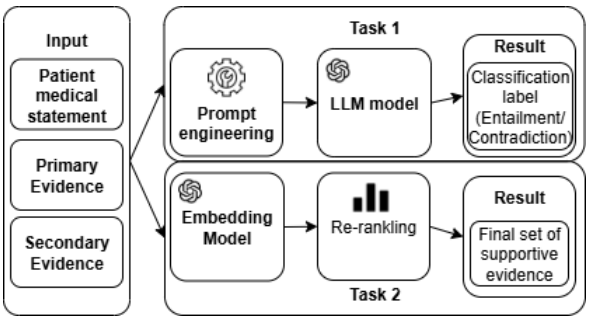

Figure 2: A simplified system overview for Task 1 and Task 2.

3.3. Experimental Setup

3.3.1. Dataset Preprocessing

In this work, we utilised the NLI4CT benchmark dataset, comprising two data categories: single-evidence instances containing only primary clinical trial (CTR) evidence, and comparison instances including both primary and secondary evidence. Each test sample was standardised into a structured JSON format with a unique identifier, section metadata (e.g., Eligibility), lists of evidence sentences, the target clinical statement, and the corresponding gold label (Entailment or Contradiction).

3.3.2. Task 1: Classification of entailment and contradiction

Prompt engineering: A zero-shot prompting strategy was employed in Task 1, in which instructional prompts guided the LLM to classify each statement–evidence pair as either entailment or contradiction. Zero-shot prompt engineering has been shown to produce accurate, task-specific reasoning from LLMs without the need for gradient-based fine-tuning (Brown et al., 2020; Wei et al., 2022). A standardised prompt template (presented in Figure 3) was applied uniformly across all evaluated models.

LLM models: The experiments included general-purpose LLMs, both instruction-tuned models LLMs such as: GPT-4o, Mistral 24B-Instruct (AI, 2025), and LLaMA 3.1–8B (Grattafiori et al., 2024), and reasoning capable LLMS such as: gpt-oss-20b (OpenAI, 2025), DeepSeek R1 (DeepSeek-AI, 2025), qwen-32B(Team, 2025), in addition

```
("system",
"You are an expert clinical NLI judge. "
"Decide whether the given Statement is supported by the provided clinical trial evidence. "
"Use ONLY the evidence given. Do NOT use outside knowledge. "
"Output must be strictly one of: Entailment, Contradiction."),

("human",
"""
Primary evidence:
{primary_evidence}

Secondary evidence:
{secondary_evidence}

Statement:
{statement}

INSTRUCTIONS:
- "Entailment": The statement is directly supported by or logically follows from the evidence
- "Contradiction": The statement directly opposes or conflicts with the evidence
- Focus on logical relationships, not just keyword matching
- Consider numerical data, statistical significance, and clinical outcomes carefully

Based on the combined evidence above, classify the relationship between the evidence and the statement.
Output only one word: Entailment OR Contradiction.
"""
)
```

Figure 3: The standardised prompt template used across all evaluated models.

to domain-specific biomedical LLMs, such as: BioMistral-7B (Labrak et al., 2024) and LLaMA3-OpenBioLLM-8B (Ankit Pal, 2024). These models were compared against a baseline consisting of general and domain-specific transformer models.

Evaluation Metrics: Model performance was initially assessed using accuracy, precision, recall, and F1 score, consistent with the NLI4CT evaluation protocol (Jansen et al., 2021).However, for simplicity and clarity of presentation, only F1 scores are reported in this paper.

3.3.3. Task 2: Clinical evidence retrieval

Embedding model: The embedding models evaluated include: Qwen3-Embedding-8B (Zhang et al., 2025), Linq-Embed-Mistral-7B (Kim et al., 2024), Salesforce/SFR-Embedding-2_R (Meng et al., 2024), bge-Large-en-v1.5 (Xiao et al., 2023), Snowflake-Arctic-Embed-L-v2.0 (Yu et al., 2024), Nomic-Embed-Text-v1.5 (Nussbaum et al., 2024), Mixedbread-MXBai-Embed-Large-v1 (Lee et al., 2024; Li and Li, 2023), as well as OpenAI's text-embedding-ada-002 (OpenAI, 2023) and text-embedding-3-large (OpenAI, 2024). These models were compared against a baseline consisting of BM5 and general transformer models.
Reranking: Reranking was performed using cosine similarity between the dense embedding of the patient statement and the embeddings of candidate evidence passages. All candidate sentences from both primary and secondary evidence were encoded, and their similarity scores with the statement were computed. Evidence items were then ranked from highest to lowest similarity, with the top-k items selected for further analysis. This reranking approach follows established methods in semantic textual similarity and information retrieval (Jansen et al., 2021).
Evaluation Metrics: Mean Average Precision (mAP) was used as the primary retrieval metric (Liu, 2011). The task was framed as ranking all possible evidence sentences, where gold evidence should be ranked higher than distractors. This metric has been standard in information retrieval evaluation and aligns with prior CTR-based inference tasks (Jansen et al., 2021).

4. Results

In this work, we aimed to improve upon the accuracy of prior methods by systematically comparing performance across two key dimensions: (1) general-purpose LLMs, domain-adapted LLMs, and domain-specific transformer models in the classification of entailment and contradiction between clinical hypotheses and CTR content, and (2) dense embedding-based retrieval models, transformer-based ranking models, and BM25 in clinical evidence retrieval. In the following, we present the results of these evaluations.

4.1. Task 1 - Results: Classification of entailment and contradiction

The comparative analysis of generative LLMs against traditional transformer-based models (as the baseline) reveals significant performance disparities across different model categories. The baseline models, including domain-specific transformer models (such as BERT-base, RoBERTa-

base, BioBERT, and BioMegatron) established the foundational performance benchmarks for the entailment task. Among these baselines, BioBERT and BioMegatron achieved the highest F1 scores of 0.644, while RoBERTa-base demonstrated the best F1 scores at 0.628, within the non domain-specific transformer models, as shown in Table 1.

The use of LLM-based approaches demonstrated substantial improvements over the established baselines, as presented in Table 1. The comparison between GPT-4o, Llama 3.1 8B, and Mistral 24B-instruct, all instruction-tuned general-purpose models specifically designed for strict instruction-following tasks (Nori et al., 2023), highlights the critical influence of model size on performance. The 24-billion parameter Mistral model consistently outperformed the 8-billion parameter Llama model across all evaluation metrics, with F1 scores of 0.6460 versus 0.6033, respectively, and both were surpassed by GPT-4o, with F1 score of 0.6997, which contains approximately 200 billion parameters. These results demonstrate a positive correlation between parameter count and model performance in instruction-following tasks.

The most remarkable performance gains emerged from the reasoning-capable LLM models, which employ chain-of-thought methodology to simulate human-like logical reasoning processes (Wei et al., 2022). This reasoning capability proves particularly relevant for the NLI4CT task, which demands sophisticated reasoning across multiple domains including common sense reasoning, biomedical reasoning, and numerical reasoning. The reasoning models demonstrated significant performance improvements compared to both baseline and instruction-based models. Qwen 32B achieved the highest overall performance with an F1 score of 0.8813, surpassing all previously presented methods, as shown in Table 1, including the previous best of 0.853 obtained by Zhou et al. (2023).

Notably, the OpenAI gpt-oss-20B model, despite having fewer parameters than the Mistral 24B-instruct model, achieved significantly superior performance with an F1 score of 0.8744 compared to 0.6460. This finding suggests that reasoning capabilities and architectural design may be more influential factors than raw parameter count in determining model effectiveness for natural language inference tasks in clinical contexts.

The evaluation also incorporated two prominent domain-adapted models: Bio-mistral7B and LLaMA3-OpenBioLLM-8B, representing different approaches to biomedical domain specialisation. Bio-mistral7B, with its 7-billion parameter architecture, achieved an F1 score of 0.6023, demonstrating modest performance improvements over the baseline models but falling short of the general-purpose Mistral 24B-instruct model. Similarly, LLaMA3-OpenBioLLM-8B recorded an F1 score of 0.6789, indicating that domain adaptation alone does not guarantee superior performance when compared to larger general-purpose models. However, the parameter-controlled comparison between models of similar scale reveals the genuine impact of domain-specific adaptation. The direct comparison between Llama 3.1 8B and LLaMA3-OpenBioLLM-8B provides crucial insights into the effectiveness of domain-specific fine-tuning. The domain-adapted LLaMA3-OpenBioLLM-8B, representing a fine-tuned version of the base Llama 3 8B model on biomedical data (Lee et al., 2020; Gu et al., 2021), significantly outperformed its general-purpose counterpart, achieving an F1 score of 0.6789 compared to 0.6033 for Llama 3.1 8B. This comparative analysis demonstrates that domain-specific adaptation provides measurable performance improvements when controlling for model architecture and parameter count. The findings suggest that while larger general-purpose models may achieve superior absolute performance through scale advantages, domain adaptation represents a valuable optimisation strategy for scenarios with computational constraints or specific deployment requirements in clinical environments. The results collectively indicate that domain adaptation effectiveness depends significantly on the comparative context and intended application. While domain-specific models may not surpass the performance of substantially larger general-purpose models, they offer important advantages in parameter efficiency and specialised knowledge representation. The consistent improvement observed in parameter-matched comparisons validates the continued relevance of domain-specific adaptation strategies for biomedical natural language processing applications, particularly in resource-constrained environments where deploying massive general-purpose models may be impractical. These findings contribute valuable insights to the ongoing discourse regarding the optimal balance between model scale, domain specialization, and task-specific performance in clinical NLP applications, providing guidance for practitioners and researchers in selecting appropriate modelling approaches for real-world deployment scenarios.

4.2. Task 2 - Results: Clinical evidence retrieval

The evaluation of evidence retrieval capabilities through Task 2 revealed the superior performance of embedding-based approaches compared to traditional transformer-based ranking methods. The baseline implementation used the following state-of-the-art general-purpose transformer-based ranking

Type	Model	F1
BL	BERT-base	0.528
BL	RoBERTa-base	0.628
BL	BioBERT	0.644
BL	BioMegatron	0.644
LLM-I	GPT4o	0.6997
LLM-I	Llama 3.1 8B	0.6033
LLM-I	Mistral 24B	0.6460
LLM-R	DeepSeek R1	0.8727
LLM-R	Qwen 32B	**0.8813**
LLM-R	OpenAI: gpt-oss-20b	0.8744
LLM-DS	Bio-mistral7B	0.6023
LLM-DS	LLaMA3-OpenBioLLM-8B	0.6789

Table 1: Results (F1 scores) for Task 1, with BL = baseline transformer models, LLM-I = instruction-tuned general-purpose LLMs, LLM-R = reasoning-capable LLMs, and LLM-DS = domain-specific LLMs.

models: DistilRoBERTa (Sanh et al., 2019), MPNet (Song et al., 2020), and MiniLM (Wang et al., 2020) variants, alongside the classical BM25 retrieval method. The transformer-based baseline models demonstrated moderate performance, with MiniLM (all-MiniLM-L6-v2) achieving the highest mean average precision (mAP) of 0.777 among the ranking models, while MPNet (all-mpnet-base-v2) and DistilRoBERTa (all-distilroberta-v1) achieved mAP scores of 0.760 and 0.762, respectively. Interestingly, the classical BM25 approach outperformed these transformer-based methods with an mAP of 0.786, suggesting that traditional lexical matching techniques retain competitive advantages in certain retrieval scenarios.

The embedding model evaluation demonstrated consistent and improvements over baseline approaches, confirming the robustness of embedding-based retrieval methodologies. The performance spectrum of embedding models ranged from 0.789 to 0.849 in mAP scores, with all tested embedding models surpassing the highest baseline performance. The bge-large-en-v1.5 model achieved the highest performance with an mAP of 0.849, closely approaching previously reported best results by Zhou et al. (2023) (mAP 0.853), followed closely by nomic-ai/nomic-embed-text-v1.5 at 0.846 and mixedbread-ai/mxbai-embed-large-v1 at 0.836. OpenAI's embedding models, embedding-ada-002 and embedding-3-large, demonstrated competitive performance with mAP scores of 0.833 and 0.836, respectively. The Snowflake embeding model (snowflake-arctic-embed-l-v2.0) achieved an mAP of 0.831, while the Qwen3-Embedding-8B and Salesforce SFR-Embedding-2_R models recorded mAP scores of 0.811 and 0.801, respectively. Even the lowest-performing embedding model, Linq-Embed-Mistral-7B with an mAP of 0.789, exceeded the performance of traditional transformer-based ranking approaches, highlighting the consistent advantage of embedding-based retrieval methodologies for evidence selection tasks in clinical contexts. Table 2 summarises the comparison.

Type	Model	mAP
BL	BM25	0.786
BL-TR	MiniLM	0.777
BL-TR	MPNet	0.760
BL-TR	DistilRoBERTa	0.762
EM	embedding-ada-002	0.833
EM	embedding-3-large	0.836
EM	Qwen3-Embedding-8B	0.811
EM	Linq-Embed-Mistral-7B	0.789
EM	Salesforce	0.801
EM	bge-large-en-v1.5	**0.849**
EM	Snowflake	0.831
EM	nomic-ainomic	0.846
EM	ixedbread-aimxbai	0.836

Table 2: Results (mAP) for Task 2, with BL = baseline models, BL-TR = baseline Transformer models, and EM = Embedding LLMs.

5. Conclusion

This paper presented a systematic evaluation of LLMs and retrieval methods for automated NLI and evidence extraction from CTRs, with the aim of improving upon previously reported results in this domain. The results demonstrated that reasoning-capable, general-purpose LLMs can substantially improve automated entailment classification with Qwen-32B achieving an F1 score of 0.8813, outperforming the other models evaluated in this study and the previously reported state-of-the-art best results. These results highlight the potential of large, reasoning-oriented LLMs to support real-time clinical evidence synthesis without task-specific fine-tuning, offering a scalable path toward reducing the evidence-to-practice gap in medical decision-making. In the evidence retrieval task from clinical trial reports, OpenAI embedding models surpassed traditional transformer-based baselines and achieved performance comparable to previously reported state-of-the-art best results.

Despite these promising results, several risks must be considered before deploying LLM-based NLI systems in real clinical environments. LLMs can produce confident but factually incorrect predictions, which is a critical concern in high-stakes medical decision-making where errors could directly impact patient care. Additionally, models operating without task-specific fine-tuning may exhibit limited

reliability on edge cases involving rare conditions or ambiguous clinical language. Regulatory requirements for clinical decision support tools, including explainability and auditability, present further deployment challenges, as current LLMs offer limited transparency into their reasoning processes.

Future work will investigate fine-tuning high-performing open-source reasoning models, such as Qwen-32B and Mistral-24B-instruct, on domain-specific medical knowledge bases and clinical NLI datasets. Both traditional and parameter-efficient fine-tuning approaches will be explored to enable optimisation in resource-constrained environments. Instruction tuning and reinforcement learning from human feedback (RLHF), adapted for clinical reasoning, represent further promising directions for improving model alignment with medical practice standards. Evaluating these approaches across broader medical domains and real-world clinical workflows will be essential to assess generalisability and robustness, with the longer-term goal of integrating validated LLM-based evidence extraction into clinical decision support systems.

6. Bibliographical References

Mistral AI. 2025. Mistral small 24b.

Malaikannan Sankarasubbu Ankit Pal. 2024. Openbiollms: Advancing open-source large language models for healthcare and life sciences. https://huggingface.co/aaditya/OpenBioLLM-Llama3-70B.

H. Bastian, P. Glasziou, and I. Chalmers. 2010. Seventy-five trials and eleven systematic reviews a day: how will we ever keep up? *PLoS Medicine*, 7(9):e1000326.

T. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901.

I. Dagan, O. Glickman, and B. Magnini. 2005. The pascal recognising textual entailment challenge. In *Proceedings of the First International Conference on Machine Learning Challenges: Evaluating Predictive Uncertainty, Visual Object Classification, and Recognizing Textual Entailment*, pages 177–190.

DeepSeek-AI. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning.

A. Fan, B. Gokkaya, M. Harman, M. Lyubarskiy, S. Sengupta, S. Yoo, and J.M. Zhang. 2024. Large language models for software engineering: survey and open problems. In *Proceedings of the IEEE/ACM 46th International Conference on Software Engineering*, pages 1–33.

J.M. Garcia-Alamino and M. Lopez-Cano. 2020. Evidence synthesis and systematic reviews in surgery: the importance, limitations, and a proposal for improvement. *Cirugía Española*, 98(5):241–244.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Y. Gu, R. Tinn, H. Cheng, M. Lucas, N. Usuyama, X. Liu, T. Naumann, J. Gao, and H. Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare*, 3(1):1–23.

J.P.T. Higgins, J. Thomas, J. Chandler, M. Cumpston, T. Li, M.J. Page, and V.A. Welch, editors. 2019. *Cochrane Handbook for Systematic Reviews of Interventions*, 2nd edition. John Wiley & Sons, Chichester.

J.P.A. Ioannidis, S. Greenland, M.A. Hlatky, M.J. Khoury, M.R. Macleod, D. Moher, K.F. Schulz, and R. Tibshirani. 2014. Increasing value and reducing waste in research design, conduct, and analysis. *The Lancet*, 383(9912):166–175.

P. Jansen, N. Balasubramanian, M. Surdeanu, and P. Clark. 2021. A study of question understanding and answering models for science exams. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 12864–12872.

K.R. Kanakarajan and M. Sankarasubbu. 2023. Saama ai research at semeval-2023 task 7: exploring the capabilities of flan-t5 for multi-evidence natural language inference in clinical trial data. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 995–1003.

Junseong Kim, Seolhwa Lee, Jihoon Kwon, Sangmo Gu, Yejin Kim, Minkyung Cho, Jy yong Sohn, and Chanyeol Choi. 2024. Linq-embed-mistral:elevating text retrieval with improved gpt

data through task-specific control and quality refinement. Linq AI Research Blog.

Yanis Labrak, Adrien Bazoge, Emmanuel Morin, Pierre-Antoine Gourraud, Mickael Rouvier, and Richard Dufour. 2024. Biomistral: A collection of open-source pretrained large language models for medical domains.

J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C.H. So, and J. Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Sean Lee, Aamir Shakir, Darius Koenig, and Julius Lipp. 2024. Open source strikes bread - new fluffy embeddings model.

Xianming Li and Jing Li. 2023. Angle-optimized text embeddings. *arXiv preprint arXiv:2309.12871*.

T.Y. Liu. 2011. Learning to rank for information retrieval. *Foundations and Trends in Information Retrieval*, 3(3):225–331.

Rui Meng, Ye Liu, Shafiq Rayhan Joty, Caiming Xiong, Yingbo Zhou, and Semih Yavuz. 2024. Sfr-embedding-2: Advanced text embedding with multi-stage training.

H. Nori, N. King, S.M. McKinney, D. Carignan, and E. Horvitz. 2023. Capabilities of gpt-4 on medical challenge problems. *arXiv preprint arXiv:2303.13375*.

Zach Nussbaum, John X Morris, Brandon Duderstadt, and Andriy Mulyar. 2024. Nomic embed: Training a reproducible long context text embedder. *arXiv preprint arXiv:2402.01613*.

OpenAI. 2023. text-embedding-ada-002. `https://platform.openai.com/docs/guides/embeddings`. Large language model.

OpenAI. 2024. text-embedding-3-large. `https://platform.openai.com/docs/guides/embeddings`. Large language model.

OpenAI. 2025. gpt-oss-120b & gpt-oss-20b model card.

B. Pahwa and B. Pahwa. 2023. Bphigh at semeval-2023 task 7: Can fine-tuned cross-encoders outperform gpt-3.5 in nli tasks on clinical trial data? In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 1936–1944.

L.N. Phan, J.T. Anibal, H. Tran, S. Chanana, E. Bahadroglu, A. Peltekian, and G. Altan-Bonnet. 2021. Scifive: a text-to-text transformer model for biomedical literature. *arXiv preprint arXiv:2106.03598*.

N. Reimers and I. Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. 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 3982–3992.

V. Sanh, L. Debut, J. Chaumond, and T. Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

K. Song, X. Tan, T. Qin, J. Lu, and T.Y. Liu. 2020. Mpnet: Masked and permuted pre-training for language understanding. In *Advances in Neural Information Processing Systems*, volume 33, pages 16857–16867.

Qwen Team. 2025. Qwen3 technical report.

J. Vladika and F. Matthes. 2023. Sebis at semeval-2023 task 7: A joint system for natural language inference and evidence retrieval from clinical trial reports. *arXiv preprint arXiv:2304.13180*.

W. Wang, F. Wei, L. Dong, H. Bao, N. Yang, and M. Zhou. 2020. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. In *Advances in Neural Information Processing Systems*, volume 33, pages 5776–5788.

J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E.H. Chi, Q.V. Le, and D. Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837.

Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighoff. 2023. C-pack: Packaged resources to advance general chinese embedding.

Puxuan Yu, Luke Merrick, Gaurav Nuti, and Daniel Campos. 2024. Arctic-embed 2.0: Multilingual retrieval without compromise. *arXiv preprint arXiv:2412.04506*.

D.A. Zarin, T. Tse, R.J. Williams, and S. Carr. 2011. Trial reporting in clinicaltrials.gov - the final rule. *New England Journal of Medicine*, 364(9):852–860.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. 2025. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*.

Y. Zhou, Z. Jin, M. Li, M. Li, X. Liu, X. You, and J. Wu. 2023. Thifly research at semeval-2023 task 7: A multi-granularity system for ctr-based textual entailment and evidence retrieval. *arXiv preprint arXiv:2306.01245*.

7. Language Resource References

M. Jullien, M. Valentini, H. Frost, P. O'Neill, Z. Lyu, D. Tran, and T. Bhowmik. 2023. Nli4ct: multi-evidence natural language inference for clinical trial reports. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16745–16764.