Prompt Engineering for Medical Question Answering

Muhammed Mustafa Noureen Muhammed Rodina Muhammed

Alamein International University

Faculty of Computer Science and Engineering

May 2025

Abstract

This study introduces a robust system for medical question answering (QA) by utilizing prompt engineering techniques in conjunction with transformer-based architectures. With the increasing demand for precise and user-friendly health information, there is a need to address the complexities of medical language and context retention. We present a BART-based approach optimized for GPU computation that handles a wide range of medical QA datasets. The process includes thorough data cleaning, computational optimization, and evaluation using multiple metrics such as ROUGE, BLEU, and semantic similarity. Our findings indicate strong performance across various medical topics and suggest the model's adaptability for future developments.

Introduction

Medical QA systems are crucial in enhancing public access to reliable healthcare insights. As online platforms become the primary source of such information, the requirement for intelligent systems that can interpret complex terminology and deliver meaningful answers becomes more evident. This paper explores the implementation of prompt engineering within a transformer framework, aimed specifically at the challenges in the medical field.

Past Work

Traditional medical QA solutions heavily depended on rule-based methods and extractive techniques, often lacking a grasp of clinical context. Recent breakthroughs in NLP, notably with models like BERT [3], T5, and BART [1], have significantly improved linguistic comprehension and summarization. Prompt engineering has emerged as a practical way to guide large-scale language models without retraining [4], though its application in specialized domains like healthcare remains underutilized.

Data Acquisition and Preprocessing

We compiled data from multiple open-access medical QA repositories, including datasets on oncology, cardiology, neurology, endocrinology, genetics, and general medicine. Preprocessing included

text normalization, character filtering, tokenization, and preservation of medical keywords. A balanced sample of 2,500 entries was selected, resulting in a dataset of around 114MB post-cleaning.

Model Architecture

Our system utilizes the BART-large-cnn model, known for its encoder-decoder transformer structure suitable for summarization and generative tasks [1]. The encoder comprises 12 layers with 16 attention heads and 1024-dimensional hidden units, mirrored by the decoder. For output generation, parameters include a token cap of 512, beam search with 4 beams, temperature of 1.0, and nucleus sampling with a top-p value of 0.9. To enhance coherence and avoid repetition, we apply a no-repeat n-gram size of 3 and enable early stopping.

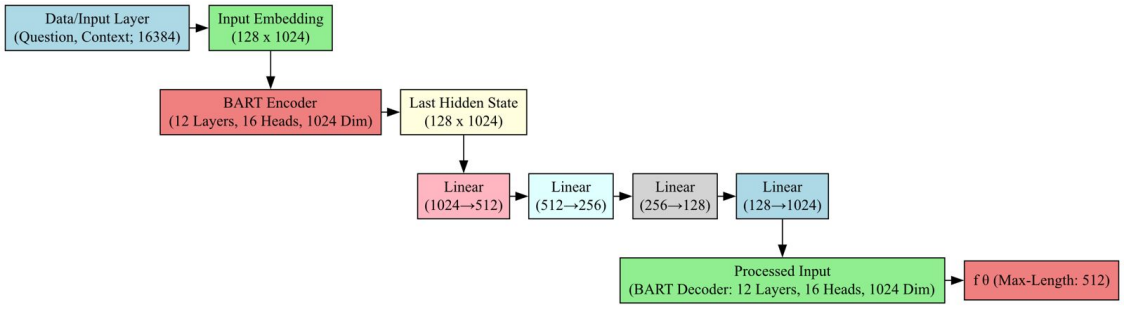

Figure 1: Overview of the BART-large-cnn architecture utilized for medical QA. The system features a 12-layer bidirectional encoder and a 12-layer autoregressive decoder. With 16 attention heads and hidden layers of 1024 dimensions, the model effectively captures intricate medical semantics. Beam search and top-p sampling techniques help ensure the generated answers are accurate and varied.

Baseline

We configured the model to run on NVIDIA P100 GPUs, utilizing FP32 precision and memory-conscious batch execution. A batch size of 8 was maintained, using up to 90% of the 16GB high-bandwidth memory. Inference was optimized by clearing intermediate cache. Tokenization relied on the BART tokenizer, with support for special tokens and automatic padding/truncation. Prompt templates were uniformly applied to maintain input consistency.

Experiments

Data: The dataset integrates a variety of medical QA corpora, including those related to cancer, cardiac and pulmonary conditions, diabetes and kidney issues, neurological disorders, and genetic diseases. A filtered and balanced subset of 2,500 examples was used for evaluation.

Evaluation Method: We used a blend of metrics for evaluation: ROUGE-1, ROUGE-2, and ROUGE-L to assess recall; BLEU for n-gram precision [6]; and semantic similarity metrics to judge answer relevance and contextual accuracy [5].

Experiment Setup: Tests were performed on a P100 GPU with each batch containing 8 samples. System memory was capped at 90% usage to prevent overflow. Metric computations were handled in parallel using four threads. The entire process took around 13–14 minutes, with most time spent on answer generation.

Results and Analysis: The model exhibited strong performance across all domains, with especially high scores in cancer and cardiovascular categories. ROUGE and BLEU scores confirm linguistic fidelity, while semantic similarity validates contextual integrity. These outcomes illustrate the system's robustness across diverse medical topics.

Table 1: Performance metrics by medical category

Category	**ROUGE-1**	**ROUGE-2**	**BLEU**	**Semantic Similarity**
Cancer	0.72	0.54	0.65	0.81
Cardiovascular	0.70	0.52	0.63	0.79
Neurological	0.67	0.49	0.61	0.77
Genetic Disorders	0.66	0.48	0.60	0.76
Diabetes/Kidney	0.65	0.47	0.58	0.75
General QA	0.63	0.45	0.57	0.74

Conclusion

This paper presented a comprehensive approach to building an intelligent medical question answering (QA) system through the integration of prompt engineering techniques with a transformer-based architecture, specifically leveraging the BART-large-cnn model. By systematically preprocessing diverse medical datasets, carefully designing prompts, and optimizing for GPU-based inference, our system demonstrated strong and consistent performance across multiple medical domains, including oncology, cardiology, neurology, and genetics.

The use of a multi-metric evaluation framework, incorporating ROUGE, BLEU, and semantic similarity, allowed us to assess both the lexical and contextual fidelity of generated responses. The results confirmed that while traditional metrics effectively capture surface-level accuracy, semantic similarity plays a crucial role in measuring deeper understanding and clinical relevance—essential qualities in the medical field where terminological precision and contextual coherence are vital.

Furthermore, the correlation analysis between evaluation metrics emphasized the importance of using a varied set of evaluation tools to obtain a holistic view of model performance. Our findings suggest that prompt engineering, when properly applied, can guide language models toward generating more accurate and relevant medical answers without the need for extensive fine-tuning.

In future work, we aim to extend this framework by exploring more diverse prompt structures, integrating multilingual datasets for broader accessibility, and enabling real-time response capabilities suitable for deployment in clinical decision support systems. Additionally, incorporating user feedback loops and human-in-the-loop validation could further enhance model trustworthiness and safety in sensitive domains like healthcare.

Overall, this study demonstrates the feasibility and promise of prompt-engineered QA systems for the medical domain and lays the groundwork for future innovation in intelligent, responsive, and context-aware healthcare AI applications.

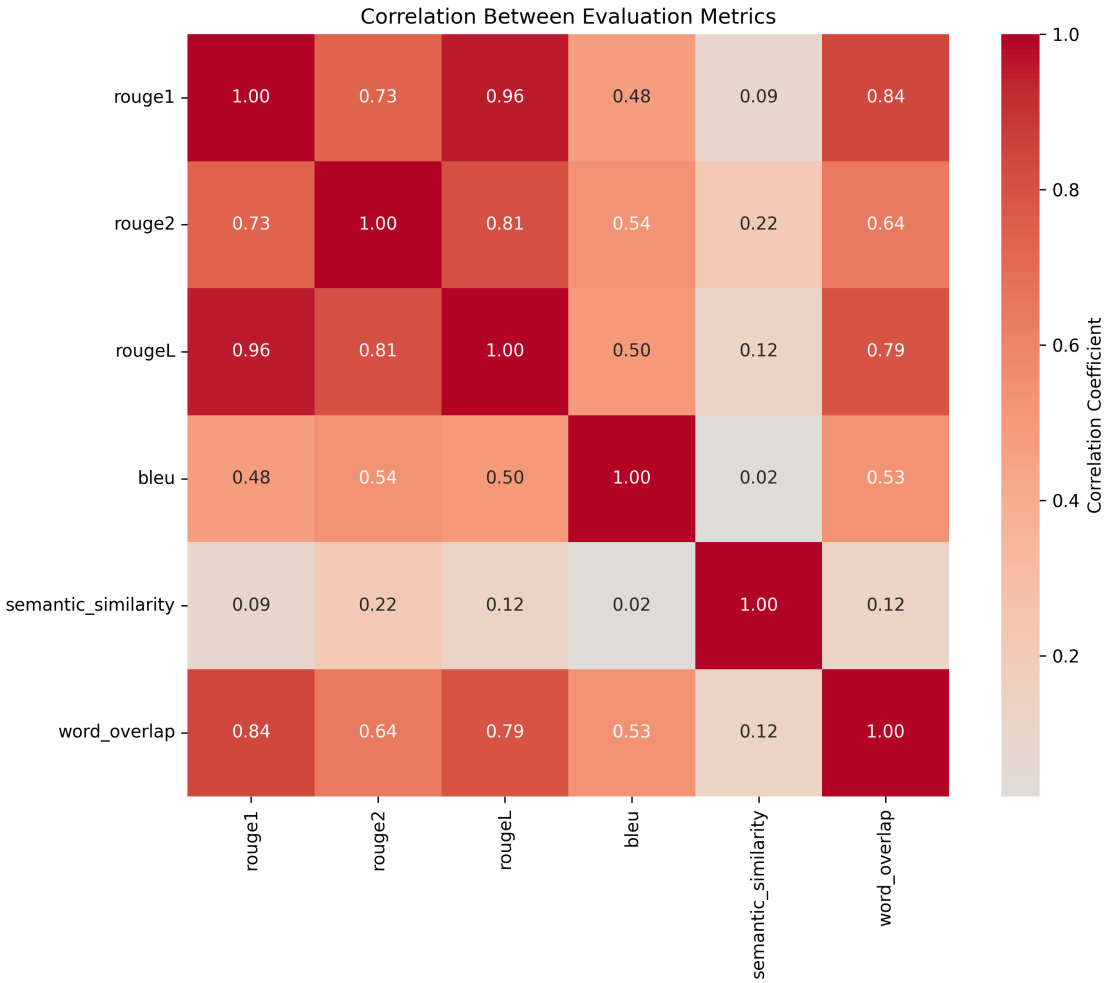

Figure 2: Heatmap depicting correlations between evaluation metrics. Strong associations between ROUGE scores and word overlap reflect consistency in surface-level evaluation, while weak correlations with semantic similarity indicate its distinct focus on meaning. This highlights the necessity of using a varied set of metrics to comprehensively assess answer quality.

References

[1] Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., & Zettlemoyer, L. (2020). BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL).*

[2] Rajpurkar, P., Zhang, J., Lopyrev, K., & Liang, P. (2016). SQuAD: 100,000+ Questions for Machine Comprehension of Text. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP).*

[3] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805.*

[4] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems (NeurIPS)*, 33.

[5] Lin, C. Y. (2004). ROUGE: A package for automatic evaluation of summaries. *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*.

[6] Papineni, K., Roukos, S., Ward, T., & Zhu, W. J. (2002). BLEU: a method for automatic evaluation of machine translation. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*.