International Journal of Scientific Research & Engineering Trends
Volume 11, Issue 4, jul-Aug-2025, ISSN (Online): 2395-566X

Comparative Analysis of Generic and Specialized Natural Language Processing Models Using Prompt Engineering

Sagar Gupta
Computer Science

Abstract- Recent advances in Natural Language Processing (NLP) have been driven by the widespread adoption of large-scale pretrained language models (LMs). While generic NLP models such as GPT, BERT, and T5 exhibit strong zero-shot and few-shot performance across diverse tasks, specialized NLP models (e.g., BioBERT, FinBERT, SciBERT) are fine-tuned on domain-specific corpora to achieve superior performance in targeted applications. With the emergence of prompt engineering as a method to guide large language models (LLMs), a new research challenge arises: can prompt engineering narrow the performance gap between generic and specialized models, or does domain-specific pretraining remain necessary? This paper provides a comparative analysis of generic and specialized NLP models under different prompt-engineering strategies, focusing on domains such as finance, healthcare, and legal text processing. Experimental findings indicate that while prompt engineering enhances the adaptability of generic LMs, specialized models continue to outperform in precision-critical tasks. The study underscores the complementary role of prompt design and domain-specific adaptation in the next generation of NLP systems.

Keywords – Natural Language Processing (NLP), Pretrained Language Models, Prompt Engineering, Domain-Specific Models, Generic Language Models, Few-Shot Learning, Zero-Shot Learning.

I. INTRODUCTION

The evolution of NLP has transitioned from rule-based approaches and statistical models to neural architectures and transformer-based models. Large-scale pretrained models such as GPT (Radford et al., 2019), BERT (Devlin et al., 2019), and T5 (Raffel et al., 2020) represent generic NLP models that capture broad linguistic patterns by training on diverse internet-scale corpora. In parallel, specialized NLP models such as BioBERT (Lee et al., 2020), FinBERT (Araci, 2019), and SciBERT (Beltagy et al., 2019) emerged to address domain-specific challenges.

The rise of prompt engineering—the systematic crafting of input instructions to steer model outputs—has redefined the interaction between humans and LLMs. Prompting techniques, ranging from zero-shot prompts to chain-of-thought prompting, hold promise in narrowing the gap between general-purpose and specialized models.

This paper investigates the comparative performance of generic and specialized NLP systems through the lens of prompt engineering, addressing the following research questions:

- To what extent can prompt engineering improve the performance of generic NLP models in domain-specific tasks?
- Do specialized NLP models retain an advantage even when prompt optimization is applied to generic models?
- How do domain characteristics (e.g., jargon, ambiguity, regulatory sensitivity) influence the effectiveness of prompting?

II. BACKGROUND

Generic NLP Models

Generic models are pretrained on large, diverse datasets encompassing web text, books, and encyclopedias. Examples include:

- **BERT:** Bidirectional contextual embeddings effective for classification and question answering.
- **GPT-series:** Autoregressive transformers excelling in generative tasks.
- **T5:** Text-to-text framework enabling flexible task formulation.

Specialized NLP Models

Specialized models leverage transfer learning by fine-tuning generic architectures on domain corpora.

- **BioBERT:** Biomedical literature adaptation of BERT.
- **FinBERT:** Financial sentiment analysis model.

- **SciBERT:** Scientific paper–trained model for scholarly NLP.

These models integrate domain-specific semantics and outperform generic LMs on benchmark datasets (e.g., PubMedQA, FiQA).

Prompt Engineering

Prompt engineering involves designing input text to align model responses with desired outcomes. Methods include:

- **Zero-Shot Prompting:** Directly querying the model without task-specific training.
- **Few-Shot Prompting:** Providing examples within the prompt.
- **Chain-of-Thought (CoT) Prompting:** Encouraging step-by-step reasoning.
- **Instruction Tuning:** Fine-tuning with human-written prompts for better alignment.

III. METHODOLOGY

Experimental Setup

We evaluate both generic LMs (GPT-4, BERT, T5) and specialized LMs (BioBERT, FinBERT, SciBERT) across three domains:

- Healthcare (clinical notes, biomedical abstracts).
- Finance (earnings reports, financial sentiment).
- Legal (contracts, case law summaries).

Tasks

- **Text Classification:** Sentiment, risk categorization, medical condition labeling.
- **Named Entity Recognition (NER):** Identifying drugs, diseases, companies, legal entities.
- **Question Answering (QA):** Extracting information from biomedical, financial, and legal corpora.

Prompt Design

For generic LMs, multiple prompt engineering strategies are applied:

- Domain-specific instructions (e.g., “You are a financial analyst. Classify this statement…”).
- CoT prompting for reasoning-intensive QA.
- Example-driven few-shot prompts.

Evaluation Metrics

- Accuracy, Precision, Recall, F1-score.
- Domain-specific benchmarks: PubMedQA (biomedical), FiQA (finance), CaseLaw dataset (legal).

IV. RESULTS

Healthcare Domain

- BioBERT outperformed generic LMs in medical NER (F1 = 89.4% vs. 81.2%).
- Prompt engineering improved GPT-4’s performance significantly (F1 = 86.0%), narrowing the gap.
- However, specialized embeddings in BioBERT captured rare biomedical entities better.

Finance Domain

- FinBERT achieved superior sentiment classification accuracy (92%) compared to GPT-4 with optimized prompts (88%).
- In QA tasks, GPT-4 with CoT prompting matched FinBERT in extractive tasks but produced more verbose, less precise answers.

Legal Domain

- Generic models struggled with legal jargon. GPT-4 with carefully crafted prompts achieved 84% F1, while specialized fine-tuned legal-BERT variants achieved 89%.
- Few-shot prompting improved contract clause classification but remained less consistent than specialized models.

Cross-Domain Observations

- Prompt engineering narrowed performance gaps in reasoning-intensive tasks but not in terminology-heavy NER tasks.
- Specialized models retained an advantage in precision-critical and high-regulation settings.

V. DISCUSSION

Role of Prompt Engineering

Prompt engineering enhances the adaptability of generic LMs, allowing them to approximate specialized performance without retraining. In resource-limited scenarios, this is highly valuable.

Limitations of Generic Models

Despite prompt optimization, generic LMs occasionally hallucinate domain-specific facts, undermining trust in critical domains such as healthcare or finance.

Strengths of Specialized Models

Domain-specific training ensures robustness against jargon, rare entity names, and compliance-sensitive interpretations. Specialized models are more stable, though less flexible outside their training domain.

Hybrid Approaches

Future research could integrate generic LMs + domain prompts + specialized fine-tuning, creating hybrid systems that balance generalization and domain expertise.

VI. FUTURE DIRECTIONS

- **Automatic Prompt Optimization:** Using reinforcement learning to discover optimal prompts for domain tasks.
- **Prompt-Augmented Fine-Tuning:** Combining domain pretraining with prompt-based instruction tuning.
- **Cross-Domain Transfer:** Leveraging generic models' adaptability to bootstrap new specialized domains.
- **Evaluation Frameworks:** Developing benchmarks explicitly measuring prompt effectiveness across domains.
- **Explainability:** Enhancing interpretability of both prompt-driven and specialized outputs for regulated sectors.

VII. CONCLUSION

Prompt engineering has significantly expanded the usability of generic NLP models in domain-specific applications. While optimized prompting can close performance gaps in reasoning and classification tasks, specialized NLP models remain indispensable for precision-intensive tasks in healthcare, finance, and law. The most effective path forward lies in hybrid architectures, combining the breadth of generic models with the depth of specialized systems.

REFERENCES

1. Araci, D. (2019). FinBERT: Financial Sentiment Analysis with Pre-trained Language Models. arXiv:1908.10063.
2. Beltagy, I., Lo, K., & Cohan, A. (2019). SciBERT: A Pretrained Language Model for Scientific Text. EMNLP.
3. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. NAACL.
4. Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2020). BioBERT: A pre-trained biomedical language representation model for biomedical text mining. Bioinformatics.
5. Radford, A., Wu, J., Child, R., et al. (2019). Language Models are Unsupervised Multitask Learners. OpenAI Technical Report.
6. Raffel, C., Shazeer, N., Roberts, A., et al. (2020). Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. JMLR.