Prompting Strategies for Maximizing Language Model Accuracy on Graduate-Level Science Questions

Assignee Research

June 6, 2026

Abstract

This report synthesises findings from 4 peer-reviewed papers addressing the following research question: What prompting strategies maximize language model accuracy on graduate-level science questions v9. 15 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.7/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Clinical information extraction for Low-resource languages with Few-shot learning using Pre-trained language models and Prompting. Research question: What prompting strategies maximize language model accuracy on graduate-level science questions v9.

2 Methodology

Systematic literature search across multiple databases yielded 4 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 3.7/10.

3 Results

4 papers retrieved. 15 claims extracted; 0 independently verified. Quality review score: 3.7/10.

4 Limitations

This report is a machine-generated literature synthesis and does not constitute original research. Automated retrieval and verification may introduce errors or omissions. Review scores reflect automated assessment, not human peer review. Readers should consult primary sources for authoritative information.

5 Extracted Claims

Claim	Verified	Confidence
The upper bound baseline, a fine-tuned sequence classifier trained on the full training corpus, exceeds 96% accuracy for	×	0.03
Further-pretrained gbert models show a statistically significant accuracy increase of 0.4-0.6 points compared to the ori	×	0.02
No statistically significant difference in accuracy is observed between further-pretrained medbertde models and the orig	×	0.03
Zero-shot prompting results for all models are below 16% accuracy, with the exception of the public medbert-base model.	×	0.06
The public medbert-base model achieves 28.3% accuracy in zero-shot prompting, while the gbert-base model achieves 7.2%.	×	0.04
Gbert models further-pretrained on both task- and domain-specific data achieve 15% accuracy in zero-shot prompting, more	×	0.06
All performance differences for gbert models in zero-shot prompting are statistically significant, except for the differ	×	0.03
PET model variants significantly outperform SC models at shot sizes less than or equal to 100 in 31 out of 32 setups whe	×	0.06
The SC medbertde-base-comb model is the only model to outperform all PET models at a shot size of 100.	×	0.03
Statistical significance in performance gains from increasing few-shot size for both PET and SC models is observed at sh	×	0.01
For gbert PET models, accuracy gradually increases with task-specific, domain-specific, and combined pretraining in that	×	0.04
Gbert SC models benefit significantly from domain-adapted models over all shot sizes, except for shot sizes of 10 and 40	×	0.04
Gbert SC models do not show significant benefits from task adaptation or the combination of task and domain adaptation.	×	0.05
The study evaluates Pattern-Exploiting Training (PET) for few-shot learning in German clinical routine settings.	×	0.07
The classification task involves categorizing paragraphs of German doctor's letters into nine section categories.	×	0.09

References

- http://arxiv.org/abs/2408.13040v1
- http://arxiv.org/abs/1211.0310v1
- http://arxiv.org/abs/2403.13369v2