ISSN: 2320-5407

International Journal of Advanced Research

Publishers Name: Jana Publication and Research LLP

www.journalijar.com

REVIEWER'S REPORT

Manuscript No.: IJAR-53246

Date: 12/08/2025

Title: Open-Source Language Models: Effects of Parameter Scale and Prompting Strategies on NLP Tasks

Recommendation:
Accept as it is ....
[✓ Accept after minor revision...........]
Accept after major revision ....
Do not accept (*Reasons below*) ....

Rating	**Excel.**	**Good**	**Fair**	**Poor**
Originality	-			
Techn. Quality	-			
Clarity	-			
Significance	-			

Reviewer Name: Dr Gulnawaz

Detailed Review Report

The paper presents a well-structured, comprehensive empirical study examining the interaction between model size and prompting strategy in open-source LLMs, with a clear focus on practical implications for NLP tasks such as NLI and summarization. The experimental design is rigorous—using identical source material to control for domain variation—and evaluation methods include both automatic metrics and human ratings, which strengthens the validity of findings.

Strengths:

- **Original Contribution:** The "capacity–context alignment" principle is a valuable conceptual addition, offering practical guidance for LLM deployment.
- **Methodological Rigor:** The study carefully controls for confounding factors, ensuring that observed performance differences stem from model size, task structure, and prompt design.
- **Balanced Evaluation:** Combines quantitative and qualitative assessments, highlighting both statistical significance and practical interpretation.
- **Clarity:** The writing is clear, logically organized, and accessible even for readers not deeply specialized in prompt engineering.

Weaknesses / Areas for Improvement:

1. **Sample Size Limitation:** While the study's controlled corpus is methodologically sound, expanding beyond 15 articles could further strengthen statistical confidence and generalizability.
2. **Model Diversity:** The analysis focuses mainly on Flan-T5 and a few other open-source models; incorporating more architectures with extended context windows would broaden applicability.

REVIEWER'S REPORT

3. **Qualitative Error Analysis:** While some post-hoc inspection is included, a deeper qualitative exploration of failure cases—especially for summarization under few-shot settings—could offer more actionable prompt engineering insights.
4. **Visual Presentation:** Figures or charts summarizing cross-task performance differences could improve accessibility for readers.

Minor Points:

- Ensure consistent reporting of effect sizes alongside significance levels in the main text, not only in supplementary tables.
- Check for typographical consistency in metric names (e.g., ROUGE-L vs Rouge-L).