Grouped-Query Attention Latency Scaling in Mistral 7B for Multi-Turn Dialogue

Assignee Research

June 6, 2026

Abstract

This report synthesises findings from 10 peer-reviewed papers addressing the following research question: How does the latency per token of Mistral 7B's grouped-query attention scale against standard multi-head attention models during multi-turn dialogue evaluation on LongBench. 8 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.2/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Mixture of Attention Heads: Selecting Attention Heads Per Token. Research question: How does the latency per token of Mistral 7B's grouped-query attention scale against standard multi-head attention models during multi-turn dialogue evaluation on LongBench?.

2 Methodology

Systematic literature search across multiple databases yielded 10 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 4.2/10.

3 Results

10 papers retrieved. 8 claims extracted; 0 independently verified. Quality review score: 4.2/10.

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
MoA base outperforms Transformer base and Admin 6L-6L by at least 0.6 BLEU on WMT14 EnDe and WMT14 EnFr datasets.	×	0.02
MoA base outperforms Transformer big on the WMT14 EnFr dataset.	×	0.02
MoA base reaches comparable results with the Mixture of Attention Experts model (MAE-7) on the WMT14 EnDe dataset.	×	0.08
MoA outperforms the standard transformer model by 0.13 perplexity on WikiText-103 test data.	×	0.04
The performance of MoA improves with the increase of number of experts E and head size D, while the number of selected h	×	0.08
MoA consists of two major components: the routing network G and a group of N attention experts {E1, ..., EN}.	×	0.04
For each query vector qt, the routing network G selects a subset of k experts G(qt) $\subseteq${Ei} based on qt and assigns a weig	×	0.03
The output of the MoA is the weighted sum of the selected experts' outputs.	×	0.04

References

- http://arxiv.org/abs/2501.01123v1
- http://arxiv.org/abs/2410.11842v3
- http://arxiv.org/abs/2210.05144v1