Scaling Hidden State Size in S4 Models with Growth Bound Matrix Regularization for WikiText-103 Inference Throughput

Assignee Research

June 12, 2026

Abstract

When using Large Language Models (LLMs) to support Knowledge Graph Engineering (KGE), one of the first indications when searching for an appropriate model is its size. According to the scaling laws, larger models typically show higher capabilities. However, in practice, resource costs are also an important factor and thus it makes sense to consider the ratio between model performance and costs. The LLM-KG-Bench framework enables the comparison of LLMs in the context of KGE tasks and assesses their capabilities of understanding and producing KGs and KG queries. Based on a dataset created in an

1 Introduction

This paper examines: How do Scaling Laws Apply to Knowledge Graph Engineering Tasks? The Impact of Model Size on Large Language Model Performance. Research question: To what extent does scaling the hidden state size of S4 models with Growth Bound Matrix regularization impact inference throughput on the WikiText-103 language modeling benchmark relative to fraternal dropout?.

2 Methodology

Systematic literature search across multiple databases yielded 12 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 8.1/10.

3 Results

12 papers retrieved. 15 claims extracted; 14 independently verified. Quality review score: 8.1/10.

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 RdfConnectionExplain task requires finding the shortest connection between two nodes in a small Knowledge Graph.	✓	0.17
The RdfConnectionExplain task requires a basic understanding of serialization formats and RDF concepts.	✓	0.19
The RdfConnectionExplain task has four variations presenting the graph in JSON-LD, N-Triples, Turtle, or RDF/XML formats	✓	0.21
The expected answer format for the RdfConnectionExplain task is a list of IRIs representing the shortest path.	✓	0.21
The listTrimF1 metric for RdfConnectionExplain computes the F1-measure on trimmed list entries without leading or traili	✓	0.20
The textHttpF1 metric for RdfConnectionExplain computes the F1-measure on IRI-like answer parts starting with 'http://'.	✓	0.21
The RdfFriendCount task presents a small Knowledge Graph with nodes of one type and edges of one type.	✓	0.21
In the RdfFriendCount task, the LLM is asked to return the node with the most incoming edges.	✓	0.21
The RdfFriendCount task has four Knowledge Graph serialization format variations: JSON-LD, N-Triples, Turtle, and RDF/XM	✓	0.28
The RdfFriendCount task computes the f1 measure on the nodes found in the answer.	✓	0.23
The RdfSyntaxFixing task provides a Knowledge Graph with syntax errors.	×	0.12
The Qwen2-Instruct model family includes versions with 0.5B, 1.5B, 7B, 57B, and 72B parameters.	✓	0.21
The Qwen2.5-Instruct model family includes versions with 0.5B, 1.5B, 3B, 7B, 14B, 32B, and 72B parameters.	✓	0.25
The Meta-LLama-3.2-Instruct model family includes versions with 1B, 3B, and 70B parameters.	✓	0.22
The Microsoft-Phi-3.5-instruct model family includes versions with 3.8B and 42B parameters.	✓	0.17

References

- http://arxiv.org/abs/1711.00066v4
- http://arxiv.org/abs/2505.16276v1
- http://arxiv.org/abs/2403.09832v1