SOVEREIGN: What is the throughput and inference efficiency trade-off of GraphMETRO's node-based alignment mechanism versu

SOVEREIGN Research Kernel
Autonomous draft — Owner review required before publication

May 28, 2026

Abstract

The discovery of deep, steerable taxonomies in large text corpora is currently restricted by a trade-off between the surface-level efficiency of topic models and the prohibitive, non-scalable assignment costs of LLM-integrated frameworks. We introduce \textbf{LogiPart}, a scalable, hypothesis-first framework for building interpretable hierarchical partitions that decouples hierarchy growth from expensive full-corpus LLM conditioning. LogiPart utilizes locally hosted LLMs on compact, embedding-aware samples to generate concise natural-language taxonomic predicates. These predicates are then eva

1 Introduction

Analysis of: LogiPart: Local Large Language Models for Data Exploration at Scale with Logical Partitioning. Research goal: What is the throughput and inference efficiency trade-off of GraphMETRO's node-based alignment mechanism versus full Bayesian inference on large-scale language model benchmarks (e.g., GLUE, SuperGLUE) under covariate shift?.

2 Methodology

Multi-query arXiv search (4 parallel queries, Relevance-sorted). TF-IDF cosine semantic verification (bigrams, threshold=0.15). NIM nv-embedqa-e5-v5 (dim=1024) for semantic indexing. Tribunal v2: 3-role parallel review (SKEPTIC/VALIDATOR/SYNTHESIZER) with revision round if score $<$ 6.5.

3 Results

12 papers retrieved. 13 claims extracted, 0 verified. Tribunal: 0.3/10 → REJECT (revision_round=0). Policy: ESCALATE_TO_OWNER.

4 Uncertainties

NIM free tier latency varies. TF-IDF verification is a weak signal. arXiv Relevance ranking is query-dependent. Tribunal consensus is LLM-based and prompt-sensitive.

5 Extracted Claims

Claim	Verified	Confidence
Models below the 14B parameter threshold (3B/8B) exhibit near-zero NMI and Accuracy scores indistinguishable from random	×	0.05
QWEN3:14B and GPT-OSS:20B show stable, non-random alignment, with the bisection-only baseline reaching an F1 of 0.80.	×	0.03
LogiPart's symbolic predicates capture a distilled logic that generalizes beyond the specific examples found in the disc	×	0.07
The predicate quality assessment uses GEMINI-2.5-FLASH as an external judge model due to its larger parameter count and	×	0.05
For each predicate, the judge is presented with a fixed set of seven yes/no questions designed to evaluate redundancy, s	×	0.02
Each question is evaluated independently across five stochastic runs with temperature set to 0.4.	×	0.01
Questions q1 and q7 are negatively oriented; all remaining questions are positively oriented, where higher values indica	×	0.01
The protocol does not aim to replace human evaluation, but serves as a scalable proxy for assessing whether the discover	×	0.04
Statistical fit often correlates negatively with human-centric interpretability.	×	0.02
Natural Language Inference (NLI) allows documents to be evaluated against explicit semantic hypotheses.	×	0.12
LLMs can be used for zero-shot classification, but that implies a significant cost per document.	×	0.04
NLI is significantly faster and cheaper than LLMs for document classification.	×	0.03
Label propagation can be efficiently implemented on GPUs.	×	0.06

References

- https://arxiv.org/abs/2603.19563
- http://arxiv.org/abs/2206.02435v2

- https://arxiv.org/abs/2509.22211