SOVEREIGN: Does Vendi-RAG's diversity optimization maintain its latency and accuracy benefits when evaluated on the MMLU

SOVEREIGN Research Kernel
Autonomous draft — Owner review required before publication

May 29, 2026

Abstract

Neural sequence-to-sequence models have provided a viable new approach for abstractive text summarization (meaning they are not restricted to simply selecting and rearranging passages from the original text). However, these models have two shortcomings: they are liable to reproduce factual details inaccurately, and they tend to repeat themselves. In this work we propose a novel architecture that augments the standard sequence-to-sequence attentional model in two orthogonal ways. First, we use a hybrid pointer-generator network that can copy words from the source text via pointing, which aids a

1 Introduction

Analysis of: Get To The Point: Summarization with Pointer-Generator Networks. Research goal: Does Vendi-RAG's diversity optimization maintain its latency and accuracy benefits when evaluated on the MMLU benchmark for open-domain QA with a 1000-document corpus, versus fusion-in-decoder and REPLUG baselines?.

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

4 papers retrieved. 8 claims extracted, 7 verified. Tribunal: 7.2/10 → APPROVE (revision_round=0). Policy: AUTO_APPROVE.

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
Neural sequence-to-sequence models have provided a viable new approach for abstractive text summarization	✓	0.36
Neural sequence-to-sequence models are liable to reproduce factual details inaccurately	✓	0.28
Neural sequence-to-sequence models tend to repeat themselves	✓	0.22
The proposed hybrid pointer-generator network can copy words from the source text via pointing	✓	0.31
The pointer-generator network retains the ability to produce novel words through the generator	✓	0.28
Coverage is used to keep track of what has been summarized	×	0.10
The model was applied to the CNN / Daily Mail summarization task	✓	0.21
The model outperforms the current abstractive state-of-the-art by at least 2 ROUGE points	✓	0.23

References

- https://openalex.org/W7113598051
- https://doi.org/10.18653/v1/p17-1099
- https://doi.org/10.1145/3560815