Executive Summary: On the Reproducibility of Superhuman Performance in Agentic RAG Systems for Scientific Question Answering

Phong-Anh Nguyen Trinh

June 27, 2025

This study investigates the reproducibility of superhuman performance claims in agentic Retrieval-Augmented Generation (RAG) systems, specifically focusing on PaperQA2's performance on the LitQA benchmark for scientific question answering. The research methodology involved collecting and validating the LitQA dataset, which contains multiple-choice questions derived from recent biology papers published after September 2021 to ensure the questions could not be answered from pre-training data. The dataset preparation revealed that 15 papers predated the specified cutoff date and were consequently excluded to maintain experimental validity, with the final dataset comprising carefully curated questions that require deep understanding and reasoning rather than simple fact retrieval. The core experimental approach utilized PaperQA, an agentic RAG system built on the paperqa Python package, which operates through a sophisticated four-stage process: dynamic paper search and ingestion using semantic search capabilities that can query external databases like Semantic Scholar, evidence gathering with vector and hybrid search that combines dense embeddings with sparse keyword matching using techniques like TF-IDF and SPLADE-style models, LLM-based re-ranking and contextual summarization (RCS) that filters and summarizes retrieved chunks by assigning relevance scores, and answer generation with citation tracking that provides verifiable sources and enables citation traversal for expanding the knowledge base. To address the challenge of obtaining structured outputs from PaperQA for automated evaluation, the researchers developed a custom multi-agent wrapper system that standardizes both input and output formats, enabling consistent evaluation through the Inspect AI framework. This wrapper system employed a secondary agent to parse PaperQA's verbose outputs and extract single-letter answers, solving the critical problem of inconsistent response formatting that would otherwise prevent reliable automated assessment. The experimental design systematically varied key hyperparameters including the choice of language model (GPT-4o-Mini, GPT-4.1, and GPT-4-Turbo), text embedding models (OpenAI's text-embedding-3-small versus Google's text-embedding-004), answer cutoff settings (max_sources ranging from 5 to 15), evidence retrieval parameters (evidence_k from 5 to 30), and the critical RCS step, with each configuration tested three times to assess performance variability and ensure statistical robustness. The results demonstrate that RAG systems consistently outperform standalone language models, with all RAG configurations achieving superhuman precision levels (minimum 77.3%, maximum 89.5%) compared to the human benchmark of 73.8%. The best-performing configuration utilized GPT-4o-Mini with Google's text-embedding-004, achieving superior accuracy and precision that exceeded both the original paper's results and human performance, suggesting that newer, more cost-effective models

can deliver exceptional results when paired with advanced embedding technology. Notably, the study revealed that newer language models have diminished the importance of the RCS step, which was previously considered crucial for superhuman performance; removing RCS led to only a 6.1% accuracy reduction compared to the original paper's 12.8% reduction, indicating that modern LLMs are more capable of processing raw retrieved information without extensive re-ranking. The research also uncovered unexpected performance variations, particularly with GPT-4.1 underperforming compared to GPT-4-Turbo despite being a newer model, and identified that the primary performance bottleneck has shifted from evidence re-ranking to the initial retrieval phase, where embedding quality and model reasoning capabilities are paramount. This shift suggests that the field should focus more on improving retrieval mechanisms rather than post-processing steps, and highlights the importance of semantic understanding in the retrieval process. However, the most significant finding concerns the reproducibility challenges inherent in these systems: performance was observed to fluctuate based on API load during peak times, likely due to the Mixture-of-Experts architecture of models like GPT-4, and even exhibited sensitivity to local hardware specifications, introducing troubling variance that questions whether these systems can be considered robustly superhuman if their performance cannot be consistently replicated. These fluctuations occurred despite setting temperature to zero for deterministic behavior, indicating that external factors beyond user control significantly impact system reliability. The study also revealed that the performance impact of hyperparameter changes was less pronounced than in the original paper, with newer models achieving near-superhuman performance even with suboptimal settings, suggesting that model improvements have made systems more robust to configuration choices. The findings have significant implications for the broader AI research community, indicating that while the raw capabilities of RAG systems are impressive, their practical deployment in scientific research requires addressing fundamental infrastructure and evaluation challenges. The study concludes that while agentic RAG systems possess the raw capability to exceed human performance on specialized scientific question-answering tasks, their practical deployment in scientific research requires more stable infrastructure and transparent evaluation methodologies, as 'superhuman' performance remains a fragile achievement heavily dependent on specific implementation conditions and external factors beyond user control. The findings suggest that the AI research community should prioritize developing more robust evaluation frameworks and infrastructure that can account for these external variables, ensuring that performance claims are not only achievable but consistently reproducible across different environments and conditions. Furthermore, the research highlights the need for standardized benchmarking approaches that can account for the inherent variability in cloud-based AI systems, and suggests that future work should focus on developing more stable and predictable AI infrastructure that can support reliable scientific applications.