



A Study: Architecture of Retrieval-Augmented Generation (RAG) Integrating Orchestration and Multi-Tiered Evaluation

DOI: <https://doi.org/10.5281/zenodo.22338947>

Ms. V. V. Thakare^{1*}, Dr. V. S. Tondre²

*Corresponding Author: ¹Brijlal Biyani Science College, Amravati.

Department of Computer Science & Application, Affiliated to SGBAU Amravati, Maharashtra, India.

thakarevaishali3@gmail.com



ORCID iD: <https://orcid.org/0009-0001-1357-4527>:

²Brijlal Biyani Science College, Amravati.

Department of Computer Science & Application, Affiliated to SGBAU Amravati, Maharashtra, India.

vstondre@gmail.com

ABSTRACT

Large Language Models (LLMs) exhibition innovative appropriate facility, their distribution in mission-critical applications is constrained by parametric vision and static data cutoff limits. Retrieval-Augmented Generation (RAG) avoids these limitations by vigorously training propagation interpreters on non-parametric suggestion removed from peripheral document stores. Though, current applications regularly smart from cascading architectural disappointment methods: coarse semantic chunking, retrieval noise propagation, prompt-level context dilution, and non-deterministic evaluation. This paper reviews an end-to-end operational framework that combines modular pipeline orchestration via LangChain with high-granularity observability and runtime execution tracing via LangSmith. By separating concerns diagonally text ingestion, vector indexing, cross-encoder reranking, and structured prompt synthesis, our architecture enables targeted component isolation. Additionally, validate a multi-tiered valuation classification that decouples retrieval diagnostics from generation reliability. The anticipated architecture offers an observed design for methodically evolving, benchmark, and continuing production-grade grounded language systems.

KEYWORDS: Retrieval-Augmented Generation (RAG), Pipeline Orchestration, LLM (Large Language Model) Observability, Distributed Tracing, Grounded Evaluation, Multi-stage retrieval, Vector database.