State of the Art (SOTA) of Retrieval-Augmented Generation (RAG)

Your Name

Department of Computer Science
University Name

October 30, 2025

Overview

What is RAG?

- **Retrieval-Augmented Generation (RAG)** combines:
  - **Retriever:** Finds relevant external documents.
  - **Generator:** Produces output using both query and retrieved text.
- Proposed by **Facebook AI (Lewis et al., 2020)**.
- Bridges gap between static LLM knowledge and dynamic information sources.

Evolution of RAG Architectures

- **RAG-Sequence / RAG-Token (2020)** – Original architecture using DPR retriever and BART generator.
- **FiD (Fusion-in-Decoder, 2021)** – Combines multiple retrieved passages in the decoder.
- **REALM, Atlas, RETRO** – Enhanced retrieval-integrated models.
- **Open-RAG systems (2023–2024)** – Modular retriever-generator pipelines for domain customization.

Advancements in Retriever Models

- **Dense Retrieval:** Uses embeddings for semantic matching (e.g., DPR, ColBERT, Contriever).
- **Hybrid Retrieval:** Combines dense + sparse (BM25 + vector-based).
- **Re-ranking Models:** Cross-encoders refine top results (e.g., monoT5, CohereRerank).
- **Long-Context Retrievers:** Handle large-scale corpora with efficient indexing (FAISS, HNSW, ScaNN).

Improvements in Generation Component

- **Grounded Generation:** LLMs explicitly cite retrieved sources.
- **Context Compression:** Summarization or distillation before generation.
- **Retrieval-Aware Fine-Tuning:** Jointly training retriever + generator.
- **Instruction-tuned LLMs:** Models like GPT-4, LLaMA 3, and Claude 3 integrated with retrieval.

Recent State-of-the-Art RAG Systems

- **Atlas (Meta, 2022)** – Joint retriever-generator training; improved open-domain QA.
- **RETRO (DeepMind, 2022)** – Retrieval during pretraining with massive database.
- **ChatGPT + Bing (OpenAI, 2023)** – Online retrieval integration for factual grounding.
- **LlamaIndex, LangChain (2023–2024)** – Frameworks for domain-specific RAG pipelines.
- **RAG-Fusion, RePlug, HyDE (2024)** – Hybrid and hallucination-reduction approaches.

Evaluation of RAG Systems

- **Metrics:**
  - Faithfulness, Groundedness, and Factual Consistency.
  - Retrieval Precision / Recall.
- **Benchmarks:** NaturalQuestions, HotpotQA, KILT, and ELI5.
- **Human Evaluation:** Still required for assessing relevance and factual grounding.

Future Research Directions

- **Neural Indexing and Memory-Augmented Models.**
- **Dynamic and Streaming Retrieval for real-time updates.**
- **Retrieval-grounded reasoning and citation verification.**
- **Domain-specific adaptive retrievers.**
- **Reducing latency and computational cost.**

Thank You!

Questions?