Speculative RAG

gollnickdata.de

Speculative RAG

Introduction

Problem:

- Baseline RAG has issues
  - to answer questions that require information from different pieces of corpus
  - when complex insights are required
  - when it should understand complicated concepts

Solution:

1. Draft Generation
   - Multiple drafts from subset of retrieved docs are created
   - Aim: increased diversity, reduced redundancy
2. Draft Verification
   - Typically uses an LLM to evaluate drafts
   - Selects best answer

Speculative RAG

RAG approaches

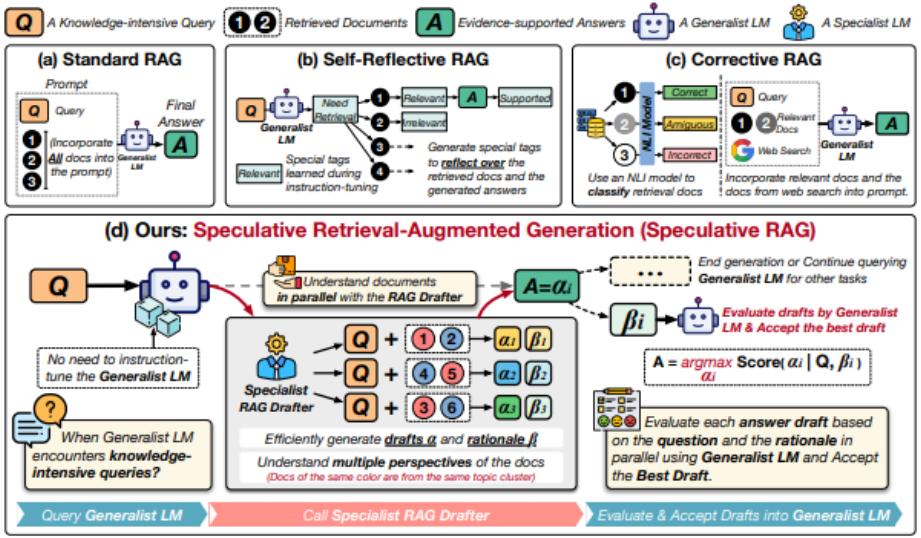


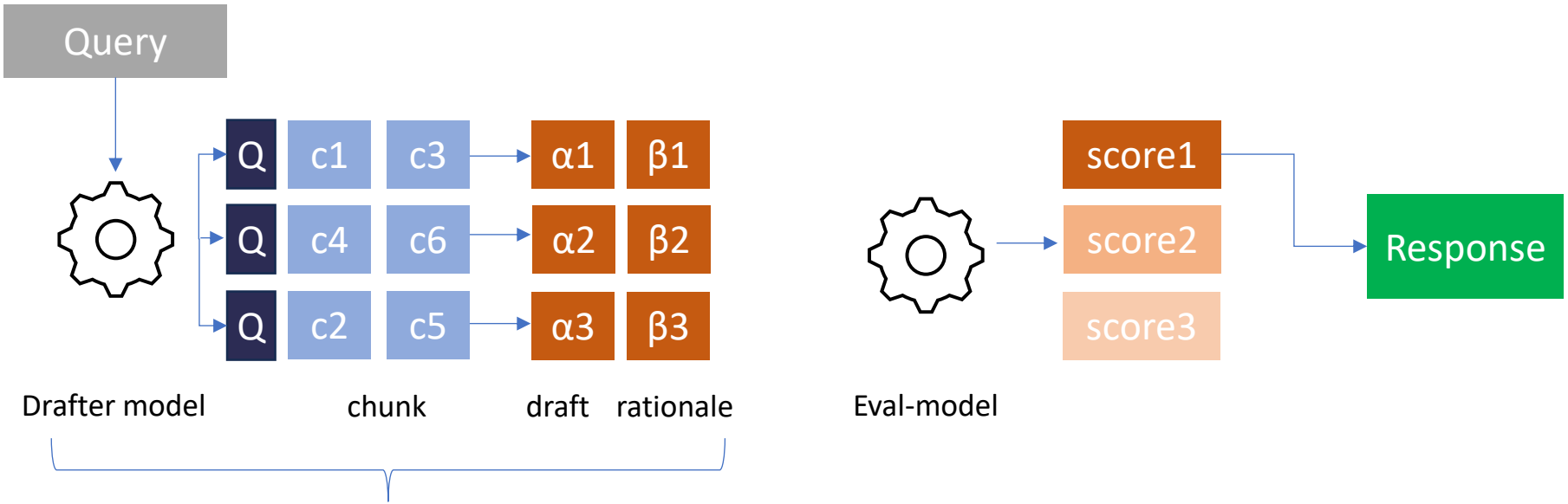
Figure 1: Illustration of different RAG approaches. Given a knowledge-intensive query Q and retrieved documents, (a) Standard RAG incorporates all documents into the prompt, increasing input length and slowing inference; (b) Self-Reflective RAG (Asai et al., 2023) requires specialized instruction-tuning of the general-purpose language model (LM) to generate specific tags for self-reflection; (c) Corrective RAG (Yan et al., 2024) employs an external retrieval evaluator to refine document quality, focusing solely on contextual information without enhancing reasoning capabilities; (d) In contrast, our proposed SPECULATIVE RAG leverages a larger generalist LM to efficiently verify multiple RAG drafts produced in parallel by a smaller, specialized LM. Each draft is generated from a distinct subset of retrieved documents, providing diverse perspectives on the evidence while minimizing the number of input tokens per draft.

SPECULATIVE RAG: ENHANCING RETRIEVAL AUGMENTED GENERATION THROUGH DRAFTING

Zilong Wang[1*] Zifeng Wang[2] Long T. Le[2] Huaixiu Steven Zheng[3]
Swaroop Mishra[3] Vincent Perot[3] Yuwei Zhang[1] Anush Mattapalli[4]
Ankur Taly[4] Jingbo Shang[1] Chen-Yu Lee[2] Tomas Pfister[2]
[1]University of California, San Diego [2]Google Cloud AI Research
[3]Google DeepMind [4]Google Cloud AI

Speculative RAG

How does it work?

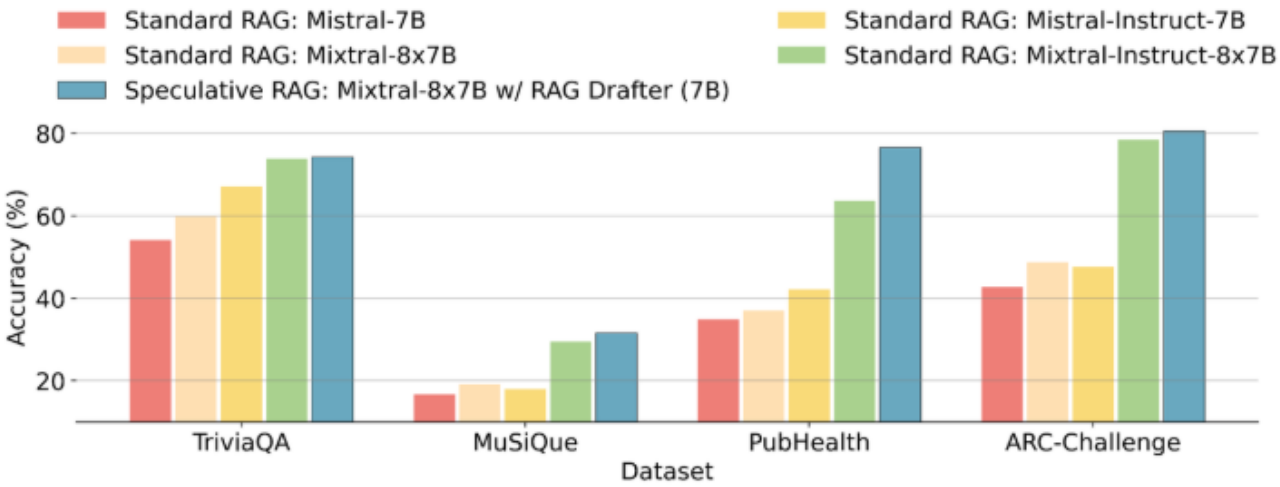


Drafter: generates drafts and rationales, Understands multiple perspectives

Evaluator: evaluates draft answers based on question and rationale, using generalist LLM, best draft accepted

Speculative RAG

Performance



Speculative RAG compared to the standard RAG with various backbone LLMs, including Mistral-7B, Mixtral-8x7B, Mistral-Instruct-7B, and Mixtral-Instruct-8x7B. On all datasets, Speculative RAG achieves the best performance.

Source: https://research.google/blog/speculative-rag-enhancing-retrieval-augmented-generation-through-drafting/