Speculative RAG: Enhancing Retrieval Augmented Generation through Drafting

Zilong Wang[1*] **Zifeng Wang**[2] **Long 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

Abstract

Retrieval augmented generation (RAG) combines the generative abilities of large language models (LLMs) with external knowledge sources to provide more accurate and up-to-date responses. Recent RAG advancements focus on improving retrieval outcomes through iterative LLM refinement or self-critique capabilities acquired through additional instruction tuning of LLMs. In this work, we introduce SPECULATIVE RAG – a framework that leverages a larger generalist LM to efficiently verify multiple RAG drafts produced in parallel by a smaller, distilled specialist LM. Each draft is generated from a distinct subset of retrieved documents, offering diverse perspectives on the evidence **while reducing input token counts per draft**. This approach enhances comprehension of each subset and mitigates potential **position bias over long context**. Our method accelerates RAG by delegating drafting to the smaller specialist LM, with the larger generalist LM performing a **single** verification pass over the drafts. Extensive experiments demonstrate that SPECULATIVE RAG achieves **state-of-the-art performance with reduced latency** on TriviaQA, MuSiQue, PubHealth, and ARC-Challenge benchmarks. It notably enhances accuracy by up to 12.97% while reducing latency by 51% compared to conventional RAG systems on PubHealth.

1 Introduction

Large language models (LLMs) have demonstrated remarkable success in question answering tasks (Brown et al., 2020; Achiam et al., 2023; Team et al., 2023). Trained on massive datasets, LLMs leverage their extensive parametric memory to generate seemingly plausible responses to user queries (Kojima et al., 2022; Kamalloo et al., 2023). However, when faced with knowledge-intensive questions demanding up-to-date information or obscure facts (Petroni et al., 2021), LLMs can struggle with factual inaccuracies and produce hallucinated content (Huang et al., 2023; Xu et al., 2024).

Retrieval Augmented Generation (RAG) has emerged as a promising solution to mitigate these issues. By incorporating information retrieved from an external database into the context (Gao et al., 2023b), RAG effectively reduces factual errors in knowledge-intensive tasks. This approach not only enables easy and efficient access to vast databases but also facilitates timely and accurate knowledge integration Due to the inherent limitations in the precision of current dense retrievers and the vastness of knowledge required to answer complex questions (Chen et al., 2022), RAG systems typically retrieve multiple documents to ensure the inclusion of all necessary information in the context (Petroni et al., 2021). This practice inevitably increases the length of the input to the LLMs,

*Work done while the author was a student researcher at Google Cloud AI Research. Correspondence to: Zilong Wang <zlwang@ucsd.edu>, Chen-Yu Lee <chenyulee@google.com>

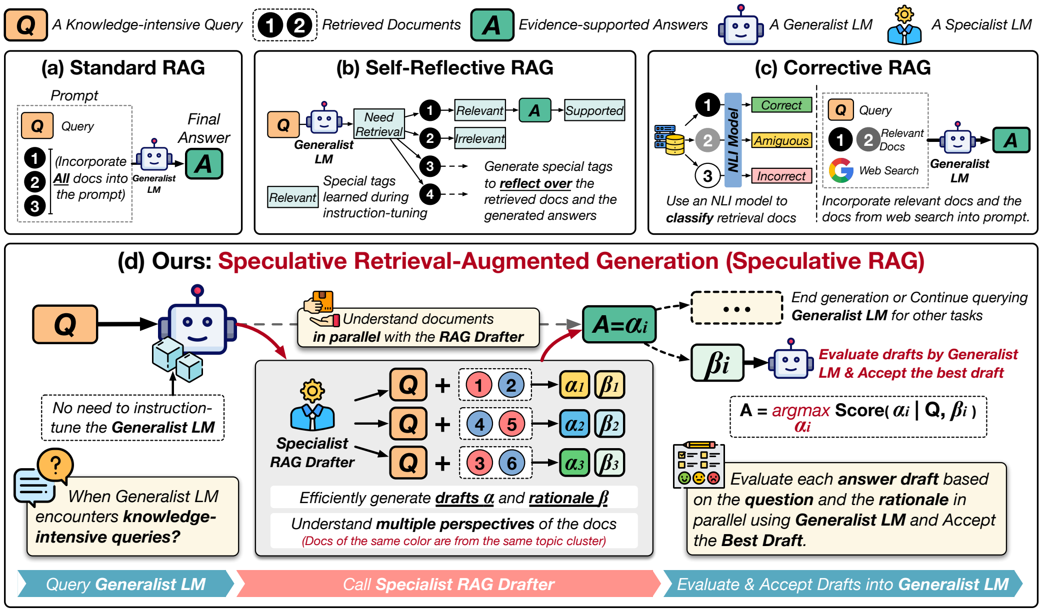

Figure 1: Illustration of different RAG approaches. Given a knowledge-intensive query Q and retrieved documents, (a) Standard RAG incorporates all documents into the prompt, increasing input length and slowing inference; (b) Self-Reflective RAG (Asai et al., 2023) requires specialized instruction-tuning 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.

presenting significant challenges, particularly since encoding lengthy retrieved documents incurs additional latency and require more complex reasoning. Recent studies have explored ways to extend the context length limit of LLMs (Ding et al., 2023; Reid et al., 2024; Ma et al., 2024), yet achieving well-grounded reasoning over extended contexts remains an open question (Liu et al., 2024; Li et al., 2024). Consequently, striking a balance between efficiency and effectiveness in RAG has become a central research question in the literature. Existing work on RAG systems primarily concentrates on improving the quality of contextual information in retrieval outcomes, but often neglecting the latency issues associated with these systems (Ma et al., 2023; Baek et al., 2023; Yan et al., 2024; Xie et al., 2023; Asai et al., 2023; Feng et al., 2023). These methods typically rely on multiple refinement iterations or customized instruction-tuning for self-critique abilities. Integrating such enhancements into generic LMs requires additional training or increased latency, posing practical challenges in real-world applications.

To this end, we introduce SPECULATIVE RAG, a RAG framework designed to offload computational burden to a smaller, specialist LM that serves as an efficient and robust RAG module for existing generalist LMs. Inspired by Speculative Decoding (Leviathan et al., 2023; Chen et al., 2023a; Xia et al., 2024), which accelerates auto-regressive LM inference by concurrently generating multiple draft tokens with a smaller model and verifying them in parallel with the base model, our approach adapts this concept to RAG.

In SPECULATIVE RAG, we partition retrieved documents into subsets for drafting answer candidates. We cluster the retrieved documents by content similarity and sample one document from each cluster to form a subset, minimizing redundancy and maximizing diversity. These document subsets are then fed to multiple instances of the RAG module, which generate draft answers with corresponding rationales in parallel. This smaller, specialized RAG module, excels at reasoning over retrieved documents and can rapidly produce accurate responses. Subsequently, the generalist LM bypasses the detailed review of potentially repetitive documents, focusing instead on validating the drafts against the rationales to determine the most accurate answer. We utilize the strong language modeling capabilities of generalist LMs, calculating the conditional generation probability of the answer drafts and rationales as a confidence score. Our key contributions are:

- We introduce a novel RAG framework that employs a smaller specialist RAG drafter to generate high-quality draft answers. Each draft is derived from a distinct subset of retrieved documents, offering diverse perspectives while reducing input token counts per draft.
- The generalist LM, operating with the RAG drafter, requires no additional tuning. It simply verifies and integrates the most promising draft into the final answer. This approach enhances comprehension of each subset and mitigates potential lost-in-the-middle (Liu et al., 2024) phenomenon.
- Our method significantly accelerates RAG by delegating drafting to the smaller specialist LM, with the larger generalist LM performing a single, unbiased verification pass over the drafts in parallel. Extensive experiments on 4 free-form question-answering and closed-set generation benchmarks demonstrate the superior effectiveness and efficiency of the method.

2 Related Works

Retrieval Augmented Generation Retrieval Augmented Generation (RAG) enhances LLMs by retrieving relevant documents from external databases and incorporating them into the generation process (Gao et al., 2023b; Lewis et al., 2020; Khandelwal et al., 2020; Izacard & Grave, 2021; Luo et al., 2023a). Recent work has primarily focused on enabling LLMs to understand when and what to retrieve (Ma et al., 2023; Chen et al., 2023b; Jiang et al., 2023b; Schick et al., 2024), or designing approaches to better utilize contexts (Yu et al., 2023; Yoran et al., 2023; Wang et al., 2023b; Sarthi et al., 2024; Baek et al., 2023; Xu et al., 2023; Kim et al., 2024). Among them, SAIL (Luo et al., 2023a) fine-tunes a pre-trained LLM on web search data to filter irrelevant contents. Self-Reflective RAG (Asai et al., 2023) introduces reflection tokens to guide retrieval and annotation in instruction-tuning datasets. However, both approaches require additional instruction-tuning of generic LLMs, which is resource-intensive and may lead to forgetting or over-fitting (Luo et al., 2023b). Furthermore, long context with retrieved documents can suffer from computational inefficiency and position bias (Liu et al., 2024). Corrective RAG (Yan et al., 2024) on the other hand proposes a lightweight retrieval evaluator, but it lacks the capability for high-level reasoning. In contrast, our proposed SPECULATIVE RAG addresses these limitations by leveraging a smaller RAG drafter model to efficiently understand diverse perspectives in retrieval results and generate drafts for the generalist LMs to verify and integrate.

Speculative Decoding Speculative decoding (Stern et al., 2018; Xia et al., 2023; Chen et al., 2023a; Leviathan et al., 2023; Xia et al., 2024) aims to reduce auto-regressive decoding latency through a draft-then-verify paradigm. This involves drafting multiple future tokens with a small model and verifying them in parallel with the target model (Xia et al., 2024). The draft model is typically either an independent model from the same series (Leviathan et al., 2023; Chen et al., 2023a) or the target model itself (Zhang et al., 2023a; Cai et al., 2024). Our approach extends this concept from token-level drafting to answer-level drafting. In contrast to traditional verification criteria (Stern et al., 2018; Xia et al., 2023; Leviathan et al., 2023; Chen et al., 2023a; Miao et al., 2024), which accept or reject tokens based on their generation probabilities, we leverage language modeling objectives to directly assess the confidence of entire answer drafts.

3 Speculative Retrieval Augmented Generation through Drafting

Problem Formulation In knowledge intensive tasks, each entry can be represented as (Q, D, A), where Q is a question or statement that requires additional knowledge; $D = \{d_1, ..., d_n\}$ is a set of n documents retrieved from the database; A is the expected answer. Particularly, in question answering tasks, Q and A are the question and the expected answer in natural language form; in the statement verification tasks, Q is a statement and $A \in \{\texttt{True}, \texttt{False}\}$ is a Boolean value indicating the statement's correctness; in the multiple choice tasks, Q is a question with a few options and $A \in \{\mathsf{A}, \mathsf{B}, \mathsf{C}, ...\}$ is the index of the correct answer. The objective of a RAG system is to generate a fluent response containing the expected answer or select the expected answer from the provided options based on the context provided by the retrieved supporting documents.

3.1 Overview

We introduce Speculative Retrieval Augmented Generation (SPECULATIVE RAG), as illustrated in Figure 1. We aim at enhancing the reasoning ability of LLMs over retrieved documents without compromising processing speed. Instead of relying on brute-force parameter scaling or instruction-tuning an entire LM to handle knowledge-intensive tasks, we propose a divide-and-conquer approach. We utilize **a smaller specialist LM**, **the RAG drafter**, to rapidly generate multiple answer drafts based on retrieved results. Then, **a larger generalist LM**, **the RAG verifier**, assesses these drafts, selects the best one based on its rationale, and integrates it into the generation results.

Algorithm 1: SPECULATIVE RAG

Data: $(Q, D = \{d_i\}_i^n)$ is the question and n retrieved documents; m subsets, each containing k documents, are sampled from D; k also corresponds to the number of clusters during clustering.

Result: $\hat{A}$ is the predicted answer to the question.

1 **Function** `Speculative RAG` (Q, D, m, k):

2 $\quad \{\boldsymbol{c}_1, \boldsymbol{c}_2, ..., \boldsymbol{c}_k\} \xleftarrow{\text{K-Means}} \mathcal{C}(d_1, ..., d_n|Q)$ ▷ Cluster the documents into k groups using an embedding model $\mathcal{C}$.

3 $\quad \Delta \leftarrow \{\}$

4 $\quad$ **repeat**

5 $\quad\quad \boldsymbol{\delta}_j \leftarrow \{\}$ ▷ Construct a subset of the retrieved documents $\boldsymbol{\delta}_j$

6 $\quad\quad$ **for** $\boldsymbol{c}_i \in \{\boldsymbol{c}_1, ..., \boldsymbol{c}_k\}$ **do**

7 $\quad\quad\quad \boldsymbol{\delta}_j = \boldsymbol{\delta}_j \cup \{\texttt{random.sample}(\boldsymbol{c}_i)\}$ ▷ Sample one document from each cluster $\boldsymbol{c}_i$ into subset $\boldsymbol{\delta}_j$.

8 $\quad\quad$ **end**

9 $\quad\quad \Delta = \Delta \cup \{\boldsymbol{\delta}_j\}$

10 $\quad$ **until** $|\Delta| = m$ ▷ Repeat the sampling until there are m unique subsets in total.

11 $\quad$ **for** $\boldsymbol{\delta}_j \in \Delta$ **do *in parallel*** ▷ Process m subsets in parallel.

12 $\quad\quad \alpha_j, \beta_j \leftarrow \mathcal{M}_{\text{Drafter}}.\texttt{generate}(Q, \boldsymbol{\delta}_j)$ ▷ Generate the **draft** α and **rationale** β with $\mathcal{M}_{\text{Drafter}}$.

13 $\quad\quad \rho_j \leftarrow \mathcal{M}_{\text{Verifier}}.\texttt{score}(\alpha_j|Q, \beta_j)$ ▷ Compute the confidence score ρ with $\mathcal{M}_{\text{Verifier}}$.

14 $\quad$ **end**

15 $\quad \hat{A} \leftarrow \arg\max_{\alpha_j} \rho_j$ ▷ Select the one with the highest score as the final answer.

16 **return** $\hat{A}$

Specifically, as shown in Algorithm 1, we first cluster the retrieved documents with regard to their relation to the posed question, where each cluster represents one perspective in the retrieval results (Line 2). Then we sample one document from each cluster into a subset so the documents in this subset covers the multiple perspectives in the retrieval results. We aim at minimizing redundancy and increase the diversity of the documents (Line 5 to 8). We denote one subset as $\boldsymbol{\delta} \subset D$ that contains retrieved documents with diverse contents and multiple perspectives in the retrieval results. Then, we distribute each subset $\boldsymbol{\delta}$ to a RAG drafter endpoint $\mathcal{M}_{\text{Drafter}}$ with the posed question Q to generate the answer draft α and the rationale β in parallel (Line 12). The RAG drafter is instruction-tuned to be a specialist in understanding the retrieved documents and produce rationales that are faithful to the input documents. It is smaller than generalist LMs, and its parallel processing further ensures high efficiency. For each draft-rationale pair (α, β) from $\mathcal{M}_{\text{Drafter}}$, we compute a confidence score with the generalist LM $\mathcal{M}_{\text{Verifier}}$ based on the question Q and corresponding rationale β (Line 13). It is worth mentioning that $\mathcal{M}_{\text{Verifier}}$ does not need to be instruction-tuned since we leverage its language modeling ability already learned during pre-training. Meanwhile, $\mathcal{M}_{\text{Verifier}}$ can verify the drafts based on the informative rationale provided by $\mathcal{M}_{\text{Drafter}}$ instead of processing tedious or possibly redundant retrieved documents. Finally, we select the answer draft with the highest confidence score as the final answer and integrate it into the generation results of the generalist LM (Line 15).

3.2 Specialist RAG Drafter

Instead of tuning a large generalist LM for the RAG scenario, we leverage a smaller specialist LM, $\mathcal{M}_{\text{Drafter}}$, to understand retrieved documents. $\mathcal{M}_{\text{Drafter}}$ is specialized in answering the given question based on the supporting documents and not expected to cope with general problems. It serves as a RAG module for the generalist LMs when solving knowledge-intensive tasks. We train $\mathcal{M}_{\text{Drafter}}$ to generate both the answer draft and the rationale to better understand the contextual documents.

Instruction Tuning Given a triplet (Q, A, D), where Q is a general query, A is the response, and D is a retrieved supporting document, we augment it with the rationale of the response A based on the document D. We denote the rationale as E which extracts essential information from the document and explains why the response is reasonable to the query concisely (Hsieh et al., 2023) so it is of shorter length and delivers information coherent with the original document. We leverage relatively strong LMs to automatically synthesize the rationale E for each triplet. Specifically, we directly query the strong LM to understand the knowledge from the document and provide the intermediate rationale between the instruction and response. Refer to Appendix A for detailed prompts. After generating the rationale, we finetune a pre-trained LM using the standard language modeling objective, maximizing the likelihood: $\mathbb{E}_{(Q,A,D,E)} \log P_{\mathcal{M}_{\text{Drafter}}}(A, E \mid Q, D)$, where (Q, A, D, E) is an augmented entry in the dataset; $P_{\mathcal{M}_{\text{Drafter}}}(A, E \mid Q, D)$ is the probability of generating the response and rationale based on the query and document. We use this instruction-tuned model as the specialist RAG drafter which learns to generate a well-grounded response and rationale given the query and relevant documents.

Multi-Perspective Sampling For each knowledge-intensive question, we retrieve a set of documents from the database using the posed question as the retrieval query. These documents may contain diverse content due to the ambiguity inherent in the query. To minimize redundancy and enhance diversity of the document subsets used for generating answer drafts, we employ a multi-perspective sampling strategy. We first cluster the documents into a few topics using an instruction-aware embedding model (Peng et al., 2024) and the K-Means clustering algorithm (Jin & Han, 2011).

$$\begin{aligned}
\mathtt{emb}(d_1), ..., \mathtt{emb}(d_n) &= \mathcal{E}(d_1, ..., d_n|Q) \\
\{\boldsymbol{c}_1, ..., \boldsymbol{c}_k\} &= \mathtt{K\text{-}Means}(\mathtt{emb}(d_1), ..., \mathtt{emb}(d_n)) \\
\boldsymbol{\delta} &= \left\{\mathtt{random.sample}(\boldsymbol{c}) \text{ for } \boldsymbol{c} \in \{\boldsymbol{c}_i\}_1^k\right\}
\end{aligned}$$

where $\mathcal{E}$ is an instruction-aware embedding model which embeds a string with regard to a provided instruction (the posed question Q); $\mathtt{emb}(d_i)$ is the embedding for the retrieved document d_i; $\boldsymbol{c}_j$ is a cluster of retrieved documents with similar topics and contents; k is a hyper-parameter that controls the number of clusters. We sample one document from each cluster into a document subset $\boldsymbol{\delta}$ so each subset contains k documents of diverse contents. In total, we construct m subsets for parallel inference with the RAG drafter.

RAG Drafting We run $\mathcal{M}_{\text{Drafter}}$ over the m document subsets and produce corresponding answer drafts. Refer to Appendix B for detailed prompt. We incorporate each document subset into the prompt and query $\mathcal{M}_{\text{Drafter}}$ for responses. We obtain m drafts as the answer candidates and each draft is grounded based on the multiple perspectives in the retrieval results. Specifically, given a document subset $\boldsymbol{\delta}_j = \{d_{j_1}, .., d_{j_k}\}$, we query $\mathcal{M}_{\text{Drafter}}$ in parallel with the following prompt for the answer draft and rationale: $Q, d_{j_1}, ..., d_{j_k} \rightarrow \alpha_j, \beta_j$, where the prompt contains the posed question Q along with the document subset; the generation result contains the answer draft α and the rationale β. We denote the conditional generation probability as $\rho_{\text{Draft},j} = P(\beta_j|Q, d_{j_1}, ..., d_{j_k}) + P(\alpha_j|Q, d_{j_1}, ..., d_{j_k}, \beta_j)$, which measures the reliability of generating rationales and the confidence in producing answer drafts.

3.3 Generalist RAG Verifier

After generating drafts and the rationale from the RAG drafter $\mathcal{M}_{\text{Drafter}}$, we evaluate them by a generalist LM $\mathcal{M}_{\text{Verifier}}$ to filter out the less reliable drafts and select the best answer. The generalist LM can be any off-the-shelf pre-trained LM. We only consider the draft-rationale pair (α, β) and skip the tedious and redundant retrieval results. We resort to the language modeling ability of the generalist LM to rank and select the draft-rationale pairs.

Evaluation Scores First, we calculate the **self-consistency score** by determining the conditional probability of generating a draft-rationale pair given the question, $\rho_{\text{Self-contain}} = P(\alpha, \beta|Q)$. This score helps assess whether the draft and rationale are self-consistent in the context of the question. Given the characteristics of language modeling, a self-consistent draft-rationale pair is expected to yield a higher probability. Furthermore, we incorporate a self-reflection statement R that prompts $\mathcal{M}_{\text{Verifier}}$ to assess the reliability of an answer draft (e.g. "Do you think the rationale supports the answer, yes or no?"). We define the **self-reflection score** as $\rho_{\text{Self-reflect}} = P(\texttt{"Yes"}|Q, \alpha, \beta, R)$ where we compute the conditional probability of the positive answer ("Yes") to the self-reflection statement.

Computation Method We can efficiently compute the self-consistency and self-reflection scores within one forward pass of $\mathcal{M}_{\text{Verifier}}$. Given a question Q and a draft-rationale pair (α, β), we construct a prompt $[Q, \alpha, \beta, R, \texttt{"Yes"}]$, where R is the self-reflection statement. We encode the prompt with $\mathcal{M}_{\text{Verifier}}$, and acquire the probability of each token conditioned on the previous tokens $P(t_i|t_{<i})$. We leverage this auto-regressive feature and aggregate the probability of the relevant

tokens to compute the self-consistent score $\rho_{\text{Self-contain}}$ and self-reflection score $\rho_{\text{Self-reflect}}$.

$$\xrightarrow{Q, \overbrace{\alpha, \beta}^{\rho_{\text{SC}}}, R, \overbrace{\texttt{"Yes"}}^{\rho_{\text{SR}}}} \Rightarrow \begin{cases} \rho_{\text{SC}} = \prod_{t_i \in \alpha} P(t_i | t_{<i}) \cdot \prod_{t_i \in \beta} P(t_i | t_{<i}) \\ \rho_{\text{SR}} = \prod_{t_i \in \texttt{"Yes"}} P(t_i | t_{<i}) \end{cases}$$

Finally, we produce the final score, $\rho_j = \rho_{\text{Draft},j} \cdot \rho_{\text{SC},j} \cdot \rho_{\text{SR},j}$, and then select the most reliable answer as the final answer to the question $\hat{A} = \arg\max_{\alpha_j} \rho_j$.

4 Experiments

We evaluate our proposed SPECULATIVE RAG on four public retrieval augmented generation benchmarks: TriviaQA (unfiltered) (Joshi et al., 2017), MuSiQue (Trivedi et al., 2022), PubHealth (Zhang et al., 2023b), and ARC-Challenge (Clark et al., 2018). TriviaQA and MuSiQue are challenging open-domain question answering datasets where RAG systems are required to answer questions on factual knowledge. TriviaQA typically requires one accurate piece of evidence from the documents, whereas MuSiQue demands multiple documents to construct a multi-hop reasoning chain. Following previous works (Guu et al., 2020; Asai et al., 2023; Yan et al., 2024), we evaluate performance of the free-form generation based on whether gold answers are contained within the generated response or not. PubHealth and ARC-Challenge are closed-set generation datasets. PubHealth is a dataset of medical claims spanning a variety of biomedical subjects and it requires the RAG system to verify a given claim based on the retrieved documents. ARC-Challenge introduces a multi-choice question answering dataset, composed of science exam questions from grade 3 to grade 9. For closed-set generation tasks, we use accuracy metrics to evaluate whether the generated answers match the ground truth.

4.1 Baselines

Standard RAG For standard RAG, we incorporate all the retrieved documents into the prompt as contextual information. Refer to Appendix C for detailed prompts. We run standard RAG experiments on off-the-shelf LLMs including Mistral$_{\text{7B}}$, Mistral-Instruct$_{\text{7B}}$ (Jiang et al., 2023a), Mixtral$_{\text{8x7B}}$, Mixtral-Instruct$_{\text{8x7B}}$ (Jiang et al., 2024), and Alpaca$_{\text{7B}}$ (Dubois et al., 2024). We also include the performance of Toolformer (Schick et al., 2024) and SAIL (Luo et al., 2023a) which are originally reported from Asai et al. (2023). Toolformer$_{\text{7B}}$ is an LM instruction-tuned to use tools including a search engine, and SAIL$_{\text{7B}}$ is an LM instruction-tuned on the Alpaca instruction tuning set augmented with search results from different sources such as DuckDuckGo and Wikipedia.

Self-Reflective RAG and Corrective RAG Self-Reflective RAG (Self-RAG) (Asai et al., 2023) and Corrective RAG (CRAG) (Yan et al., 2024) are more advanced RAG systems that enhances the quality of contextual information in the retrieval results. CRAG introduces an external evaluator to assess the quality of retrieved documents, and to refine them before the response generation. Self-RAG instruction-tunes an LM to generate special self-refection tags. These tags guides the LM to dynamically retrieve documents when necessary, critique the retrieved documents relevance before generating responses. Self-CRAG is to apply the Self-RAG approach on the refined documents of CRAG. We adopt the same backbone LLMs across all methods as our proposed SPECULATIVE RAG for fair comparisons.

4.2 Experiment Settings

In our experiments, we utilize Mistral$_{\text{7B}}$ (v0.1) as our base LM for the RAG drafter. For RAG verifier, we employ either Mistral$_{\text{7B}}$ (v0.1) or Mixtral$_{\text{8x7B}}$ (v0.1) without any fine-tuning, denoted as $\mathcal{M}_{\text{Verifier-7B}}$ or $\mathcal{M}_{\text{Verifier-8x7B}}$. We pre-compute embeddings of retrieved documents using a lightweight instruction-aware embedding model InBedder$_{\text{Roberta}}$ (Peng et al., 2024) as part of the retrieval process. Inference is conducted using the vLLM framework (Kwon et al., 2023) with greedy decoding (temperature = 0). We adopt the same experiment settings from Asai et al. (2023) and include a more challenging benchmark, MuSiQue (Trivedi et al., 2022). Our focus is on RAG reasoning rather than evidence citation, so we omit the other two long-form generation benchmarks, Biography (Min et al., 2023) and ALCE-ASQA (Gao et al., 2023a). On TriviaQA, PubHealth, and ARC-Challenge, we retrieve top 10 documents and generate 5 drafts per query ($m = 5$), with each draft based on a subset of 2 documents ($k = 2$). For the MuSiQue dataset, we retrieve top 15 documents and generate 10

Table 1: Retrieval augmentation generation results on TriviaQA, MuSiQue, PubHealth, and ARC-Challenge (ARC-C). (*We use the RAG drafter's generation probability ρ_{Draft} as the confidence score for selecting drafts when we use it alone; † indicates numbers reported in Asai et al. (2023); − denotes numbers that are not reported by the original papers or are not applicable; ‡we use Mistral$_{7B}$ or Mixtral$_{8x7B}$ as the RAG verifier, and denote them as $\mathcal{M}_{\text{Verifier-7B}}$ or $\mathcal{M}_{\text{Verifier-8x7B}}$.)

RAG Method	Free-form		Closed-set	
	TriviaQA	MuSiQue	PubHealth	ARC-C
Standard RAG				
Mistral$_{7B}$ (Jiang et al., 2023a)	54.15	16.71	34.85	42.75
Mixtral$_{8x7B}$ (Jiang et al., 2024)	59.85	19.16	37.08	48.72
Mistral-Instruct$_{7B}$ (Jiang et al., 2023a)	67.11	17.99	42.15	47.70
Mixtral-Instruct$_{8x7B}$ (Jiang et al., 2024)	73.91	29.42	63.63	78.41
Alpaca$_{7B}$ (Dubois et al., 2024)†	64.1	-	40.2	48.1
Toolformer$_{6B}$ (Schick et al., 2024)†	48.8	-	-	-
SAIL$_{7B}$ (Luo et al., 2023a)†	-	-	69.2	48.4
Self-Reflective RAG & Corrective RAG				
CRAG$_{\text{Mistral-7B}}$ (Yan et al., 2024)	-	-	59.04	74.87
Self-RAG$_{\text{Mistral-7B}}$ (Asai et al., 2023)	64.84	21.72	72.44	74.91
Self-CRAG$_{\text{Mistral-7B}}$ (Yan et al., 2024)	-	-	72.85	75.26
Our Speculative RAG				
$\mathcal{M}_{\text{Drafter-7B}}$*	71.11	27.89	75.58	74.49
$\mathcal{M}_{\text{Verifier-7B}}$‡ + $\mathcal{M}_{\text{Drafter-7B}}$	73.91	31.03	75.79	76.19
$\mathcal{M}_{\text{Verifier-8x7B}}$‡ + $\mathcal{M}_{\text{Drafter-7B}}$	**74.24**	**31.57**	**76.60**	**80.55**

drafts for each query ($m = 10$), each using a subset of 6 documents due to more complex reasoning. Further details regarding instruction-tuning can be found in Appendix E.

4.3 Main Results

We compare SPECULATIVE RAG with standard RAG approaches, as well as the more advanced Self-Reflective RAG and Corrective RAG on four datasets: TriviaQA, MuSiQue, PubHealth, and ARC-Challenge. We report the performance of $\mathcal{M}_{\text{Drafter-7B}}$ when used alone or paired with the RAG verifier (e.g. $\mathcal{M}_{\text{Verifier-7B}}$, $\mathcal{M}_{\text{Verifier-8x7B}}$). Following prior work (Asai et al., 2023; Yan et al., 2024), we report accuracy as the performance metric.

Superior Performance over Baselines Table 1 demonstrates that SPECULATIVE RAG consistently outperforms all baselines across all four benchmarks. Particularly, $\mathcal{M}_{\text{Verifier-8x7B}}$ + $\mathcal{M}_{\text{Drafter-7B}}$ surpasses the most competitive standard RAG model, Mixtral-Instruct$_{8x7B}$, by 0.33% on TriviaQA, 2.15% on MuSiQue, 12.97% on PubHealth, and 2.14% on ARC-Challenge. With a comparable number of instruction-tuned parameters, $\mathcal{M}_{\text{Verifier-7B}}$ + $\mathcal{M}_{\text{Drafter-7B}}$ outperforms all Self-Reflective and Corrective RAG methods, and $\mathcal{M}_{\text{Drafter}}$ alone can surpass these baselines in most settings.

Effective Instruction Tuning for RAG Drafter Our instruction tuning is effective in enhancing the reasoning ability of the drafter model (Hsieh et al., 2023), as we observe a remarkable performance improvement comparing Mistral$_{7B}$ and $\mathcal{M}_{\text{Drafter-7B}}$. Moreover, the performance of Mixtral$_{8x7B}$ significantly improves when paired with the instruction-tuned RAG drafter $\mathcal{M}_{\text{Drafter-7B}}$, showing gains of 14.39% on TriviaQA, 12.41% on MuSiQue, 39.52% on PubHealth, and 31.83% on ARC-Challenge. Similar improvements are observed with Mistral$_{7B}$ as well. For Mistral$_{7B}$, we observed improvements of 19.76% on TriviaQA, 14.32% on MuSiQue, 40.94% on PubHealth, and 33.44% on ARC-Challenge. We attribute these improvements to the superior reasoning capabilities of the RAG drafter over the retrieved documents in SPECULATIVE RAG. By minimizing the redundancy in the sampled documents, the RAG drafter generates higher quality answer drafts based on diverse perspectives from the retrieval results.

Reliable Scoring by RAG Verifier The reliable draft verification by the generalist LM also contributes to the enhanced performance. The performance improves remarkably comparing $\mathcal{M}_{\text{Drafter-7B}}$ and $\mathcal{M}_{\text{Verifier-7B}}$ + $\mathcal{M}_{\text{Drafter-7B}}$. The instruction-tuned RAG drafter is specialized in generating answer drafts based on the retrieved documents while the language modeling capabilities of generic LMs are leveraged to validate each draft in light of its rationale. This method is both effective and easy to implement, showcasing the effectiveness of this verification approach.

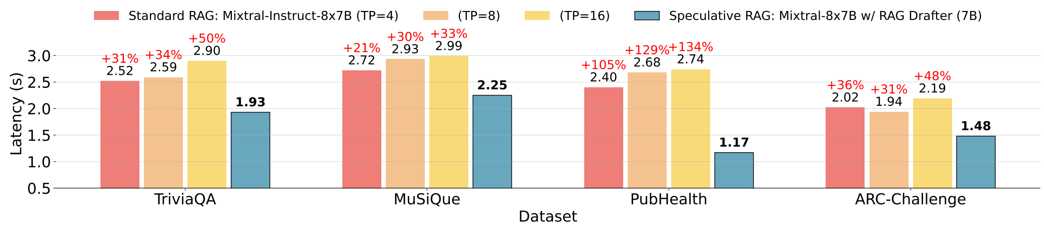

Figure 2: Latency analysis of Standard RAG: Mixtral-Instruct$_{8x7B}$ with tensor parallelism and SPECULATIVE RAG: $\mathcal{M}_{\text{Verifier-8x7B}}$ + $\mathcal{M}_{\text{Drafter-7B}}$ on TriviaQA, MuSiQue, PubHealth, and ARC-Challenge. The latency difference between Standard RAG and SPECULATIVE RAG is highlighted in *red* (+x%). TP indicates the tensor parallelism size when running Mixtral-Instruct$_{8x7B}$ for Standard RAG. The latency varies across different datasets due to different retrieved document lengths. SPECULATIVE RAG encodes the retrieved documents in parallel and generates answer drafts with a smaller RAG drafter. This significantly improves the efficiency over Standard RAG.

4.4 Latency Analysis

We analyze the latency of Standard RAG and our SPECULATIVE RAG on TriviaQA, MuSiQue, PubHealth, and ARC-Challenge. We randomly sample 100 cases from each dataset and report the average time cost for each case, as shown in Figure 2. To simulate real-world application scenarios, we process cases individually without batching. As representative example, we run $\mathcal{M}_{\text{Verifier-8x7B}}$ + $\mathcal{M}_{\text{Drafter-7B}}$ for SPECULATIVE RAG and Mixtral-Instruct$_{8x7B}$ for Standard RAG, as these demonstrate the highest performance among competitive baselines (see Table 1). We launch 5 endpoints of $\mathcal{M}_{\text{Drafter-7B}}$ for parallel drafting on TriviaQA, PubHealth, and ARC-Challenge. We launch 10 endpoints for MuSiQue due to more drafts. We use tensor parallelism to fit Mixtral-Instruct$_{8x7B}$ into the GPU memory. We report the latency of Mixtral-Instruct$_{8x7B}$ under tensor parallelism sizes of 4, 8, 16. Increasing tensor parallelism does not improve efficiency due to overheads in tensor aggregation and communication. In contrast, SPECULATIVE RAG, with its smaller RAG drafter and parallel draft generation, consistently achieves the lowest latency across all datasets. Particularly, it reduces latency by up to 23.41% on TriviaQA, 17.28% on MuSiQue, 51.25% on PubHealth, and 26.73% on ARC-Challenge. This highlights the advantage of our approach in reducing processing time while maintaining high performance.

4.5 Ablation Studies

We conduct ablation studies on the key components of SPECULATIVE RAG during both drafting and verification stages on TriviaQA and PubHealth in Table 2. We use $\mathcal{M}_{\text{Verifier-8x7B}}$ + $\mathcal{M}_{\text{Drafter-7B}}$ as a running configuration. Same as the main results, we report the accuracy as performance metrics.

Diversity and reduced redundancy in retrieval improves draft quality significantly. In the first set of experiments, we evaluate the impact of multi-perspective sampling during the drafting. Recall that SPECULATIVE RAG clusters retrieved documents into distinct perspectives and sample one document from each cluster to reduce redundancy for the draft generation. We compare this against two alternative sampling strategies: (1) Random sampling without clustering, where we randomly select a document subset as context, and (2) Sampling from the same cluster, where we select all documents from a single cluster. Our results indicate that our proposed sampling method yields the best performance thanks to its ability to leverage diverse context. Particularly, it improves the accuracy up to 1.88% on TriviaQA and 2.23% on PubHealth. While random sampling without clustering introduces diversity, it is prone to including redundant documents, degrading draft quality. Sampling from the same cluster significantly underperforms due to a lack of diverse perspectives.

Scoring method on self-consistency and self-reflection refines draft quality effectively. In the second set of experiments, we examine the scoring method during verification. We remove each of the specific confidence scores, ρ_{Draft}, $\rho_{\text{Self-contain}}$, or $\rho_{\text{Self-reflect}}$ in turn. Performance drops are observed when any score is removed. Particularly, removing ρ_{Draft} leads to a minimal decline, 0.19% on TriviaQA and 1.12% on PubHealth, likely due to the limited verification capability of the smaller RAG drafter. Removing either $\rho_{\text{Self-contain}}$ or $\rho_{\text{Self-reflect}}$ results in similar performance decreases, around 2.0% on TriviaQA and around 0.8% on PubHealth, indicating that both self-containment and self-reflection capture different key aspects of reasoning and are crucial during verification. Random

Appendix

A Prompt of Rationale Generation

```
======================================== Prompt ========================================

# Memorize this piece of evidence in mind and use it as if you already know it.
# Evidence: State religion
Despite enjoying considerable popular support, Christianity was still not the official
state religion in Rome, although it was in some neighboring states such as Armenia, Iberia,
 and Aksum. Roman Religion (Neoplatonic Hellenism) was restored for a time by the Emperor
Julian from 361 to 363. Julian does not appear to have reinstated the persecutions of the
earlier Roman emperors. Catholic Christianity, as opposed to Arianism and other ideologies
deemed heretical, was declared to be the state religion of the Roman Empire on 27 February
380 by the decree "De fide catolica".

# Instruction: When did christianity become official religion of Rome?

# Response: 380 AD

# Rationale: (complete the following sentence with details from the evidence; you can only
use the information from the evidence)

====================================== Completion ======================================

Catholic Christianity was declared the official state religion of the Roman Empire on
February 27th, 380, through the decree "De fide catolica".
```

Figure 4: Prompt of Rationale Generation for Gemini-Ultra

B Prompt of RAG Drafting

```
======================================== Prompt ========================================

Response to the instruction. Also provide rationale for your response.
## Instruction: In Buddhism, what is the state of blissful repose or absolute existence by
someone relieved of the necessity of rebirth?
## Evidence:
[1] Buddhism
Nirvana literally means "blowing out, quenching, becoming extinguished". In early Buddhist
texts, it is the state of restraint and self-control that leads to the "blowing out" and
the ending of the cycles of sufferings associated with rebirths and redeaths. Many later
Buddhist texts describe nirvana as identical with "anatta" with complete "emptiness,
nothingness". In some texts, the state is described with greater detail, such as passing
through the gate of emptiness ("sunyata") realizing that there"
[2] Salvation
It includes a variety of disciplines, such as yoga and meditation. Nirvana is the profound
peace of mind that is acquired with moksha (liberation). In Buddhism and Jainism, it is the
 state of being free from suffering. In Hindu philosophy, it is union with the Brahman (
Supreme Being). The word literally means "blown out" (as in a candle) and refers, in the
Buddhist context, to the blowing out of the fires of desire, aversion, and delusion, and
the imperturbable stillness of mind acquired thereafter. In Theravada Buddhism the emphasis
 is on one's

====================================== Completion ======================================
## Rationale: Nirvana literally means 'blowing out, quenching, becoming extinguished'. It
is described as a state of "restraint and self-control" that leads to the "blowing out" and
 the ending of the cycles of sufferings associated with rebirths and redeaths.
## Response: In Buddhism, the state of blissful repose or absolute existence by someone
relieved of the necessity of rebirth is called Nirvana.
```

Figure 5: Prompt of RAG Drafting

C Prompt of Standard RAG

```
======================================== Prompt ========================================

Below is an instruction that describes a task. Write a response that appropriately
completes the request.

### Evidence:
[1] Britain (place name)
Britain, after which "Britain" became the more commonplace name for the island called Great
 Britain. After the Anglo-Saxon period, "Britain" was used as a historical term only.
Geoffrey of Monmouth in his pseudohistorical "Historia Regum Britanniae" ...

[2] Great Britain
The peoples of these islands of "Prettanike" were called the "Priteni" or "Pretani". "
Priteni" is the source of the Welsh language term Prydain, "Britain", which has the same
source as the Goidelic term Cruithne used to refer to the early Brythonic-speaking
inhabitants of Ireland. The latter were later called Picts or Caledonians ...

...

[10] Albion
Albion is an alternative name for Great Britain. The oldest attestation of the toponym
comes from the Greek language. It is sometimes used poetically and generally to refer to
the island, but is less common than 'Britain' today. The name for Scotland in most of the
Celtic languages is related to Albion: "Alba" in Scottish Gaelic, "Albain" ...

### Instruction: What was Britain called - before it was Britain?

### Response:
```

Figure 6: Prompt of Standard RAG for Non-instruction-tuned LM

```
======================================== Prompt ========================================

[INST] Below is an instruction that describes a task. Write a response for it and state
your explanation supporting your response.

### Instruction: What was Britain called - before it was Britain?
### Evidence:
[1] Britain (place name)
Britain, after which "Britain" became the more commonplace name for the island called Great
 Britain. After the Anglo-Saxon period, "Britain" was used as a historical term only.
Geoffrey of Monmouth in his pseudohistorical "Historia Regum Britanniae" ...

[2] Great Britain
The peoples of these islands of "Prettanike" were called the "Priteni" or "Pretani". "
Priteni" is the source of the Welsh language term Prydain, "Britain", which has the same
source as the Goidelic term Cruithne used to refer to the early Brythonic-speaking
inhabitants of Ireland. The latter were later called Picts or Caledonians ...

...

[10] Albion
Albion is an alternative name for Great Britain. The oldest attestation of the toponym
comes from the Greek language. It is sometimes used poetically and generally to refer to
the island, but is less common than 'Britain' today. The name for Scotland in most of the
Celtic languages is related to Albion: "Alba" in Scottish Gaelic, "Albain" ...
[/INST] The response is:
```

Figure 7: Prompt of Standard RAG for Instruction-tuned LM

D Case Study

Figure 8 shows two drafts generated for the same question. We observe that our RAG drafter can well understand the multiple perspectives in the retrieval results and generate high-quality drafts. Our RAG verifier can also help filter out the unreliable drafts as we observe a relatively low scores in the first draft in Figure 8.

```
================================== Prompt of Drafter ===================================

Response to the instruction. Also provide rationale for your response.
## Instruction: Which actress/singer starred as Doralee Rhodes in the 1980 film, "Nine to
Five"?
## Evidence:
[1] Diana DeGarmo
Sheila until the show's closure on June 27, 2010. DeGarmo played the role of Doralee Rhodes
 in the national tour of "9 to 5", which was launched in Nashville on September 21, 2010.
She ended her run as Doralee after the July 2011 Minneapolis tour stop. DeGarmo appeared on
 musical theatre composer Scott Alan's third album, ...
[2] 9 to 5 (musical)
Pippa Winslow as Violet Newstead, Louise Olley as Doralee Rhodes and Leo Sene as Franklin
Hart Jr, with Samantha Giffard as Roz, Matthew Chase as Joe and Mark Houston, Rachel Ivy,
and Blair Anderson. "9 to 5" will play in the West End at the Savoy Theatre from January 29
 to August 31, 2019. The production stars Amber Davies (Judy), ...

===================================== Completion =======================================
## Rationale: Diana DeGarmo played the role of Doralee Rhodes in the national tour of "9 to
 5", which began in September 2010.
## Response: Diana DeGarmo

===================================== Verification =====================================
draft score: 0.6594
self-containment score: 0.3417
self-reflection score: 0.5238

----------------------------------------------------------------------------------------

================================== Prompt of Drafter ===================================

Response to the instruction. Also provide rationale for your response.
## Instruction: Which actress/singer starred as Doralee Rhodes in the 1980 film, "Nine to
Five"?
## Evidence:
[1] Cynthia Rhodes
Rhodes garnered her first non-dance related role in Michael Crichton's 1984 science fiction
 thriller "Runaway" with Tom Selleck, Kirstie Alley and Gene Simmons. Her most notable role
 was as dance instructor Penny Johnson in the hit 1987 motion picture "Dirty Dancing" with
Jennifer Grey and Patrick Swayze. Rhodes' final motion picture role was the character of
Vickie Phillips opposite actor Jameson Parker in the sleeper action-adventure movie ...
[2] Dolly Parton
Parton's feature film "9 to 5" (1980). The musical ran at the Ahmanson Theatre, Los Angeles
 in late 2008. It opened on Broadway at the Marquis Theatre in New York City, on April 30,
2009, to mixed reviews. The title track of her 2008 "Backwoods Barbie" album was written
for the musical's character Doralee. Although her score (as well as the musical debut of
actress Allison Janney) were praised, ...

===================================== Completion =======================================
## Rationale: Dolly Parton starred in the 1980 film "9 to 5" as Doralee Rhodes.
Additionally, Parton's 2008 musical "9 to 5" was based on the film, with Parton reprising
her role as Doralee.
## Response: Dolly Parton starred as Doralee Rhodes in the 1980 film, "Nine to Five".

===================================== Verification =====================================
draft score: 0.71
self-containment score: 0.4346
self-reflection score: 0.7449
```

Figure 8: Case study of SPECULATIVE RAG from TriviaQA where *Dolly Parton* is the corrrect answer.