

SPECULATIVE RAG: ENHANCING RETRIEVAL AUGMENTED GENERATION THROUGH DRAFTING

Anonymous authors

Paper under double-blind review

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, PopQA, PubHealth, and ARC-Challenge benchmarks. It notably enhances accuracy by up to 12.97% while reducing latency by 50.83% 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, 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. Integrat-

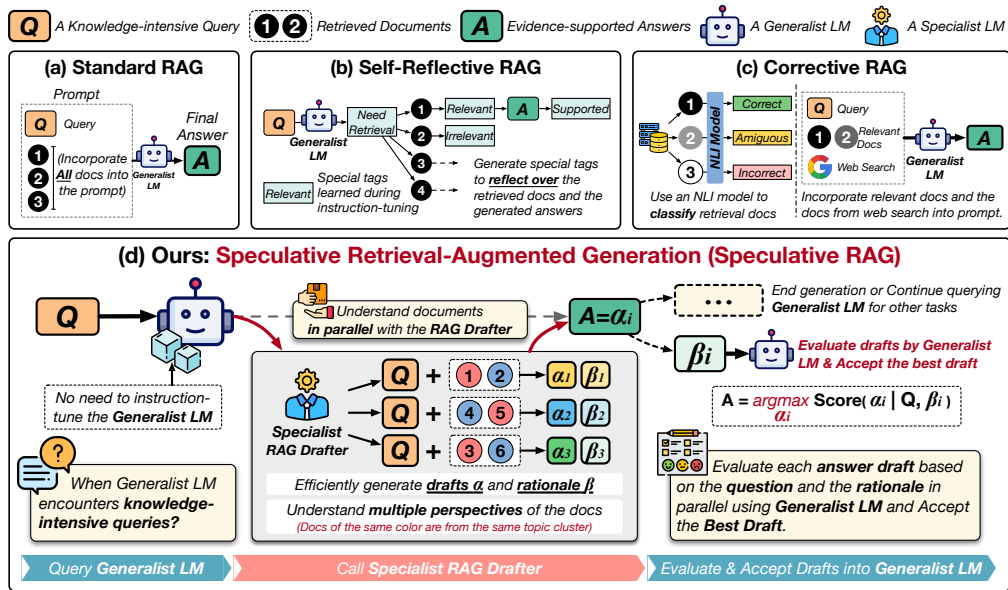


Figure 1: Illustration of different RAG approaches. Given a knowledge-intensive query Q and retrieved documents, (a) Standard RAG incorporates all documents into the prompt, increasing input length and slowing inference; (b) Self-Reflective RAG (Asai et al., 2023) requires specialized instruction-tuning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

ing 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., 2024a), 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 5 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., 2024a) 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., 2024a). 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; $D = \{d_1, \dots, d_n\}$ 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 $A \in \{\text{True}, \text{False}\}$ is a Boolean value indicating the statement’s correctness; in the multiple choice tasks, Q is a question with a few options and $A \in \{A, B, C, \dots\}$ 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

162 drafts based on retrieved results. Then, **a larger generalist LM, the RAG verifier**, assesses these
 163 drafts, selects the best one based on its rationale, and integrates it into the generation results.
 164

Algorithm 1: SPECULATIVE RAG

165 **Data:** $(Q, D = \{d_i\}_i^n)$ is the question and n retrieved documents; m subsets, each containing k
 166 documents, are sampled from D ; k also corresponds to the number of clusters during clustering.
 167 **Result:** \hat{A} is the predicted answer to the question.
 168 **Function Speculative RAG** (Q, D, m, k) :
 169 **1** $\{c_1, c_2, \dots, c_k\} \xleftarrow{\text{K-Means}} \mathcal{C}(d_1, \dots, d_n | Q)$ \triangleright Cluster the documents into k groups using an embedding model \mathcal{C} .
 170 $\Delta \leftarrow \{\}$
 171 **2** **repeat**
 172 $\delta_j \leftarrow \{\}$ \triangleright Construct a subset of the retrieved documents δ_j
 173 **3** **for** $c_i \in \{c_1, \dots, c_k\}$ **do**
 174 $\delta_j = \delta_j \cup \{\text{random.sample}(c_i)\}$ \triangleright Sample one document from each cluster c_i into subset δ_j .
 175 **4** **end**
 176 $\Delta = \Delta \cup \{\delta_j\}$
 177 **5** **until** $|\Delta| = m$ \triangleright Repeat the sampling until there are m unique subsets in total.
 178 **6** **for** $\delta_j \in \Delta$ **do in parallel** \triangleright Process m subsets in parallel.
 179 $\alpha_j, \beta_j \leftarrow \mathcal{M}_{\text{Drafter}} \cdot \text{generate}(Q, \delta_j)$ \triangleright Generate the **draft** α and **rationale** β with $\mathcal{M}_{\text{Drafter}}$.
 180 $\rho_j \leftarrow \mathcal{M}_{\text{Verifier}} \cdot \text{score}(\alpha_j | Q, \beta_j)$ \triangleright Compute the confidence score ρ with $\mathcal{M}_{\text{Verifier}}$.
 181 **7** **end**
 182 $\hat{A} \leftarrow \arg \max_{\alpha_j} \rho_j$ \triangleright Select the one with the highest score as the final answer.
 183 **8** **return** \hat{A}

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

3.2 SPECIALIST RAG DRAFTER

202 Instead of tuning a large generalist LM for the RAG scenario, we leverage a smaller specialist LM,
 203 $\mathcal{M}_{\text{Drafter}}$, to understand retrieved documents. $\mathcal{M}_{\text{Drafter}}$ is specialized in answering the given question
 204 based on the supporting documents and not expected to cope with general problems. It serves as a
 205 RAG module for the generalist LMs when solving knowledge-intensive tasks. We train $\mathcal{M}_{\text{Drafter}}$ to
 206 generate both the answer draft and the rationale to better understand the contextual documents.

207 **Instruction Tuning** Given a triplet (Q, A, D) , where Q is a general query, A is the response, and D
 208 is a retrieved supporting document, we augment it with the rationale of the response A based on the
 209 document D . We denote the rationale as E which extracts essential information from the document
 210 and explains why the response is reasonable to the query concisely (Hsieh et al., 2023) so it is of
 211 shorter length and delivers information coherent with the original document. We leverage relatively
 212 strong LMs to automatically synthesize the rationale E for each triplet. Specifically, we directly
 213 query the strong LM to understand the knowledge from the document and provide the intermediate
 214 rationale between the instruction and response. Refer to Appendix F for detailed prompts. After
 215 generating the rationale, we finetune a pre-trained LM using the standard language modeling objec-
 tive, maximizing the likelihood: $\mathbb{E}_{(Q,A,D,E)} \log P_{\mathcal{M}_{\text{Drafter}}}(A, E | Q, D)$, where (Q, A, D, E) is an

augmented entry in the dataset; $P_{\mathcal{M}_{\text{Drafter}}}(A, E | Q, D)$ 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 (Jin & Han, 2011).

$$\begin{aligned} \text{emb}(d_1), \dots, \text{emb}(d_n) &= \mathcal{E}(d_1, \dots, d_n | Q) \\ \{c_1, \dots, c_k\} &= \text{K-Means}(\text{emb}(d_1), \dots, \text{emb}(d_n)) \\ \delta &= \{\text{random.sample}(c) \text{ for } c \in \{c_i\}_1^k\} \end{aligned}$$

where \mathcal{E} is an instruction-aware embedding model which embeds a string with regard to a provided instruction (the posed question Q); $\text{emb}(d_i)$ is the embedding for the retrieved document d_i ; 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 δ 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 G 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 $\delta_j = \{d_{j_1}, \dots, d_{j_k}\}$, we query $\mathcal{M}_{\text{Drafter}}$ in parallel with the following prompt for the answer draft and rationale: $Q, d_{j_1}, \dots, d_{j_k} \rightarrow \alpha_j, \beta_j$, 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 $\rho_{\text{Draft},j} = P(\beta_j | Q, d_{j_1}, \dots, d_{j_k}) + P(\alpha_j | Q, d_{j_1}, \dots, d_{j_k}, \beta_j)$, 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, $\rho_{\text{Self-contains}} = P(\alpha, \beta | Q)$. 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 $\rho_{\text{Self-reflect}} = P(\text{"Yes"} | Q, \alpha, \beta, R)$ 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 $[Q, \alpha, \beta, R, \text{"Yes"}]$, 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 $\rho_{\text{Self-contains}}$ and self-reflection score $\rho_{\text{Self-reflect}}$.

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

270 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
 271 answer as the final answer to the question $\hat{A} = \arg \max_{\alpha_j} \rho_j$.
 272

273 4 EXPERIMENTS

274
 275 We evaluate our proposed SPECULATIVE RAG on five public retrieval augmented generation bench-
 276 marks: TriviaQA (unfiltered) (Joshi et al., 2017), MuSiQue (Trivedi et al., 2022), PopQA (Mallen
 277 et al., 2023), PubHealth (Zhang et al., 2023b), and ARC-Challenge (Clark et al., 2018). We provide
 278 representative examples for case study in Appendix I. TriviaQA, MuSiQue, PopQA are challenging
 279 open-domain question answering datasets where RAG systems are required to answer questions on
 280 factual knowledge. TriviaQA and PopQA typically require one accurate piece of evidence from
 281 the documents, whereas MuSiQue demands multiple documents to construct a multi-hop reasoning
 282 chain. More detailed experiments on multi-hop reasoning can be found in Appendix E. Following
 283 previous works (Guu et al., 2020; Asai et al., 2023; Yan et al., 2024), we evaluate performance of the
 284 free-form generation based on whether gold answers are contained within the generated response or
 285 not. PubHealth and ARC-Challenge are closed-set generation datasets. PubHealth is a dataset of
 286 medical claims spanning a variety of biomedical subjects and it requires the RAG system to verify a
 287 given claim based on the retrieved documents. ARC-Challenge introduces a multi-choice question
 288 answering dataset, composed of science exam questions from grade 3 to grade 9. For closed-set gen-
 289 eration tasks, we use accuracy metrics to evaluate whether the generated answers match the ground
 290 truth.

291 4.1 BASELINES

292
 293 **Standard RAG** For standard RAG, we incorporate all the retrieved documents into the prompt
 294 as contextual information. Refer to Appendix H for detailed prompts. We run standard RAG
 295 experiments on off-the-shelf LLMs including Mistral_{7B}, Mistral-Instruct_{7B} (Jiang et al., 2023a),
 296 Mixtral_{8x7B}, Mixtral-Instruct_{8x7B} (Jiang et al., 2024), and Alpaca_{7B} (Dubois et al., 2024). We also
 297 include the performance of Toolformer (Schick et al., 2024) and SAIL (Luo et al., 2023a) which are
 298 originally reported from Asai et al. (2023). Toolformer_{7B} is an LM instruction-tuned to use tools
 299 including a search engine, and SAIL_{7B} is an LM instruction-tuned on the Alpaca instruction tuning
 300 set augmented with search results from different sources such as DuckDuckGo and Wikipedia.

301
 302 **Self-Reflective RAG and Corrective RAG** Self-Reflective RAG (Self-RAG) (Asai et al., 2023)
 303 and Corrective RAG (CRAG) (Yan et al., 2024) are more advanced RAG systems that enhances the
 304 quality of contextual information in the retrieval results. CRAG introduces an external evaluator to
 305 assess the quality of retrieved documents, and to refine them before the response generation. Self-
 306 RAG instruction-tunes an LM to generate special self-reflection tags. These tags guides the LM to
 307 dynamically retrieve documents when necessary, critique the retrieved documents relevance before
 308 generating responses. Self-CRAG is to apply the Self-RAG approach on the refined documents
 309 of CRAG. We adopt the same backbone LLMs across all methods as our proposed SPECULATIVE
 310 RAG for fair comparisons.

311 4.2 EXPERIMENT SETTINGS

312
 313 In our experiments, we utilize Mistral_{7B} (v0.1) as our base LM for the RAG drafter. For RAG
 314 verifier, we employ either Mistral_{7B} (v0.1) or Mixtral_{8x7B} (v0.1) without any fine-tuning, de-
 315 noted as $\mathcal{M}_{\text{Verifier-7B}}$ or $\mathcal{M}_{\text{Verifier-8x7B}}$. We pre-compute embeddings of retrieved documents using
 316 a lightweight instruction-aware embedding model InBedder_{Roberta} (Peng et al., 2024) as part of the
 317 retrieval process. Inference is conducted using the vLLM framework (Kwon et al., 2023) with
 318 greedy decoding (temperature = 0). We adopt the same experiment settings from Asai et al. (2023)
 319 and include a more challenging benchmark, MuSiQue (Trivedi et al., 2022). Our focus is on RAG
 320 reasoning rather than evidence citation, so we omit the other two long-form generation benchmarks,
 321 Biography (Min et al., 2023) and ALCE-ASQA (Gao et al., 2023a). On TriviaQA, PopQA, Pub-
 322 Health, and ARC-Challenge, we retrieve top 10 documents and generate 5 drafts per query ($m = 5$),
 323 with each draft based on a subset of 2 documents ($k = 2$). For MuSiQue, we retrieve top 15 docu-
 ments and generate 10 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 A.

Table 1: Retrieval augmentation generation results on TriviaQA, MuSiQue, PopQA, 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; \dagger indicates numbers reported in [Asai et al. \(2023\)](#); – denotes numbers that are not reported by the original papers or are not applicable; \ddagger 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	PopQA	PubHealth	ARC-C
<i>Standard RAG</i>					
Mistral _{7B} (Jiang et al., 2023a)	54.15	16.71	31.38	34.85	42.75
Mixtral _{8x7B} (Jiang et al., 2024)	59.85	19.16	34.02	37.08	48.72
Mistral-Instruct _{7B} (Jiang et al., 2023a)	67.11	17.99	42.17	42.15	47.70
Mixtral-Instruct _{8x7B} (Jiang et al., 2024)	73.91	29.42	53.68	63.63	78.41
Alpaca _{7B} (Dubois et al., 2024) [†]	64.1	-	46.7	40.2	48.1
Toolformer _{6B} (Schick et al., 2024) [†]	48.8	-	-	-	-
SAIL _{7B} (Luo et al., 2023a) [†]	-	-	-	69.2	48.4
<i>Self-Reflective RAG & Corrective RAG</i>					
CRAG _{Mistral-7B} (Yan et al., 2024)	-	-	-	59.04	74.87
Self-RAG _{Mistral-7B} (Asai et al., 2023)	64.84	21.72	52.68	72.44	74.91
Self-CRAG _{Mistral-7B} (Yan et al., 2024)	-	-	-	72.85	75.26
<i>Our Speculative RAG</i>					
$\mathcal{M}_{\text{Drafter-7B}}$ *	71.11	27.89	56.40	75.58	74.49
$\mathcal{M}_{\text{Verifier-7B}}^{\ddagger} + \mathcal{M}_{\text{Drafter-7B}}$	73.91	31.03	56.75	75.79	76.19
$\mathcal{M}_{\text{Verifier-8x7B}}^{\ddagger} + \mathcal{M}_{\text{Drafter-7B}}$	74.24	31.57	57.54	76.60	80.55

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 five datasets: TriviaQA, MuSiQue, PopQA, 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 our method consistently outperforms all baselines across all five 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, 3.86% on PopQA, 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 surpasses the 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 $\mathcal{M}_{\text{Drafter-7B}}$. Additionally, we further investigate the performance of $\mathcal{M}_{\text{Drafter-7B}}$ when we directly feed all documents to the RAG drafter and generate one draft, with detailed results provided in Appendix B. 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, 23.52% on PopQA, 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, 25.37% on PopQA, 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.

4.4 EFFECTS OF GENERATED RATIONALE FOR VERIFICATION

In SPECULATIVE RAG, we utilize the generated rationale β from the RAG drafter as an indicator of the trustworthiness of answer drafts α .

Shortened length compared to the retrieved documents. The rationales highlight relevant points, omit redundant information, and bridge logical gaps between drafts and their supporting documents. We compare the number of tokens in the generated rationale and the retrieved documents, and plot them in Figure 2. We find that the generated rationale is significantly shorter than the retrieved documents.

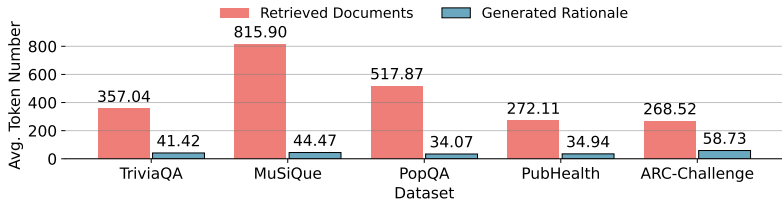


Figure 2: Average number of tokens in the generated rationale and the retrieved documents in TriviaQA, MuSiQue, PopQA, PubHealth, and ARC-Challenge. The generated rationale is of much shorter length than the original retrieved documents.

Table 2: Performance and latency analysis of SPECULATIVE RAG on TriviaQA and PubHealth using $\mathcal{M}_{\text{Verifier-8x7B}} + \mathcal{M}_{\text{Drafter-7B}}$. We add the original document subset δ to the context or replace the generated rationale β with the original retrieved document subset δ during verification, i.e. we compute the self-containment score as $\rho_{\text{Self-contain}} = P(\alpha, \delta|Q)$ or $\rho_{\text{Self-contain}} = P(\alpha, \delta, \beta|Q)$, and compute the self-reflection score as $\rho_{\text{Self-reflect}} = P(\text{"Yes"}|Q, \alpha, \delta, R)$ or $\rho_{\text{Self-reflect}} = P(\text{"Yes"}|Q, \alpha, \delta, \beta, R)$, where Q is the query; α is the answer draft; R is the self-reflection statement.

	TriviaQA		PubHealth	
	Accuracy (%)	Latency (s)	Accuracy (%)	Latency (s)
$\mathcal{M}_{\text{Verifier-8x7B}} + \mathcal{M}_{\text{Drafter-7B}}$				
$\rho = \text{Score}(\alpha Q, \beta)$	74.24	1.93	76.60	1.17
$\rho = \text{Score}(\alpha Q, \delta)$	74.08 (-0.16)	2.13 (+10.36%)	76.09 (-0.51)	1.31 (+11.97%)
$\rho = \text{Score}(\alpha Q, \beta, \delta)$	74.32 (+0.08)	2.17 (+12.44%)	76.29 (-0.31)	1.33 (+13.68%)

Comparable performance with retrieved documents and lower latency. To evaluate the effectiveness of the rationales, we create two alternative scoring methods: (a) replacing rationale with retrieved documents ($\rho = \text{Score}(\alpha|Q, \delta)$), or (b) adding retrieved documents to rationale ($\rho = \text{Score}(\alpha|Q, \beta, \delta)$). We compare these alternatives to the scoring method used in SPECULATIVE RAG ($\rho = \text{Score}(\alpha|Q, \beta)$) in Table 2. The results show that incorporating longer retrieved documents does not consistently improve performance and tends to increase latency. This suggests that the generated rationale is already of high quality and serves as an effective bridge between the supporting documents and the generated answer drafts. By leveraging this rationale, we can efficiently verify drafts using a generic LM, leading to accurate final results. We further validate the rationale generation in the instruction-tuning stage. See Appendix D for more details.

4.5 LATENCY ANALYSIS WITH BASELINES

We analyze the latency of Standard RAG, Self-RAG, and our SPECULATIVE RAG on TriviaQA, MuSiQue, PopQA, 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 3. 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 $\text{Mixtral-Instruct}_{8x7B}$ for Standard RAG, as these demonstrate the highest performance among competitive baselines (see Table 1). We also include the analysis for Standard RAG: $\text{Mixtral-Instruct}_{7B}$ and Self-RAG: $\text{Mixtral-Instruct}_{7B}$ in this study. For SPECULATIVE RAG, we launch 5 endpoints of $\mathcal{M}_{\text{Drafter-7B}}$ for parallel drafting on TriviaQA, PopQA, PubHealth, and ARC-Challenge. We launch 10 endpoints for MuSiQue due to more drafts. We use tensor parallelism of 4 to fit $\text{Mixtral-Instruct}_{8x7B}$ into the GPU memory. We use the same tensor parallelism setting for the other methods for a fair comparison.

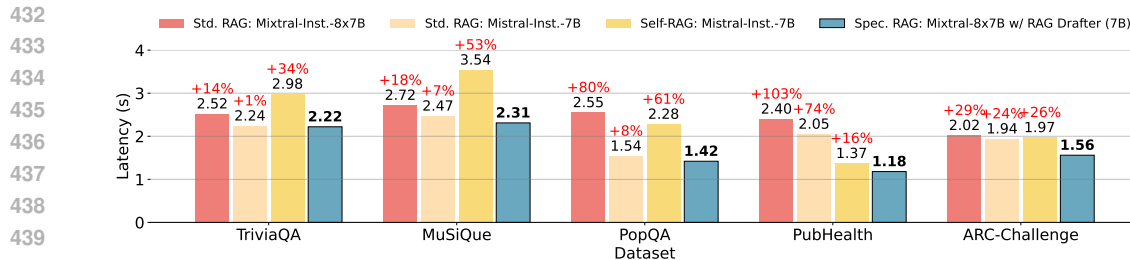


Figure 3: Latency analysis of Standard RAG, Self-RAG, and SPECULATIVE RAG on TriviaQA, MuSiQue, PopQA, PubHealth, and ARC-Challenge. The latency difference between Standard RAG/Self-RAG and SPECULATIVE RAG is highlighted in red (+x%). The latency varies across different datasets due to different retrieved document lengths. SPECULATIVE RAG encodes the retrieved documents in parallel and generates answer drafts with a smaller RAG drafter. This significantly improves the efficiency.

Reducing processing time while maintaining high performance As the results demonstrate, SPECULATIVE RAG consistently achieves the lowest latency compared to all other methods. This advantage comes from its utilization of fewer documents needed per draft and parallel drafting. Particularly, compared with the most competitive baseline, Standard RAG: $\mathcal{M}_{\text{Verifier-8x7B}} + \mathcal{M}_{\text{Drafter-7B}}$, our proposed SPECULATIVE RAG reduces latency by up to 11.90% on TriviaQA, 15.07% on MuSiQue, 44.31% on PopQA, 50.83% on PubHealth, and 22.77% on ARC-Challenge. Furthermore, a direct comparison between Standard RAG: Mistral-Instruct_{7B} and our method reveals that the higher latency of Standard RAG: Mistral-Instruct_{7B} is due to its longer context length which contains all retrieved documents. Self-RAG: Mistral-Instruct_{7B} also exhibits higher latency due to the generation of longer answers with self-reflection tags and the additional overhead associated with evidence selection. These findings highlight the advantage of our approach in reducing processing time while maintaining high performance.

4.6 ABLATION STUDIES

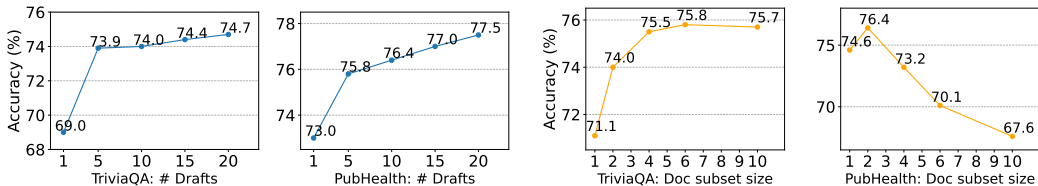
We conduct ablation studies on the multi-perspective sampling (Section 3.2) and the evaluation scores (Section 3.3) of SPECULATIVE RAG during the drafting or the verification stages on TriviaQA and PubHealth in Table 3. We use $\mathcal{M}_{\text{Verifier-8x7B}} + \mathcal{M}_{\text{Drafter-7B}}$ as 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 multi-perspective clustering, where we randomly select a document subset as context, and (2) Always 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 selection without verification leads to substantial underperformance, resulting in a performance decline of 5.69% on TriviaQA and 5.37% on PubHealth.

Table 3: Ablation study of SPECULATIVE RAG in the drafting and verification stages on TriviaQA and PubHealth.

	TriviaQA	PubHealth
$\mathcal{M}_{\text{Verifier-8x7B}} + \mathcal{M}_{\text{Drafter-7B}}$	74.24	76.60
<i>Drafting Stage</i>		
Random sampling w/o multi-perspective clustering	73.02 (-1.22)	75.38 (-1.22)
Always sampling from the same perspective cluster	72.36 (-1.88)	74.37 (-2.23)
<i>Verification Stage</i>		
w/o ρ_{Draft} ($\rho = \rho_{\text{Self-retain}} \cdot \rho_{\text{Self-reflect}}$)	74.05 (-0.19)	75.48 (-1.12)
w/o $\rho_{\text{Self-retain}}$ ($\rho = \rho_{\text{Draft}} \cdot \rho_{\text{Self-reflect}}$)	72.04 (-2.20)	75.89 (-0.71)
w/o $\rho_{\text{Self-reflect}}$ ($\rho = \rho_{\text{Draft}} \cdot \rho_{\text{Self-retain}}$)	72.36 (-1.88)	75.68 (-0.92)
Random selection w/o verification	68.55 (-5.69)	71.23 (-5.37)



(a) We include 1, 5, 10, 15, 20 drafts and sample 2 supporting documents for each draft. (b) We sample 1, 2, 4, 6, 10 supporting documents for each draft and we generate 10 answer drafts.

Figure 4: Performance analysis of SPECULATIVE RAG with (a) different numbers of drafts, and (b) different supporting document subset size on TriviaQA and PubHealth.

4.7 EFFECTS OF DRAFT NUMBER AND DOCUMENT SUBSET SIZE

Increasing the number of drafts improves performance without adding latency. We investigate the performance of SPECULATIVE RAG under varying numbers of drafts. Using $\mathcal{M}_{\text{Verifier-7B}} + \mathcal{M}_{\text{Drafter-7B}}$ with 1, 5, 10, 15, 20 drafts on TriviaQA and PubHealth. We sample two documents as context per draft. The results are illustrated in Figure 4(a). Since we retrieve top 10 documents in total, we sample up to 20 drafts in these experiments. The results indicate that incorporating more drafts can further improve performance, likely thanks to higher coverage of diverse perspective of documents. Importantly, in SPECULATIVE RAG, we can launch multiple RAG drafter instances to generate drafts in parallel without additional latency.

Increasing the document subset size doesn’t always lead to better performance. We also examine the effect of document subset size. By varying the number of documents (1, 2, 4, 6, or 10) sampled for draft generation on TriviaQA and PubHealth (Figure 4(b)), we find that including more documents in the context does not always lead to consistent performance improvement. While TriviaQA queries may benefit from more supporting documents due to their complexity, $\mathcal{M}_{\text{Verifier-7B}} + \mathcal{M}_{\text{Drafter-7B}}$ can surpass Mistral-Instruct_{7B} even with a single supporting document per draft. Furthermore, with two or more documents per draft, $\mathcal{M}_{\text{Verifier-7B}} + \mathcal{M}_{\text{Drafter-7B}}$ can even surpass Mixtral-Instruct_{8x7B}. This further demonstrates the effectiveness of our drafting design.

5 CONCLUSION

Our proposed SPECULATIVE RAG decomposes RAG tasks into two separate steps of drafting followed by verification. SPECULATIVE RAG delegates the heavy lifting of drafting to a small specialized RAG drafter, while verification is done using a large generalist LM. The parallel generation of multiple drafts from diverse document subsets provides high quality answer candidates while reducing input token counts and the potential risk of position-bias-over-long-context, resulting in substantial improvements in both the quality and speed of the final output generation. We demonstrate the effectiveness of SPECULATIVE RAG with accuracy gains up to 12.97% while reducing latency by 50.83% compared to conventional RAG systems. SPECULATIVE RAG sheds new light on the potential of collaborative architectures for enhancing RAG performance through task decomposition.

REFERENCES

- 540
541
542 Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Ale-
543 man, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical
544 report. *arXiv preprint arXiv:2303.08774*, 2023.
- 545 Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. Self-rag: Learning to
546 retrieve, generate, and critique through self-reflection. *arXiv preprint arXiv:2310.11511*, 2023.
- 547 Jinheon Baek, Soyeong Jeong, Minki Kang, Jong C Park, and Sung Hwang. Knowledge-augmented
548 language model verification. In *Proceedings of the 2023 Conference on Empirical Methods in*
549 *Natural Language Processing*, pp. 1720–1736, 2023.
- 550 Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal,
551 Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are
552 few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.
- 553 Tianle Cai, Yuhong Li, Zhengyang Geng, Hongwu Peng, Jason D Lee, Deming Chen, and Tri
554 Dao. Medusa: Simple llm inference acceleration framework with multiple decoding heads. *arXiv*
555 *preprint arXiv:2401.10774*, 2024.
- 556 Charlie Chen, Sebastian Borgeaud, Geoffrey Irving, Jean-Baptiste Lespiau, Laurent Sifre, and John
557 Jumper. Accelerating large language model decoding with speculative sampling. *arXiv preprint*
558 *arXiv:2302.01318*, 2023a.
- 559 Tong Chen, Hongwei Wang, Sihao Chen, Wenhao Yu, Kaixin Ma, Xinran Zhao, Hongming Zhang,
560 and Dong Yu. Dense x retrieval: What retrieval granularity should we use? *arXiv preprint*
561 *arXiv:2312.06648*, 2023b.
- 562 Wei Chen, Yu Liu, Weiping Wang, Erwin M Bakker, Theodoros Georgiou, Paul Fieguth, Li Liu, and
563 Michael S Lew. Deep learning for instance retrieval: A survey. *IEEE Transactions on Pattern*
564 *Analysis and Machine Intelligence*, 2022.
- 565 Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and
566 Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge.
567 *arXiv preprint arXiv:1803.05457*, 2018.
- 568 Jiayu Ding, Shuming Ma, Li Dong, Xingxing Zhang, Shaohan Huang, Wenhui Wang, Nanning
569 Zheng, and Furu Wei. Longnet: Scaling transformers to 1,000,000,000 tokens. *arXiv preprint*
570 *arXiv:2307.02486*, 2023.
- 571 Yann Dubois, Chen Xuechen Li, Rohan Taori, Tianyi Zhang, Ishaan Gulrajani, Jimmy Ba, Carlos
572 Guestrin, Percy S Liang, and Tatsunori B Hashimoto. AlpacaFarm: A simulation framework for
573 methods that learn from human feedback. *Advances in Neural Information Processing Systems*,
574 36, 2024.
- 575 Shangbin Feng, Weijia Shi, Yuyang Bai, Vidhisha Balachandran, Tianxing He, and Yulia Tsvetkov.
576 Knowledge card: Filling llms’ knowledge gaps with plug-in specialized language models. In *The*
577 *Twelfth International Conference on Learning Representations*, 2023.
- 578 Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. Enabling large language models to generate
579 text with citations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural*
580 *Language Processing*, pp. 6465–6488, 2023a.
- 581 Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, and
582 Haofen Wang. Retrieval-augmented generation for large language models: A survey. *arXiv*
583 *preprint arXiv:2312.10997*, 2023b.
- 584 Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. Retrieval augmented
585 language model pre-training. In *International conference on machine learning*, pp. 3929–3938.
586 PMLR, 2020.
- 587 Cheng-Yu Hsieh, Chun-Liang Li, Chih-Kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alexander Rat-
588 ner, Ranjay Krishna, Chen-Yu Lee, and Tomas Pfister. Distilling step-by-step! outperforming
589 larger language models with less training data and smaller model sizes. In *ACL*, 2023.
- 590
591
592
593

- 594 Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong
595 Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. A survey on hallucination in large language
596 models: Principles, taxonomy, challenges, and open questions. *arXiv preprint arXiv:2311.05232*,
597 2023.
- 598
599 Gautier Izacard and Edouard Grave. Leveraging passage retrieval with generative models for open
600 domain question answering. In Paola Merlo, Jorg Tiedemann, and Reut Tsarfaty (eds.), *Proceed-*
601 *ings of the 16th Conference of the European Chapter of the Association for Computational Lin-*
602 *guistics: Main Volume*, pp. 874–880, Online, April 2021. Association for Computational Linguis-
603 tics. doi: 10.18653/v1/2021.eacl-main.74. URL [https://aclanthology.org/2021.](https://aclanthology.org/2021.eacl-main.74)
604 [eacl-main.74](https://aclanthology.org/2021.eacl-main.74).
- 605 Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand
606 Joulin, and Edouard Grave. Unsupervised dense information retrieval with contrastive learning.
607 *arXiv preprint arXiv:2112.09118*, 2021.
- 608
609 Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot,
610 Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al.
611 Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023a.
- 612
613 Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bam-
614 ford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al.
615 Mixtral of experts. *arXiv preprint arXiv:2401.04088*, 2024.
- 616
617 Zhengbao Jiang, Frank Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang,
618 Jamie Callan, and Graham Neubig. Active retrieval augmented generation. In Houda Bouamor,
619 Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Meth-*
620 *ods in Natural Language Processing*, pp. 7969–7992, Singapore, December 2023b. Associa-
621 tion for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.495. URL [https://](https://aclanthology.org/2023.emnlp-main.495)
622 aclanthology.org/2023.emnlp-main.495.
- 623
624 Li Jiapeng, Liu Runze, Li Yabo, Zhou Tong, Li Mingling, and Chen Xiang. Tree of reviews: A tree-
625 based dynamic iterative retrieval framework for multi-hop question answering. *arXiv preprint*
626 *arXiv:2404.14464*, 2024.
- 627
628 Xin Jin and Jiawei Han. K-means clustering. *Encyclopedia of machine learning*, pp. 563–564, 2011.
- 629
630 Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. Triviaqa: A large scale distantly
631 supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meet-*
632 *ing of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1601–1611,
633 2017.
- 634
635 Ehsan Kamalloo, Nouha Dziri, Charles Clarke, and Davood Rafiei. Evaluating open-domain ques-
636 tion answering in the era of large language models. In *Proceedings of the 61st Annual Meeting of*
637 *the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5591–5606, 2023.
- 638
639 Urvashi Khandelwal, Omer Levy, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. Generalization
640 through Memorization: Nearest Neighbor Language Models. In *International Conference on*
641 *Learning Representations (ICLR)*, 2020.
- 642
643 Jaehyung Kim, Jaehyun Nam, Sangwoo Mo, Jongjin Park, Sang-Woo Lee, Minjoon Seo, Jung-Woo
644 Ha, and Jinwoo Shin. Sure: Summarizing retrievals using answer candidates for open-domain qa
645 of llms. *arXiv preprint arXiv:2404.13081*, 2024.
- 646
647 Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large
648 language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems*,
649 2022.
- 650
651 Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E.
652 Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model
653 serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating*
654 *Systems Principles*, 2023.

- 648 Yaniv Leviathan, Matan Kalman, and Yossi Matias. Fast inference from transformers via speculative
649 decoding. In *International Conference on Machine Learning*, pp. 19274–19286. PMLR, 2023.
650
- 651 Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal,
652 Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented genera-
653 tion for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:
654 9459–9474, 2020.
- 655 Tianle Li, Ge Zhang, Quy Duc Do, Xiang Yue, and Wenhui Chen. Long-context llms struggle with
656 long in-context learning. *arXiv preprint arXiv:2404.02060*, 2024.
657
- 658 Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and
659 Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the*
660 *Association for Computational Linguistics*, 12, 2024.
661
- 662 Hongyin Luo, Yung-Sung Chuang, Yuan Gong, Tianhua Zhang, Yoon Kim, Xixin Wu, Danny Fox,
663 Helen Meng, and James Glass. Sail: Search-augmented instruction learning. *arXiv preprint*
664 *arXiv:2305.15225*, 2023a.
- 665 Yun Luo, Zhen Yang, Fandong Meng, Yafu Li, Jie Zhou, and Yue Zhang. An empirical study
666 of catastrophic forgetting in large language models during continual fine-tuning. *arXiv preprint*
667 *arXiv:2308.08747*, 2023b.
668
- 669 Xinbei Ma, Yeyun Gong, Pengcheng He, Hai Zhao, and Nan Duan. Query rewriting in retrieval-
670 augmented large language models. In *Proceedings of the 2023 Conference on Empirical Methods*
671 *in Natural Language Processing*, pp. 5303–5315, 2023.
672
- 673 Xuezhe Ma, Xiaomeng Yang, Wenhan Xiong, Beidi Chen, Lili Yu, Hao Zhang, Jonathan May, Luke
674 Zettlemoyer, Omer Levy, and Chunting Zhou. Megalodon: Efficient llm pretraining and inference
675 with unlimited context length. *arXiv preprint arXiv:2404.08801*, 2024.
- 676 Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi.
677 When not to trust language models: Investigating effectiveness of parametric and non-parametric
678 memories. In *Proceedings of the 61st Annual Meeting of the Association for Computational*
679 *Linguistics (Volume 1: Long Papers)*, pp. 9802–9822, 2023.
680
- 681 Xupeng Miao, Gabriele Oliaro, Zhihao Zhang, Xinhao Cheng, Zeyu Wang, Zhengxin Zhang, Rae
682 Ying Yee Wong, Alan Zhu, Lijie Yang, Xiaoxiang Shi, et al. Specinfer: Accelerating large lan-
683 guage model serving with tree-based speculative inference and verification. In *Proceedings of the*
684 *29th ACM International Conference on Architectural Support for Programming Languages and*
685 *Operating Systems, Volume 3*, pp. 932–949, 2024.
- 686 Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. Can a suit of armor conduct elec-
687 tricity? a new dataset for open book question answering. In *Proceedings of the 2018 Conference*
688 *on Empirical Methods in Natural Language Processing*, pp. 2381–2391, 2018.
689
- 690 Sewon Min, Kalpesh Krishna, Xinxu Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke
691 Zettlemoyer, and Hannaneh Hajishirzi. Factscore: Fine-grained atomic evaluation of factual pre-
692 cision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods*
693 *in Natural Language Processing*, pp. 12076–12100, 2023.
- 694 Letian Peng, Yuwei Zhang, Zilong Wang, Jayanth Srinivasa, Gaowen Liu, Zihan Wang, and Jingbo
695 Shang. Answer is all you need: Instruction-following text embedding via answering the question.
696 *arXiv preprint arXiv:2402.09642*, 2024.
697
- 698 Fabio Petroni, Aleksandra Piktus, Angela Fan, Patrick Lewis, Majid Yazdani, Nicola De Cao, James
699 Thorne, Yacine Jernite, Vladimir Karpukhin, Jean Maillard, et al. Kilt: a benchmark for knowl-
700 edge intensive language tasks. In *Proceedings of the 2021 Conference of the North American*
701 *Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp.
2523–2544, 2021.

- 702 Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. Deepspeed: System opti-
703 mizations enable training deep learning models with over 100 billion parameters. In *Proceedings*
704 *of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*,
705 pp. 3505–3506, 2020.
- 706 Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy Lillicrap, Jean-
707 baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, et al. Gem-
708 ini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint*
709 *arXiv:2403.05530*, 2024.
- 710 Parth Sarthi, Salman Abdullah, Aditi Tuli, Shubh Khanna, Anna Goldie, and Christopher D Man-
711 ning. Raptor: Recursive abstractive processing for tree-organized retrieval. *arXiv preprint*
712 *arXiv:2401.18059*, 2024.
- 713 Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Eric Hambro,
714 Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language models can
715 teach themselves to use tools. *Advances in Neural Information Processing Systems*, 36, 2024.
- 716 Zhengliang Shi, Shuo Zhang, Weiwei Sun, Shen Gao, Pengjie Ren, Zhumin Chen, and Zhaochun
717 Ren. Generate-then-ground in retrieval-augmented generation for multi-hop question answering.
718 *arXiv preprint arXiv:2406.14891*, 2024.
- 719 Ivan Stelmakh, Yi Luan, Bhuwan Dhingra, and Ming-Wei Chang. Asqa: Factoid questions meet
720 long-form answers. In *Proceedings of the 2022 Conference on Empirical Methods in Natural*
721 *Language Processing*, pp. 8273–8288, 2022.
- 722 Mitchell Stern, Noam Shazeer, and Jakob Uszkoreit. Blockwise parallel decoding for deep autore-
723 gressive models. *Advances in Neural Information Processing Systems*, 31, 2018.
- 724 Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu,
725 Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly
726 capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.
- 727 Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. Musique: Multihop
728 questions via single-hop question composition. *Transactions of the Association for Computational*
729 *Linguistics*, 10:539–554, 2022.
- 730 Yizhong Wang, Hamish Ivison, Pradeep Dasigi, Jack Hessel, Tushar Khot, Khyathi Raghavi
731 Chandu, David Wadden, Kelsey MacMillan, Noah A Smith, Iz Beltagy, et al. How far can camels
732 go. *Exploring the state of instruction tuning on open resources*, 2023a.
- 733 Zhiruo Wang, Jun Araki, Zhengbao Jiang, Md Rizwan Parvez, and Graham Neubig. Learning to
734 filter context for retrieval-augmented generation. *arXiv preprint arXiv:2311.08377*, 2023b.
- 735 Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi,
736 Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Huggingface’s transformers:
737 State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*, 2019.
- 738 Heming Xia, Tao Ge, Peiyi Wang, Si-Qing Chen, Furu Wei, and Zhifang Sui. Speculative decod-
739 ing: Exploiting speculative execution for accelerating seq2seq generation. In *Findings of the*
740 *Association for Computational Linguistics: EMNLP 2023*, pp. 3909–3925, 2023.
- 741 Heming Xia, Zhe Yang, Qingxiu Dong, Peiyi Wang, Yongqi Li, Tao Ge, Tianyu Liu, Wenjie Li, and
742 Zhifang Sui. Unlocking efficiency in large language model inference: A comprehensive survey
743 of speculative decoding. *arXiv preprint arXiv:2401.07851*, 2024a.
- 744 Sirui Xia, Xintao Wang, Jiaqing Liang, Yifei Zhang, Weikang Zhou, Jiayi Deng, Fei Yu, and Yanghua
745 Xiao. Ground every sentence: Improving retrieval-augmented llms with interleaved reference-
746 claim generation. *arXiv preprint arXiv:2407.01796*, 2024b.
- 747 Jian Xie, Kai Zhang, Jiangjie Chen, Renze Lou, and Yu Su. Adaptive chameleon or stubborn sloth:
748 Revealing the behavior of large language models in knowledge conflicts. In *The Twelfth Interna-*
749 *tional Conference on Learning Representations*, 2023.

756 Fangyuan Xu, Weijia Shi, and Eunsol Choi. Recomp: Improving retrieval-augmented lms with
757 compression and selective augmentation. *arXiv preprint arXiv:2310.04408*, 2023.
758

759 Ziwei Xu, Sanjay Jain, and Mohan Kankanhalli. Hallucination is inevitable: An innate limitation of
760 large language models. *arXiv preprint arXiv:2401.11817*, 2024.

761 Shi-Qi Yan, Jia-Chen Gu, Yun Zhu, and Zhen-Hua Ling. Corrective retrieval augmented generation.
762 *arXiv preprint arXiv:2401.15884*, 2024.
763

764 Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov,
765 and Christopher D Manning. Hotpotqa: A dataset for diverse, explainable multi-hop question
766 answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language*
767 *Processing*, pp. 2369–2380, 2018.

768 Ori Yoran, Tomer Wolfson, Ori Ram, and Jonathan Berant. Making retrieval-augmented language
769 models robust to irrelevant context. *arXiv preprint arXiv:2310.01558*, 2023.
770

771 Wenhao Yu, Hongming Zhang, Xiaoman Pan, Kaixin Ma, Hongwei Wang, and Dong Yu.
772 Chain-of-note: Enhancing robustness in retrieval-augmented language models. *arXiv preprint*
773 *arXiv:2311.09210*, 2023.

774 Jun Zhang, Jue Wang, Huan Li, Lidan Shou, Ke Chen, Gang Chen, and Sharad Mehrotra. Draft &
775 verify: Lossless large language model acceleration via self-speculative decoding. *arXiv preprint*
776 *arXiv:2309.08168*, 2023a.

777 Tianhua Zhang, Hongyin Luo, Yung-Sung Chuang, Wei Fang, Luc Gaitskell, Thomas Hartvigsen,
778 Xixin Wu, Danny Fox, Helen Meng, and James Glass. Interpretable unified language checking.
779 *arXiv preprint arXiv:2304.03728*, 2023b.
780
781
782
783
784
785
786
787
788
789
790
791
792
793
794
795
796
797
798
799
800
801
802
803
804
805
806
807
808
809

APPENDIX

A INSTRUCTION-TUNING SETTINGS

We construct our training dataset for the RAG drafter from diverse instruction-following pairs. We sample instances from Open-Instruct processed data (Wang et al., 2023a) and knowledge-intensive datasets (Petroni et al., 2021; Stelmakh et al., 2022; Mihaylov et al., 2018). We augment the instruction-following pairs with retrieved documents and generated rationale. We use the off-the-shelf dense retriever Contriever-MS MARCO (Izacard et al., 2021) to retrieve up to 10 documents for each pair and use Gemini-Ultra (Team et al., 2023) to generate rationale. In total, we acquire a dataset of 40k instances. We use Mistral_{7B} (v0.1) as our base LM for the RAG drafter. We reproduce the performance of Self-RAG (Asai et al., 2023) and CRAG (Yan et al., 2024) with Mistral_{7B} (v0.1) for a fair comparison. We implement the training scripts using the Transformers library from Hugging Face (Wolf et al., 2019). We employ DeepSpeed (Rasley et al., 2020) to accelerate the training process. All experiments are conducted on a Linux server equipped with 16 Nvidia A100-SXM4-40GB GPUs.

Additionally, we replace Gemini-Ultra (Team et al., 2023) with GPT-4o (Achiam et al., 2023) when curating the instruction-tuning data for our RAG drafter to investigate the effects of different LLMs. These results demonstrate that SPECULATIVE RAG maintains its performance advantage, even when trained on data curated by GPT-4o. It consistently outperforms the baselines from Standard RAG, SelfRAG, and CRAG, further validating the effectiveness of our approach.

Table 4: RAG results on TriviaQA, PubHealth, ARC-Challenge with the RAG drafter trained on instruction-tuning data curated by GPT-4o and Gemini-Ultra.

	TriviaQA	PubHealth	ARC-C
$\mathcal{M}_{\text{Verifier-7B}} + \mathcal{M}_{\text{Drafter-7B}}^{\text{Gemini-U}}$	73.91	75.79	76.19
$\mathcal{M}_{\text{Verifier-8x7B}} + \mathcal{M}_{\text{Drafter-7B}}^{\text{Gemini-U}}$	74.24	76.60	80.55
$\mathcal{M}_{\text{Verifier-7B}} + \mathcal{M}_{\text{Drafter-7B}}^{\text{GPT-4o}}$	72.24	73.05	76.54
$\mathcal{M}_{\text{Verifier-8x7B}} + \mathcal{M}_{\text{Drafter-7B}}^{\text{GPT-4o}}$	73.58	73.35	80.63

B EFFECTS OF INSTRUCTION TUNING

In SPECULATIVE RAG, we introduce a framework that combines the RAG drafter and the verifier. In this ablation study, we directly feed all documents to the RAG drafter and generate one draft ($m = 1$, $k = \text{total \# of documents}$). As shown in Table 5, we observe that instruction tuning effectively enhances the document understanding capability of the RAG drafter, as it outperforms both Mistral_{7B} and Mistral-Instruct_{7B}. However, there remains a gap compared to SPECULATIVE RAG, showing the effectiveness of the drafting and verification framework.

Table 5: RAG results on TriviaQA and PubHealth ($m = 1$, $k = \text{total \# of docs}$)

	TriviaQA	PubHealth
Mistral _{7B}	54.15	34.85
Mistral-Instruct _{7B}	67.11	42.15
$\mathcal{M}_{\text{Drafter-7B}} (m = 1, k = 10)$	73.23	65.25
$\mathcal{M}_{\text{Verifier-7B}} + \mathcal{M}_{\text{Drafter-7B}} (m = 5, k = 2)$	73.91	75.79

C EFFECTS OF SELF-REFLECTION STATEMENT

We use “Do you think the explanation supports the answers? (Yes or No)” as the self-reflection statement in our main results. In this study, we replace it with other alternatives to see how the self-reflection statement affects the accuracy. The results are reported in Table 6. We observe that

the performance does not change a lot given different self-reflection statements, which shows the stable verification capability of the generalist LMs by language modeling objective.

Table 6: Performance analysis of SPECULATIVE RAG with different self-reflection statements R when computing the self-reflection score $\rho_{\text{Self-reflect}} = P(\text{"Yes"}|Q, \alpha, \beta, R)$, where Q is the query, α, β are the generated answer draft and rationale.

Reflection Statement	TriviaQA	PubHealth
<i>Do you think the explanation supports the answers? (Yes or No)</i>	74.24	76.60
<i>Does the rationale support the answer? (Yes or No)</i>	74.22	76.09
<i>What do you think about the rationale? A good one? (Yes or No)</i>	74.25	75.79
<i>Is the rationale good enough to support the answer? (Yes or No)</i>	74.39	76.29

D EFFECTS OF RATIONALE GENERATION

We acknowledge that the generation of rationale potentially increases the inference cost during the drafting stage while this is crucial for the verifier in our method to assess the quality and reliability of generated drafts. And, the potential overhead can be mitigated through efficient parallel inference.

To further study the impact of rationale generation, we finetune the RAG drafter without rationale. We denote this setting as: *without rationale in drafting*. Similarly, *with rationale/doc in verification* indicates that we use the generated rationale or the retrieved documents as reference during the verification stage. We use $\mathcal{M}_{\text{Verifier-8x7B}} + \mathcal{M}_{\text{Drafter-7B}}$ as a running example. The results are shown in Table 7.

Table 7: Ablation study on the draft generation in the drafting and verification stages on TriviaQA and PubHealth.

Drafting	Verification	TrivialQA	PubHealth	ARC-C
<i>w/o rationale</i>	<i>w/ doc</i>	70.86	75.28	78.67
<i>w/ rationale</i>	<i>w/ doc</i>	74.08	76.09	80.46
<i>w/ rationale</i>	<i>w/ rationale</i>	74.24	76.60	80.55

Better answer drafting As explored in Hsieh et al. (2023), incorporating rationale generation during instruction-tuning can lead to the RAG drafter producing higher-quality answer drafts. The results in Table 7 clearly demonstrate this. We observe a significant performance drop across all three benchmarks when the RAG drafter is finetuned without the rationale component,

Lower latency and cost in verification We verify each draft against the rationale instead of the retrieved documents. From the ablation results, these generated rationales serve as high-quality grounding facts, improving verification performance compared to using the retrieved documents.

E EFFICACY OF SPECULATIVE RAG IN MULTI-HOP REASONING

We further validate SPECULATIVE RAG in the scenario of multi-hop reasoning. One of the key challenges of multi-hop reasoning is to effectively combine multiple pieces of evidence to arrive at the final answer. Indeed, the ability to verify or contrast information across documents is crucial to solve complex questions. We compare the performance of SPECULATIVE RAG with baselines on MuSiQue (Trivedi et al., 2022) and HotpotQA (Yang et al., 2018), two multi-hop reasoning benchmarks. We randomly sample 500 examples from the validation set of HotpotQA as the test set in our experiment. We adopt the same setting as MuSiQue on HotpotQA. The results are in Table 8. We find that our SPECULATIVE RAG achieves the best performance. Specifically, SPECULATIVE RAG improves accuracy by 2.15% on MuSiQue and by a substantial 5.4% on HotpotQA.

Our approach tackles this challenge by multi-perspective sampling when selecting documents for each draft (Section 3.2). We cluster the retrieved documents into distinct topics using an instruction-aware embedding model (Peng et al., 2024). Then, we sample one document from each cluster to

Table 8: RAG results on MuSiQue and HotpotQA

	MuSiQue	HotpotQA
Mixtral-Instruct _{8x7B}	29.42	43.60
Self-RAG _{Mixtral-7B}	21.72	27.20
$\mathcal{M}_{\text{Verifier-7B}} + \mathcal{M}_{\text{Drafter-7B}}$	31.03	47.60
$\mathcal{M}_{\text{Verifier-8x7B}} + \mathcal{M}_{\text{Drafter-7B}}$	31.57	49.00

form a diverse document subset, ensuring each drafter receives a variety of perspectives from the retrieval results. To validate the efficacy of this strategy, we further conduct an ablation study on MuSiQue and HotpotQA in Table 9. From the table, our sampling strategy effectively guarantees the diversity of information within the supporting document subsets, leading to improved performance of SPECULATIVE RAG on these tasks.

Table 9: Ablation study of multi-perspective sampling on multi-hop reasoning benchmarks: MuSiQue, HotpotQA.

	MuSiQue	HotpotQA
Random sampling	29.33	48.2
Multi-perspective sampling	31.57	49.00

E.1 PERFORMANCE BREAKDOWN ON HOTPOTQA

HotpotQA includes two types of questions: bridge-type questions in HotpotQA require a two-step reasoning process where the answer to the first step is crucial for answering the second. For example:

- *"When was the singer and songwriter of Radiohead born?"*
 - Step 1: Who is the singer and songwriter of Radiohead? → Thom Yorke
 - Step 2: When was [Thom Yorke](answer of step 1) born? → October 7, 1968
 - Final answer: October 7, 1968

In contrast, comparison-type questions also involve two steps, but the answers to each step are independent of each other. For example:

- *"Who was born first, Morgan Llywelyn or Robert Jordan?"*
 - Step 1: What's Morgan Llywelyn's DOB? → December 3, 1937
 - Step 2: What's Robert Jordan's DOB? → October 17, 1948
 - Final answer: Morgan Llywelyn

Table 10: Performance of SPECULATIVE RAG for different question types

Question Type	# of Questions	SPECULATIVE RAG
Bridge-type	400	41.75
Comparison-type	100	78.00
Overall	500	49.00

We report the performance breakdown of SPECULATIVE RAG on HotpotQA in Table 10. The results demonstrate a superior performance on comparison-type questions with multi-perspective sampling. This aligns with our expectations, as multi-perspective sampling ensures the document subset covers the diverse topics necessary for answering comparison-type questions. Revisiting the example above, *"Who was born first, Morgan Llywelyn or Robert Jordan?"*, with $k = 4$, our approach clusters retrieved documents into four groups. Group 0 and 3 focus on Morgan, while group 1 and 2 focus on Robert. As we sample one document from each group for the drafters,

972 this clustering result ensures each drafter receives documents about both individuals. This balanced
973 information distribution is crucial for the comparison-type questions. In contrast, random sampling
974 risks providing a drafter with information about only one person, yielding a suboptimal draft.

975 On the other hand, we also observe that the multi-perspective sampling is less helpful for bridge-type
976 questions. These questions require the LLM to first identify the "*bridge entity*" (e.g., Thom Yorke in
977 the earlier example), a task our current work isn't explicitly designed for. While multi-perspective
978 sampling effectively covers different topics in the drafts and the map-reduce approach accelerates
979 inference, they might not directly contribute to pinpointing the "bridge entity" - the key to answering
980 bridge-type questions.

981 We believe our framework could be effectively combined with other techniques specifically designed
982 for bridge-type questions, such as those proposed in [Xia et al. \(2024b\)](#); [Jiapeng et al. \(2024\)](#); [Shi](#)
983 [et al. \(2024\)](#). For instance, the Tree-of-Reviews (ToR) framework, introduced in [Jiapeng et al.](#)
984 [\(2024\)](#), addresses multi-hop reasoning problems by dynamically initiating new searches based on
985 previously retrieved documents and constructing various reasoning paths. This dynamic searching
986 strategy can be integrated into our SPECULATIVE RAG, enabling each drafter to answer bridge-type
987 questions more effectively.

988

989

990

991

992

993

994

995

996

997

998

999

1000

1001

1002

1003

1004

1005

1006

1007

1008

1009

1010

1011

1012

1013

1014

1015

1016

1017

1018

1019

1020

1021

1022

1023

1024

1025

1026 F PROMPT OF RATIONALE GENERATION

1027

1028

1029

1030

1031

1032

1033

1034

1035

1036

1037

1038

1039

1040

1041

1042

1043

1044

1045

1046

1047

1048

1049

1050

1051

1052

1053

1054

1055

1056

1057

1058

1059

1060

1061

1062

1063

1064

1065

1066

1067

1068

1069

1070

1071

1072

1073

1074

1075

1076

1077

1078

1079

```
===== 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 5: Prompt of Rationale Generation for Gemini-Ultra

G 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 6: Prompt of RAG Drafting

H PROMPT OF STANDARD RAG

1080
1081
1082
1083
1084
1085
1086
1087
1088
1089
1090
1091
1092
1093
1094
1095
1096
1097
1098
1099
1100
1101
1102
1103
1104
1105
1106
1107
1108
1109
1110
1111
1112
1113
1114
1115
1116
1117
1118
1119
1120
1121
1122
1123
1124
1125
1126
1127
1128
1129
1130
1131
1132
1133

```

===== 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 7: 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 8: Prompt of Standard RAG for Instruction-tuned LM

I CASE STUDY

Figure 9 and 10 are two representative cases from TriviaQA and PubHealth. They show the 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 9 and 10.

```

===== 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 9: Case study of SPECULATIVE RAG from TriviaQA where *Dolly Parton* is the correct answer.

```

1188
1189 ===== Prompt of Drafter =====
1190 Response to the instruction. Also provide rationale for your response.
1191 ## Instruction: A lot of the problems with forest fires is because of bad policy (not to
1192 clear out the forests) that was put in place by environmentalists.
1193 ## Evidence:
1194 [1] Forest restoration
1195 That can be because of provoked deforestation in order to substitute forests by crop areas,
1196 or in dry areas, because of wild fires occurring naturally or intentionally. A whole
1197 section of forest landscape restoration is linked to this particular problem, as in many
1198 cases, the net loss of ecosystem value is very high and can open the drop to an accelerated
1199 further degradation of the soil conditions through erosion and desertification. This
1200 indeed has dire consequences on both the quality of the habitats and their related fauna.
1201 ...
1202 [2] Wildfire
1203 Attempts to thin out the fuel load may be met with opposition due to the desirability of
1204 forested areas. Wildland goals may be further resisted because of endangered species
1205 protections and habitat preservation. The ecological benefit of fire is often overridden by
1206 the economic benefits of protecting structures and lives. Additionally, federal policies
1207 that cover wildland areas usually differ from local and state policies that govern urban
1208 lands. ...
1209
1210 ===== Completion =====
1211 ## Rationale: Inadequate forest management can contribute to the severity of wildfires.
1212 ## Response: True
1213
1214 ===== Verification =====
1215 draft score: 0.8161
1216 self-containment score:0.1360
1217 self-reflection score: 0.8877
1218
1219 -----
1220 ===== Prompt of Drafter =====
1221 Response to the instruction. Also provide rationale for your response.
1222 ## Instruction: A lot of the problems with forest fires is because of bad policy (not to
1223 clear out the forests) that was put in place by environmentalists.
1224 ## Evidence:
1225 [1] 1997 Indonesian forest fires
1226 The 1997 Indonesian forest fires were caused by changing land use which made the tropical
1227 forest vulnerable to fire during a drought associated with that year's El Nino. Indonesian
1228 forests have historically been resistant to burning even during long dry seasons and
1229 despite the use of fire to clear land for swidden agriculture. The land use changes that
1230 led to the fires were a combination of industrial-scale logging, draining peatlands for
1231 conversion to oil palm and fast-growing tree plantations, and a massive government program
1232 to drain swamps and convert them to rice paddies. ...
1233 [2] Wildfire
1234 Attempts to thin out the fuel load may be met with opposition due to the desirability of
1235 forested areas. Wildland goals may be further resisted because of endangered species
1236 protections and habitat preservation. The ecological benefit of fire is often overridden by
1237 the economic benefits of protecting structures and lives. Additionally, federal policies
1238 that cover wildland areas usually differ from local and state policies that govern urban
1239 lands. ...
1240
1241 ===== Completion =====
1242 ## Rationale: The 1997 Indonesian forest fires were caused by **land use changes**,
1243 including industrial-scale logging, draining peatlands.
1244 ## Response: False
1245
1246 ===== Verification =====
1247 draft score: 0.8288
1248 self-containment score: 0.1325
1249 self-reflection score: 0.9533
1250
1251 -----
1252 ===== Prompt of Drafter =====
1253 Response to the instruction. Also provide rationale for your response.
1254 ## Instruction: A lot of the problems with forest fires is because of bad policy (not to
1255 clear out the forests) that was put in place by environmentalists.
1256 ## Evidence:
1257 [1] 1997 Indonesian forest fires
1258 The 1997 Indonesian forest fires were caused by changing land use which made the tropical
1259 forest vulnerable to fire during a drought associated with that year's El Nino. Indonesian
1260 forests have historically been resistant to burning even during long dry seasons and
1261 despite the use of fire to clear land for swidden agriculture. The land use changes that
1262 led to the fires were a combination of industrial-scale logging, draining peatlands for
1263 conversion to oil palm and fast-growing tree plantations, and a massive government program
1264 to drain swamps and convert them to rice paddies. ...
1265 [2] Wildfire
1266 Attempts to thin out the fuel load may be met with opposition due to the desirability of
1267 forested areas. Wildland goals may be further resisted because of endangered species
1268 protections and habitat preservation. The ecological benefit of fire is often overridden by
1269 the economic benefits of protecting structures and lives. Additionally, federal policies
1270 that cover wildland areas usually differ from local and state policies that govern urban
1271 lands. ...
1272
1273 ===== Completion =====
1274 ## Rationale: The 1997 Indonesian forest fires were caused by **land use changes**,
1275 including industrial-scale logging, draining peatlands.
1276 ## Response: False
1277
1278 ===== Verification =====
1279 draft score: 0.8288
1280 self-containment score: 0.1325
1281 self-reflection score: 0.9533
1282
1283 -----
1284 ===== Prompt of Drafter =====
1285 Response to the instruction. Also provide rationale for your response.
1286 ## Instruction: A lot of the problems with forest fires is because of bad policy (not to
1287 clear out the forests) that was put in place by environmentalists.
1288 ## Evidence:
1289 [1] 1997 Indonesian forest fires
1290 The 1997 Indonesian forest fires were caused by changing land use which made the tropical
1291 forest vulnerable to fire during a drought associated with that year's El Nino. Indonesian
1292 forests have historically been resistant to burning even during long dry seasons and
1293 despite the use of fire to clear land for swidden agriculture. The land use changes that
1294 led to the fires were a combination of industrial-scale logging, draining peatlands for
1295 conversion to oil palm and fast-growing tree plantations, and a massive government program
1296 to drain swamps and convert them to rice paddies. ...
1297 [2] Wildfire
1298 Attempts to thin out the fuel load may be met with opposition due to the desirability of
1299 forested areas. Wildland goals may be further resisted because of endangered species
1300 protections and habitat preservation. The ecological benefit of fire is often overridden by
1301 the economic benefits of protecting structures and lives. Additionally, federal policies
1302 that cover wildland areas usually differ from local and state policies that govern urban
1303 lands. ...
1304
1305 ===== Completion =====
1306 ## Rationale: The 1997 Indonesian forest fires were caused by **land use changes**,
1307 including industrial-scale logging, draining peatlands.
1308 ## Response: False
1309
1310 ===== Verification =====
1311 draft score: 0.8288
1312 self-containment score: 0.1325
1313 self-reflection score: 0.9533
1314
1315 -----
1316 ===== Prompt of Drafter =====
1317 Response to the instruction. Also provide rationale for your response.
1318 ## Instruction: A lot of the problems with forest fires is because of bad policy (not to
1319 clear out the forests) that was put in place by environmentalists.
1320 ## Evidence:
1321 [1] 1997 Indonesian forest fires
1322 The 1997 Indonesian forest fires were caused by changing land use which made the tropical
1323 forest vulnerable to fire during a drought associated with that year's El Nino. Indonesian
1324 forests have historically been resistant to burning even during long dry seasons and
1325 despite the use of fire to clear land for swidden agriculture. The land use changes that
1326 led to the fires were a combination of industrial-scale logging, draining peatlands for
1327 conversion to oil palm and fast-growing tree plantations, and a massive government program
1328 to drain swamps and convert them to rice paddies. ...
1329 [2] Wildfire
1330 Attempts to thin out the fuel load may be met with opposition due to the desirability of
1331 forested areas. Wildland goals may be further resisted because of endangered species
1332 protections and habitat preservation. The ecological benefit of fire is often overridden by
1333 the economic benefits of protecting structures and lives. Additionally, federal policies
1334 that cover wildland areas usually differ from local and state policies that govern urban
1335 lands. ...
1336
1337 ===== Completion =====
1338 ## Rationale: The 1997 Indonesian forest fires were caused by **land use changes**,
1339 including industrial-scale logging, draining peatlands.
1340 ## Response: False
1341
1342 ===== Verification =====
1343 draft score: 0.8288
1344 self-containment score: 0.1325
1345 self-reflection score: 0.9533

```

Figure 10: Case study of SPECULATIVE RAG from PubHealth where *False* is the correct answer.