

# Plan-and-Refine in RAG: Generating Diverse and Comprehensive Responses through Global Exploration and Local Exploitation

Alireza Salemi  
Center for Intelligent Information  
Retrieval  
University of Massachusetts Amherst  
Amherst, MA, USA  
asalemi@cs.umass.edu

Chris Samarinas  
Center for Intelligent Information  
Retrieval  
University of Massachusetts Amherst  
Amherst, MA, USA  
csamarinas@cs.umass.edu

Hamed Zamani  
Center for Intelligent Information  
Retrieval  
University of Massachusetts Amherst  
Amherst, MA, USA  
zamani@cs.umass.edu

## Abstract

This paper studies the limitations of (retrieval-augmented) large language models (LLMs) in generating diverse and comprehensive responses, and introduces the Plan-and-Refine (P&R) framework based on a two phase system design. In the global exploration phase, P&R generates a diverse set of plans for the given input, where each plan consists of a list of diverse query aspects with corresponding additional descriptions. This phase is followed by a local exploitation phase that generates a response proposal for the input query conditioned on each plan and iteratively refines the proposal for improving the proposal quality. Finally, a reward model is employed as a utility function to select the proposal with the highest factuality and coverage scores. We conduct our experiments based on the ICAT evaluation methodology—a recent approach for answer factuality and comprehensiveness evaluation. Experiments on the two diverse information seeking benchmarks adopted from non-factoid question answering and TREC search result diversification tasks demonstrate that P&R significantly outperforms baselines, achieving up to a 13.1% improvement on the ANTIQUE dataset and a 15.41% improvement on the TREC dataset. Furthermore, a smaller scale user study confirms the substantial efficacy of the P&R framework.

## CCS Concepts

• **Information systems** → **Information retrieval diversity**; *Question answering*; **Information retrieval query processing**; • **Computing methodologies** → **Natural language generation**.

## Keywords

Retrieval-augmented generation, planning, iterative refinement, diverse text generation, comprehensive answer generation, factuality

## ACM Reference Format:

Alireza Salemi, Chris Samarinas, and Hamed Zamani. 2026. Plan-and-Refine in RAG: Generating Diverse and Comprehensive Responses through Global Exploration and Local Exploitation. In *Proceedings of the 2026 International ACM SIGIR Conference on Innovative Concepts and Theories in Information Retrieval (ICTIR) (ICTIR '26)*, July 25, 2026, Melbourne, VIC, Australia. ACM, New York, NY, USA, 13 pages. <https://doi.org/10.1145/3805713.3820415>



This work is licensed under a Creative Commons Attribution International 4.0 License.

ICTIR '26, Melbourne, VIC, Australia

© 2026 Copyright held by the owner/author(s). Publication rights licensed to ACM.  
ACM ISBN 979-8-4007-2600-2/2026/07  
<https://doi.org/10.1145/3805713.3820415>

## 1 Introduction

LLMs have shown strong performance in text generation by producing fluent, coherent, engaging, and contextually related responses to their inputs [18, 21, 24, 70]. To address the well-known hallucination issue and deal with non-stationary and up-to-date information, state-of-the-art question answering systems as well as generative and conversational search engines enhance LLMs through retrieval augmentation [78], commonly referred to as retrieval-augmented generation (RAG). However, recent studies reveal that the text generated by RAG models still contains non-factual content [44], and more importantly, often lacks diversity and comprehensiveness [60]. This limitation is critical because diversity in responses allows users to explore multiple perspectives, uncover less obvious yet relevant information, and make more informed decisions. In applications such as non-factoid question answering, exploratory search, healthcare, legal assistance, education, research, and other information-driven decision-making systems, the absence of diverse and comprehensive responses can result in incomplete understanding, overlooked alternatives, and reduced user trust. Ensuring accuracy, diversity, and comprehensiveness is therefore not just a desirable feature, but a necessity for these real-world scenarios.

This paper bridges this gap by developing long-form generation methods with the following defined desiderata: (1) **diversity and comprehensiveness**: LLM responses must address all diverse aspects of the input, and (2) **factuality**: claims made by the LLM must be factually accurate. While the concept of novelty and diversity in retrieval has been explored within the information retrieval community [8, 37, 42, 61], training LLMs and RAG systems to generate diverse and comprehensive responses is relatively underexplored.

Recent work by Samarinas et al. [60] shows that highly capable open-source and proprietary LLMs not only sometimes generate non-factual content, but also do not perform well in generating diverse and comprehensive responses, even if they are specifically asked to in their prompts. This observation has also been validated in our experiments. We also observe that diversifying the retrieval results in the RAG pipelines does not improve response coverage and completeness. This deficiency in generating comprehensive and factual responses arises from several factors. First, the pre-training objectives for sequence-to-sequence models [67] and post-training techniques [23, 46, 64] are not specifically designed to encourage the generation of diverse outputs. We observe that techniques like Chain-of-Thought (CoT) prompting [40, 72, 75] that perform well in mathematical reasoning tasks, fall short in improving response diversity and completeness. Second, the prevalent autoregressive

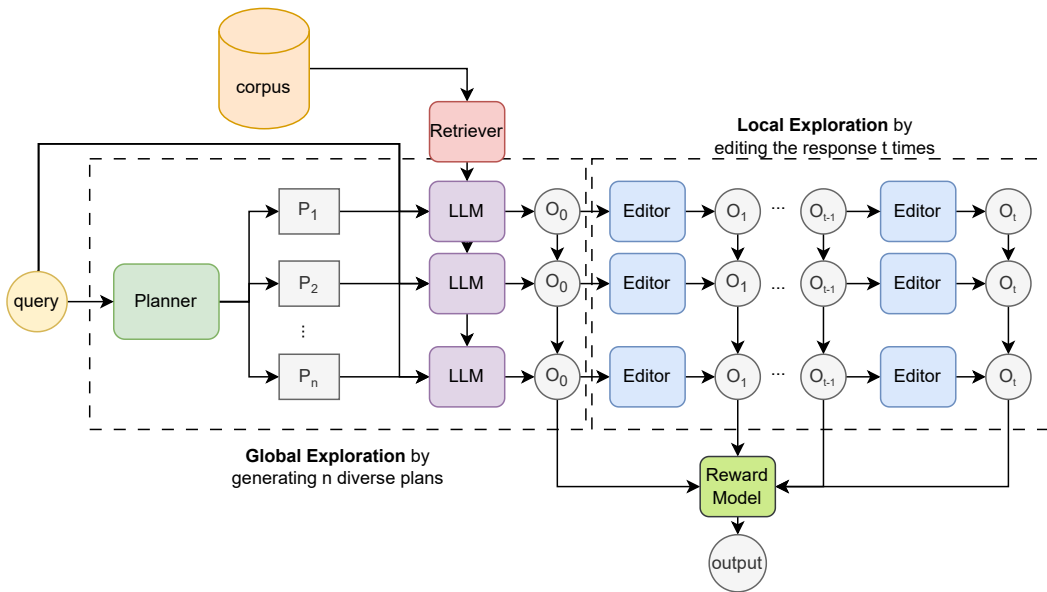


Figure 1: An overview of the P&R framework.

generation paradigm, which relies on greedy decoding or sampling-based token selection, is inherently limited. It tends to favor locally optimum token predictions, often overlooking factual and comprehensive completions that diverge from the initial token prefix. This token-by-token generation process exacerbates the influence of early poor token choices, potentially distorting the response structure and leaving critical elements inadequately addressed.

This paper introduces the *Plan-and-Refine* (P&R) framework to address both of these issues. P&R is generic and can be applied to any RAG pipelines, regardless of their retrieval, reranking, and generation approaches.<sup>1</sup> An overview of this framework is presented in Figure 1. The framework consists of two main phases. We refer to the first phase as “*planning*,” which generates a set of diverse plans for *global exploration*. Each plan includes a list of diverse query aspects essential for creating a comprehensive response to the query, the reasoning behind the value and relevance of each aspect, and the corresponding query formulation of each aspect for retrieving diverse and relevant content. A diverse set of plans are created by a *planner* that we optimize via self-training, enabling it to identify diverse key aspects and later structure comprehensive responses effectively. Using each generated plan and the retrieved information for each reformulated query aspect in the plan, an LLM generates a detailed response to the query. Therefore, each plan results in a potential response for the query. This global exploration phase is followed by a refining phase as *local exploitation*. This second phase, refines the LLM response multiple times to improve its comprehensiveness and factuality conditioned on the given plan. Finally, P&R utilizes a trained reward model to evaluate all the generated refinements. The reward model selects the refinement with the

highest factuality and coverage, ensuring the final output provides the most accurate and comprehensive answer to the input query.

We conduct our experiments on two diverse information seeking tasks that benefit from comprehensive responses. In the first set of experiments, we adopt ANTIQUE [25]—the largest public non-factoid question answering dataset with complete manual relevance judgments. In the second set of experiments, we adopt the description queries in the TREC Web Track data from 2009 to 2012 [12–15] which is based on ClueWeb09 English documents. TREC Web Track ran a successful search result diversification task during this period, meaning that the queries in the dataset have multiple aspects and benefit from diverse perspectives. We used the recent ICAT evaluation methodology [60] for evaluating factuality and information coverage in the generated text in response to queries in these datasets. Our results demonstrate that the P&R framework outperforms a competitive and diverse set of open-source and proprietary baselines across both datasets, achieving a statistically significant relative improvement of 13.1% on the ANTIQUE dataset and 15.4% on the TREC datasets. We further conducted a small user study to demonstrate user’s preferences over the best baseline model. A follow-up side-by-side user study suggests that in 63% of cases, annotators prefer P&R’s responses over the ones produced by our best performing RAG baseline with the same LLM. To foster research in this area, we release our codebase.<sup>2</sup>

## 2 Related Work

**Retrieval-Augmented Generation.** RAG [30, 38, 57] is a framework that integrates information retrieval and text generation to enhance the quality and relevance of generated content by incorporating external knowledge during generation [4, 59, 65]. In contrast

<sup>1</sup>P&R is not limited to RAG and can be applied to any LLM, however, since retrieval is essential in generating factual up-to-date responses, we solely focus on P&R in RAG.

<sup>2</sup>Available at: <https://github.com/alirezasailemi7/PR-RAG>

to traditional LLMs that rely solely on knowledge acquired during pre-training, RAG systems retrieve information from external knowledge bases using a retriever, allowing them to produce contextually and factually accurate outputs [33, 56, 78]. The versatility of RAG makes it applicable to various domains, including knowledge-grounding in textual [30, 38, 47, 77] and multi-modal [9, 22, 50, 55], personalization [35, 51–54, 58], and reducing hallucinations in generated content [1, 63]. This paper uses RAG to enhance factuality and coverage of responses generated by LLMs.

**Planning & Reasoning in Text Generation.** Addressing complex problems often involves breaking them down into smaller subproblems and solving each independently [28, 29, 31]. This process can be viewed as planning a sequence of simpler steps to tackle a larger challenge. These subproblems may, in turn, require reasoning to solve effectively. Reasoning refers to a model’s capability to process information step-by-step, often referred to as chain-of-thought (CoT). CoT reasoning enhances performance of LLMs on tasks requiring mathematical, logical, and commonsense reasoning [40, 72, 75]. This reasoning ability has recently been applied in areas such as evaluation [32], code generation [10], improving alignment [49], and personalization [53]. Research on reasoning in free-form text generation has been relatively limited, focusing primarily on logical and complex tasks [76]. However, recent studies have shown its effectiveness in generating high-quality free-form text, including tasks requiring emotional and personalized generation [39, 53]. This paper uses planning with reasoning to improve comprehensiveness and factuality of generated responses by LLMs. Specifically, we leverage planning to explore the global space of complete and factual solutions to the input prompt prior to generating the final response. Following this, reasoning is applied to assess and execute each step in the solution, ensuring factuality and comprehensiveness of responses.

**Diversity & Coverage in Text Generation.** Diversity and coverage have been extensively studied in retrieval community [7, 42, 73], with several TREC tracks dedicated to them [12–15]. Traditionally, these concepts have been approached as syntactical problems in text generation, where diversity is often evaluated based on the variety of words and phrases using n-gram metrics [3], with less attention given to content diversity [68]. Consequently, much of the prior work has focused on improving syntactical diversity [62, 74]. In parallel, ICAT [60] has been introduced as a metric that evaluates diversity, completeness, and factuality of generated responses based on their content rather than syntax. In this paper, we focus on enhancing the diversity, coverage, and factuality of LLM-generated response to the question, specifically in terms of content.

**Scaling Test-Time Compute.** Recent advancements in the reasoning capabilities of LLMs for logical and mathematical tasks have shown that increasing the compute budget during the inference phase is an effective approach to improve performance [11, 66]. This method allows LLMs to utilize additional inference resources to explore the response space, enabling providing more accurate answers in tasks such as logical reasoning, mathematical problem-solving, and code generation [5, 6, 79]. While most prior research has studied this concept in domains such as math, code, and logic, its use in free-form text generation remains relatively unexplored.

We extend this to free-form generation, utilizing increased inference compute to search the response space more effectively and generate responses that are more comprehensive and factual.

### 3 Problem Formulation

A generative language model  $M_G$  takes a prompt  $x$  and produces  $\bar{y}$  as the response. The quality of the generated output can be assessed based on various factors; coherence, factual accuracy, relevance, fluency, and alignment. One aspect that has received relatively little attention is *comprehensiveness* while maintaining *factuality* [60]. In this context, the response should offer a comprehensive and thorough coverage of topics related to the input  $x$ , ensuring the output remains factually accurate and minimizes hallucinated or incorrect information based on a reference knowledge corpus  $C$ . This corpus can take various forms; unstructured text, an encyclopedia, or even the entire web. The only requirement is that it must be a trusted source of information. The primary goal of this paper is to enhance the capability of LLMs to generate responses that are both highly factual and comprehensive. We assume that this quality can be quantified using a utility function or evaluation metric  $\mu$ . Specifically, we employ ICAT [60] as the evaluation metric to assess the coverage of diverse factual information in long-form text generation. We focus on improving the ability of LLMs to achieve higher ICAT scores, thus advancing the quality of their output in terms of coverage of diverse factual information. We assume access to a set of training queries that benefit from comprehensive and diverse responses  $D_{\text{train}} = \{x_i\}_{i=1}^{|D_{\text{train}}|}$  and a set of validation queries  $D_{\text{test}} = \{x_i\}_{i=1}^{|D_{\text{test}}|}$ , both consist solely of input prompts without any corresponding reference outputs. Such queries can be obtained from non-factoid question answering datasets, community question answering websites, and discussion forums. In this setup, we propose methods to improve  $\mu$  in a reference-free setting.

### 4 The P&R Framework

Ensuring accurate and complete LLM responses is vital to prevent misinformation, cover key aspects of prompts comprehensively, and build user trust [19, 60]. As discussed in Section 1, prior research highlights that LLMs struggle to consistently produce complete and accurate responses. Even very capable models like GPT-4 [45] cover less than 50% of relevant subtopics on average for a given prompt [60]. Furthermore, while RAG enhances factuality, we show it reduces the coverage of generated responses (see Section 5.2 for details). Several factors contribute to this deficiency in LLMs. Current pre-training sequence-to-sequence [67] and post-training objectives [23, 46, 64] do not effectively encourage factual and comprehensive responses. Even techniques like Chain-of-Thought prompting [40, 72, 75], designed for mathematical reasoning, fail to improve response completeness. Moreover, the token-by-token text generation approach, whether greedy or sampled, is sub-optimal. It can often overlook factual and complete responses that deviate from the prefix of generated tokens, leading to incomplete outputs. In essence, the LLM’s initial token selection influences output structure, causing key aspects to be missed or underrepresented.

A straightforward solution to these issues could be to explicitly ask LLMs to generate complete and factual responses that consider all aspects of the question. Furthermore, post-training techniques

such as RLHF [46] or Self-Training [23, 64] can be used to optimize a reward model or metric that accounts for the completeness of the response. However, these approaches do not address the inherent problem of sampling responses from LLMs, where the structure of the output can still be influenced by the initial tokens, potentially leading to incomplete or inaccurate responses. To address the two aforementioned challenges, we introduce P&R, a novel approach that first generates a set of plans outlining the aspects that need to be covered, along with the rationale about why each aspect is important for ensuring a complete and factual response to the prompt and the query to retrieve information about each aspect. The model then generates responses based on each plan and retrieved documents and iteratively refines them through multiple editing steps. Finally, a reward model is employed to select the response with the highest score as the final output. The following subsections provide a detailed explanation of this approach.

**Overview.** The overview of P&R is shown in Figure 1. We assume the existence of a planner model  $M_P(x)$  that takes the input prompt  $x$  and returns a plan  $p$  for generating factual and complete responses to the prompt. The plan  $p = \{(a_i, q_i, r_i)\}_{i=1}^{|p|}$  consists of a set of aspects  $a_i$  about the prompt, a query  $q_i$  to gather information about the respective aspect, and a reason  $r_i$  explaining why this aspect is important for generating a complete and factual response to the prompt  $x$ . We assume the existence of a retrieval model  $R$  and a retrieval budget  $k$  to collect the necessary information for improving the factuality of the claims in the response. To gather the necessary information for executing a plan  $p$ , for each  $(a_i, q_i, r_i) \in p$ , we retrieve  $\frac{k}{|p|}$  documents for the query  $q_i$  from the corpus  $C$ . The retrieved documents form the context  $I_p = \bigcup_{(a_i, q_i, r_i) \in p} R(q_i, \frac{k}{|p|}, C)$  that can be used during response generation. We assume the existence of a generative model  $M_G(x, p, I_p)$  that, given a prompt  $x$ , a plan  $p$ , and the associated context  $I_p$ , generates an output response  $o_p$  to the prompt  $x$  with the given plan  $p$  as the steps it should take. To explore diverse solutions to the problem, we sample  $n$  distinct plans using the planner model  $M_P$ , resulting in a set of plans  $P = \{p_i\}_{i=1}^n$ . These plans provide a range of strategies that can be utilized to address the problem effectively. This step can be seen as sampling and searching through the space of all potential solutions to the problem, a process we refer to as **Global Exploration**. Then, the generative model  $M_G$  is applied to each plan  $p \in P$ , producing an initial set of proposed responses  $O_0 = \{M_G(x, p, I_p) \mid p \in P\}$ .

While global exploration generates a diverse set of solutions to the prompt, it often falls short in meeting specific requirements with precision. To address this, we introduce the concept of **Local Exploitation**, which focuses on refining these solutions through targeted adjustments. This approach enhances and ensures higher-quality responses. For this, we assume the existence of an editing model  $M_E$  that, given the input prompt  $x$ , a plan  $p$ , and a previously generated response  $o_{t-1}$  for this prompt and plan, improves the response to generate  $o_t = M_E(x, p, o_{t-1})$ . Using this iterative approach, we can refine the initial set of generated responses. At each step, the updated responses are represented as  $O_t = \{M_E(x, p, o_{t-1}) \mid o_{t-1} \in O_{t-1}, p \in P\}$ . By repeating this editing process  $T$  times, we obtain a final set of response proposals, denoted as  $O_F = \bigcup_{t=0}^T O_t$ , which encompasses all the initial

Plan Generation	<p>Your task is to convert the following search query into maximum 5 diverse aspects and perspectives that cover all aspects of the original query. The aspects and perspectives should be non-overlapping and should not be redundant. The aspects and perspectives should cover all aspects that a comprehensive response to the original search query should cover.</p> <p># your input: - query: the original search query</p> <p># your output: Your output should be a valid json list of maximum 5 items enclosed in ""json"" block that contains the following fields:</p> <ul style="list-style-type: none"> <li>- aspect: the aspect that covers a specific aspect of the original search query</li> <li>- query: the query that should be used to cover the specific aspect</li> <li>- reason: the reason why this aspect and query is important to cover in a comprehensive response to the original search query</li> </ul> <p>query: {query} output: ""json</p>
Response Generation with Plan & context	<p>Your task is to generate a comprehensive and factual response to the given query. You can use the information provided in the context to generate a more comprehensive and factual response. Your response should cover the following aspects and perspectives that cover all aspects of the original query. You can use the following plan to generate a comprehensive response to the query.</p> <p>query: {query} plan: To answer the query, you should cover the following aspects and perspectives:</p> <ul style="list-style-type: none"> <li>- aspect: {aspect_1}</li> <li>- reason: {reason_1}</li> <li>...</li> <li>- aspect: {aspect_n}</li> <li>- reason: {reason_n}</li> </ul> <p>context: {retrieved_docs}</p> <p>response:</p>
Response Editing	<p>Your task is to improve the comprehensiveness and accuracy of the response generated for the query. To achieve this, provide a more detailed and factually accurate response, using the provided plan as a guide to ensure the response is both thorough and precise.</p> <p>query: {query} plan: To answer the query, you should cover the following aspects and perspectives:</p> <ul style="list-style-type: none"> <li>- aspect: {aspect_1}</li> <li>- reason: {reason_1}</li> <li>...</li> <li>- aspect: {aspect_n}</li> <li>- reason: {reason_n}</li> </ul> <p>context: {retrieved_docs} generated response: {initial_response}</p> <p>improved response:</p>

Figure 2: The prompt templates used with P&R.

set of responses and refined outputs generated in iterations. Finally, to identify the most suitable response among all proposed candidates, we need to employ a mechanism to select the one that best meets the prompt's requirements, prioritizing completeness and factuality—key objectives of this problem. We assume the existence of a reward model  $M_R(x, o)$  that assigns a score to each generated output  $o \in O_F$  based on the input prompt  $x$ . The final response to the prompt is selected as the output that achieves the highest score according to the reward model, formally as:  $o_f = \arg \max_{o \in O_F} M_R(x, o)$ . This ensures the chosen response is the most complete and factual among the generated candidates.

#### 4.1 Global Exploration through Planning

We define a plan for responding to a prompt  $x$  as a set of steps, each consisting of three key components: 1) a title that identifies an aspect to be addressed in order to provide a complete and factual response, 2) a justification or reasoning that explains why this particular aspect is important and how it contributes to addressing the prompt, and 3) a retrieval query designed to gather information about the specified aspect from a corpus. To obtain a plan, we sample it from a planner  $M_P$ , which is an LLM guided by the prompt shown in Figure 2. This prompt guides the LLM to analyze the input  $x$  and generate the aspects that should be included in a complete and factual response. Next, using the queries in the generated plan  $p$ , we employ the retriever  $R$  within a defined retrieval budget  $k$  to gather a supporting context. For each component  $(a_i, q_i, r_i) \in p$ , we

retrieve  $\frac{k}{|p|}$  documents from the corpus  $C$ . The resulting context for the plan  $p$  is denoted as  $I_p = \bigcup_{(a_i, q_i, r_i) \in p} R(q_i, \frac{k}{|p|}, C)$ . To produce a response for the prompt  $x$ , we use the generative model  $M_G$ . This model, which is an LLM, takes the input  $x$ , the generated plan  $p$ , and the corresponding context  $I_p$  to generate a response. The process uses the “response generation with plan and context” prompt in Figure 2. This prompt guides the model to incorporate the generated aspects and their associated reasoning from the plan, along with the provided context, to produce a comprehensive and factual response.

To sample a plan from  $M_p$ , various strategies can be employed. The most common is greedy sampling, which returns the most probable plan for the given input  $x$ . While effective, the most probable plan may not always yield the most complete and factual response. Alternatively, nucleus sampling [27], which introduces randomness, can generate diverse plans but risks reducing performance when only one plan is sampled. To balance these trade-offs, we propose sampling  $N$  plans using high-temperature nucleus sampling, denoted as  $P = \{p_i | p_i \sim M_p(x), \text{ for } i = 1, \dots, N\}$ . This allows us to explore multiple potential strategies for addressing  $x$ , effectively conducting a global search across the response space to identify diverse and potentially better plans for response generation. Finally, for each plan  $p \in P$ , we use the generative model  $M_G$  to generate a response. This results in an initial set of responses, denoted as  $O_0 = \{M_G(x, p, I_p) | p \in P\}$ , serving as the starting point for further refinement and selection in response to the input  $x$ .

**Optimization.** To optimize this step, we employ Self-Training [23, 64] as the optimization approach. Importantly, during this process, we only optimize the planner model  $M_p$ , while keeping the generative model  $M_G$  frozen. For this purpose, for each input  $x \in D_{\text{train}}$ , we sample  $B = 32$  plans using a high temperature  $\tau = 0.7$ . We then generate a response for each plan and its corresponding context. To select high-quality plans that resulted in high-quality responses, we use the evaluation metric  $\mu$  and retain only those plans that their corresponding responses achieved a score higher than the input-dependent threshold  $\alpha_x$ , as follows:

$$D_{\text{plan}} = \{(x, p) | x \sim D_{\text{train}}; p \sim M_p(x); \mu(x, M_G(x, p, I_p)) \geq \alpha_x\}$$

to form the training dataset  $D_{\text{plan}}$  for training the the planner model. We set the  $\alpha_x$  parameter based on the score of the generated responses. Specifically,  $\alpha_x$  is chosen as the score corresponding to the top  $Z$ -percentile of the generated responses (we use  $z = 95$  by default in this paper, unless otherwise noted). This ensures that only the highest-scoring responses, as determined by the evaluation metric  $\mu$ , are retained for training the planner model. Finally, we train the planner using a sequence-to-sequence loss function [67] with  $D_{\text{plan}}$ , where for each  $(x, p) \in D_{\text{plan}}$  the model to generates the plan  $p$  as the output for the input  $x$ .

## 4.2 Local Exploitation through Refining

While global exploration generates a diverse set of solutions to the prompt, it often lacks the precision required to meet specific requirements. In fact, we observed that refining the response generated by a plan, using the same plan, leads to improved results. This suggests that focusing on enhancing existing solutions rather than exploring new ones can also yield to more accurate and complete responses to the prompt. This iterative process of improving the

Vanilla LLM	Your task is to generate a comprehensive and factual response to the following query: query: {query} response:
RAG LLM	Your task is to generate a comprehensive and factual response to the given query. You can use the information provided in the context to generate a more comprehensive and factual response. query: {query} context: {retrieved_docs} response:
LLM with CoT	Your task is to generate a comprehensive and factual response to the following query. You should first think step by step about the information that is needed to be present in the answer to the query and then generate a response that is both comprehensive and factually accurate. You should start your thinking by "thought:" and your final response to the query by "response:". query: {query} thought:
RAG LLM with CoT	Your task is to generate a comprehensive and factual response to the following query. You should first think step by step about the information that is needed to be present in the answer to the query and then generate a response that is both comprehensive and factually accurate. You should start your thinking by "thought:" and your final response to the query by "response:". You can use the information provided in the context to generate a more comprehensive and factual response. query: {query} context: {retrieved_docs} thought:

Figure 3: The prompts used by the baselines.

response generated for a prompt  $x$  by a plan  $p$ , using the same plan and refining the solution based on previous outputs, can be viewed as a local exploitation over the response space. Unlike global exploration, where both the plan and the output can vary, in local exploitation, the plan—the general instruction for the model in responding to the prompt—remains the same. It is the output that evolves through successive edits, refining the response according to the same guiding plan. To perform iterative refinement, we use the editing model  $M_E$ , an LLM that uses the response editing prompt shown in Figure 2. The model takes the input  $x$ , the plan  $p$ , and the previous output  $o_{t-1}$  generated using this plan as input, and produces the refined output  $o_t = M_E(x, p, o_{t-1})$ . This iterative process allows us to refine the initial set of responses  $O_0$  from the global exploration phase. At each step, the updated responses are represented as  $O_t = \{M_E(x, p, o_{t-1}) | o_{t-1} \in O_{t-1}, p \in P\}$ . By repeating this editing process  $T$  times, we obtain a final set of proposals, denoted as  $O_F = \bigcup_{t=0}^T O_t$ , encompassing both the initial responses and the refined outputs after each editing step generated through the iterative steps. Therefore, the final response to the input  $x$  can be selected from the set of proposed responses  $O_F$ , resulting from both global exploration using diverse plans and multiple rounds of local exploitation through iterative editing of responses.

**Optimization.** To optimize the editing model  $M_E$ , we first sample a plan  $p$  from the optimized planner model  $M_p$  for each input  $x \in D_{\text{train}}$ . For each plan  $p$ , we generate  $B = 8$  pairs of outputs from the generative model  $M_G$  using a high sampling temperature  $\tau = 0.7$ . These pairs are selected such that the difference in their scores, as evaluated by the metric  $\mu$ , is at least  $\beta$ . This ensures that the training dataset for  $M_E$  includes meaningful differences in response quality, so the model can learn how to improve the previous response and generate a new one. Then, we form the training dataset  $D_{\text{edit}}$  for training the editing model  $M_E$  as:

$$D_{\text{edit}} = \{(x, p, o_0, o_1) | x \sim D_{\text{train}}; p \sim M_p(x); o_0, o_1 \sim M_G(x, p, I_p); \mu(x, o_1) - \mu(x, o_0) \geq \beta\}$$

where we set  $\beta = 0.1$ . To optimize the editing model  $M_E$ , we employ the sequence-to-sequence loss function [67]. For each training example  $(x, p, o_0, o_1) \in D_{\text{edit}}$ , the model takes the input  $x$ , the plan  $p$ , and the lower-quality output  $o_0$  as input and is trained to generate the higher-quality output  $o_1$ . This objective aligns the editing model’s predictions with outputs that demonstrate improved quality, as defined by the evaluation metric  $\mu$ .

### 4.3 Response Selection through Ranking

Global and local exploitation steps produce a set of proposed responses  $O_F$ , rather than a single response. To generate a final response  $o_f$  to the prompt  $x$ , a selection mechanism is required to identify the most suitable response. For this purpose, we use a reward model  $M_R$ , which evaluates each candidate response based on the prompt  $x$  and assigns it a score between 0 and 1. To implement  $M_R$ , we employ a text encoder model  $Enc$ . The reward model computes the score as follows:  $M_R(x, o) = \sigma(Enc([x.o]) \cdot W)$  where  $W \in \mathbb{R}^{d \times 1}$  is a trainable weight matrix,  $d$  represents the dimension of the encoder’s output representations,  $\sigma$  is the sigmoid activation function, and  $[.]$  is the concatenation with separate token function. We use ModernBERT-base<sup>3</sup> [71] as the encoder for our reward model. This formulation allows  $M_R$  to evaluate the relevance and quality of a response  $o$  to the prompt  $x$ , enabling the selection of the final response  $o_f$  as:  $o_f = \arg \max_{o \in O_F} M_R(x, o)$ .

**Optimization.** To optimize the reward model  $M_R$ , we construct a training dataset by sampling  $B = 8$  pairs of plans  $p_0$  and  $p_1$  from the optimized planner model  $M_P$  with a high temperature  $\tau = 0.7$  for each input  $x \in D_{\text{train}}$ . The corresponding outputs  $o_0 = M_G(x, p_0, I_{p_0})$  and  $o_1 = M_G(x, p_1, I_{p_1})$ , generated using the generative model  $M_G$ , are included in the dataset if the difference in their evaluation scores  $\mu(x, o_1) - \mu(x, o_0)$  is at least  $\gamma$ . Formally, the training dataset for reward model is defined as:

$$D_{\text{reward}} = \{(x, o_0, o_1) \mid x \sim D_{\text{train}}; p_0, p_1 \sim M_P(x); \\ o_0 \sim M_G(x, p_0, I_{p_0}); o_1 \sim M_G(x, p_1, I_{p_1}); \mu(x, o_1) - \mu(x, o_0) \geq \gamma\}$$

where we set  $\gamma = 0.1$ . To train the reward model  $M_R$ , following Ouyang et al. [46], we minimize the following loss function:

$$L = \mathbb{E}_{(x, o_0, o_1) \sim D_{\text{reward}}} [-\log(\sigma(M_R(x, o_1) - M_R(x, o_0)))]$$

where  $\sigma$  is the sigmoid function. This pairwise loss function is designed to ensure that the reward model assigns higher scores to preferred outputs relative to less preferred ones, as determined by the evaluation metric  $\mu$ . By enforcing this, the model learns to differentiate between higher-quality and lower-quality responses, aligning its predictions with the preferences encoded in  $\mu$ .

## 5 Experiments

### 5.1 Experimental Setup

**Datasets.** We use ANTIQUE [25], a retrieval dataset designed for non-factoid question answering, and TREC Web Track Diversity tasks from 2009 to 2012 [12–15]. These datasets do not include predefined gold responses to questions, but provide a corpus containing the necessary information to answer them. The ANTIQUE dataset consists of 2,426 training questions and 200 test questions.

<sup>3</sup>Available at: <https://hf.co/answerdotai/ModernBERT-base>

For pre-processing, we filter out documents with fewer than 50 words from the corpus to ensure the quality and richness of the documents in the knowledge source. This filtering results in a corpus consisting of 97,327 documents. For the TREC Web Track Diversity tasks, there is no training dataset available, but the query set consists of 200 queries. We exclude queries that seek information about a specific webpage (navigational), reducing the set to 179 queries. For the corpus, we use the ClueWeb09 corpus [69]. Note that we only use this dataset to evaluate the P&R framework under the zero-shot setting, as it does not include any training query set.

**Evaluation.** We evaluate the factuality and coverage of the generated responses using the ICAT metric [60], which is specifically designed for this purpose. ICAT offers three levels of annotation for evaluating responses: 1) ICAT<sup>M</sup>: Requires a predefined set of subtopics for each query, along with annotations specifying which subtopics are addressed by each document in the corpus, 2) ICAT<sup>S</sup>: Similar to ICAT<sup>M</sup>, but leverages an LLM to determine which subtopics are covered by a document, eliminating the need for manual document-level annotations, and 3) ICAT<sup>A</sup>: Extends ICAT<sup>S</sup> by using an LLM to generate the subtopics for a query, removing the dependency on predefined subtopic annotations. ICAT also employs natural language inference (NLI) to fact-check the claims in the generated response. The final score is calculated using the F-measure, balancing the factuality of the response with its coverage of the subtopics. For more details, we refer the reader to Samarinas et al. [60]. For the LM backbone, we follow Samarinas et al. [60] and use an instruction-tuned LLaMA 3.1 model with 8 billion parameters.<sup>4</sup> For extracting atomic claims, we leverage the trained version of this model provided by ICAT.<sup>5</sup> For NLI and fact verification, we employ a trained DeBERTa v3 [26, 36] model suggested by ICAT.<sup>6</sup> As the knowledge source, we use the corresponding corpus in each of the evaluation datasets, i.e., the ANTIQUE corpus and the ClueWeb09-Category B English corpus for the TREC Web Track queries. Spam documents were removed from the ClueWeb corpus using the Waterloo Spam Scorer with the 70% threshold.<sup>7</sup>

**Training & Inference Configurations.** We use the Adam optimizer [34] with a learning rate of  $5 \times 10^{-5}$  for training the LLMs and  $1 \times 10^{-5}$  for training the reward model. Gradient clipping is applied with a value of 1, and the training is conducted for a maximum of 2000 steps. A warmup phase is set for 2.5% of the training steps, following a linear learning rate scheduler. Models are evaluated every 100 steps using 10% of the training set as a randomly sampled validation subset, and the checkpoint with the best performance is selected. We set the combined maximum input and output length to 4096 tokens. We use the instruction-tuned Gemma v2 [20] with 2 billion parameters as the LLM for planner and editing model and ModernBERT-base [71] with 150 million parameters as the reward model. For the generative model, which remains frozen in all experiments, we use instruction-tuned Gemma v2 [20] with 2 billion parameters,<sup>8</sup> LLaMA 3.2 [2] with 1.2 billion parameters,<sup>9</sup>

<sup>4</sup>Available at: <https://hf.co/meta-llama/Llama-3.1-8B-Instruct>

<sup>5</sup>Available at: <https://hf.co/algoprog/fact-generation-llama-3.1-8b-instruct-lora>

<sup>6</sup>Available at: <https://hf.co/MoritzLaurer/DeBERTa-v3-base-mnli-fever-anli>

<sup>7</sup>Available at: <https://plg.uwaterloo.ca/~gvcormac/clueweb09spam/>.

<sup>8</sup>Available at: <https://hf.co/google/gemma-2-2b-it>

<sup>9</sup>Available at: <https://hf.co/meta-llama/Llama-3.2-1B-Instruct>

**Table 1: Performance of P&R compared to baselines on ANTIQUE. The † and ‡ show statistically significant improvements over the best open-source and proprietary baselines, respectively, as determined by a t-test ( $p < 0.05$ ).**

Method	ICAT <sub>Coverage</sub>	ICAT <sub>Factuality</sub>	ICAT-A <sub>1</sub>
<b>Proprietary LLMs</b>			
1 Gemini 2.0 Flash	<b>0.7057</b>	0.4488	0.5214
2 GPT-4o mini	0.6551	0.4934	0.5376
<b>Retrieval-Augmented Proprietary LLMs</b>			
3 RAG Gemini 2.0 Flash	0.6499	0.5474	0.5640
4 RAG GPT-4o mini	0.6439	0.5354	0.5576
<b>Open-Source LLMs</b>			
5 Llama 3.2	0.3959	0.3201	0.3251
6 - w/ CoT	0.3523	0.3444	0.3207
7 - w/ Best-of-N	0.4521	0.3995	0.3924
8 Phi 3 mini	0.5483	0.4433	0.4511
9 - w/ CoT	0.4973	0.4219	0.4116
10 - w/ Best-of-N	0.5489	0.4754	0.4741
11 Gemma v2	0.6064	0.4936	0.5143
12 - w/ CoT	0.5257	0.4890	0.4659
13 - w/ Best-of-N	0.5789	0.4787	0.4952
14 - w/ Self-Training	0.5839	0.5268	0.5243
<b>Retrieval-Augmented Open-Source LLMs</b>			
15 RAG Llama 3.2	0.3162	0.3295	0.2872
16 - w/ CoT	0.3112	0.3243	0.2878
17 - w/ Best-of-N	0.3564	0.3712	0.3363
18 - MMR Reranking	0.3005	0.2830	0.2751
19 RAG Phi 3 mini	0.5369	0.5557	0.5022
20 - w/ CoT	0.5173	0.5635	0.5071
21 - w/ Best-of-N	0.5493	0.5386	0.5021
22 - MMR Reranking	0.5541	0.5656	0.4758
23 RAG Gemma v2	0.5457	0.5904	0.5256
24 - w/ CoT	0.5028	0.5655	0.4880
25 - w/ Best-of-N	0.4873	0.5809	0.4901
26 - w/ Self-Training	0.5382	0.6054	0.5310
27 - MMR Reranking	0.5162	0.5977	0.5006
<b>P&amp;R</b>			
28 Llama 3.2	0.6145 <sup>†</sup>	0.6333 <sup>‡</sup>	0.5920 <sup>†‡</sup>
29 Phi 3 mini	0.6077 <sup>†</sup>	<b>0.6387<sup>‡</sup></b>	0.5868 <sup>†‡</sup>
30 Gemma v2	0.6318 <sup>†</sup>	0.6237 <sup>‡</sup>	<b>0.6010<sup>†‡</sup></b>

and Phi-3 [48] with 3.8 billion parameters.<sup>10</sup> The batch size for all experiments is set to 64. Experiments use 4 NVIDIA A100 GPUs (80GB VRAM) and 128GB of RAM. For sampling from the generative model  $M_G$ , we use nucleus sampling [27] with a temperature of  $\tau = 0.1$ . For the editing model  $M_E$ , nucleus sampling is applied with  $\tau = 0$ . When sampling plans with the planner  $M_P$ , we use a nucleus sampling temperature of  $\tau = 0.7$  for global exploration and  $\tau = 0$  otherwise. We define the exploration budget as the total number of responses generated and edited during the process of responding to an input.<sup>11</sup> We perform  $N = 4$  global and  $T = 4$  local exploitation steps to achieve a total generation budget of 16, unless stated otherwise. As a retriever, we use a BERT model<sup>12</sup> [16] pre-trained on retrieval tasks [43]. For indexing, we employ the Faiss library [17] to construct a hybrid IVF-HNSW index [41] for ANTIQUE and a flat index for TREC, chosen based on the corpus size. The total retrieval budget for P&R is set to  $k = 40$  for the ANTIQUE dataset

<sup>10</sup> Available at: <https://hf.co/microsoft/Phi-3-mini-4k-instruct>

<sup>11</sup> P&R’s average generated output length in our experiments is  $316.4 \pm 144.7$  words.

<sup>12</sup> Available at: <https://hf.co/Snowflake/snowflake-arctic-embed-l>

**Table 2: Effect of different components of P&R on ANTIQUE with Gemma v2 as the generator.**

Method	ICAT <sub>Coverage</sub>	ICAT <sub>Factuality</sub>	ICAT-A <sub>1</sub>
30 <b>P&amp;R (Gemma v2)</b>	0.6318	0.6237	0.6010
31 - w/o Global	0.6423	0.6073	0.5961
32 - w/o Local	0.6554	0.6017	0.5960
33 - w/o Local & Global	0.6543	0.5808	0.5832
34 - w/o Local & Global & Self-Training	0.6318	0.5512	0.5556

and  $k = 5$  for the TREC dataset. These are chosen based on the document length in each corpus and the context size of the LLMs.

**Baselines.** We leverage a variety of baseline LLMs of different sizes, both open-source and proprietary, with and without retrieval augmentation. The prompts used for the baselines are provided in Figure 3. For retrieval augmentation, we use the same retriever P&R. For each baseline, we set the retrieval budget based on the performance on the validation set, ranging between 10 and 40, similar to the configuration used for P&R. These baselines include:

- **Open-Source:** We utilize three open-source instruction-tuned LLMs as the backbone for baselines: Llama 3.2 [2], with 1.2 billion parameters, Gemma V2 [20], with 2.6 billion parameters, and Phi 3 [48], with 3.8 billion parameters. For CoT models, we evaluate only the final response and do not assess the intermediate reasoning steps. For Best-of-N, we generate  $N = 16$  outputs for each LLM with a temperature of 0.7 using nucleus sampling [27], rerank them using an off-the-shelf reranking model,<sup>13</sup> and select the top-ranked output as the final response. We also train Gemma v2 using self-training with ICAT, in the same setting as P&R. We leverage the high-scoring outputs of the model to train the model, enabling it to learn how to generate similar high-quality responses. This allows us to assess the potential improvements self-training can contribute to baseline models. Finally, we use Maximal Marginal Relevance (MMR) [8] with  $\lambda = 0.1$ <sup>14</sup> to rerank top 1,000 documents retrieved, investigating whether diversity in retrieval can enhance coverage of responses.
- **Proprietary:** For proprietary LLMs, we use two highly capable models with strong reasoning abilities: GPT-4o-mini<sup>15</sup> from OpenAI and Gemini 2 Flash<sup>16</sup> from Google. These models inherently perform CoT, so we do not explicitly prompt them for this. Additionally, due to the high cost associated with the Best-of-N approach, we do not apply this method to the proprietary LLMs.

## 5.2 Main Findings

**How does P&R perform compared to the baselines?** We compare P&R against different baselines under different experimental conditions. The results of these experiments on the ANTIQUE dataset are presented in Table 1. These results demonstrate that P&R statistically significantly outperforms both open-source and

<sup>13</sup> Available at: <https://hf.co/cross-encoder/ms-marco-MiniLM-L-12-v2>

<sup>14</sup> We select this value as it yields the best performance among the candidates  $\{0.05, 0.1, 0.15, 0.2, \dots, 0.5\}$ .

<sup>15</sup> Available at: <https://platform.openai.com/docs/models/gpt-4o-mini>

<sup>16</sup> Available at: <https://ai.google.dev/gemini-api/docs/models/gemini-v2>

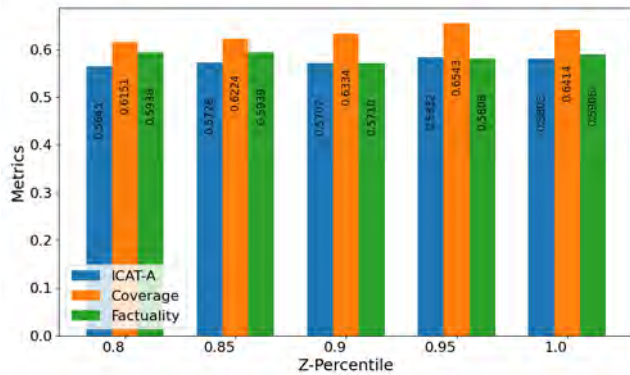


Figure 4: Effect of generated plan selection threshold for self-training planner for the ANTIQUE dataset (Gemma v2).

proprietary LLMs on the ICAT-A metric, emphasizing its superior performance in generating complete and factual responses. Specifically, P&R achieves a 13.1% relative improvement over the best open-source baseline (row 26 in Table 1) and a 6.5% improvement over the best proprietary baseline (row 3). This highlights the effectiveness of P&R in improving factuality, coverage, and their aggregation (ICAT-A).

The results in Table 1 suggest that RAG enhances the performance of LLMs on this task. Specifically, RAG helps generate more factual responses by incorporating relevant retrieved documents. However, it may lead to a reduction in coverage, as the retrieved documents tend to be more similar to one another, limiting the diversity of the generated content. Another interesting observation regarding the baselines is that the CoT approach tends to negatively affect the performance of LLMs in most cases on this task. We believe this occurs because LLMs are typically trained to apply CoT for reasoning and mathematical tasks. However, the task of generating factual and complete responses is inherently different from these types of reasoning tasks. Thus, CoT may not be as effective for this task, leading to suboptimal performance in generating accurate and comprehensive outputs. In contrast, the Best-of-N approach generally enhances the performance of LLMs; however, it remains less effective for the Gemma 2 models. Moreover, self-training proves to be the most effective strategy for training the baselines, though it still significantly lags behind P&R in terms of overall performance (row 26 vs 30). Given this, P&R shows the best and most promising results for this task. Finally, we find that using MMR to diversify the retrieval results does not yield improvement in coverage and factuality of the generated responses in most cases; instead, it leads to a drop in performance (rows 18, 22, and 27).

**How do global and local exploitation affect performance?** For this, we evaluate global and local exploitation separately, each using the same generation budget as P&R with both strategies combined (i.e., 16 generations). On the ANTIQUE dataset, we conduct experiments using only local exploitation, where a single plan is sampled greedily (with a temperature of  $\tau = 0.0$ ) and refined through 16 editing steps, and only global exploration, where 16 plans are sampled using a higher temperature of  $\tau = 0.7$  from the planner. The results

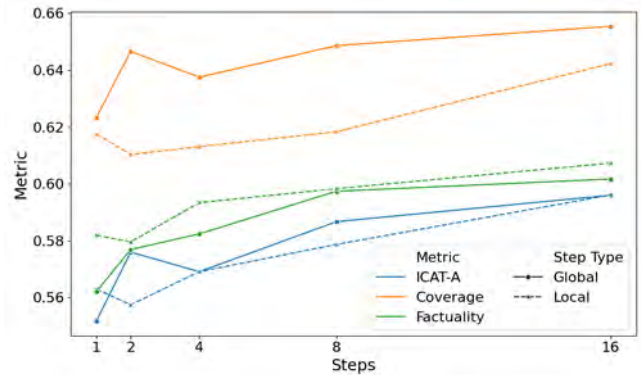


Figure 5: Effect of local and global exploration steps on P&R's performance for the ANTIQUE dataset (Gemma v2).

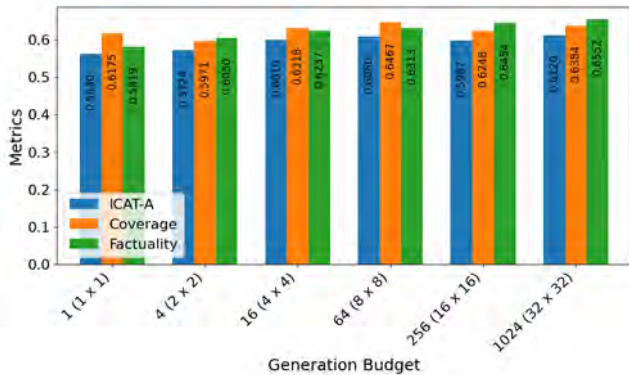
are reported in Table 2 (row 31 for local exploitation only and row 32 for global exploration only). The findings indicate that while using either local or global exploration achieves nearly identical ICAT-A scores, both are suboptimal compared to using all components, which combines both approaches. However, both methods outperform the planning-only configurations with (row 33) and without (row 34) self-training. Additionally, they achieve statistically significant improvements over all baselines in Table 1. These results highlight the effectiveness of global and local exploitation and show their complementary strengths in P&R.

**How does planning with and without self-training affect performance?** We focus on evaluating the planner without any global or local exploitation. Instead, we sample a single plan greedily (temperature  $\tau = 0.0$ ) to generate responses. We test both the untrained and trained planner under these conditions. The results in Table 2 for the ANTIQUE dataset indicate that the trained planner (row 33) alone is suboptimal compared to full pipeline, but it achieves a 4.9% relative improvement on the ICAT-A metric compared to the untrained planner (row 34). This shows the effectiveness of self-training in improving the planner's ability to generate better plans. Interestingly, the untrained planner (row 34) outperforms the best-performing open-source baseline (row 26 in Table 1) with a 4.6% relative improvement, showing the value of planning even without training. To explore further, we compare P&R without a trained planner, global, and local exploitation to the best RAG baseline on the TREC dataset, which does not include a training set. Since the TREC dataset includes human annotations for subtopics that need to be addressed for each query, we report all variations of ICAT on this dataset. As reported in Table 3, P&R with an untrained planner and no exploration achieves a statistically significant improvement over the baseline, with a 9.4%, 36.3%, and 15.4% relative gain on ICAT-M, ICAT-S, and ICAT-A metrics, respectively. Our findings show that P&R can significantly improve performance compared to the best baseline across different levels of annotated data availability.

**How does planner's self-training threshold affect performance?** An important hyperparameter in the proposed self-training approach for training the planner is the top Z-Percentile of the generated plans to be used for training. To investigate this, we train the

**Table 3: Performance of P&R (Gemma v2) without self-training and exploration compared to RAG baseline on TREC using different variations of ICAT metric. The † shows statistically significant improvements over the baseline using t-test ( $p < 0.05$ ).**

Metric	Factuality	Manual		Semi-Automatic		Automatic	
		Coverage	ICAT-M <sub>1</sub>	Coverage	ICAT-S <sub>1</sub>	Coverage	ICAT-A <sub>1</sub>
RAG Gemma v2	<b>0.6720</b>	0.2980	0.3203	0.1970	0.2294	0.5079	0.5148
P&R (w/o self-training & exploration)	0.6325	<b>0.3523<sup>†</sup></b>	<b>0.3507<sup>†</sup></b>	<b>0.2819<sup>†</sup></b>	<b>0.3129<sup>†</sup></b>	<b>0.6665<sup>†</sup></b>	<b>0.5943<sup>†</sup></b>

**Figure 6: Effect of generation budget on the performance of P&R on the ANTIQUE dataset (Gemma v2).**

model using different values for  $Z$  and evaluate the planner with them. In this experiment on the ANTIQUE dataset, we generate a single plan greedily (temperature  $\tau = 0.0$ ) and produce a response. The results, shown in Figure 4, indicate that as  $Z$  increases, the results improve, as it implies that the model is trained on higher-quality outputs. We observed that the best performance occurs at  $Z = 0.95$ . However, setting  $Z = 1$ , which means only the output with highest score being selected for training, leads to missing high-quality outputs that could aid training. Therefore, optimizing this hyperparameter is crucial for achieving optimal results.

**How does global and local exploitation budget affect performance?** To separately study the impact of local and global exploration steps, we conduct experiments on the ANTIQUE dataset using only local exploitation with a single plan sampled greedily (with a temperature of  $\tau = 0.0$ ) and edited  $N$  times, and only global exploration with  $N$  plans sampled using a higher temperature of  $\tau = 0.7$  from the planner. The results of this experiment are shown in Figure 5 (B). The results suggest that increasing the number of local and global exploration steps leads to improvements across all metrics, with the ICAT-A metric being nearly identical for both approaches with 16 generated outputs. However, it can be observed that increasing the number of global plans results in higher coverage, while increasing local exploitation steps leads to higher factuality. This observation indicates that sampling multiple plans produces outputs that cover more topics, but may lack factual accuracy. In contrast, sampling a single plan and applying multiple local exploitations and editing steps results in lower coverage but higher factual accuracy. Given this, we show that the primary contribution of global exploration is to enhance coverage, while the main contribution of local exploitation is to improve factuality.

**How does exploration budget affect performance?** In this experiment on the ANTIQUE dataset, we evaluate the performance of our approach under different exploration budgets: 1, 4, 16, 64, 256, and 1024 responses per input, where the budget is allocated equally to global and local exploitation. The results are visualized in Figure 6, illustrating the impact of increasing the exploration budget on the performance of P&R. The results indicate that increasing the exploration budget leads to improved performance on the ICAT-A metric. A general trend of improvement in topic coverage is also observed, though with some fluctuations. In contrast, factuality shows a consistent increase as the budget grows. These findings suggest that higher exploration budgets enhance both factuality and topic coverage, with a particularly pronounced effect on factuality.

**How does P&R align with human preferences?** We randomly selected 50 queries from the ANTIQUE test set and generated outputs using P&R and the best-performing open-source baseline from Table 1, *RAG Gemma 2 w/ Self-Training*. We use this baseline not only due to its strong performance, but also due to the fact that it uses the same LLM as P&R, for a fair comparison. Two human annotators evaluated the outputs based on three criteria: coverage of topics related to the input, factual accuracy of the generated responses, and overall quality. The inter-annotator agreement, measured using Cohen’s  $\kappa$ , was 0.6189, indicating substantial agreement. The results are presented in Table 4. For coverage, annotators preferred P&R 64% of the time, compared to 26% for the baseline. For factuality, the outputs of both models were rated equally in 56% of cases, but in the remaining instances, P&R was preferred 35% of the time, while the baseline was chosen in only 9% of cases. Overall, P&R was selected as the preferred output in 63% of cases, compared to 29% for the baseline. These findings indicate that P&R aligns more closely with human preferences and produces higher-quality outputs.

### 5.3 Case Study

To provide a clearer understanding of how P&R works, we present an output example for a query from the ANTIQUE dataset in Figure 7. Here, we generate two plans for global exploration, and for local exploitation, we iteratively edit the responses up to a maximum of 32 steps. As illustrated in Figure 7, the two plans share several steps in addressing the query while also considering unique aspects (unique aspects are highlighted in different colors, and shared steps are shown in the same color). For instance, second plan emphasizes the economic, philosophical, and ethical reasons behind depression following a school change, whereas the first plan focuses on mentioning individual experiences, examples, and social groups that can help alleviate such challenges. This difference in the plans resulted in two distinct initial responses in terms of both content and style. Next, the initial generated responses are refined by the editing

Why is switching schools so depressing?		
<p><b>Plan 1</b></p> <p>aspect: Emotional Impact query: psychological effects of school changes reason: Understanding the emotional impact of school changes, such as anxiety, stress, and depression, is crucial for a comprehensive response.</p> <p>aspect: Social Dynamics query: social impact of school transitions on students reason: Addressing the social dynamics of school transfers, including the impact on students' social relationships and the potential for negative interactions, is crucial for a complete picture.</p> <p>aspect: Social Support Systems query: resources for students during school transitions reason: Providing access to resources, such as guidance counselors, support groups, and online platforms, is vital for addressing the potential for depression during school transitions.</p> <p>aspect: Educational System query: the role of education in shaping individuals reason: Understanding the role of education in shaping individuals' emotional and mental qualities, including the potential for both positive and negative influences, is essential for a complete response.</p> <p>aspect: Individual Experiences query: personal stories of school transfers reason: Sharing personal stories, such as those of students who have experienced school transfers, can provide valuable insights and perspectives on the topic.</p>	<p><b>Initial Output 1</b></p> <p>Switching schools can be a stressful and emotionally challenging experience, especially for students who are already dealing with anxiety or depression. The emotional impact of school changes can be significant, as students may experience feelings of loneliness, isolation, and a loss of identity.</p> <p>Social dynamics play a crucial role in this transition. Students may struggle to adjust to new social groups, navigate unfamiliar relationships, and cope with potential negative interactions. The lack of familiarity with the school environment, teachers, and peers can lead to feelings of insecurity and anxiety.</p> <p>Schools can provide valuable support systems to address these challenges. Guidance counselors, support groups, and online platforms can offer students a safe space to share their feelings, seek advice, and connect with others who understand their experiences.</p> <p>The educational system itself can also influence the emotional well-being of students. While education can be a positive force for personal growth and development, it can also contribute to stress and pressure, particularly if students feel overwhelmed by academic expectations or lack of support.</p> <p>Individual experiences vary greatly, and some students may find the transition easier than others. However, it's important to acknowledge that school changes can be emotionally challenging and to provide students with the necessary resources and support to navigate this transition successfully.</p>	<p><b>Edited Output 1</b></p> <p>The transition to a new school can be a challenging experience, both emotionally and socially. While it can be a positive opportunity for growth and learning, it can also lead to feelings of anxiety, stress, and depression in some of the key aspects. Addressing this requires a multifaceted approach that considers individual experiences, social dynamics, and the broader educational system. Here's a breakdown of the key aspects:</p> <p><b>Emotional Impact:</b></p> <ul style="list-style-type: none"> <li><b>Anxiety and Stress:</b> Students experiencing anxiety or stress during school transitions might find the change overwhelming. This can lead to feelings of being judged, pressured, and lacking a sense of control.</li> <li><b>Depression:</b> Some students may feel depressed by the new school environment, especially if they've experienced a negative past. This can be exacerbated by social isolation, lack of support, and difficulty adjusting to new routines.</li> <li><b>Social Anxiety:</b> Social anxiety can become more pronounced in unfamiliar settings. Students may feel self-conscious, worry about social interactions, and feel pressured to conform to a new social hierarchy.</li> </ul> <p><b>Social Dynamics:</b></p> <ul style="list-style-type: none"> <li><b>Social Pressure:</b> The social hierarchy at new schools can create pressure to conform to a specific role or group. This can lead to social anxiety, peer pressure, and feelings of being excluded.</li> <li><b>Social Support Systems:</b> Students need access to resources, such as guidance counselors, support groups, and online platforms, to help them navigate the transition. These resources can provide valuable advice, support, and a sense of community.</li> <li><b>Peer Relationships:</b> The new school environment can impact existing friendships and relationships. Students may face challenges adjusting to new social norms and navigating new dynamics within their new school.</li> </ul> <p><b>Educational System:</b></p> <ul style="list-style-type: none"> <li><b>Academic Expectations:</b> Students need to be academically supported during the transition. This can involve providing them with the resources and guidance they need to succeed in their new academic environment.</li> <li><b>Social Skills and Development:</b> The transition can be an opportunity for students to develop social skills and learn to navigate social interactions in diverse settings. This can involve teaching them about empathy, communication, and conflict resolution.</li> <li><b>Career Counseling:</b> Students may need guidance on career counseling as they navigate their new school and explore their career aspirations.</li> </ul> <p><b>Individual Experiences:</b></p> <ul style="list-style-type: none"> <li><b>Personal Stories:</b> Hearing from students about their experiences can provide valuable insights and perspectives on the transition.</li> <li><b>Challenges and Struggles:</b> Students may face challenges during the transition, such as financial worries, social isolation, and academic difficulties.</li> <li><b>Coping Mechanisms:</b> Students need to develop coping mechanisms to deal with the stress and anxiety of the transition.</li> </ul> <p>By addressing these aspects, schools can create a more supportive and effective learning environment for students. This can involve providing them with the resources, support, and guidance they need to navigate the transition successfully.</p>
<p><b>Plan 2</b></p> <p>aspect: Emotional impact of school changes query: the psychological effects of changing schools reason: This aspect explores the potential emotional and psychological changes associated with school transitions, such as anxiety, stress, and a sense of belonging.</p> <p>aspect: Social and economic implications of school changes query: the social and economic consequences of school changes reason: This aspect examines the social and economic factors that influence school transitions, including things like cost of living, housing options, and access to higher education.</p> <p>aspect: Educational opportunities and challenges of school changes query: the educational benefits and challenges of school changes reason: This aspect focuses on the educational opportunities and challenges associated with school shifts, such as academic expectations, curriculum changes, and a focus on future-proof skills.</p> <p>aspect: Impact on individual students and families query: the impact of school changes on students and families reason: This aspect emphasizes the individual and family-centric perspectives on school changes, considering the impact on student development, family dynamics, and overall well-being.</p> <p>aspect: Philosophical and ethical considerations of school changes query: the philosophical and ethical implications of school changes reason: This aspect explores the ethical and philosophical considerations associated with school transitions, including issues like privacy concerns, data security, and the potential for exploitation.</p>	<p><b>Initial Output 2</b></p> <p>To understand why switching schools might be depressing, we need to consider the following factors:</p> <p><b>Emotional Impact</b></p> <ul style="list-style-type: none"> <li><b>Loss of familiarity and routine:</b> Leaving behind a familiar environment, friends, and teachers can be unsettling and lead to feelings of loneliness and insecurity.</li> <li><b>Social adjustment:</b> Adapting to a new school environment, navigating new social dynamics, and building new relationships can be challenging and stressful.</li> <li><b>Academic pressure:</b> The transition to a new school can involve a new curriculum, different teaching styles, and potentially higher academic expectations. This can lead to anxiety and stress.</li> </ul> <p><b>Social and Economic Implications</b></p> <ul style="list-style-type: none"> <li><b>Cost of living:</b> Moving to a new school district can involve significant financial costs, such as housing, transportation, and other expenses.</li> <li><b>Access to higher education:</b> The quality of schools in different districts can vary, impacting access to higher education opportunities.</li> </ul> <p><b>Educational Opportunities and Challenges</b></p> <ul style="list-style-type: none"> <li><b>Curriculum differences:</b> Schools may have different curriculums, teaching methods, and learning materials, which can be challenging to adjust to.</li> <li><b>Academic expectations:</b> The academic expectations and rigor of a new school can be different from the previous one, leading to a sense of being overwhelmed or unprepared.</li> </ul> <p><b>Impact on Individual Students and Families</b></p> <ul style="list-style-type: none"> <li><b>Family dynamics:</b> The transition to a new school can impact family dynamics, leading to stress and tension.</li> <li><b>Student development:</b> The emotional and social challenges of switching schools can affect a student's overall development and well-being.</li> </ul> <p><b>Philosophical and Ethical Considerations</b></p> <ul style="list-style-type: none"> <li><b>Privacy concerns:</b> Schools collect and use student data, raising ethical concerns about privacy and data security.</li> <li><b>Exploitation:</b> There is a potential for schools to exploit students' vulnerabilities and personal information.</li> </ul> <p>It's important to note that the experience of switching schools can vary greatly depending on the individual student, their support system, and the specific circumstances of the transition.</p>	<p><b>Edited Output 2</b></p> <p>The transition to a new school can be a challenging and emotional process, and there is no single answer that can capture the full picture. While some students may find the change liberating, others may feel anxious or depressed. Several factors contribute to this emotional impact, including:</p> <p><b>Emotional Impact of School Changes:</b></p> <ul style="list-style-type: none"> <li><b>Loss of Control:</b> Students may feel a sense of loss of control, especially if they are not able to predict or prevent changes. This can lead to feelings of powerlessness and anxiety.</li> <li><b>Social Anxiety:</b> The pressure to socialize in new situations and navigate potential challenges can contribute to anxiety and fear.</li> <li><b>Social Pressure:</b> The academic expectations and curriculum changes can create a sense of academic anxiety and stress.</li> <li><b>Family Dynamics:</b> Changes in family structure or relationships can create emotional support issues for students.</li> <li><b>Cost of Living:</b> The cost of living, particularly in terms of housing options and access to higher education, can be a significant factor for students.</li> <li><b>Access to Education:</b> The availability of quality education and resources can impact a student's success in their new school.<sup>1</sup> <b>Social Support:</b> The lack of social support from peers or family can lead to feelings of isolation and stress.</li> </ul> <p><b>Individual Opportunities and Challenges of School Changes:</b></p> <ul style="list-style-type: none"> <li><b>Academic Expectations:</b> Students may not be able to meet the academic expectations set for their new school, leading to stress and anxiety.</li> <li><b>Curriculum Changes:</b> The changes in curriculum can be challenging for students to navigate, especially if they are not familiar with the new system.</li> <li><b>Focus on Future Skills:</b> The emphasis on future skills and career preparation can create pressure on students to succeed in their new school.</li> <li><b>Impact on Individual Students and Families:</b> <ul style="list-style-type: none"> <li><b>Family Structure and Relationships:</b> Changes in family structure or relationships can create emotional support issues for students.</li> <li><b>Academic Performance:</b> Students may struggle to adapt to the demands of a new school, impacting their academic performance.</li> <li><b>Financial Stability:</b> The cost of living and access to resources can contribute to anxiety and stress for students.</li> </ul> </li> <li><b>Philosophical and Ethical Considerations of School Changes:</b> <ul style="list-style-type: none"> <li><b>Privacy Concerns:</b> Students may have privacy concerns about their personal information and how their data is used.</li> <li><b>Data Security:</b> The security of student data and preventing unauthorized access to it is paramount.</li> <li><b>The Right to Transition:</b> Students have the right to a smooth and transition into their new school environment.</li> <li><b>The Impact on Teachers:</b> Teachers may need to adapt to the changes in their teaching environment and develop new strategies for engaging students.</li> </ul> </li> </ul> <p>It is important to remember that the transition to a new school is a complex and multifaceted process. Addressing the concerns and challenges associated with school changes requires a collaborative effort from students, teachers, and families. Providing support and guidance for students and families can help them navigate the transition and achieve success in their new school environment.</p>

Figure 7: Case study on generated plans, responses, and edited responses by P&R (Gemma v2). The aspects that differ between the two plans are highlighted using different colors. The selected response is marked by a cup symbol.

Table 4: Human alignment comparison of P&R and RAG Gemma 2 with Self-Training (best-performing baseline).

Winner	Coverage (%)	Factuality (%)	Overall (%)
P&R	64	35	63
Baseline	26	9	29
Tie	10	56	8

model over multiple steps to produce the edited outputs depicted in Figure 7. An interesting observation is that the edited responses exhibit greater depth in categorizing various aspects and provide more detailed and structured explanations. This structuring is particularly noticeable in the first output. Initially, the first response was presented as paragraphs without utilizing markdown formatting or hierarchical organization for different aspects. However, the edited output introduces markdown elements and restructures the response, enhancing its coverage and factuality. Finally, the reward model selected the second edited output as the final response to the question. This choice reflects its superior coverage and factual accuracy, as evidenced by its ability to address a broader range of aspects while maintaining a high degree of factual correctness.

## 6 Conclusion, Limitations, and Future Work

This paper introduced P&R, a novel approach for improving the factuality and coverage of generated responses by LLMs. P&R starts by generating a diverse set of plans for responding to a prompt and retrieves documents from a corpus to gather the necessary information to execute each plan. It then generates a response for each plan

and iteratively refines the response to enhance their factuality and comprehensiveness. Finally, a reward model selects the most factual and complete response from the set of generated proposals. Experiments on the ANTIQUE and TREC datasets demonstrated that P&R significantly outperforms both open-source and proprietary baselines, achieving up to a 13.1% improvement over open-source models and a 6.5% improvement over proprietary models. Furthermore, human evaluations reveal that P&R has substantially higher agreement with human preferences compared to baselines.

While the primary goal of this paper is to introduce a framework for improving the factuality and coverage of responses, one limitation of P&R is that it belongs to the family of methods that increase test-time computation to enhance performance (similar to the best-of-N method used in this paper). Although P&R remains faster than most proprietary LLMs, future work could explore more efficient resource allocation between global and local exploitation to reduce latency, as well as adaptive strategies to decide—on a per-query basis—when to use P&R versus a simpler RAG approach.

## Acknowledgment

This work was supported in part by the Center for Intelligent Information Retrieval, in part by NSF grants #2143434 and #2402873, and with support from Google.org. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect those of the sponsor.

## References

- [1] Garima Agrawal, Tharindu Kumarage, Zeyad Alghami, and Huan Liu. 2023. Can Knowledge Graphs Reduce Hallucinations in LLMs? : A Survey. arXiv:2311.07914 [cs.CL]
- [2] Meta AI. 2024. The Llama 3 Herd of Models. arXiv:2407.21783 [cs.AI] <https://arxiv.org/abs/2407.21783>
- [3] Danial Alihosseini, Ehsan Montahaei, and Mahdieh Soleymani Baghshah. 2019. Jointly Measuring Diversity and Quality in Text Generation Models. In *Proceedings of the Workshop on Methods for Optimizing and Evaluating Neural Language Generation*, Antoine Bosselut, Asli Celikyilmaz, Marjan Ghazvininejad, Srinivasan Iyer, Urvashi Khandelwal, Hannah Rashkin, and Thomas Wolf (Eds.). Association for Computational Linguistics, Minneapolis, Minnesota, 90–98. doi:10.18653/v1/W19-2311
- [4] Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2024. Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection. In *The Twelfth International Conference on Learning Representations*. <https://openreview.net/forum?id=hSyW5go0v8>
- [5] Zhenni Bi, Kai Han, Chuanjian Liu, Yehui Tang, and Yunhe Wang. 2024. Forest-of-Thought: Scaling Test-Time Compute for Enhancing LLM Reasoning. arXiv:2412.09078 [cs.CL] <https://arxiv.org/abs/2412.09078>
- [6] Bradley Brown, Jordan Juravskj, Ryan Ehrlich, Ronald Clark, Quoc V. Le, Christopher Ré, and Azalia Mirhoseini. 2024. Large Language Monkeys: Scaling Inference Compute with Repeated Sampling. arXiv:2407.21787 [cs.LG] <https://arxiv.org/abs/2407.21787>
- [7] Rodrigo Tripodi Calumby. 2016. Diversity-oriented Multimodal and Interactive Information Retrieval. *SIGIR Forum* 50 (2016), 86. <https://api.semanticscholar.org/CorpusID:16816864>
- [8] Jaime Carbonell and Jade Goldstein. 1998. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (Melbourne, Australia) (*SIGIR '98*). Association for Computing Machinery, New York, NY, USA, 335–336. doi:10.1145/290941.291025
- [9] Wenhui Chen, Hexiang Hu, Xi Chen, Pat Verga, and William Cohen. 2022. MuRAG: Multimodal Retrieval-Augmented Generator for Open Question Answering over Images and Text. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 5558–5570. doi:10.18653/v1/2022.emnlp-main.375
- [10] Yongchao Chen, Harsh Jhamtani, Srinagesh Sharma, Chuchu Fan, and Chi Wang. 2024. Steering Large Language Models between Code Execution and Textual Reasoning. arXiv:2410.03524 [cs.CL] <https://arxiv.org/abs/2410.03524>
- [11] Yanxi Chen, Xuchen Pan, Yaliang Li, Bolin Ding, and Jingren Zhou. 2024. A Simple and Provable Scaling Law for the Test-Time Compute of Large Language Models. arXiv:2411.19477 [cs.CL] <https://arxiv.org/abs/2411.19477>
- [12] Charles L. A. Clarke, Nick Craswell, and Ian Soboroff. 2009. Overview of the TREC 2009 Web Track. In *Text Retrieval Conference*. <https://api.semanticscholar.org/CorpusID:13369675>
- [13] Charles L. A. Clarke, Nick Craswell, Ian Soboroff, and Gordon V. Cormack. 2010. Overview of the TREC 2010 Web Track. In *Text Retrieval Conference*. <https://api.semanticscholar.org/CorpusID:16213318>
- [14] Charles L. A. Clarke, Nick Craswell, Ian Soboroff, and Ellen M. Voorhees. 2011. Overview of the TREC 2011 Web Track. In *Text Retrieval Conference*. <https://api.semanticscholar.org/CorpusID:30284167>
- [15] Charles L. A. Clarke, Nick Craswell, and Ellen M. Voorhees. 2012. Overview of the TREC 2012 Web Track. In *Text Retrieval Conference*. <https://api.semanticscholar.org/CorpusID:11517775>
- [16] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Jill Burstein, Christy Doran, and Thamar Solorio (Eds.). Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186. doi:10.18653/v1/N19-1423
- [17] Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. 2024. The Faiss library. *ArXiv abs/2401.08281* (2024). <https://api.semanticscholar.org/CorpusID:267028372>
- [18] Ronen Eldan and Yuanzhi Li. 2023. TinyStories: How Small Can Language Models Be and Still Speak Coherent English? arXiv:2305.07759 [cs.CL] <https://arxiv.org/abs/2305.07759>
- [19] Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. 2023. Enabling Large Language Models to Generate Text with Citations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 6465–6488. doi:10.18653/v1/2023.emnlp-main.398
- [20] Gemma-Team. 2024. Gemma 2: Improving Open Language Models at a Practical Size. arXiv:2408.00118 [cs.CL] <https://arxiv.org/abs/2408.00118>
- [21] Carlos Gómez-Rodríguez and Paul Williams. 2023. A Confederacy of Models: a Comprehensive Evaluation of LLMs on Creative Writing. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 14504–14528. doi:10.18653/v1/2023.findings-emnlp.966
- [22] Liangke Gui, Borui Wang, Qiuyuan Huang, Alexander Hauptmann, Yonatan Bisk, and Jianfeng Gao. 2022. KAT: A Knowledge Augmented Transformer for Vision-and-Language. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Seattle, United States, 956–968. doi:10.18653/v1/2022.naacl-main.70
- [23] Caglar Gulcehre, Tom Le Paine, Srivatsan Srinivasan, Ksenia Konyushkova, Lotte Weerts, Abhishek Sharma, Aditya Siddhant, Alex Ahern, Miaosen Wang, Chenjie Gu, Wolfgang Macherey, Arnaud Doucet, Orhan Firat, and Nando de Freitas. 2023. Reinforced Self-Training (ReST) for Language Modeling. arXiv:2308.08998 [cs.CL] <https://arxiv.org/abs/2308.08998>
- [24] Muhammad Usman Hadi, al tashi, Rizwan Qureshi, Abbas Shah, Muhammad Irfan, Anas Zafar, Muhammad Bilal Shaikh, Naveed Akhtar, Jia Wu, Seyyedal Mirjalili, Qasem Al-Tashi, Amgad Muneer, Mohammed Ali Al-garadi, Gru Cnn, and T5 RoBERTa. [n. d.]. Large Language Models: A Comprehensive Survey of its Applications, Challenges, Limitations, and Future Prospects. <https://api.semanticscholar.org/CorpusID:266378240>
- [25] Helia Hashemi, Mohammad Aliannejadi, Hamed Zamani, and W. Bruce Croft. 2020. ANTIQUE: A Non-factoid Question Answering Benchmark. In *Advances in Information Retrieval: 42nd European Conference on IR Research, ECIR 2020, Lisbon, Portugal, April 14–17, 2020, Proceedings, Part II* (Lisbon, Portugal). Springer-Verlag, Berlin, Heidelberg, 166–173. doi:10.1007/978-3-030-45442-5\_21
- [26] Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. DeBERTaV3: Improving DeBERTa using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing. arXiv:2111.09543 [cs.CL] <https://arxiv.org/abs/2111.09543>
- [27] Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The Curious Case of Neural Text Degeneration. In *International Conference on Learning Representations*. <https://openreview.net/forum?id=rygGQYrFvH>
- [28] Mengkang Hu, Pu Zhao, Can Xu, Qingfeng Sun, Jianguang Lou, Qingwei Lin, Ping Luo, and Saravan Rajmohan. 2024. AgentGen: Enhancing Planning Abilities for Large Language Model based Agent via Environment and Task Generation. arXiv:2408.00764 [cs.CL] <https://arxiv.org/abs/2408.00764>
- [29] Xu Huang, Weiwu Liu, Xiaolong Chen, Xingmei Wang, Hao Wang, Defu Lian, Yasheng Wang, Ruiming Tang, and Enhong Chen. 2024. Understanding the planning of LLM agents: A survey. arXiv:2402.02716 [cs.AI] <https://arxiv.org/abs/2402.02716>
- [30] Gautier Izacard and Edouard Grave. 2021. Leveraging Passage Retrieval with Generative Models for Open Domain Question Answering. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. Association for Computational Linguistics, Online, 874–880. doi:10.18653/v1/2021.eacl-main.74
- [31] Xue Jiang, Yihong Dong, Lecheng Wang, Zheng Fang, Qiwei Shang, Ge Li, Zhi Jin, and Wenpin Jiao. 2024. Self-Planning Code Generation with Large Language Models. *ACM Trans. Softw. Eng. Methodol.* 33, 7, Article 182 (Sept. 2024), 30 pages. doi:10.1145/3672456
- [32] Masahiro Kaneko, Danushka Bollegala, Naoaki Okazaki, and Timothy Baldwin. 2024. Evaluating Gender Bias in Large Language Models via Chain-of-Thought Prompting. arXiv:2401.15585 [cs.CL] <https://arxiv.org/abs/2401.15585>
- [33] To Eun Kim, Alireza Salemi, Andrew Drozdz, Fernando Diaz, and Hamed Zamani. 2024. Retrieval-Enhanced Machine Learning: Synthesis and Opportunities. arXiv:2407.12982 [cs.LG] <https://arxiv.org/abs/2407.12982>
- [34] Diederik P. Kingma and Jimmy Ba. 2014. Adam: A Method for Stochastic Optimization. *CoRR abs/1412.6980* (2014). <https://api.semanticscholar.org/CorpusID:6628106>
- [35] Ishita Kumar, Snigdha Viswanathan, Sushrita Yerra, Alireza Salemi, Ryan A. Rossi, Franck Dernoncourt, Hanieh Deilamsalehy, Xiang Chen, Ruiyi Zhang, Shubham Agarwal, Nedim Lipka, Chien Van Nguyen, Thien Huu Nguyen, and Hamed Zamani. 2024. LongLaMP: A Benchmark for Personalized Long-form Text Generation. arXiv:2407.11016 [cs.CL] <https://arxiv.org/abs/2407.11016>
- [36] Moritz Laurer, Wouter van Atteveldt, Andreu Casas, and Kasper Welbers. 2024. Less Annotating, More Classifying: Addressing the Data Scarcity Issue of Supervised Machine Learning with Deep Transfer Learning and BERT-NLI. *Political Analysis* 32, 1 (2024), 84–100. doi:10.1017/pan.2023.20
- [37] Teerapong Leelanupab. 2012. A ranking framework and evaluation for diversity-based retrieval. <https://api.semanticscholar.org/CorpusID:29355641>
- [38] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems (Vancouver, BC, Canada) (NIPS'20)*. Curran Associates Inc., Red Hook, NY, USA, Article 793, 16 pages.
- [39] Zaijing Li, Gongwei Chen, Rui Shao, Yuquan Xie, Dongmei Jiang, and Liqiang Nie. 2024. Enhancing Emotional Generation Capability of Large Language Models

- via Emotional Chain-of-Thought. arXiv:2401.06836 [cs.CL] <https://arxiv.org/abs/2401.06836>
- [40] Hanmeng Liu, Zhiyang Teng, Leyang Cui, Chaoli Zhang, Qiji Zhou, and Yue Zhang. 2023. LogiCoT: Logical Chain-of-Thought Instruction Tuning. In *The 2023 Conference on Empirical Methods in Natural Language Processing*. <https://openreview.net/forum?id=qlCtkvgQJH>
- [41] Yu A. Malkov and D. A. Yashunin. 2020. Efficient and Robust Approximate Nearest Neighbor Search Using Hierarchical Navigable Small World Graphs. *IEEE Trans. Pattern Anal. Mach. Intell.* 42, 4 (April 2020), 824–836. doi:10.1109/TPAMI.2018.2889473
- [42] David McSherry. 2002. Diversity-Conscious Retrieval. In *Proceedings of the 6th European Conference on Advances in Case-Based Reasoning (ECCBR '02)*. Springer-Verlag, Berlin, Heidelberg, 219–233.
- [43] Luke Merrick. 2024. Embedding And Clustering Your Data Can Improve Contrastive Pretraining. arXiv:2407.18887 [cs.LG] <https://arxiv.org/abs/2407.18887>
- [44] Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. FActScore: Fine-grained Atomic Evaluation of Factual Precision in Long Form Text Generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 12076–12100. doi:10.18653/v1/2023.emnlp-main.741
- [45] OpenAI. 2024. GPT-4 Technical Report. arXiv:2303.08774 [cs.CL] <https://arxiv.org/abs/2303.08774>
- [46] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (Eds.), Vol. 35. Curran Associates, Inc., 27730–27744. [https://proceedings.neurips.cc/paper\\_files/paper/2022/file/b1efde53be364a73914f58805a001731-Paper-Conference.pdf](https://proceedings.neurips.cc/paper_files/paper/2022/file/b1efde53be364a73914f58805a001731-Paper-Conference.pdf)
- [47] Fabio Petroni, Aleksandra Piktus, Angela Fan, Patrick Lewis, Majid Yazdani, Nicola De Cao, James Thorne, Yacine Jernite, Vladimir Karpukhin, Jean Mailard, Vassilis Plachouras, Tim Rocktäschel, and Sebastian Riedel. 2021. KILT: a Benchmark for Knowledge Intensive Language Tasks. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Online, 2523–2544. doi:10.18653/v1/2021.naacl-main.200
- [48] Phi-Team. 2024. Phi-3 Technical Report: A Highly Capable Language Model Locally on Your Phone. arXiv:2404.14219 [cs.CL] <https://arxiv.org/abs/2404.14219>
- [49] Leonardo Ranaldi and Andre Freitas. 2024. Aligning Large and Small Language Models via Chain-of-Thought Reasoning. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, Yvette Graham and Matthew Purver (Eds.). Association for Computational Linguistics, St. Julian's, Malta, 1812–1827. <https://aclanthology.org/2024.eacl-long.109/>
- [50] Alireza Salemi, Juan Altmayer Pizzorno, and Hamed Zamani. 2023. A Symmetric Dual Encoding Dense Retrieval Framework for Knowledge-Intensive Visual Question Answering. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval (Taipei,Taiwan) (SIGIR '23)*. Association for Computing Machinery, New York, NY, USA, 110–120. doi:10.1145/3539618.3591629
- [51] Alireza Salemi, Surya Kallumadi, and Hamed Zamani. 2024. Optimization Methods for Personalizing Large Language Models through Retrieval Augmentation. In *Proceedings of the 47th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (Washington, DC, USA) (SIGIR '24)*. (to appear).
- [52] Alireza Salemi, Julian Killingback, and Hamed Zamani. 2025. ExPerT: Effective and Explainable Evaluation of Personalized Long-Form Text Generation. In *Findings of the Association for Computational Linguistics: ACL 2025*. Association for Computational Linguistics, Vienna, Austria, 17516–17532. doi:10.18653/v1/2025.findings-acl.900
- [53] Alireza Salemi, Cheng Li, Mingyang Zhang, Qiaozhu Mei, Weize Kong, Tao Chen, Zhuowan Li, Michael Bendersky, and Hamed Zamani. 2025. Reasoning-Enhanced Self-Training for Long-Form Personalized Text Generation. arXiv:2501.04167 [cs.CL] <https://arxiv.org/abs/2501.04167>
- [54] Alireza Salemi, Sheshera Mysore, Michael Bendersky, and Hamed Zamani. 2024. LaMP: When Large Language Models Meet Personalization. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Bangkok, Thailand, 7370–7392. doi:10.18653/v1/2024.acl-long.399
- [55] Alireza Salemi, Mahta Rafiee, and Hamed Zamani. 2023. Pre-Training Multi-Modal Dense Retrievers for Outside-Knowledge Visual Question Answering. In *Proceedings of the 2023 ACM SIGIR International Conference on Theory of Information Retrieval (Taipei, Taiwan) (ICTIR '23)*. Association for Computing Machinery, New York, NY, USA, 169–176. doi:10.1145/3578337.3605137
- [56] Alireza Salemi and Hamed Zamani. 2024. Evaluating Retrieval Quality in Retrieval-Augmented Generation. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval (Washington DC, USA) (SIGIR '24)*. Association for Computing Machinery, New York, NY, USA, 2395–2400. doi:10.1145/3626772.3657957
- [57] Alireza Salemi and Hamed Zamani. 2024. Towards a Search Engine for Machines: Unified Ranking for Multiple Retrieval-Augmented Large Language Models. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval (Washington DC, USA) (SIGIR '24)*. Association for Computing Machinery, New York, NY, USA, 741–751. doi:10.1145/3626772.3657733
- [58] Alireza Salemi and Hamed Zamani. 2025. Comparing Retrieval-Augmentation and Parameter-Efficient Fine-Tuning for Privacy-Preserving Personalization of Large Language Models. In *Proceedings of the 2025 International ACM SIGIR Conference on Innovative Concepts and Theories in Information Retrieval (ICTIR) (Padua, Italy) (ICTIR '25)*. Association for Computing Machinery, New York, NY, USA, 286–296. doi:10.1145/3731120.3744595
- [59] Alireza Salemi and Hamed Zamani. 2025. Learning to Rank for Multiple Retrieval-Augmented Models through Iterative Utility Maximization. In *Proceedings of the 2025 International ACM SIGIR Conference on Innovative Concepts and Theories in Information Retrieval (ICTIR) (Padua, Italy) (ICTIR '25)*. Association for Computing Machinery, New York, NY, USA, 183–193. doi:10.1145/3731120.3744584
- [60] Chris Samarinas, Alexander Krubner, Alireza Salemi, Youngwoo Kim, and Hamed Zamani. 2025. Beyond Factual Accuracy: Evaluating Coverage of Diverse Factual Information in Long-form Text Generation. In *Findings of the Association for Computational Linguistics: ACL 2025*. Association for Computational Linguistics, Vienna, Austria, 13468–13482. doi:10.18653/v1/2025.findings-acl.693
- [61] Scott Sanner, Shengbo Guo, Thore Graepel, Sadegh Kharazmi, and Sarvnaz Karimi. 2011. Diverse retrieval via greedy optimization of expected 1-call@k in a latent subtopic relevance model. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management (Glasgow, Scotland, UK) (CIKM '11)*. Association for Computing Machinery, New York, NY, USA, 1977–1980. doi:10.1145/2063576.2063869
- [62] Zhihong Shao, Minlie Huang, Jiangtao Wen, Wenfei Xu, and Xiaoyan Zhu. 2019. Long and Diverse Text Generation with Planning-based Hierarchical Variational Model. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (Eds.). Association for Computational Linguistics, Hong Kong, China, 3257–3268. doi:10.18653/v1/D19-1321
- [63] Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. 2021. Retrieval Augmentation Reduces Hallucination in Conversation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*. Association for Computational Linguistics, Punta Cana, Dominican Republic, 3784–3803. doi:10.18653/v1/2021.findings-emnlp.320
- [64] Avi Singh, John D. Co-Reyes, Rishabh Agarwal, Ankesh Anand, Piyush Patil, Xavier Garcia, Peter J. Liu, James Harrison, Jaehoon Lee, Kelvin Xu, Aaron Parisi, Abhishek Kumar, Alex Alemi, Alex Rizkowsky, Azade Nova, Ben Adlam, Bernd Bohnet, Gamaleldin Elsayed, Hanie Sedghi, Igor Mordatch, Isabelle Simpson, Izzeddin Gur, Jasper Snoek, Jeffrey Pennington, Jiri Hron, Kathleen Kenealy, Kevin Swersky, Kshitij Mahajan, Laura Culp, Lechao Xiao, Maxwell L. Bileschi, Noah Constant, Roman Novak, Rosanne Liu, Tris Warkentin, Yundi Qian, Yamini Bansal, Ethan Dyer, Behnam Neyshabur, Jascha Sohl-Dickstein, and Noah Fiedel. 2024. Beyond Human Data: Scaling Self-Training for Problem-Solving with Language Models. arXiv:2312.06585 [cs.LG] <https://arxiv.org/abs/2312.06585>
- [65] Shamane Siriwardhana, Rivindu Weerasekera, Elliott Wen, Tharindu Kaluarachchi, Rajib Rana, and Suranga Nanayakkara. 2023. Improving the Domain Adaptation of Retrieval Augmented Generation (RAG) Models for Open Domain Question Answering. *Transactions of the Association for Computational Linguistics* 11 (2023), 1–17. doi:10.1162/tacl\_a\_00530
- [66] Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. 2024. Scaling LLM Test-Time Compute Optimally can be More Effective than Scaling Model Parameters. arXiv:2408.03314 [cs.LG] <https://arxiv.org/abs/2408.03314>
- [67] Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2 (Montreal, Canada) (NIPS'14)*. MIT Press, Cambridge, MA, USA, 3104–3112.
- [68] Guy Tevet and Jonathan Berant. 2021. Evaluating the Evaluation of Diversity in Natural Language Generation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, Paola Merlo, Jorg Tiedemann, and Reut Tsarfaty (Eds.). Association for Computational Linguistics, Online, 326–346. doi:10.18653/v1/2021.eacl-main.25
- [69] The Lemur Project. 2009. The ClueWeb09 Dataset. <https://lemurproject.org/clueweb09/> Accessed: 2024-06-24.
- [70] Longyue Wang, Chenyang Lyu, Tianbo Ji, Zhirui Zhang, Dian Yu, Shuming Shi, and Zhaopeng Tu. 2023. Document-Level Machine Translation with Large Language Models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.).

- Association for Computational Linguistics, Singapore, 16646–16661. doi:10.18653/v1/2023.emnlp-main.1036
- [71] Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Nathan Cooper, Griffin Adams, Jeremy Howard, and Iacopo Poli. 2024. Smarter, Better, Faster, Longer: A Modern Bidirectional Encoder for Fast, Memory Efficient, and Long Context Finetuning and Inference. arXiv:2412.13663 [cs.CL] <https://arxiv.org/abs/2412.13663>
- [72] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2024. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems (New Orleans, LA, USA) (NIPS '22)*. Curran Associates Inc., Red Hook, NY, USA, Article 1800, 14 pages.
- [73] Bo Xu, Hongfei Lin, Liang Yang, Kan Xu, Yijia Zhang, Dongyu Zhang, Zhihao Yang, Jian Wang, Yuan Lin, and Fuliang Yin. 2018. Improve Diversity-oriented Biomedical Information Retrieval using Supervised Query Expansion. In *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. 556–559. doi:10.1109/BIBM.2018.8621577
- [74] Jingjing Xu, Xuancheng Ren, Junyang Lin, and Xu Sun. 2018. Diversity-Promoting GAN: A Cross-Entropy Based Generative Adversarial Network for Diversified Text Generation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii (Eds.). Association for Computational Linguistics, Brussels, Belgium, 3940–3949. doi:10.18653/v1/D18-1428
- [75] Han Yin, Jianxing Yu, Miaopei Lin, and Shiqi Wang. 2024. Answering Spatial Commonsense Questions Based on Chain-of-Thought Reasoning with Adaptive Complexity. In *Web and Big Data*, Wenjie Zhang, Anthony Tung, Zhonglong Zheng, Zhengyi Yang, Xiaoyang Wang, and Hongjie Guo (Eds.). Springer Nature Singapore, Singapore, 186–200.
- [76] Fei Yu, Hongbo Zhang, Prayag Tiwari, and Benyou Wang. 2024. Natural Language Reasoning, A Survey. *ACM Comput. Surv.* 56, 12, Article 304 (Oct. 2024), 39 pages. doi:10.1145/3664194
- [77] Hamed Zamani and Michael Bendersky. 2024. Stochastic RAG: End-to-End Retrieval-Augmented Generation through Expected Utility Maximization. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval (Washington DC, USA) (SIGIR '24)*. Association for Computing Machinery, New York, NY, USA, 2641–2646. doi:10.1145/3626772.3657923
- [78] Hamed Zamani, Fernando Diaz, Mostafa Dehghani, Donald Metzler, and Michael Bendersky. 2022. Retrieval-Enhanced Machine Learning. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval (Madrid, Spain) (SIGIR '22)*. Association for Computing Machinery, New York, NY, USA, 2875–2886. doi:10.1145/3477495.3531722
- [79] Kexun Zhang, Shang Zhou, Danqing Wang, William Yang Wang, and Lei Li. 2024. Scaling LLM Inference with Optimized Sample Compute Allocation. arXiv:2410.22480 [cs.CL] <https://arxiv.org/abs/2410.22480>