Writing Like the Best: Exemplar-Based Expository Text Generation

Yuxiang Liu Kevin Chen-Chuan Chang
University of Illinois at Urbana-Champaign, USA
{yuxiang, kcchang}@illinois.edu

Abstract

We introduce the *Exemplar-Based Expository Text Generation* task, aiming to generate an expository text on a new topic using an exemplar on a similar topic. Current methods fall short due to their reliance on extensive exemplar data, difficulty in adapting topic-specific content, and issues with long-text coherence. To address these challenges, we propose the concept of *Adaptive Imitation* and present a novel RECURRENT PLAN-THEN-ADAPT (**REPA**) framework. REPA leverages large language models (LLMs) for effective adaptive imitation through a fine-grained plan-then-adapt process. REPA also enables recurrent segment-by-segment imitation, supported by two memory structures that enhance input clarity and output coherence. We also develop task-specific evaluation metrics–*imitativeness*, *adaptiveness*, and *adaptive-imitativeness*–using LLMs as evaluators. Experimental results across our collected three diverse datasets demonstrate that REPA surpasses existing baselines in producing factual, consistent, and relevant texts for this task.

1 Introduction

The increasing demand for digital content creation necessitates advanced natural language generation techniques to produce high-quality text at scale (Gatt and Krahmer, 2018; Xu et al., 2024). A system that can reliably generate *expository texts*[1] that introduce or summarize topics–such as concepts, entities, or subjects–is highly desirable (Jiang et al., 2024; Shao et al., 2024a). This is particularly valuable in domains requiring *consistent* style, like faculty profiles, university overviews, product descriptions, or event introductions, where large volumes of expository texts must be produced from limited examples. Such a system can significantly reduce human effort by automating the writing process, enhance consistency by maintaining a uniform style, and enable scalability by rapidly generating large volumes of text.

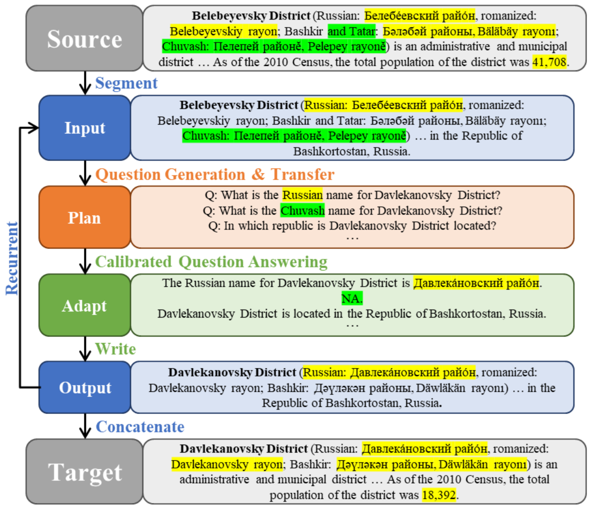

Figure 1: An illustration of REPA for *Exemplar-Based Expository Text Generation*, where yellow text indicates adapted facts and green text indicates discarded facts.

For expository text generation, prior methods that rely on extensive domain corpora (Balepur et al., 2023; Shao et al., 2024b) struggle when such data is scarce, while approaches for open-ended generation without domain-specific data (Yang et al., 2022, 2023a) often fail to produce structurally consistent content that adheres to domain conventions. To this end, we introduce *Exemplar-Based Expository Text Generation*, a novel task that generates an expository text on a new topic by leveraging an exemplar from a similar topic, avoiding reliance on large corpora or unconstrained generation methods. As shown in Figure 1, given a text on "Belebeyevsky District", the task is to generate a text on "Davlekanovsky District", preserving the source's overall structure while incorporating topic-specific details. Since expository texts are typically long-form[2], this task focuses on generating infor-

[1]Expository text is writing that presents factual information with the purpose of informing, explaining, or describing a topic for readers, rather than entertaining or persuading them.

[2]Long-form text refers to content that extends beyond the

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 25739–25764
July 27 - August 1, 2025

mative, topic-centric long-form content to better address practical needs.

The goal is twofold: to preserve the structure and content of the exemplar, ensuring *cross-topic consistency*, while maintaining accuracy and relevance to the new topic by addressing *cross-topic variability*. However, existing methods that rely on extensive exemplar data are impractical (Balepur et al., 2023; Shao et al., 2024b), making it difficult to achieve *cross-topic consistency* due to limited data for generalization. Additionally, maintaining structural consistency with the exemplar while ensuring factual accuracy for the new topic is particularly challenging due to *cross-topic variability*, which often leads to hallucinations and inaccuracies (Ji et al., 2023; Rawte et al., 2023).

To address these challenges, we introduce a new concept called **Adaptive Imitation**, inspired by how humans learn to write through studying exemplars (Vijayakumar, 2024; Chen, 2024; Carter et al., 2018; Wu, 2019; Wette, 2014). This involves *imitating* the organizational structure of an exemplar while *adapting* the content to fit specific topics and writing requirements. Specifically, we propose a *plan-then-adapt* approach for fine-grained step-by-step control of LLMs with an **imitative planner** and an **adaptive generator**. The imitative planner leverages LLMs and generates topic-centric outlines framed as questions based on the exemplar. These outlines are then transferred to the new topic through straightforward topic token substitution, ensuring *cross-topic consistency* (PLAN: Section 3.2.1). The adaptive generator incorporates retrieval augmentation (Ram et al., 2023a; Shi et al., 2023) and confidence calibration (Xiong et al., 2023; Tian et al., 2023) to realize outlines by answering questions factually and discriminatively, effectively adapting topic-specific content and thus addressing *cross-topic variability* (ADAPT: Section 3.2.2).

Moreover, to scale to longer texts, we adopt a recurrent, segment-by-segment processing strategy, where both input and output are handled incrementally. To address coreference issues between input segments, we introduce a *short-term memory* that retains recent input segments during the planning phase. To reduce output redundancy and maintain coherence, we incorporate a *long-term memory* that summarizes all previously generated segments during the adaptation phase. This recurrent framework, integrated with the two memory structures, enables our model to process arbitrarily long texts while preserving information integrity. In summary, we propose a novel RECURRENT PLAN-THEN-ADAPT (**REPA**) framework, which integrates planning, adaptation, and recurrent processing with the usage of both short-term and long-term memory.

Finally, as established metrics do not adequately assess stylistic fidelity and knowledge transfer in imitation and adaptation, failing to capture the nuances of this task, we additionally develop task-specific metrics *imitativeness*, *adaptiveness*, and *adaptive-imitativeness* using LLM-as-a-Judge (Zheng et al., 2024). Human evaluations further confirm that the LLM-based judgments align well with human judgments. We collect three diverse datasets covering both open-domain and domain-specific scenarios. To evaluate our approach, we conduct extensive experiments, including comparisons with strong baselines and ablation studies. The results show that REPA significantly outperforms the baselines by producing factual texts that achieve cross-topic consistency and handle cross-topic variability, with high level of imitativeness and adaptiveness. Additionally, our analysis reveals that each module within RePA contributes effectively to its overall performance.

Our contributions can be summarized as follows:
(1) We are the first to study *Exemplar-Based Expository Text Generation* task, addressing a practical yet under-explored area with broad applications.
(2) We present a novel RECURRENT PLAN-THEN-ADAPT (**REPA**) framework with two memory structures, achieving fine-grained control of LLMs.
(3) We develop task-specific metrics *imitativeness*, *adaptiveness*, and *adaptive-imitativeness* using LLM-as-a-Judge for comprehensive evaluation.
(4) We collect three diverse datasets and our extensive experiments demonstrate the superior performance of our proposed method for this task[3].

2 Related Work

2.1 Long Input Processing

Our task is distinguished by its long input. Previous work on long input processing has focused on tasks like outline generation using hierarchical decoders (Zhang et al., 2019), summarization with

sentence level, such as one or more paragraphs or even longer passages (Shen et al., 2019; Guan et al., 2021; Hu et al., 2022; Min et al., 2023; Liang et al., 2023).

[3] Our codes and datasets are available in our repository: `https://github.com/liuyuxiang512/RePA.git`.

extract-then-generate (Mao et al., 2022), divide-and-conquer (Zhang et al., 2022), graph-based methods (Hua et al., 2023), and dialogue response generation using retrieval (Kumari et al.) or memory augmentation (Lu et al., 2023; Lee et al., 2023; Wang et al., 2023). Some works aim to improve efficiency with sparse attention (Beltagy et al., 2020; Ivgi et al., 2023; Jin et al., 2024). However, these methods primarily generate outputs that are significantly shorter and structurally different from their inputs, containing only a subset of the input's information. In contrast, our task requires outputs comparable in length and structure to the inputs, while incorporating additional relevant knowledge.

2.2 Long-Form Text Generation

Prior research on long-form text generation (LFTG) (Köksal et al., 2023; You et al., 2023; Zhou et al., 2023; Liang et al., 2023; Adewoyin et al., 2022) has addressed tasks such as story (Yang et al., 2022, 2023a), data-to-text (Moryossef et al., 2019; Bai et al., 2021), script (Mirowski et al., 2023), and expository text generation (Balepur et al., 2023), where *plan-then-generate* framework is commonly applied to improve coherence of generating long-form text in one go. Various formats of "plan" have been proposed, including key phrases (Hu et al., 2022), events (Goldfarb-Tarrant et al., 2020), data records (Moryossef et al., 2019), and section titles (Shao et al., 2024b). Some approaches incorporate discourse guidance (Adewoyin et al., 2022) or combine planning with discourse for improved fluency (Sun et al., 2022). Other methods enhance planning with retrieval to address logic conflicts (Guan et al., 2020) or leverage memory to retain key information (Zhou et al., 2023). Unlike existing LFTG methods that use brief inputs and favor open-ended, creative generation with minimal constraints, our task focuses on producing factual, informative outputs from long, detailed inputs, which requires strict adherence to the source text and limits open-endedness. Additionally, unlike methods that rely on an external or learned "bank of plans", our task necessitates deriving the plan exclusively from a single, lengthy input, ensuring alignment with its content and structure.

2.3 Confidence Calibration of LLMs

Confidence calibration–the ability to produce accurate confidence scores indicating the correctness of generated text–is crucial for the trustworthiness of real-world systems. Previous methods required white-box access to model architectures or fine-tuning (Kadavath et al., 2022; Jiang et al., 2021; Lin et al., 2022; Yang et al., 2023b; Zhang et al., 2023; Slobodkin et al., 2023), which is infeasible for proprietary Large Language Model (LLM) APIs. Recently, Xiong et al. (2023) and Tian et al. (2023) explored black-box LLM uncertainty estimation in reasoning and factual short-answer, finding that verbalized confidences emitted as output tokens are typically better calibrated than the model's conditional probabilities. Building on this, we integrate confidence calibration into our retrieval-augmented generation (RAG) process, allowing for explicit assessment of the accuracy of generated texts.

3 Method

3.1 Task Definition

Exemplar-Based Expository Text Generation task is formally defined as: Given an expository text $\mathbf{X} = \{x_i\}_{i=1}^m$ of a sequence of sentences on a source topic $\mathbf{t_x}$, a target topic $\mathbf{t_y}$, and external knowledge sources denoted as $\mathbf{K}$, the objective is to produce a new expository text $\mathbf{Y} = \{y_j\}_{j=1}^n$ on the target topic $\mathbf{t_y}$ that imitates the content and structure of $\mathbf{X}$ while adapting topic-specific content.

Our goal is to develop an instructed imitative content creator, a system capable of generating a long text on a target topic given a provided long exemplar text on a source topic. This task, essentially knowledge-intensive one-shot long-form text generation, involves comprehending the structure and content of the source document, extracting pertinent information about the source topic, and then adapting this information to ensure relevance and accuracy for the target topic. Crucially, the challenge lies in preserving the coherence and quality of the source text while ensuring the accuracy and relevance of the generated output to the target topic.

3.2 REPA: RECURRENT PLAN-THEN-ADAPT

We propose a novel RECURRENT PLAN-THEN-ADAPT (**REPA**) model, outlined in Algorithm 1. REPA first segments[4] the input text $\mathbf{X}$ into a sequence of text segments $\{x_t\}_{t=1}^T$, then recurrently processes each input segment x_t and generates an output segment y_t. At each recurrence step, REPA employs a PLAN-THEN-ADAPT process. The PLAN stage involves learning outlines q_t from

[4]We opt for sentences as the recurrence basis as they are semantic units that strike a balance between granularity and coherence for imitation and adaptation.

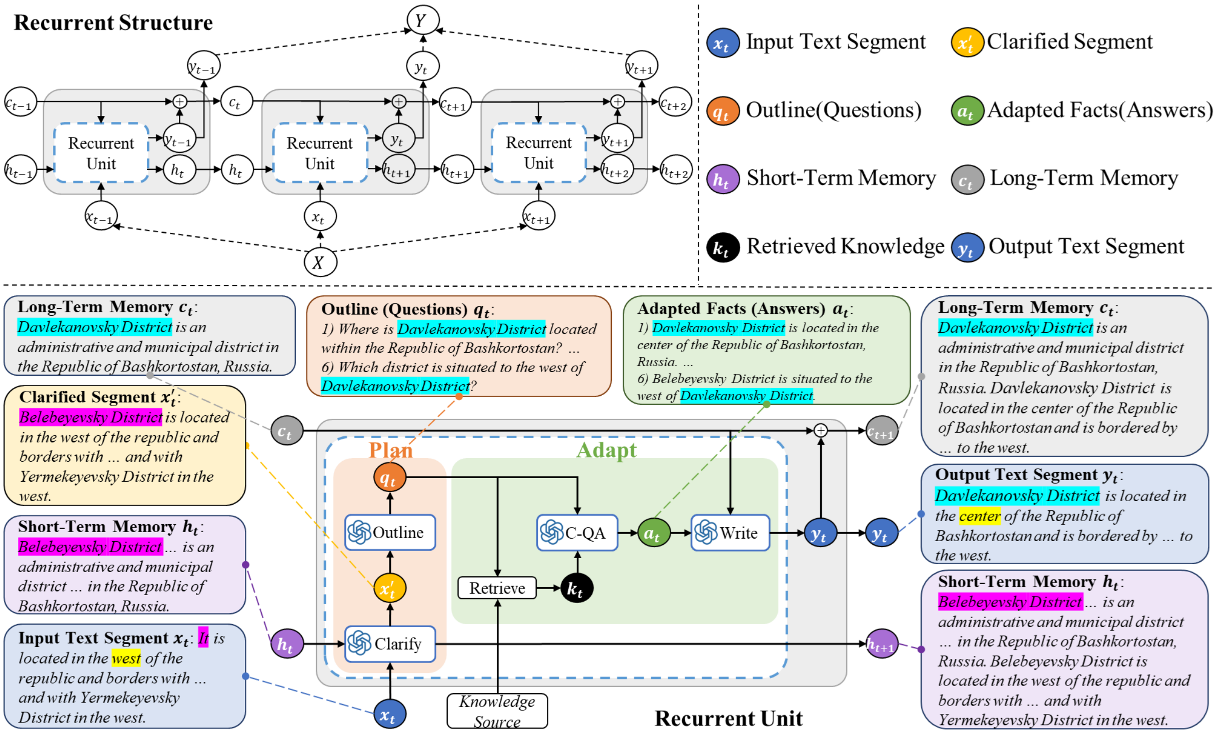

Figure 2: Overview of REPA. Top left shows the high-level recurrent structure for sequential processing. Bottom details the recurrent unit with a running example: "Clarify" and "Outline" in PLAN, and "Calibrated-QA" (C-QA) and "Write" in ADAPT, with memory usages in "Clarify" (short-term) and "Write" (long-term).

input x_t as plans for generation. The ADAPT stage is to realize outlines q_t flexibly and effectively with knowledge $\mathbf{K}$ to generate the output y_t adaptively. Specifically, we introduce two on-the-fly memory structures to retain essential information from previous recurrent steps: a short-term memory h_t for the history input segments and a long-term memory c_t for the history output segments. This design ensures our model's capability to handle arbitrarily long text without sacrificing information.

Algorithm 1 RECURRENT PLAN-THEN-ADAPT

Initialize $\mathbf{Y}$ as an empty string
for $x_t \in \{x_t\}_{t=1}^{T} \in \mathbf{X}$ **do**
  $q_t, h_{t+1} \leftarrow \text{PLAN}(x_t, h_t)$
  $y_t, c_{t+1} \leftarrow \text{ADAPT}(q_t, c_t, \mathbf{K})$
  Append y_t to $\mathbf{Y}$
end for
return $\mathbf{Y}$

As showed in Figure 2, REPA mirrors the recurrence structure in LSTMs, but a closer comparison highlights several distinct features: 1) it employs a text-based representation for input x_t, output y_t, short-term memory h_t, and long-term memory c_t; 2) it leverages prompting of general-purpose LLMs $\mathcal{M}$ such as GPT-4 to perform computations; and 3) it integrates the LLM-driven PLAN-THEN-ADAPT process within each recurrent unit. Formally:

$$y_t, c_{t+1}, h_{t+1} = \text{Recurrent}(x_t, c_t, h_t, \mathcal{M}, \mathbf{K}),$$

where x_t, y_t, h_t, c_t denote input segment, output segment, shot-term memory, long-term memory at time step t, respectively. $\mathbf{K}$ denotes external knowledge. The final output is a concatenation of all output segments, i.e., $\mathbf{Y} = y_1 y_2 \cdots y_T$. Figure 2 also presents a running example, with full examples available in Appendix A.

3.2.1 The PLAN Module

Content planning plays a crucial role in guiding long-form text generation. However, unlike traditional approaches that rely on external "bank of plans" for planning (Hua et al., 2019; Goldfarb-Tarrant et al., 2020; Balepur et al., 2023), our task necessitates deriving a content plan exclusively from a single lengthy input text. This constraint imposes strict requirements on maintaining fidelity to the original input's content and structure. Drawing inspiration from the "Questions under Discussion" (QUD) theory, recent works have conceptualized text plans as sequences of question-answer pairs (Narayan et al., 2023; Huot et al., 2023) for query-focused summarization. However, such approaches often require exhaustive annotation for question generation (Liu et al., 2023). To grasp the content and structure of the input text without plan annotation, we aim to extract key information and frame it into questions by prompting LLMs, which serve as outlines to guide subsequent generation, helping to retain *cross-topic consistency*. Our PLAN module comprises two sequential components: Clarify and Outline.

Clarify Segmenting long inputs into smaller segments can introduce coreference ambiguities, where pronouns may lack clear antecedents in a current segment x_t, *e.g.*, "it" in the input text segment (Figure 2) is unclear. This ambiguity complicates the understanding of the current input. To address this, we introduce the Clarify component, which uses short-term memory, h_t, to resolve coreference issues, e.g., replacing "it" with "Belebeyevsky District" for better comprehension. The memory h_t stores key information from *recent* input segments and is updated each time a clarified segment is generated. Essentially, this short-term memory acts as a sliding window of context, helping to *clarify* the current segment by replacing pronouns with their corresponding antecedents (Figure 4). Formally:

$$x'_t, h_{t+1} = \text{LLM}_{\text{Clarify}}(x_t, h_t),$$

Outline The Outline component then frames key points into outlines, initially focusing on the source topic $\mathbf{t_x}$ before transferring to the target topic $\mathbf{t_y}$. The outlines are conceptualized as a set of topic-centric questions on topic $\mathbf{t_y}$, achieved by first generating questions from clarified input segment and replacing source topic tokens with the target ones in the questions. Essentially, the Outline component performs question generation and transfer (Figure 5), as examples in Figure 3. The advantage of using questions as outlines lies in their conciseness for summarizing key points and their transitivity for transferring key points from source topic to target topic with simply topic token substitution. Thus, we have:

$$q_t = \text{LLM}_{\text{Outline}}(x'_t, t_x, t_y)$$

3.2.2 The ADAPT Module

Previous *plan-then-generate* frameworks often assume plans will seamlessly translate into effective outputs. However, in our task, plans for the target topic are derived from those for the source topic. Ideally, talking points would align perfectly between the source and target topics, but in reality, they may vary, making the outlines less suitable for the target topic, *e.g.*, "Chuvash name" does not exist for target topic in Figure 1. Since our goal is to generate informative texts which maintain factual correctness, it is essential to *adapt* the outlines for the target topic. Therefore, we propose an ADAPT module to handle outlines flexibly and gracefully, accommodating *cross-topic variability*. The intuition of this module is that "an imperfect outline for the target topic is acceptable if handled correctly". Specifically, there are two components: Calibrated Question Answering (Calibrated-QA) and Write.

Calibrated-QA A straightforward approach to handling outlines in the form of questions is retrieval-augmented question answering (QA). However, outlines may not perfectly align with the target topic, which can result in unsuitable or unanswerable questions. To address this, we introduce a *refusal* mechanism inspired by recent works on confidence calibration (Xiong et al., 2023; Tian et al., 2023), which employ verbalized confidence to calibrate black-box LLMs and have demonstrated success in verbalized confidence calibration for factoid short-answer QA. Our approach involves prompting LLMs to generate confidence in their answers (Figure 6), and to refuse to answer questions deemed unanswerable by filtering out those with low-confidence answers, as an example in Figure 1 where "NA" corresponds to a low-confidence answer. This method shifts the focus from generating a perfect outline to handling it flexibly and appropriately. For retrieval, we conduct both per-topic retrieval using the target topic and per-query retrieval using the current question. We apply in-context retrieval-augmented language models (RALM) (Ram et al., 2023b) for calibrated QA. Essentially, Calibrated-QA involves retrieval-augmented QA with verbalized confidence calibration, denoted as:

$$\begin{aligned} k_t &= \text{Retriever}(t_y, q_t), \\ a_t &= \text{LLM}_{\text{Calibrated-QA}}(q_t, k_t, \theta), \end{aligned}$$

where k_t is knowledge, and a_t are answers with verbalized confidences higher than threshold θ (or adapted facts as shown in Figure 2).

Write Based on the adapted facts, the Write component aims to generate an output segment which aligns with the content and structure of the input, maintains coherency and avoids repetition in output. It first generates a draft output segment that is consistent with the adapted facts on the target topic, then revises it to remove redundant content from previous output segments (Figure 7), with a complete example in Figure 3. To support this process, we introduce a long-term memory c_t to store key information from *all* previous output segments, and c_t is updated to summarize the latest output segment once the current target segment is generated. Formally:

$$y_t = \text{LLM}_{\text{Write}}(a_t, c_t),$$
$$c_{t+1} = \text{LLM}_{\oplus}(c_t, y_t),$$

where $\oplus$ denotes summarization (Figure 8).

4 Experimental Setup

4.1 Datasets

To ensure effective imitation and adaptation, source and target topics must belong to the same domain. Such "correspondence" demands a high level of similarity, though "variations" usually exist. To this end, we collect three diverse datasets[5] to cover both open-domain and domain-specific scenarios, including Wikipedia, RoleEE, and USNews.

Wikipedia The Wikipedia dataset is domain-agnostic. We collected the overview sections of open-domain Wikipedia articles, with Wikipedia titles serving as topics. Specifically, we used the English Wikipedia dump as of April 1st, 2024[6], and trained title embeddings using Wikipedia2Vec[7] (Yamada et al., 2020). We then employed cosine similarity to pair Wikipedia topics and texts, ensuring a similarity score higher than 0.95. To enhance text quality and similarity, we filtered out texts with fewer than 3 sentences or fewer than 60 words, as well as topic pairs with significant divergence in their categories[8], indicated by a percentage of common category tags higher than 0.3. We totally collected 1000 samples and a random manual check of the collected Wikipedia dataset showed its suitability for our task.

RoleEE The RoleEE dataset is a multi-domain event dataset introduced by Jiao et al. (2022), featuring 50 impactful hot event types. We selected three event categories–academic award ceremony, music award ceremony, and satellite launch–and took the corresponding events as topics. After manually removing poor-quality texts within each category, we paired them using the Hugging Face text embedding tool[9] to obtain the top 500 topic/text pairs. Specifically, for each event, we retained the first paragraph from its text to ensure text similarity.

USNews The USNews dataset is a domain-specific dataset from U.S. News best colleges[10]. We crawled the overviews of 420 best national universities, with universities serving as topics. These were then paired based on cosine similarity using the same Hugging Face text embedding tool to obtain the top 500 topic/text pairs. Specifically, for each university, we extracted the first paragraph from its overview section, and the first sentence of the second paragraph if available, considering the high similarity of first paragraphs across universities and the significant divergence in subsequent paragraphs of their overviews.

4.2 Baselines

We employ the following baselines, and LLMs denote GPT-4 (Achiam et al., 2023) or LLaMA 3[11].
LLM is to directly prompt an LLM to generate a target text given the source text and the target topic.
RollingLLM (RoM) is to divide input text into segments and recurrently prompt an LLM to generate a target segment until reaching the last input segment, similar to using a sliding window approach for generating long texts with Transformers.
o1 is a new OpenAI model[12] that excel at solving complex problems by spending more time thinking before responding.
Self-Refine (SR) (Madaan et al., 2024) is to improve initial outputs through iterative feedback and refinement. We divide the input text into segments and perform 4 iterations of Self-Refine on each segment for comparable inference steps with REPA.
Default is to simply replace source topic in the source text with target topic, serving as an approximation for understanding cross-topic variability.

We also developed retrieval-augmented generation (RAG) variants of the aforementioned LLM-based baselines to integrate retrieved knowledge

[5]Our datasets consist of long texts, though not as extensive as those spanning thousands of words as we prioritize high-quality, similar text pairs with acceptable variance; we also tested longer texts, as shown in Appendix C.
[6]https://dumps.wikimedia.org/enwiki/20240401/
[7]https://wikipedia2vec.github.io/wikipedia2vec/
[8]https://en.wikipedia.org/wiki/Wikipedia:Contents/Categories
[9]https://huggingface.co/jinaai/jina-embeddings-v2-base-en
[10]https://www.usnews.com/best-colleges
[11]https://LLaMA.meta.com/LLaMA3/
[12]https://openai.com/index/introducing-openai-o1-preview/

on the target topic during generation, specifically **LLM+Retr**, **RoM+Retr**, **o1+Retr**, and **SR+Retr**.

4.3 Implementation Details

REPA is built upon two latest general-purpose Language Models (LLMs) in our experiments: OpenAI GPT-4 (Achiam et al., 2023) and Meta LLaMA 3[13]. Specifically, we employed the models "gpt-4-0125-preview", "meta/meta-LLaMA-3-8b-instruct", and "meta/meta-LLaMA-3-70b-instruct". o1 is actually the OpenAI model "o1-preview-2024-09-12". For these LLMs, we configure the following parameters: temperature as 0.3, frequency penalty as 0.3, max generation tokens as 256, and the number of response choices as 1. Additionally, for calibrated QA, we set the confidence threshold as 0.7, and aimed to make the process as deterministic as possible by setting the seed. The prompts used in each module are provided in Appendix B.

For retrieval, we employed both per-topic and per-query retrieval. For topic retrieval, we utilized the Bing API as the retriever, with the topics as the input. This applies to both our model and the concerned baselines. Given that datasets derived from Wikipedia articles require knowledge from the open web, we excluded Wikipedia domains when using the Bing API[14]. Similarly, for datasets sourced from U.S. News, we excluded U.S. News domains[15]. For query retrieval, we employed bi-encoder in DPR (Karpukhin et al., 2020) as the retriever to retrieve from a fixed knowledge base, comprising sentence-level facts. Specifically, we used the "multiset" versions of question and context encoders, respectively, and used their cosine similarity for retrieval. When incorporating knowledge pieces into the prompts, we selected the top 10 Bing results and the top 3 DPR results.

4.4 Metrics

Basic Metrics To provide a basic assessment of this generation task, we use well-established metrics commonly used in previous research, namely ROUGE (Lin, 2004), BLEU (Papineni et al., 2002), and METEOR (Banerjee and Lavie, 2005).

Task-Specific Metrics Established metrics cannot capture the nuances of our task. We leverage LLM-as-a-Judge (Zheng et al., 2024) to evaluate *Imitativeness* and *Adaptiveness*. Imitativeness represents the model's ability to faithfully replicate the structure and content of the exemplar, ensuring cross-topic consistency. Adaptiveness assesses how effectively the model adapts the exemplar content to a new topic, addressing cross-topic variability while ensuring consistency. We also introduce *Adaptive-Imitativeness*, which is a F1-score of Imitativeness and Adaptiveness. Additional details and discussion on addressing the known limitations of LLM-Judge are provided in Appendix D.

Factuality Metrics Given the knowledge-intensive nature of our task, maintaining factual accuracy is crucial. Inspired by FActScore (Min et al., 2023), a metric for evaluating factual accuracy of long-form text, we decompose output text into sentence-level facts, then take ground truth as knowledge source to calculate the percentages of entailment (**NLI-E**) and contradiction (**NLI-C**) sentences in outputs. Our human evaluation of the effectiveness of NLI-based metrics are included in Appendix E. We also compute the percentage of hallucinated tokens compared to both inputs and ground truths, named **Halluc**.

5 Experimental Results

5.1 Comparison with Baselines

We present the evaluation results of models built on GPT-4 in Tables 1 and 2. Results of models built on LLaMA 3 are included in Appendix F. We also conducted a case study as in Appendix G.

Considering basic generation metrics, REPA outperforms baselines on almost all metrics, although LLaMA 3-based baselines may achieve higher scores on metrics such as ROUGE and Meteor. Notably, the Default baseline, which simply replaces the source topic with the target topic in the source text, achieves a Meteor score of 0.7667 on the Wikipedia dataset, higher than all GPT-4 based models, suggesting that **established basic metrics might not be a reliable indicator of model performance for our task**. For factuality metrics including Halluc, NLI-E and NLI-C, REPA significantly outperforms baselines including retrieval-augmented baselines, indicating that **REPA generates more factual content**.

For task-specific metrics *Imitativeness*, *Adaptiveness*, and *Adaptive-Imitativeness*, we find that prompting LLMs yields high Imitativeness, showing **LLMs are strong imitators regardless of output factuality**. However, their extremely low Adaptiveness indicates poor cross-topic variabil-

[13]https://LLaMA.meta.com/LLaMA3/

[14]Wikipedia domains: "wikipedia.org", "wikiwand.com", "wiki2.org", and "wikimedia.org".

[15]U.S. News domains: "usnews.com"

Datasets	Models	R1↑	R2↑	R↑L	RLsum↑	Meteor↑	BLEU↑	Halluc↓	NLI-E↑	NLI-C↓
Wikipedia	REPA	**0.8112**	**0.7146**	**0.7600**	**0.7625**	0.7368	**0.6672**	**6.5714**	**0.7927**	**0.0439**
	LLM	0.6855	0.4835	0.6154	0.6133	0.6984	0.4236	23.5794	0.3604	0.4561
	LLM+Retr	0.7583	0.6121	0.7027	0.7057	0.7595	0.5887	15.8273	0.5343	0.3106
	RoM	0.7300	0.5563	0.6785	0.6774	0.7448	0.5190	17.3810	0.3640	0.5223
	RoM+Retr	0.7343	0.6073	0.6889	0.6906	0.7328	0.5408	14.9102	0.5240	0.3536
	o1	0.7658	0.6345	0.7223	0.7210	0.7547	0.6018	14.4050	0.4954	0.3720
	o1+Retr	0.7421	0.6258	0.6993	0.7002	0.6646	0.5804	13.2928	0.6227	0.2194
	SR	0.7387	0.5628	0.6782	0.6748	0.6957	0.5225	21.6280	0.3701	0.4653
	SR+Retr	0.7839	0.6247	0.7187	0.7202	0.7599	0.6011	10.6257	0.6032	0.2751
	Default	0.7520	0.6042	0.7129	0.7127	**0.7667**	0.5705	16.4128	0.1311	0.7152
RoleEE	REPA	**0.9184**	**0.8691**	**0.9029**	**0.9027**	**0.9170**	**0.7548**	**5.2604**	**0.9067**	**0.0653**
	LLM	0.6780	0.5329	0.6399	0.6526	0.7200	0.3784	26.9641	0.2696	0.5487
	LLM+Retr	0.8848	0.8197	0.8692	0.8726	0.9093	0.6924	9.4551	0.7109	0.1833
	RoM	0.7910	0.6572	0.7735	0.7747	0.8094	0.6096	15.3346	0.2543	0.7003
	RoM+Retr	0.8842	0.8191	0.8715	0.8715	0.8954	0.7117	6.4164	0.7400	0.2053
	o1	0.8306	0.7273	0.8165	0.8136	0.8334	0.6932	13.1344	0.3793	0.5457
	o1+Retr	0.8849	0.8233	0.8706	0.8671	0.8812	0.7814	9.0228	0.7583	0.1673
	SR	0.8126	0.7856	0.7892	0.7821	0.8255	0.6241	18.0331	0.2527	0.6415
	SR+Retr	0.9041	0.8306	0.8652	0.8624	0.9083	0.7262	7.7313	0.7981	0.1376
	Default	0.8090	0.6722	0.7898	0.7893	0.8351	0.6003	15.7650	0.1200	0.8301
USNews	REPA	**0.8922**	**0.8396**	**0.8642**	**0.8653**	**0.8971**	**0.8129**	**4.6048**	**0.8085**	**0.0441**
	LLM	0.7651	0.6258	0.7258	0.7270	0.8043	0.5427	18.3919	0.3749	0.5257
	LLM+Retr	0.5842	0.4303	0.5234	0.5249	0.5994	0.3037	33.9516	0.3836	0.2853
	RoM	0.7390	0.6178	0.7156	0.7153	0.8008	0.5585	15.7913	0.3621	0.4873
	RoM+Retr	0.6561	0.5127	0.6150	0.6172	0.6866	0.4312	23.4262	0.3322	0.4047
	o1	0.8327	0.7404	0.8138	0.8144	0.8548	0.6970	12.1111	0.4200	0.4827
	o1+Retr	0.7265	0.6082	0.6880	0.6875	0.7273	0.5572	19.4480	0.3970	0.3372
	SR	0.7381	0.6137	0.7019	0.6973	0.8126	0.5495	19.5238	0.3527	0.4562
	SR+Retr	0.8458	0.7239	0.8182	0.8139	0.8639	0.7236	11.5269	0.5338	0.1757
	Default	0.7245	0.5808	0.6999	0.7013	0.7768	0.5494	18.1090	0.1667	0.5700

Table 1: Evaluation results on basic and factuality metrics. LLM denotes GPT-4 specifically.

Datasets	Models	*I.*↑	*A.*↑	*A.-I.*↑
Wikipedia	REPA	4.16	**3.90**	**3.93**
	LLM	4.52	2.44	3.06
	LLM+Retr	4.46	2.78	3.25
	RoM	4.58	2.32	3.00
	RoM+Retr	4.08	2.56	3.00
	o1	4.34	2.94	3.40
	o1+Retr	4.32	3.02	3.42
	SR	4.56	2.54	3.25
	SR+Retr	4.22	3.04	3.50
	Default	**5.00**	1.08	1.73
RoleEE	REPA	4.80	**4.30**	**4.46**
	LLM	4.70	2.76	3.23
	LLM+Retr	4.80	4.26	4.39
	RoM	4.62	1.94	2.55
	RoM+Retr	4.70	4.04	4.21
	o1	4.64	2.66	3.12
	o1+Retr	4.68	4.24	4.34
	SR	4.70	2.62	3.30
	SR+Retr	4.74	4.22	3.33
	Default	**5.00**	1.24	1.87
USNews	REPA	4.22	**4.32**	**4.22**
	LLM	4.20	3.06	3.45
	LLM+Retr	4.02	2.86	3.25
	RoM	4.58	2.74	3.30
	RoM+Retr	4.08	2.40	2.96
	o1	4.20	3.18	3.53
	o1+Retr	4.14	2.98	3.37
	SR	4.44	2.98	3.49
	SR+Retr	4.12	3.74	3.84
	Default	**5.00**	1.00	1.67

Table 2: Evaluation results on task-specific metrics. LLM denotes GPT-4 specifically. ***I.*** denotes *Imitativeness*, ***A.*** denotes *Adaptiveness*, and ***A.-I.*** denotes *Adaptive-Imitativeness*.

ity recognition and weak topic-specific adaptation. Moreover, the Default baseline achieves the highest Imitativeness, suggesting that **Imitativeness alone is insufficient** and Adaptive-Imitativeness is needed for comprehensive model evaluation. In contrast, our proposed REPA prioritizes Adaptiveness and Adaptive-Imitativeness over Imitativeness, striking a balance that leads to the best overall performance for this task. Its task-specific design enables superior adaptive imitation and ensures consistent results across diverse datasets.

Additionally, our collected datasets have varying degrees of correspondence and variation. The RoleEE dataset reflects high correspondence (low variation), USNews medium, and Wikipedia relatively low (high variation). Results in Table 1 and 2 show better model performance with higher correspondence and lower variation. Overall, the results highlight REPA's superior ability to generate factual texts with high imitativeness and adaptiveness.

5.2 Ablation Study

To study how different components of REPA contribute to its overall performance, we conduct an ablation study with the following variations:

1) w/o Clarify-STM removes the Clarify component in PLAN, as well as the *short-term memory*

	R1↑	R2↑	RL↑	RLsum↑	Meteor↑	BLEU↑	Halluc↓	NLI-E↑	NLI-C↓	*I.*↑	*A.*↑	*A.-I.*↑
Full	**0.8112**	**0.7146**	**0.7600**	**0.7625**	**0.7368**	**0.6672**	**6.5714**	**0.7927**	**0.0439**	**4.16**	**3.90**	**3.93**
- C	0.8000	0.6958	0.7369	0.7395	0.7149	0.6341	6.7293	0.7826	0.0539	**4.16**	3.72	3.83
- O	0.5879	0.4574	0.5215	0.5218	0.5928	0.3415	25.3955	0.5847	0.0850	3.82	2.88	3.15
- F	0.8039	0.7121	0.7581	0.7571	0.7204	0.6543	7.0788	0.7481	0.0859	4.08	3.54	3.69
- R	0.7775	0.6634	0.7154	0.7167	0.7065	0.6037	6.5915	0.7898	0.0510	4.10	3.74	3.84
- S	0.7482	0.6365	0.6627	0.6679	0.6562	0.5770	8.2648	0.7135	0.0683	4.06	3.43	3.74

Table 3: Evaluation results across all metrics for ablation study on Wikipedia dataset. LLM denotes GPT-4 specifically. - C, - O, - F, - R, -S denote model variants w/o Clarify-STM, w/o Outline, w/o Refusal, w/o Revise-LTM, w/o Segment, respectively.

for retaining history processed input segments.

2) w/o Outline removes the Outline component in PLAN. Consequently, Calibrated-QA is also removed, and retrieval is based solely on the clarified input segment without any generated questions.

3) w/o Refusal removes confidence calibration in QA, resulting in no refusal in answering questions.

4) w/o Revise-LTM removes the Revise component in Write, as well as the *long-term memory* for storing history generated output.

5) w/o Segment removes text segmentation, and the input exemplar text was processed as a single block in the PLAN-THEN-ADAPT process, bypassing recurrent steps.

We evaluate these variants with the Wikipedia dataset on all metrics. As the GPT 4-based results shown in Table 3 (additional results on LLaMA 3-based variants are included in Appendix F), we find that the Outline (or PLAN) component significantly contributes to the overall model performance, demonstrating the effectiveness of using questions as a format of outlines for guiding factual adaptive-imitative generation. Additionally, the Refusal mechanism enhances the generation of factual content; removing it results in a decrease in factuality. Overall, the full REPA model exhibits the best performance across all metrics on the Wikipedia dataset, and each component of REPA plays a crucial role in improving its overall performance.

5.3 Human Evaluation of LLM Judge

Following the methodology of Zheng et al. (2024), we measured the agreement between the LLM-judge and human annotators, calculated as the probability that both parties would select the same model output from a randomly chosen pair. As shown in Table 4, there is **a strong correlation between LLM and human judgments**, with mean agreement scores of 79.0% for Imitativeness and 82.9% for Adaptiveness. These results demonstrate that our LLM-judge metrics align closely with human judgments, validating their reliability. Moreover, the LLM-judge's agreement rates were comparable to or exceeded inter-annotator agreement among humans, which was 79.2% for Imitativeness and 80.8% for Adaptiveness. Additional details are included in Appendix H.

Dataset	*I.* - w/o tie	*A.* - w/o tie
LLM-Human	79.0%	82.9%
Human-Human	79.2%	80.8%

Table 4: Agreements on Imitativeness (*I.*) and Adaptiveness (*A.*) metrics.

6 Conclusion and Discussion

In summary, we introduce a new, practical yet under-explored task: *Exemplar-Based Expository Text Generation*. To ensure *cross-topic consistency*, address *cross-topic variability*, and scale to long-form text, we propose **REPA** (RECURRENT PLAN-THEN-ADAPT), a model incorporating two memory structures–a short-term memory and a long-term memory. To address the limitations of existing evaluation metrics, we employ LLM-as-a-Judge to develop task-specific evaluators alongside established metrics. Extensive comparisons and ablation studies on three diverse, newly collected datasets demonstrate the effectiveness of our model across basic generation metrics, factuality metrics, and task-specific metrics, including *imitativeness*, *adaptiveness*, and *adaptive-imitativeness*.

Limitations

While our REPA model demonstrates promising results, several limitations remain. Firstly, our proposed *Exemplar-Based Expository Text Generation* task is explicitly designed to generate expository texts using **a high-quality exemplar from a similar topic**. The goal of our task is to "*write like the best*", which inherently assumes that a well-crafted and topically relevant exemplar is provided. The quality and topical alignment of the exemplar are fundamental premises of our approach, and our study is scoped around these assumptions. However, scenarios involving low-quality or dissimilar exemplars pose challenges to the model's performance, and future work might explore these scenarios. Additionally, although our recurrent prompting pipeline enhances performance, it may introduce inefficiencies compared to certain baseline methods. Moreover, the reliance on large language models (LLMs) such as GPT-4 and LLaMA 3 limits applicability in resource-constrained environments.

Future work could address these limitations in several ways. First, exploring methods for on-demand retrieval of relevant knowledge could mitigate the occasional inaccuracies generated by LLMs and improve overall efficiency. Second, incorporating multiple exemplars, rather than relying on a single one, may enhance generalization by broadening the diversity and scope of the output. Using exemplars from related topics could provide a more comprehensive perspective, enabling richer and more varied content generation.

Acknowledgments

This material is based upon work supported by the National Science Foundation IIS 16-19302 and IIS 16-33755, Zhejiang University ZJU Research 083650, IBM-Illinois Center for Cognitive Computing Systems Research (C3SR) and IBM-Illinois Discovery Accelerator Institute (IIDAI), grants from eBay and Microsoft Azure, UIUC OVCR CCIL Planning Grant 434S34, UIUC CSBS Small Grant 434C8U, and UIUC New Frontiers Initiative. Any opinions, findings, conclusions, or recommendations expressed in this publication are those of the author(s) and do not necessarily reflect the views of the funding agencies.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Rilwan Adewoyin, Ritabrata Dutta, and Yulan He. 2022. RSTGen: Imbuing fine-grained interpretable control into long-FormText generators. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1822–1835, Seattle, United States. Association for Computational Linguistics.

Yang Bai, Ziran Li, Ning Ding, Ying Shen, and Hai-Tao Zheng. 2021. Infobox-to-text generation with tree-like planning based attention network. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, pages 3773–3779.

Nishant Balepur, Jie Huang, and Kevin Chang. 2023. Expository text generation: Imitate, retrieve, paraphrase. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11896–11919, Singapore. Association for Computational Linguistics.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Rebekah Carter, Yenna Salamonson, Lucie M Ramjan, and Elizabeth Halcomb. 2018. Students use of exemplars to support academic writing in higher education: An integrative review. *Nurse education today*, 65:87–93.

Jun Chen. 2024. Exploring imitative learning in a blended efl writing class.

Albert Gatt and Emiel Krahmer. 2018. Survey of the state of the art in natural language generation: Core tasks, applications and evaluation. *Journal of Artificial Intelligence Research*, 61:65–170.

Seraphina Goldfarb-Tarrant, Tuhin Chakrabarty, Ralph Weischedel, and Nanyun Peng. 2020. Content planning for neural story generation with aristotelian rescoring. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4319–4338, Online. Association for Computational Linguistics.

Jian Guan, Fei Huang, Zhihao Zhao, Xiaoyan Zhu, and Minlie Huang. 2020. A knowledge-enhanced pretraining model for commonsense story generation. *Transactions of the Association for Computational Linguistics*, 8:93–108.

Jian Guan, Xiaoxi Mao, Changjie Fan, Zitao Liu, Wenbiao Ding, and Minlie Huang. 2021. Long text generation by modeling sentence-level and discourse-level coherence. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics* and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6379–6393, Online. Association for Computational Linguistics.

Zhe Hu, Hou Pong Chan, Jiachen Liu, Xinyan Xiao, Hua Wu, and Lifu Huang. 2022. PLANET: Dynamic content planning in autoregressive transformers for long-form text generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2288–2305, Dublin, Ireland. Association for Computational Linguistics.

Xinyu Hua, Zhe Hu, and Lu Wang. 2019. Argument generation with retrieval, planning, and realization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2661–2672, Florence, Italy. Association for Computational Linguistics.

Yilun Hua, Zhaoyuan Deng, and Kathleen McKeown. 2023. Improving long dialogue summarization with semantic graph representation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13851–13883, Toronto, Canada. Association for Computational Linguistics.

Fantine Huot, Joshua Maynez, Shashi Narayan, Reinald Kim Amplayo, Kuzman Ganchev, Annie Priyadarshini Louis, Anders Sandholm, Dipanjan Das, and Mirella Lapata. 2023. Text-blueprint: An interactive platform for plan-based conditional generation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 105–116.

Maor Ivgi, Uri Shaham, and Jonathan Berant. 2023. Efficient long-text understanding with short-text models. *Transactions of the Association for Computational Linguistics*, 11:284–299.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

Yucheng Jiang, Yijia Shao, Dekun Ma, Sina Semnani, and Monica Lam. 2024. Into the unknown unknowns: Engaged human learning through participation in language model agent conversations. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 9917–9955, Miami, Florida, USA. Association for Computational Linguistics.

Zhengbao Jiang, Jun Araki, Haibo Ding, and Graham Neubig. 2021. How can we know when language models know? on the calibration of language models for question answering. *Transactions of the Association for Computational Linguistics*, 9:962–977.

Yizhu Jiao, Sha Li, Yiqing Xie, Ming Zhong, Heng Ji, and Jiawei Han. 2022. Open-vocabulary argument role prediction for event extraction. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5404–5418, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Hongye Jin, Xiaotian Han, Jingfeng Yang, Zhimeng Jiang, Zirui Liu, Chia-Yuan Chang, Huiyuan Chen, and Xia Hu. 2024. Llm maybe longlm: Self-extend llm context window without tuning. *arXiv preprint arXiv:2401.01325*.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, et al. 2022. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics.

Abdullatif Köksal, Timo Schick, Anna Korhonen, and Hinrich Schütze. 2023. Longform: Optimizing instruction tuning for long text generation with corpus extraction. *arXiv preprint arXiv:2304.08460*.

Lilly Kumari, Usama Bin Shafqat, and Nikhil Sarda. Retrieval augmented generation for dialog modeling.

Gibbeum Lee, Volker Hartmann, Jongho Park, Dimitris Papailiopoulos, and Kangwook Lee. 2023. Prompted LLMs as chatbot modules for long open-domain conversation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4536–4554, Toronto, Canada. Association for Computational Linguistics.

Xiaobo Liang, Zecheng Tang, Juntao Li, and Min Zhang. 2023. Open-ended long text generation via masked language modeling. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 223–241, Toronto, Canada. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Teaching models to express their uncertainty in words. *arXiv preprint arXiv:2205.14334*.

Yuxiang Liu, Jie Huang, and Kevin Chang. 2023. Ask to the point: Open-domain entity-centric question generation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2703–2716, Singapore. Association for Computational Linguistics.

Junru Lu, Siyu An, Mingbao Lin, Gabriele Pergola, Yulan He, Di Yin, Xing Sun, and Yunsheng Wu. 2023. Memochat: Tuning llms to use memos for consistent long-range open-domain conversation. *arXiv preprint arXiv:2308.08239*.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. 2024. Self-refine: Iterative refinement with self-feedback. *Advances in Neural Information Processing Systems*, 36.

Ziming Mao, Chen Henry Wu, Ansong Ni, Yusen Zhang, Rui Zhang, Tao Yu, Budhaditya Deb, Chenguang Zhu, Ahmed Awadallah, and Dragomir Radev. 2022. DYLE: Dynamic latent extraction for abstractive long-input summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1687–1698, Dublin, Ireland. Association for Computational Linguistics.

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*, pages 12076–12100, Singapore. Association for Computational Linguistics.

Piotr Mirowski, Kory W Mathewson, Jaylen Pittman, and Richard Evans. 2023. Co-writing screenplays and theatre scripts with language models: Evaluation by industry professionals. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–34.

Amit Moryossef, Yoav Goldberg, and Ido Dagan. 2019. Step-by-step: Separating planning from realization in neural data-to-text generation. 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)*, pages 2267–2277, Minneapolis, Minnesota. Association for Computational Linguistics.

Shashi Narayan, Joshua Maynez, Reinald Kim Amplayo, Kuzman Ganchev, Annie Louis, Fantine Huot, Anders Sandholm, Dipanjan Das, and Mirella Lapata. 2023. Conditional generation with a question-answering blueprint. *Transactions of the Association for Computational Linguistics*, 11:974–996.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Ori Ram, Yoav Levine, Itay Dalmedigos, Dor Muhlgay, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. 2023a. In-context retrieval-augmented language models. *Transactions of the Association for Computational Linguistics*, 11:1316–1331.

Ori Ram, Yoav Levine, Itay Dalmedigos, Dor Muhlgay, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. 2023b. In-context retrieval-augmented language models. *Transactions of the Association for Computational Linguistics*, 11:1316–1331.

Vipula Rawte, Amit Sheth, and Amitava Das. 2023. A survey of hallucination in large foundation models. *arXiv preprint arXiv:2309.05922*.

Yijia Shao, Yucheng Jiang, Theodore Kanell, Peter Xu, Omar Khattab, and Monica Lam. 2024a. Assisting in writing Wikipedia-like articles from scratch with large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6252–6278, Mexico City, Mexico. Association for Computational Linguistics.

Yijia Shao, Yucheng Jiang, Theodore A Kanell, Peter Xu, Omar Khattab, and Monica S Lam. 2024b. Assisting in writing wikipedia-like articles from scratch with large language models. *arXiv preprint arXiv:2402.14207*.

Dinghan Shen, Asli Celikyilmaz, Yizhe Zhang, Liqun Chen, Xin Wang, Jianfeng Gao, and Lawrence Carin. 2019. Towards generating long and coherent text with multi-level latent variable models. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2079–2089, Florence, Italy. Association for Computational Linguistics.

Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Rich James, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. 2023. Replug: Retrieval-augmented black-box language models. *arXiv preprint arXiv:2301.12652*.

Aviv Slobodkin, Omer Goldman, Avi Caciularu, Ido Dagan, and Shauli Ravfogel. 2023. The curious case of hallucinatory (un)answerability: Finding truths in the hidden states of over-confident large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3607–3625, Singapore. Association for Computational Linguistics.

Xiaofei Sun, Zijun Sun, Yuxian Meng, Jiwei Li, and Chun Fan. 2022. Summarize, outline, and elaborate: Long-text generation via hierarchical supervision from extractive summaries. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6392–6402, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher Manning. 2023. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5433–5442, Singapore. Association for Computational Linguistics.

Chintalapalli Vijayakumar. 2024. Exemplification in student essay writing: A study of learner corpus of essay writing (lcew). *International Journal of Applied Linguistics*, 34(4):1514–1532.

Qingyue Wang, Liang Ding, Yanan Cao, Zhiliang Tian, Shi Wang, Dacheng Tao, and Li Guo. 2023. Recursively summarizing enables long-term dialogue memory in large language models. *arXiv preprint arXiv:2308.15022*.

Rosemary Wette. 2014. Teachers' practices in eap writing instruction: Use of models and modeling. *System*, 42:60–69.

Zhiwei Wu. 2019. Understanding students' mimicry, emulation and imitation of genre exemplars: An exploratory study. *English for Specific Purposes*, 54:127–138.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. 2023. Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms. *arXiv preprint arXiv:2306.13063*.

Han Xu, Xingyuan Wang, and Haipeng Chen. 2024. Towards real-time and personalized code generation. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 5568–5569.

Ikuya Yamada, Akari Asai, Jin Sakuma, Hiroyuki Shindo, Hideaki Takeda, Yoshiyasu Takefuji, and Yuji Matsumoto. 2020. Wikipedia2Vec: An efficient toolkit for learning and visualizing the embeddings of words and entities from Wikipedia. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 23–30, Online. Association for Computational Linguistics.

Kevin Yang, Dan Klein, Nanyun Peng, and Yuandong Tian. 2023a. DOC: Improving long story coherence with detailed outline control. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3378–3465, Toronto, Canada. Association for Computational Linguistics.

Kevin Yang, Yuandong Tian, Nanyun Peng, and Dan Klein. 2022. Re3: Generating longer stories with recursive reprompting and revision. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4393–4479, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yuqing Yang, Ethan Chern, Xipeng Qiu, Graham Neubig, and Pengfei Liu. 2023b. Alignment for honesty. *arXiv preprint arXiv:2312.07000*.

Wang You, Wenshan Wu, Yaobo Liang, Shaoguang Mao, Chenfei Wu, Maosong Cao, Yuzhe Cai, Yiduo Guo, Yan Xia, Furu Wei, et al. 2023. Eipe-text: Evaluation-guided iterative plan extraction for long-form narrative text generation. *arXiv preprint arXiv:2310.08185*.

Hanning Zhang, Shizhe Diao, Yong Lin, Yi R Fung, Qing Lian, Xingyao Wang, Yangyi Chen, Heng Ji, and Tong Zhang. 2023. R-tuning: Teaching large language models to refuse unknown questions. *arXiv preprint arXiv:2311.09677*.

Ruqing Zhang, Jiafeng Guo, Yixing Fan, Yanyan Lan, and Xueqi Cheng. 2019. Outline generation: Understanding the inherent content structure of documents. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 745–754.

Yusen Zhang, Ansong Ni, Ziming Mao, Chen Henry Wu, Chenguang Zhu, Budhaditya Deb, Ahmed Awadallah, Dragomir Radev, and Rui Zhang. 2022. Summ^n: A multi-stage summarization framework for long input dialogues and documents. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1592–1604, Dublin, Ireland. Association for Computational Linguistics.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2024. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36.

Wangchunshu Zhou, Yuchen Eleanor Jiang, Peng Cui, Tiannan Wang, Zhenxin Xiao, Yifan Hou, Ryan Cotterell, and Mrinmaya Sachan. 2023. Recurrentgpt: Interactive generation of (arbitrarily) long text. *arXiv preprint arXiv:2305.13304*.

A A Complete Example

We present a complete running example of the inputs and outputs for each module in a recurrent step in Figure 3.

B Prompts for Model

In our proposed REPA model, we use five prompts in each PLAN-THEN-ADAPT recurrent step to achieve fine-grained control of LLMs: the Clarify prompt (Figure 4), Outline prompt (Figure 5), Calibrated-QA prompt (Figure 6), Write prompt (Figure 7), and Summarize prompt (Figure 8).

Clarify Module

- **Inputs:**
  - Short-term memory (recent history input): "Belebeyevsky District: Belebeyevsky District (Russian: Белебе́евский райо́н, romanized: Belebeyevskiy rayon; Bashkir and Tatar: Бәләбәй районы, Bäläbäy rayonı; Chuvash: Пелепей районĕ, Pelepey rayonĕ) is an administrative and municipal district (raion), one of the fifty-four in the Republic of Bashkortostan, Russia."
  - Input text segment: "It is located in the west of the republic and borders with Tuymazinsky and Buzdyaksky Districts in the north, Davlekanovsky District in the east, Alsheyevsky District in the southeast, Bizhbulyaksky District in the south, and with Yermekeyevsky District in the west."
- **Outputs:**
  - Clarified text segment: "Belebeyevsky District is located in the west of the Republic of Bashkortostan and borders with Tuymazinsky and Buzdyaksky Districts in the north, Davlekanovsky District in the east, Alsheyevsky District in the southeast, Bizhbulyaksky District in the south, and with Yermekeyevsky District in the west."
  - Updated short-term memory: "Belebeyevsky District: Belebeyevsky District (Russian: Белебе́евский райо́н, romanized: Belebeyevskiy rayon; Bashkir and Tatar: Бәләбәй районы, Bäläbäy rayonı; Chuvash: Пелепей районĕ, Pelepey rayonĕ) is an administrative and municipal district (raion), one of the fifty-four in the Republic of Bashkortostan, Russia. Belebeyevsky District is located in the west of the Republic of Bashkortostan and borders with Tuymazinsky and Buzdyaksky Districts in the north, Davlekanovsky District in the east, Alsheyevsky District in the southeast, Bizhbulyaksky District in the south, and with Yermekeyevsky District in the west."

Outline Module

- **Inputs:**
  - Clarified text segment: "Belebeyevsky District is located in the west of the Republic of Bashkortostan and borders with Tuymazinsky and Buzdyaksky Districts in the north, Davlekanovsky District in the east, Alsheyevsky District in the southeast, Bizhbulyaksky District in the south, and with Yermekeyevsky District in the west."
  - Target topic: "Davlekanovsky District"
- **Outputs:**
  - Plans (a set of questions on target topic):
    - "Where is Davlekanovsky District located within the Republic of Bashkortostan?"
    - "Which districts border Davlekanovsky District to the north?"
    - "What district borders Davlekanovsky District to the east?"
    - "Which district is located to the southeast of Davlekanovsky District?"
    - "What district lies to the south of Davlekanovsky District?"
    - "Which district is situated to the west of Davlekanovsky District?"

Calibrated-QA Module

- **Inputs:**
  - Plans: "Where is Davlekanovsky District located within the Republic of Bashkortostan?", "Which districts border Davlekanovsky District to the north?", etc.
  - Retrieved Knowledge: per-topic retrieval with search engine and per-question retrieval with dense index
- **Outputs:**
  - Answers:
    - "Davlekanovsky District is located in the center of the Republic of Bashkortostan, Russia."
    - "Blagovarsky and Chishminsky Districts border Davlekanovsky District to the north."
    - "Aurgazinsky District borders Davlekanovsky District to the east."
    - "The Aurgazinsky District is located to the southeast of Davlekanovsky District."
    - "Alsheyevsky District lies to the south of Davlekanovsky District."
    - "Belebeyevsky District is situated to the west of Davlekanovsky District."

Write Module

- **Inputs:**
  - Answers: "Davlekanovsky District is located in the center of the Republic of Bashkortostan, Russia.", etc.
  - Long-term memory: "Davlekanovsky District is an administrative and municipal district in the Republic of Bashkortostan, Russia."
  - Input text segment: "It is located in the west of the republic and borders with Tuymazinsky and Buzdyaksky Districts in the north, Davlekanovsky District in the east, Alsheyevsky District in the southeast, Bizhbulyaksky District in the south, and with Yermekeyevsky District in the west."
- **Outputs:**
  - Output text segment: "It is located in the center of the Republic of Bashkortostan and is bordered by Blagovarsky and Chishminsky Districts to the north, Aurgazinsky District to the east and southeast, Alsheyevsky District to the south, and Belebeyevsky District to the west."
  - Updated long-term memory: "Davlekanovsky District is an administrative and municipal district in the Republic of Bashkortostan, Russia. It is situated in the center of Bashkortostan and shares borders with Blagovarsky and Chishminsky Districts to the north, Aurgazinsky District to the east and southeast, Alsheyevsky District to the south, and Belebeyevsky District to the west."

Figure 3: A complete running example.

```
Clarify the provided TEXT about the given TOPIC by substituting every pronoun
present with its corresponding antecedent, as identified from its preceding
context provided in SHORT TERM MEMORY. The antecedents should be discerned
based on their latest reference in SHORT TERM MEMORY. In addition, replace
the abbreviated forms of the TOPIC in the TEXT with their full names. Then,
return the revised version of TEXT that includes these updates.

TOPIC:
${SOURCE_TOPIC}

SHORT TERM MEMORY:
${SHORT_TERM_MEMORY}

TEXT:
${SEGMENT}

Respond in this format:

Clarified TEXT: <The revised TEXT.>
```

Figure 4: Prompt of the Clarify component in PLAN stage.

```
Generate topic-centric questions focusing on the given TOPIC from the
provided TEXT. Follow the steps below:
1. Decompose the TEXT into different facts of the given TOPIC without loss of
facts.
2. For each fact, formulate a topic-centric question centered on the given
TOPIC asking about the fact. Ensure that the TOPIC appears in the question.
3. Combine questions by eliminating duplicates and return the remaining
questions in a list.

TOPIC:
${SOURCE_TOPIC}

TEXT:
${CLARIFIED_SEGMENT}

Format your response as:

Q: <First question>
...
Q: <Last question>

If you cannot generate any question, return "${ERROR_TOKEN}".
```

Figure 5: Prompt of the Outline component in PLAN stage.

```
Generate a brief answer to the QUERY concerning the TOPIC, based on the
KNOWLEDGE provided. The answer should be a complete sentence with its context
in the QUERY. If the QUERY cannot be answered by the provided KNOWLEDGE,
provide your best guess for the QUERY. After answering, assess the
probability that your response is accurate, ranging from 0.0 (completely
uncertain) to 1.0 (completely certain). If you cannot give a guess, generate
"${ERROR_TOKEN}" with probability 0.0.

KNOWLEDGE:
${KNOWLEDGE}

TOPIC:
${TARGET_TOPIC}

QUERY:
${TARGET_QUERY}

Respond in this format:

Response Text: <Your answer to QUERY according to KNOWLEDGE, or your best
guess if necessary, or "${ERROR_TOKEN}". Note that the answer should be a
complete sentence with its context in the QUERY.>

Confidence Probability: <Your estimation of how likely your response is
correct, on a scale from 0.0 to 1.0.>
```

Figure 6: Prompt of the Calibrated-QA component in ADAPT stage.

```
You need to generate a SEGMENT from the given TEXT following three steps
below.
First, remove negative facts of TOPIC in the provided TEXT.
Next, re-write the TEXT centered on the given TOPIC to strictly align with
the EXEMPLAR's pronoun usage, content and structure. Eliminate any
unnecessary details in TEXT, refraining from adding new information (talking
points) beyond those in EXEMPLAR. Keep the revision accurate and relevant
about TOPIC according to TEXT, and you now get a DRAFT SEGMENT.
Finally, compare the DRAFT SEGMENT against the given MEMORY (a brief summary
of all previous segments). If the DRAFT SEGMENT contains exact content that
directly overlaps with the MEMORY, revise the SEGMENT by removing these
redundancies, and return the revised FINAL SEGMENT. Otherwise, return the
DRAFT SEGMENT as FINAL SEGMENT without revision.

TOPIC:
${TARGET_TOPIC}

EXEMPLAR:
${SEGMENT}

TEXT:
${TARGET_ANSWERS}

MEMORY:
${LONG_TERM_MEMORY}

Respond in this format:

DRAFT SEGMENT: <The draft segment after the first step>

FINAL SEGMENT: <The final segment after the second step>
```

Figure 7: Prompt of the Write component in ADAPT stage.

```
You are provided with a MEMORY (a brief summary of all previous segments), a
current SEGMENT, and a TOPIC (the focus of the MEMORY and the current
SEGMENT). MEMORY serves as a brief repository of key content from all
previous segments. Update MEMORY by integrating essential details from the
SEGMENT, ensuring it aligns with the existing MEMORY, and remains concise and
coherent. Keep the updated MEMORY focused on the TOPIC and within five
sentences. Finally, return the updated MEMORY.

TOPIC:
${TARGET_TOPIC}

SEGMENT:
${TARGET_SEGMENT}

MEMORY:
${LONG_TERM_MEMORY}
```

Figure 8: Prompt of the post-Write summarization step in ADAPT stage.

C Case Study on Longer Texts

Given the lack of extensive, high-quality, paired longer-text datasets, we evaluated our model's capacity for generating extended outputs using a manually curated example. Specifically, we tasked the model with generating a lengthy text on a target topic, "Beyoncé" (Table 6) based on an exemplar on "Taylor Swift" (Table 5). In future research, larger-scale evaluations with additional longer-text datasets could provide more comprehensive assessments.

D Details on Task-Specific LLM-Judge

D.1 Implementation Details

Recent research (Zheng et al., 2024) suggests that Large Language Models (LLMs) such as GPT-4 or LLaMA perform comparably to humans in evaluating model performance. Consequently, we leverage LLMs to evaluate *Imitativeness* and *Adaptiveness*. Specifically, we compare model outputs with exemplars for *Imitativeness* (prompt in Figure 9) and with both exemplars and ground truths for *Adaptiveness* (prompt in Figure 10), and to provide ratings from 1 to 5 for each aspect. Built upon **Imitativeness** and **Adaptiveness**, we propose **Adaptive-Imitativeness**, a composite score that measures the model's performance in handling both cross-topic consistency and variability, akin to the calculation of an F1 score.

D.2 Discussion on Known Limitations

There are known limitations of LLM-as-a-judge such as verbosity and self-enhancement biases, and we'd like to clarify in this section.

Verbosity bias means LLM judges favor longer, verbose responses compared to shorter alternatives. However, our proposed REPA, despite having the best performance, usually generates the shortest output length, as shown in Table 7 below. Therefore, REPA's performance is still convincing given the verbosity bias.

Self-enhancement bias implies LLM judges may favor the answers generated by themselves. However, we do not have comparisons between methods built on different backbone LLMs, as we compare all methods on GPT-4 (or on LLaMA 3) only. Therefore, self-enhancement bias does not influence our evaluation.

E Details on NLI-based Metrics

Inspired by FActScore (Min et al., 2023), we use NLI-based metrics to assess factuality by decomposing model outputs into sentence-level facts and classifying whether each fact entails or contradicts the ground truths, which serve as the knowledge source of the target topic given the context of source text's content. Specifically, we use the public HuggingFace checkpoint "*geckos/bart-fined-tuned-on-entailment-classification*". Its accuracy on SNLI corpus (Bowman et al., 2015) is 85.9% on training set and 86.1% on testing set, which is satisfactory for classifying entailment.

To validate the correlation of NLI-based metrics with human evaluations for hallucination/correctness assessment in our datasets, we conducted a human evaluation study. Specifically, we engaged three expert human annotators, independent of the paper, with experience in NLP and fact-checking. We randomly selected 50 task data samples from each dataset, totally 150 task data samples, each associated with two model outputs: LLM+Retr and RoM, to ensure a relatively balanced evaluation dataset.

For each model output, we decomposed it into sentence-level facts, resulting in 1032 facts (evaluation samples). Each fact was paired with its ground truth target text as the knowledge source. The three expert annotators were then asked to classify each fact as *Supported*, *Not-supported*, or *Irrelevant*, corresponding to *Entail*, *Contradict*, or *Neutral* in the NLI-based assessment.

The human evaluation results indicated 43% *Supported* and 46% *Not-supported* evaluation samples. The Fleiss's Kappa for inter-annotator agreement is 0.78, demonstrating high reliability. The accuracy of the NLI-based assessment is 83.7%, confirming its effectiveness for evaluating the correctness of model outputs.

F Additional Results and Discussions

Evaluation results for REPA and baselines built on LLaMA 3 are presented in Table 8 and Table 9, where the former covers basic and factuality metrics and the latter covers task-specific metrics. The results revealed that REPA consistently outperformed baselines in factuality and task-specific metrics. While some baselines demonstrated strong performance on basic generation metrics (e.g., R1, R2, RL, Meteor, BLEU) and imitativeness, they exhibited significantly higher hallucination rates

Taylor Alison Swift (born December 13, 1989) is an American singer-songwriter and director. Taylor Swift is regarded as an influential cultural figure of the 21st century. Taylor Swift is regarded as an influential cultural figure of the 21st century. Throughout her career, Taylor Swift has been recognized for her heartfelt lyrics and catchy melodies. She rose to fame following the release of her self-titled debut album in October 2006.

Starting her career as a solo artist, Taylor Alison Swift has achieved global superstardom, including winning the Grammy Award for album of the year for Midnights (2022), suggesting she is among the best-selling artists of all time. Taylor Swift's self-titled debut album was released in October 2006. She then followed with the U.S. number-one solo albums "Taylor Swift" (2006), "Fearless" (2008), and "Speak Now" (2011). After creating her own management company, 13 Management, Taylor Swift achieved critical acclaim for her self-titled debut album "Taylor Swift," which explored themes such as love, dreams, and personal experiences as a teenager.

Taylor Swift's most successful songs on the Billboard Hot 100 include her numerous hits that have defined her career. Outside of music, Taylor Swift has starred as an actress in films such as The Pink Panther (2006), Obsessed (2009), and The Lion King (2019). Taylor Swift has made a significant name for herself in the music industry. Her accolades include a record 32 Grammy Awards (including the 2010 Song of the Year), as well as 26 MTV Video Music Awards (including the 2014 Michael Jackson Video Vanguard Award) – all of which are more than any other artist in the music industry. She is known for her significant impact on contemporary music and her successful tours.

Taylor Alison Swift was born on December 13, 1989, at the hospital in West Reading, Pennsylvania to Andrea Gardner Swift (née Finlay), a mutual fund marketing executive, and Scott Kingsley Swift, a stockbroker for Merrill Lynch. Taylor Swift's mother, Andrea Gardner Swift, worked in mutual fund marketing, and her father, Scott Kingsley Swift, was a stockbroker. Taylor Swift is a distinguished artist with no familial ties to Giselle Knowles-Carter or Solange Knowles. Taylor Swift's immediate family includes her parents, Scott Kingsley Swift and Andrea Gardner Swift.

Taylor Swift's interest in music and performing was evident from a young age, showcasing her talents in a school talent show. Taylor Swift's voice captivates audiences with her heartfelt lyrics and catchy melodies, marking her as a significant figure in contemporary music. Her vocal abilities distinguish her as a significant figure in contemporary music.

Taylor Swift's music is generally pop and country, captivating audiences with her heartfelt lyrics and catchy melodies. Taylor Swift almost exclusively releases her songs in English. Taylor Swift's early career was centered on country music, showcasing her storytelling abilities and personal experiences through her songs. She then transitioned to pop music, marking a significant shift in her musical style and broadening her appeal to a global audience. She is tied with American lyricist Diane Warren at third with nine songwriting credits on number-one singles. The controversy surrounding Taylor Swift's songwriting credits began with interviews in which she attributed herself as the songwriter for songs in which she was a co-writer or for which her contributions were marginal.

Taylor Swift's global superstardom and history of winning prestigious awards, such as the Grammy for album of the year for Midnights in 2024, showcase her strong and dedicated fan base that likely contributes to her high sales figures. Taylor Swift's success has made her a cultural icon, captivating audiences with heartfelt lyrics and catchy melodies. She has become one of the most influential artists in contemporary music. She was the highest-paid musician of 2016, underscoring her significant influence in the music industry.

Writing for a leading entertainment outlet, it could be said that Taylor Swift has been a defining figure in music since 2010. Her approach to narrative songwriting and personal connection with audiences showcases the power of storytelling in music. Taylor Swift has been credited with reviving the album as an art form in an era dominated by singles and streaming.

Table 5: REPA's source text on "Taylor Swift".

Beyoncé Giselle Knowles-Carter (born September 4, 1981) is an American singer-songwriter and businesswoman. Beyoncé is regarded as an influential cultural figure of the 21st century. Throughout her career, Beyoncé has been recognized for her distinctive vocal range and live performances. She rose to fame in the late 1990s as the lead singer of Destiny's Child.

Starting her career in the music industry, Beyoncé formed the singing-rapping girl group Destiny's Child in 1990 with childhood friends. She has achieved global superstardom, recognized for her distinctive vocal range, live performances, and as an influential cultural figure of the 21st century. Her success extends to music, business, and a significant social media presence with millions of followers. It is reasonable to infer that she is among the best-selling artists of all time. She then followed with the U.S. number-one solo albums "B'Day" (2006) and "4" (2011). After achieving fame as the lead singer of Destiny's Child, Beyoncé embarked on a solo career with her self-titled debut album "Beyoncé."

Beyoncé's most successful songs on the Billboard Hot 100 include her numerous hits that have defined her career. Beyoncé has made a significant name for herself in the music industry. Her accolades include the 2010 Song of the Year Grammy Award for "Single Ladies (Put a Ring on It)." She is known for her significant impact on contemporary music and her successful tours.

Beyoncé Giselle Knowles-Carter was born on September 4, 1981, at the hospital in Houston, Texas to Tina Knowles, a businesswoman in the fashion industry, and Mathew Knowles, involved in her music career. Beyoncé's mother, Tina Knowles, worked in the fashion industry, and her father, Mathew Knowles, was involved in her music career. Beyoncé's immediate family includes her parents, Tina Knowles (née Beyincé) and Mathew Knowles.

Beyoncé's interest in music and performing was evident from a young age, showcasing her talents by forming the singing-rapping girl group Destiny's Child with childhood friends when she was nine years old. Beyoncé's voice captivates audiences with her distinctive vocal range and live performances, marking her as a significant figure in contemporary music. Her vocal abilities distinguish her as a significant figure in contemporary music.

Beyoncé's music is generally pop, R&B, and hip hop, captivating audiences with her heartfelt lyrics and catchy melodies. Beyoncé almost exclusively releases her songs in English. Beyoncé's early career focused on R&B and pop music, highlighting her storytelling skills and personal experiences in her songs. She transitioned to pop music, broadening her appeal to a global audience. She is tied with American lyricist Diane Warren at third with nine songwriting credits on number-one singles.

Beyoncé's history of winning prestigious awards, such as the Grammy Awards, showcases her global superstardom and strong, dedicated fan base that likely contributes to her high sales figures. Beyoncé's success has made her a cultural icon, captivating audiences with her distinctive vocal range and live performances. She has become one of the most influential artists in contemporary music. She was the highest-paid musician of 2016, underscoring her significant influence in the music industry.

Writing for a leading entertainment outlet, it could be said that Beyoncé has been a defining figure in music since 2010. Her approach to narrative songwriting and personal connection with audiences showcases the power of storytelling in music. Beyoncé has effectively navigated the shift towards streaming and singles, maintaining her relevance and success in the music industry.

Table 6: REPA's generated text for "Beyoncé".

```
Please act as an impartial judge and evaluate the quality of the text
generated by an AI writing assistant to the imitative writing task, which is
to generate a text on a target topic given an exemplar on a source topic.

Your evaluation should focus on the imitativeness of the generated text of AI
writing assistant, i.e, how well does it mimic the exemplar by retaining its
structure and cross-topic content consistency, and refraining from adding new
content.

Note that for content, you should consider only the talking points without
topic-specific details of these talking points.

Be as objective as possible. Provide your explanation, and rate the generated
text on a scale of 1 to 5 by strictly following this format: "[[rating]]",
for example: "Rating: [[3]]". Please first return the rating and then the
explanation in your response.

[The Start of Exemplar]
${SOURCE_TEXT}
[The End of Exemplar]

[The Start of Assistant's Writing]
${OUTPUT_TEXT}
[The End of Assistant's Writing]
```

Figure 9: Prompt of Imitativeness evaluation.

```
Please act as an impartial judge and evaluate the quality of the text
generated by an AI writing assistant to the imitative writing task, which is
to generate a text on a target topic given an exemplar on a source topic. You
will also be provided with a reference target text so that you can compare it
with the exemplar for understanding cross-topic variability.

Your evaluation should focus on the adaptiveness of the generated text of AI
writing assistant, i.e., how effectively does it adapt exemplar's content to
the target topic by handling cross-topic variability while retaining cross-
topic consistency.

Note that for content, you should consider only the talking points without
topic-specific details of these talking points. If the content appears in the
exemplar but doesn't appear in the reference target text, it means that such
content is unnecessary for the target topic and thus shouldn't appear in the
Assistant's text.

Begin your evaluation by comparing the exemplar and the reference text for
difference to understand cross-topic content variability, then comparing the
AI assistant's text with the reference target text to determine the
adaptiveness according to the definition above.

Be as objective as possible. Provide your explanation, and rate the generated
text on a scale of 1 to 5 by strictly following this format: "[[rating]]",
for example: "Rating: [[3]]". Please first return the rating and then the
explanation in your response.

[The Start of Exemplar]
${SOURCE_TEXT}
[The End of Exemplar]

[The Start of Reference Text]
${TARGET_TEXT}
[The End of Reference Text]

[The Start of Assistant's Writing]
${OUTPUT_TEXT}
[The End of Assistant's Writing]
```

Figure 10: Prompt of Adaptiveness evaluation.

and reduced adaptiveness. These findings reinforce the limitation of basic generation metrics for this task and highlight the importance of factuality and task-specific evaluation.

Though REPA exhibited minor instability in basic generation metrics with smaller backbones, it remained robust on critical metrics for the zero-shot task (e.g., NLI, A., A.-I.), demonstrating

Method	Length Ratio	Avg Length
RePA	0.9497	109.55
LLM	1.2076	140.75
LLM+Retr	1.0438	121.65
RoM	1.0928	127.15
RoM+Retr	1.1113	128.9
IRP	1.0481	121.35
Default	1.0754	110.74
Gold	1.0	116.0

Table 7: Output length comparison of REPA and baselines on Wikipedia dataset with GPT-4 backbone.

its robustness even with constrained computational resources. Additionally, larger LLMs consistently achieved better performance across both task-specific and basic generation metrics, further supporting the necessity of strong backbone LLMs for achieving optimal performance with REPA.

Additional ablation study results on LLaMA 3-based models are shown in Table 10.

Moreover, since our proposed REPA framework involves multiple intermediate steps, there is a risk of cascading errors. Specifically, our observations show that when the Clarify component makes mistakes–such as incorrectly identifying the antecedents of pronouns–the Outline component is likely to follow suit, producing inaccurate topic-centric outlines and misguiding the subsequent ADAPT stage. Additionally, the QA component tends to encounter challenges due to the limitations of retrievers for topic retrieval and the DPR model for question-based retrieval, which can result in retrieving irrelevant information. This, in turn, affects the Write component, leading to omissions of crucial facts and an incomplete final output.

Although multi-step pipelines inherently pose a risk of error propagation, we have implemented several mechanisms to minimize this risk. For example, in the final "*Write*" step, which follows the "*Calibrated-QA*" process in **Adapt**, the original input segment is included in the prompt to regularize the generated output (Figure 7). This design helps mitigate potential errors from earlier stages. Furthermore, as demonstrated in our ablation study (Table 3, 10), removing any step in the pipeline degrades performance. This finding underscores that every stage contributes positively to the overall efficacy of REPA. Therefore, while the risk of error propagation is a theoretical consideration, its practical impact on performance is minimal.

G Case Study

We further conduct case studies to examine the outputs from different models compared with target text and show an example in Table 11. We find that both LLM and RollingLLM achieve good imitativeness, as they cover all talking points from the source text. However, these “adapted facts" are not correct for the target topic, indicating that both models struggle with adaptive imitation – failing to generate well-adapted content that is relevant and factual to the target topic.

For LLM+Retr and RoM+Retr, which are equipped with retrieval from the same knowledge sources as described in Section 4, the factuality improves. However, there are still facts that are incorrect for the target topic. This shows that adaptiveness remains an issue, as these models fail to retrieve the best knowledge and correctly incorporate it into the generated texts.

In contrast, our proposed model can generate text that is both imitative of the source text and perfectly adapted to the target topic. It takes into account cross-topic consistency and variability, demonstrating that our proposed model significantly improves upon the baselines for our task. This illustrates the effectiveness of our approach in generating content that is not only imitative to the source text but also well-suited to the target topic.

H Human Evaluation on LLM Judge

To assess the effectiveness of our proposed LLM-as-a-Judge metrics, we conducted a human evaluation study by comparing LLM-judged outputs with human judgements. We recruited three expert annotators, all graduate students with specialized knowledge in NLP, to evaluate model outputs using the same instructions given to the LLMs (Figure 9 and 10). The study involved 50 randomly selected samples from each of three datasets (a total of 150 samples). Each sample comprised task inputs paired with outputs from nine models, including our proposed RePA model and eight baselines described in Section 4.2, with GPT-4 used for LLM-based baselines. This process resulted in 1,350 outputs evaluated per LLM-judge metric. Each model outputs was evaluated based on two criteria: Imitativeness (Figure 9) and Adaptiveness (Figure 10). All three annotators independently assessed each output. We then follow (Zheng et al., 2024) to first convert the single-answer grading into pairwise comparison results (5.4k votes), then calculate the probabil-

Datasets	Models	R1↑	R2↑	RL↑	RLsum↑	Meteor↑	BLEU↑	Halluc↓	NLI-E↑	NLI-C↓
Wikipedia	*LLaMA 3 70B*									
	RePA	**0.8164**	0.6937	0.7727	**0.7736**	0.7888	**0.6598**	**7.5361**	**0.7992**	**0.0741**
	LLM	0.8158	0.6912	**0.7733**	0.7709	**0.8161**	0.6468	11.6726	0.4181	0.4352
	LLM+Retr	0.8038	**0.6997**	0.7693	0.7720	0.7740	0.6531	10.4543	0.6108	0.2811
	RoM	0.7983	0.6690	0.7566	0.7537	0.8024	0.6289	12.3465	0.4066	0.4669
	RoM+Retr	0.7843	0.6668	0.7492	0.7505	0.7645	0.6001	10.7982	0.5678	0.3182
	LLaMA 3 8B									
	RePA	0.6105	0.4489	0.5295	0.5253	0.5196	0.3344	13.6244	**0.7509**	**0.0963**
	LLM	**0.7329**	0.5870	**0.6916**	**0.6899**	**0.7098**	**0.5729**	15.3616	0.3950	0.4766
	LLM+Retr	0.7073	**0.6079**	0.6849	0.6809	0.6560	0.5606	13.0367	0.5580	0.3434
	RoM	0.5711	0.4280	0.5316	0.5309	0.4617	0.3540	**12.7861**	0.3190	0.5181
	RoM+Retr	0.6286	0.5119	0.5915	0.5830	0.5624	0.4634	15.6738	0.5835	0.2497
RoleEE	*LLaMA 3 70B*									
	RePA	0.9178	**0.8730**	**0.9079**	**0.9084**	0.9211	**0.7873**	**4.5693**	**0.9022**	**0.0245**
	LLM	0.8585	0.7586	0.8421	0.8421	0.8820	0.6936	11.1463	0.2957	0.6443
	LLM+Retr	0.8859	0.8058	0.8541	0.8546	0.8704	0.6932	7.0544	0.7787	0.1390
	RoM	0.8427	0.7367	0.8265	0.8261	0.8678	0.6725	11.6951	0.2253	0.7393
	RoM+Retr	**0.9223**	0.8724	0.9070	0.9063	**0.9259**	0.7817	6.2451	0.7253	0.2090
	LLaMA 3 8B									
	RePA	0.6063	0.5181	0.5236	0.5236	0.7112	0.3461	**9.9206**	**0.8366**	**0.0496**
	LLM	**0.8137**	0.6913	**0.7975**	**0.7990**	**0.8340**	0.6225	14.0902	0.1467	0.8070
	LLM+Retr	0.7807	**0.6966**	0.7654	0.7664	0.7629	**0.6311**	12.0810	0.6970	0.2127
	RoM	0.7950	0.6762	0.7759	0.7778	0.8028	0.6178	13.5541	0.1313	0.8290
	RoM+Retr	0.7374	0.6556	0.7172	0.7184	0.7223	0.5381	13.7997	0.6190	0.1958
USNews	*LLaMA 3 70B*									
	RePA	**0.8846**	**0.8268**	**0.8543**	**0.8547**	**0.9062**	**0.7759**	**5.8306**	**0.7038**	**0.0282**
	LLM	0.8296	0.7404	0.8062	0.8073	0.8599	0.7056	11.8293	0.3527	0.5633
	LLM+Retr	0.7888	0.6888	0.7676	0.7691	0.8040	0.6136	13.8425	0.4020	0.4473
	RoM	0.7818	0.6885	0.7611	0.7618	0.8446	0.6395	14.5091	0.3413	0.5153
	RoM+Retr	0.7765	0.6791	0.7572	0.7579	0.8089	0.6239	13.4760	0.3246	0.4407
	LLaMA 3 8B									
	RePA	0.6520	0.5296	0.6122	0.6119	0.7223	0.3832	16.9316	**0.6519**	**0.0512**
	LLM	**0.8011**	**0.7040**	**0.7874**	**0.7880**	**0.8389**	**0.6309**	**12.5432**	0.3480	0.5593
	LLM+Retr	0.7387	0.6392	0.7215	0.7215	0.7497	0.5543	13.5612	0.4338	0.4046
	RoM	0.5259	0.4243	0.4933	0.4926	0.4806	0.3956	16.7682	0.3050	0.3474
	RoM+Retr	0.6420	0.5242	0.6147	0.6116	0.6751	0.4182	20.7881	0.3146	0.3481

Table 8: Evaluation results on basic and factuality metrics. LLM denotes LLaMA 3.

ity of both–LLM-judge and a randomly selected human judge, or two randomly selected human judges–agreeing on a randomly selected pairwise comparison, to calculate LLM-human agreement or human-human agreement. Additional results are shown in Table 12.

I Practical Application Latency and Cost

To estimate latency and costs, we used a subset of the Wikipedia dataset, and calculated latency in terms of API calls and the cost per output token, based on x=$15/1M tokens (OpenAI GPT-4 pricing). The results are summarized in Table 13.

Datasets	Models	*I.*↑	*A.*↑	*A.-I.*↑
Wikipedia	*LLaMA 3 70B*			
	REPA	4.14	**3.70**	**3.80**
	LLM	**4.72**	2.60	3.25
	LLM+Retr	4.22	2.74	3.15
	RoM	4.60	2.38	3.07
	RoM+Retr	4.16	2.58	3.06
	LLaMA 3 8B			
	REPA	3.88	**2.72**	**3.06**
	LLM	**4.74**	2.00	2.71
	LLM+Retr	4.12	2.42	2.89
	RoM	3.44	1.68	2.18
	RoM+Retr	3.78	1.80	2.37
RoleEE	*LLaMA 3 70B*			
	REPA	4.76	**4.58**	**4.64**
	LLM	4.78	2.56	3.12
	LLM+Retr	4.64	3.86	4.03
	RoM	4.84	1.94	2.61
	RoM+Retr	4.62	3.68	3.93
	LLaMA 3 8B			
	REPA	3.96	**3.22**	**3.38**
	LLM	**4.90**	1.92	2.63
	LLM+Retr	4.54	2.52	2.87
	RoM	4.48	1.50	2.17
	RoM+Retr	4.12	2.96	3.27
USNews	*LLaMA 3 70B*			
	REPA	4.32	**4.22**	**4.22**
	LLM	4.24	2.58	3.10
	LLM+Retr	4.22	2.36	2.94
	RoM	4.68	2.58	3.22
	RoM+Retr	4.10	2.16	2.80
	LLaMA 3 8B			
	REPA	4.08	**3.64**	**3.77**
	LLM	**4.42**	2.32	2.94
	LLM+Retr	4.14	1.88	2.52
	RoM	3.08	1.56	2.04
	RoM+Retr	3.92	1.94	2.59

Table 9: Evaluation results on task-specific metrics. LLM denotes LLaMA 3. ***I.*** denotes *Imitativeness*, ***A.*** denotes *Adaptiveness*, and ***A.-I.*** denotes *Adaptive-Imitativeness*.

	R1↑	R2↑	RL↑	RLsum↑	Meteor↑	BLEU↑	Halluc↓	↑NLI-E	↓NLI-C	*I.*↑	*A.*↑	*A.-I.*↑
Full	**0.8164**	**0.6937**	**0.7727**	**0.7736**	**0.7888**	**0.6598**	**7.5361**	**0.7992**	**0.0741**	4.14	**3.70**	**3.7996**
- C	0.7945	0.6806	0.7409	0.7411	0.6997	0.6057	8.6106	0.7416	0.0974	**4.18**	3.56	3.7403
- O	0.7489	0.6375	0.6932	0.6969	0.7167	0.5573	10.3022	0.5971	0.2569	4.04	2.46	2.9267
- F	0.7918	0.6640	0.7239	0.7263	0.7010	0.5986	8.7846	0.7039	0.1442	4.14	3.28	3.5330
- R	0.7391	0.5958	0.6708	0.6746	0.6445	0.5191	8.3785	0.7658	0.0916	4.06	3.68	3.7758

Table 10: Evaluation results across all metrics for ablation study on Wikipedia dataset. LLM denotes LLaMA 3 70B specifically. ***I.*** denotes *Imitativeness*, ***A.*** denotes *Adaptiveness*, and ***A.-I.*** denotes *Adaptive-Imitativeness*. - C, - O, - F, - R denote model variants w/o Clarify, w/o Outline, w/o Refusal, w/o Revise, respectively.

Type	Text
Source	Belebeyevsky District (Russian: Белебеéевский райоéн, romanized: Belebeyevskiy rayon; Bashkir and Tatar: Бәләбәй районы, Bäläbäy rayonı; Chuvash: Пелепей районĕ, Pelepey rayonĕ) is an administrative and municipal district (raion), one of the fifty-four in the Republic of Bashkortostan, Russia. It is located in the west of the republic and borders with Tuymazinsky and Buzdyaksky Districts in the north, Davlekanovsky District in the east, Alsheyevsky District in the southeast, Bizhbulyaksky District in the south, and with Yermekeyevsky District in the west. The area of the district is 1,911.2 square kilometers (737.9 sq mi). Its administrative center is the town of Belebey (which is not administratively a part of the district). As of the 2010 Census, the total population of the district was 41,708.
Target	Davlekanovsky District (Russian: Давлекаéновский райоéн; Bashkir: Дәүләкән районы, Däwläkän rayonı) is an administrative and municipal district (raion), one of the fifty-four in the Republic of Bashkortostan, Russia. It is located in the center of the republic and borders with Blagovarsky and Chishminsky Districts in the north, Karmaskalinsky District in the northeast, Aurgazinsky District in the east, Alsheyevsky District in the south, Belebeyevsky District in the west, and with Buzdyaksky District in the northwest. The area of the district is 1,907 square kilometers (736 sq mi). Its administrative center is the town of Davlekanovo (which is not administratively a part of the district). As of the 2010 Census, the total population of the district was 18,392.
REPA	Davlekanovsky District (Russian: Давлекаéновский райоéн, romanized: Davlekanovsky rayon; Bashkir: Дәүләкән районы, Däwläkän rayonı) is an administrative and municipal district (raion), one of the fifty-four in the Republic of Bashkortostan, Russia. It is located in the center of the Republic of Bashkortostan and is bordered by Blagovarsky and Chishminsky Districts to the north, Aurgazinsky District to the east and southeast, Alsheyevsky District to the south, and Belebeyevsky District to the west. The area of the district is 1,907 square kilometers (736 sq mi). Its administrative center is the town of Davlekanovo. As of the 2010 Census, the total population of the district was 18,392.
LLM	Davlekanovsky District (Russian: Давлекановский райоéн, romanized: Davlekanovskiy rayon; Bashkir and Tatar: Дәүләкән районы, Däwläkän rayonı; Chuvash: Давлекан районĕ, Davlekan rayonĕ) is an administrative and municipal district (raion), one of the fifty-four in the Republic of Bashkortostan, Russia. It is situated in the southeast of the republic and shares borders with Belebeyevsky and Alsheyevsky Districts in the west, Sterlitamaksky District in the north, Ishimbaysky District in the northeast, Gafuriysky District in the east, and with Kugarchinsky District in the south. The district spans an area of 1,540 square kilometers (594.6 sq mi). The administrative center of the district is the town of Davlekanovo (which is not administratively a part of the district). According to the 2010 Census, the total population of the district was 36,950.
LLM+Retr	Davlekanovsky District (Russian: Давлекановский район, romanized: Davlekanovskiy rayon; Bashkir and Tatar: Дәүләкән районы, Däwläkän rayonı; Chuvash: Давлекан районĕ, Davlekan rayonĕ) is an administrative and municipal district (raion), one of the fifty-four in the Republic of Bashkortostan, Russia. It is located in the center of the republic and borders with Blagovarsky and Chishminsky Districts in the north, Karmaskalinsky District in the northeast, Aurgazinsky District in the east, Alsheyevsky District in the south, Belebeyevsky District in the west, and with Yermekeyevsky District in the northwest. The area of the district is 2,300 square kilometers (888 sq mi). Its administrative center is the town of Davlekanovo (which is not administratively a part of the district). As of the 2010 Census, the total population of the district was 46,500.
RoM	Davlekanovsky District (Russian: Давлекановский райоéн, romanized: Davlekanovskiy rayon; Bashkir: Дәүләкән районы, Däwläkän rayonı; Tatar: Дәүләкан районы, Däwläkan rayonı) is an administrative and municipal district (raion), one of the fifty-four in the Republic of Bashkortostan, Russia. It is situated in the eastern part of the republic and shares borders with Belebeyevsky and Alsheyevsky Districts in the west, Tuymazinsky District in the north, Karmaskalinsky District in the northeast, Blagovarsky District in the east, and with Chekmagushevsky District in the south. The area of the district is 1,502 square kilometers (580 sq mi). Its administrative center is the town of Davlekanovo (which is not administratively a part of the district). As of the 2010 Census, the total population of the district was 46,857.
RoM+Retr	Davlekanovsky District (Russian: Давлекановский район, romanized: Davlekanovskiy rayon; Bashkir: Дәүләкән районы, Däwläkän rayonı) is an administrative and municipal district (raion), one of the fifty-four in the Republic of Bashkortostan, Russia. It is located in the center of the republic and borders with Blagovarsky and Chishminsky Districts in the north, Karmaskalinsky District in the northeast, Aurgazinsky District in the east, Alsheyevsky District in the south, Belebeyevsky District in the west, and... The area of the district is 2,300 square kilometers (888 sq mi). Its administrative center is the town of Davlekanovo (which is not administratively a part of the district). As of the 2010 Census, the total population of the district was 24,073.

Table 11: Case study of a Wikipedia example comparing proposed model with GPT-4 related baselines.

Dataset	***I.* - w/ tie**	***I.* - w/o tie**	***A.* - w/ tie**	***A.* - w/o tie**
Wikipedia	57.3%	78.0%	62.1%	82.7%
RoleEE	58.7%	78.3%	59.7%	81.3%
USNews	58.6%	80.7%	62.7%	84.8%
Mean	58.2%	**79.0%**	61.5%	**82.9%**

Table 12: Agreement between LLM-judge and human-judge on Imitativeness (*I.*) and Adaptiveness (*A.*) metrics, including w/ and w/o tie votes for calculating agreement.

Methods	Mean API calls (times)	Costs per output token (x=$15/1M)
RePA	39	50.1
LLM	1	1.5
LLM+Retr	1	2.7
RoM	5	1.7
RoM+Retr	5	8.2

Table 13: Model latency and costs.