Analysis of Plan-based Retrieval for Grounded Text Generation

Ameya Godbole◇†∗, **Nicholas Monath**♠∗, **Seungyeon Kim**♣, **Ankit Singh Rawat**♣
Andrew McCallum♠, **Manzil Zaheer**♠
◇University of Southern California   ♠Google DeepMind   ♣Google Research
ameyagod@usc.edu   nmonath@google.com

Abstract

In text generation, hallucinations refer to the generation of seemingly coherent text that contradicts established knowledge. One compelling hypothesis is that hallucinations occur when a language model is given a generation task outside its parametric knowledge (due to rarity, recency, domain, etc.). A common strategy to address this limitation is to infuse the language models with retrieval mechanisms, providing the model with relevant knowledge for the task. In this paper, we leverage the planning capabilities of instruction-tuned LLMs and analyze how planning can be used to guide retrieval to further reduce the frequency of hallucinations. We empirically evaluate several variations of our proposed approach on long-form text generation tasks. By improving the coverage of relevant facts, plan-guided retrieval and generation can produce more informative responses while providing a higher rate of attribution to source documents.

1 Introduction

Large, parametric language models (LLMs) provide highly fluent text for many applications such as summarization, dialogue, and translation (Devlin et al., 2019; Brown et al., 2020; Thoppilan et al., 2022; Chowdhery et al., 2024; Anil et al., 2023, inter alia). Among the errors made by these models, producing generations with factual and/or grounding errors, often referred to as *hallucinations*, limit the broader applicability and capability of language models (Gao et al., 2023a; Manakul et al., 2023; Min et al., 2023; Ji et al., 2023a; Peng et al., 2023). Hallucinations differ from other kinds of errors in that the generated text is syntactically correct and semantically plausible. These hallucinations are generations that, were they factually accurate, would be satisfactory model output.

As an example of such hallucinations, we consider the following generation from the Falcon 180B model demo[1] for the query *write a bio about the author lorrie moore.*

> lorrie moore is an american writer **born in glasgow, kentucky** in 1957. she graduated from saint lawrence university in canton, new york in 1978, and earned her mfa from the **university of wisconsin-madison** in 1982. moore has written numerous works of fiction, including short story collections such as self help, like life, and birds of america, as well as novels like anagrams, who will run the frog hospital?, and a gate at the stairs. moore has won many awards, including **the national book critics circle award for fiction in 1990**, the pen/malamud award for excellence in short fiction in **1994**, and the rea award for the short story in **2018**. her writing style is often described as sharp, funny, and introspective, with themes centered around family relationships, love, loss, and the complexities of modern life.

While the model accurately generates the book titles, it hallucinates many facts **(red, bold-faced)** about Moore such as the place of her birth (she was born in Glens Falls, New York), the university of her MFA (she received an MFA from Cornell University). The model also confuses all of the awards given to the author (incorrect years, award names, missing awards).

Recent work on detecting and mitigating such hallucinations has included modeling uncertainty (Lee et al., 2024), constrained decoding (Lu et al., 2022), specially trained or auxiliary model-based decoding (Zhou et al., 2023; Roit et al., 2023), question-answering planning (Narayan et al., 2023; Huot et al., 2023), and retrieval-based verification of claims (Gao et al., 2023a). These methods demonstrate success by leveraging the model's ability to accurately produce factual information via

---

†Work done as a Student Researcher at Google.
∗Equal contribution.

[1]https://huggingface.co/spaces/tiiuae/falcon-180b-demo

its parameters or through text provided in-context. However, understanding and analyzing where and why such hallucinations appear and why providing evidence in-context can successfully mitigate hallucinations has only begun to be studied recently (Li et al., 2023b; Das et al., 2022; Sadat et al., 2023).

In this work, we study how planning can be used to guide retrieval to improve factual text generation. We hypothesize that language models hallucinate when they are required to generate certain facts based on the prompt but do not have the information either memorized or in their context. Thus, we hypothesize that with specific, comprehensive facts in-context can mitigate hallucinations. Our investigation considers the following questions:

Q1. How to effectively discover the comprehensive collection of facts needed to generate text about a particular subject? (§3.2,§3.3, Table 1, 2, 7)
Finding: We observe that generating search queries based on the LLM plans lets the retrieval system gather fine-grained facts that the LLM needs to write its final response.

Q2. How can we effectively retrieve and represent these needed facts? (§3.3, §4.2, Table 3, 4 5)
Finding: We observe that the groundedness of the final model response is heavily influenced by the retrieved information and not as much by the wording of the prompts. We also show that including unanswerable search queries (and explicitly marking them as unanswerable) is influential in reducing the generation of ungrounded outputs.

We demonstrate the generality of the approach in writing grounded text by evaluating model responses in two domains: writing biographies and writing event descriptions. In addition to entities and events that may have a significant internet presence, we specifically consider 2 settings that would challenge LLMs: (1) writing biographies for people (researchers) who do not have Wikipedia profiles (long-tail, low-frequency entities) and (2) writing about current news events (outside of the parametric knowledge of the LLMs).

2 Related Work

The focus of our paper is to provide analysis as to how retrieval augmented language models can reduce hallucinations by conditioning on relevant facts in-context. This work abuts many related areas in text generation (factuality-focused and otherwise), in-context learning, retrieval, and more. Refer to other related work in App A.3.
Verifying Attribution. Rashkin et al. (2023) in-

troduce formal definitions of what it means for a piece of text to be attributable to a given source, referred to as *AIS* (Attributable to Identified Sources). This definition has been used to design datasets and models to automatically predict whether a particular source document supports a claim. Verification models have considered a variety of formulations such as question-answering based (Honovich et al., 2021), lexical alignment based (Goyal and Durrett, 2020) and NLI-based (Honovich et al., 2022; Gekhman et al., 2023). Among these, NLI-based models are a commonly used automated AIS metrics to measure whether a particular output is grounded in provided sources (Bohnet et al., 2023; Gao et al., 2023a, inter alia).

By incorporating web or corpus search, the AIS models can be used to check whether a particular claim is true with respect to general world knowledge (Min et al., 2023; Peng et al., 2023). FactScore (Min et al., 2023) prompts an LLM to break down claims into atomic facts to be verified. We use a simplified version of the FactScore metric by compute sentence-level entailment scores. These search-enabled verification models have enabled the study of how language models can generate attributed text along with citations and study accuracy of such deployed approaches (Liu et al., 2023; Gao et al., 2023b).

Retrieval for Generating Attributable Text. It is well known that language models store a large amount of knowledge in their parameters (Petroni et al., 2019, 2020). Retrieving evidence documents allows LLMs to synthesize text about facts outside their parametric knowledge (Mallen et al., 2023), including using up-to-date (Zhu et al., 2020; Vu et al., 2024) and proprietary (Min et al., 2024) information. Past work has incorporated retrieval for post-hoc verification and editing (Gao et al., 2023a; Chen et al., 2023a). Other approaches incorporate search results throughout the generation process either based on heuristics (Trivedi et al., 2023; Amplayo et al., 2023; Jiang et al., 2023b; Press et al., 2023, inter alia) or based on trigger tokens generated by the model (Asai et al., 2024).

Planning of Long-form Generation. Past work has demonstrated that planning (outlining the output) before generation can help to improve the factuality and quality of the model output. QA blueprint-based methods (Narayan et al., 2023; Huot et al., 2023) use question-answer pairs to change the LLM distribution over output text to make summarization models more grounded. Huot et al. (2024) extend this approach to cross-lingual
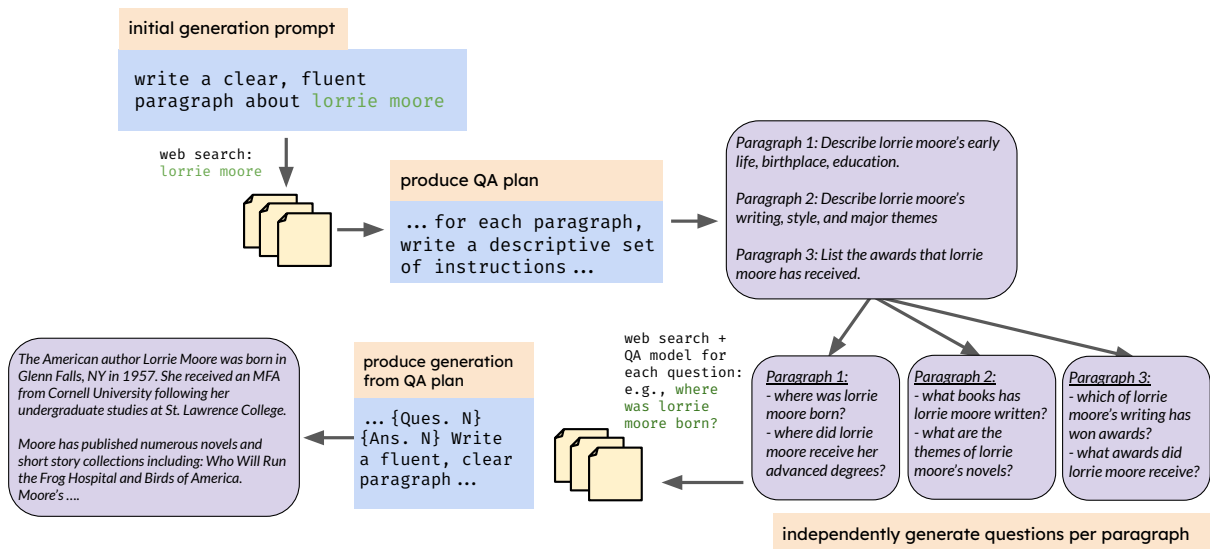
Figure 1: **Summary of Planning and Retrieval used to generate text.** Given an initial prompt, a plan is first generated that outlines the segments to be written. Next, search queries are generated for each segment which are then used for fine-grained retrieval retrieval of source documents. The final response is generated conditioned on the plan, the queries and the retrieved documents.

summarization. Akash et al. (2023) use a trained model to generate a keyword plan before generating a each new sentence and use the plan to inform search. **Our contribution:** We study whether zero-shot LLM generated plans can be used to retrieve relevant information beyond simple search and how the plans along with the new, diverse search results can guide the final generation to be more grounded. Concurrently with this work, Shao et al. (2024) also demonstrate the importance of pre-writing (outlining, iterative search and outline refinement) for long-form expository writing, further supporting the findings of our work.

3 Empirical Study

Our goal is to understand the effect of providing more complete contextual information to language models on hallucination in text generation. We compare different strategies for performing retrieval and placing facts into context.

We consider the task of generating descriptive text about an `entity`. This task effectively constitutes writing a bio about an individual (such as `author Lorrie Moore` in the previous example) or a summary about an event. This text will then be verified against web sources for attribution using automated methods (Honovich et al., 2022).

3.1 Direct Generation without Retrieval

The simplest approach is to directly prompt the LLM to generate text about the given `entity`. This approach relies on information known by the parametric language model without any additional re-

trieval. We provide examples of the prompts used in §A.2.1, which use entity name and possibly a disambiguating attribute (occupation, location, etc.).

3.2 Incorporating Retrieval

Towards providing the relevant facts needed for the LLM to write about a particular `entity`, we simply perform web search with the `entity` name as the query and provide the search results (titles and text) in the model context to allow retrieval augmented generation (Gao et al., 2023b) (see § A.2.2).

3.3 Planning and Blueprints

Past work has demonstrated that LLMs can generate reasoning chains to improve their task performance (Wei et al., 2024, interalia). Additionally, Narayan et al. (2023) have shown that question-answer based plans allow models to generate more faithful text. We combine these abilities of LLMs to guide retrieval and provide the information that the model may need to generate a response without (or with fewer) hallucinations.

We first prompt the model to write a list describing what paragraphs should be produced (see §A.2.3). Each item in this outline provides short sentence or two about the content for each paragraph. The outline is conditioned on the initial search results (using a query of the `entity` name) to guide the LLM outline to generate targeted, entity-specific outlines rather than generic plans.

Next, we generate search queries from the paragraph descriptions. We take the content description for each paragraph and the initial search results

13103

as context and generate web search queries (see §A.2.4). These queries allow the search engine to gather information about specific aspects of the entity. The result of this prompt is a list of questions such as those shown in Figure 1 for each paragraph. These questions will be used to gather additional web search results.

We consider two variants for incorporating these newly gathered web-search results. The first, simply places the new results in context. The second uses a separate off-the-shelf span-selection QA model to provide answers for each question. We use a confidence threshold to decide how many answers to include in the prompt for the final generation. If there are no answers that the QA model is confident about, we keep the question but mark the answer as 'unanswerable'. We combine the questions for all paragraphs together to form the prompt we use to generate the final output (see §A.2.5).

4 Empirical Analysis

We will measure quantitatively the ability of models to provide attributed generations. We will analyze how producing attributed text is affected by the way in which evidence is retrieved and used. In particular, we aim to investigate the following:

Q1: Does providing more comprehensive collections of facts yield more attributable generations?

    §4.4 Observations: 1-3

Q2: Can planning and additional web-searches improve the collection of facts used for generations?

    §4.4 Observations: 4, 7, 8

Q3: Which representations of facts in-context are most effective for producing grounded texts?

    §4.4 Observations: 5, 6

4.1 Datasets

As described in §3.1, the task we consider is writing a biography of an entity or summary of an event. We select this task because it likely necessitates gathering of knowledge more so than tasks such as summarization in which most (if not all) of the required information is provided. These datasets used for the task are meant to cover both entities and events from the head and the tail of the parametric knowledge of LLMs.

The setting we consider is one where the system is presented only with an entity or event name and (possibly) an adjective for entity disambiguation (e.g., Gerhard Fischer (inventor)). Since our evaluation focus is about hallucination, we do not consider defining a "ground-truth" biography or

event summary, but rather determine if the generated text is grounded. We collect entity and event names using the following:

Researcher. We consider writing bios for the organizing boards and committee for: NeurIPS 2023, 2023 organizing committee for ASTMH (American Society for Tropical Medicine and Hygiene), the 27th Nordic Particle Physics Meeting (2023), and the American Comparitive Literature Association (ACLA). This results in a list of 106 entity names. We observe that there is sufficient information on the internet to write a biography for these entities, yet most do not have a Wikipedia page. We hypothesize that the entities in this dataset are likely in the 'tail' of the distribution of entity mentions.

Wiki-Ent. To select a collection of entities closer to the 'head' of the distribution of entity mentions, we select a subset of 81 entities from the list of entities used by (Min et al., 2023). Each entity has a Wikipedia page (though we do not treat the Wikipedia page in any special way, it may or may not be retrieved by search-based methods).

Wiki-Event. We collect a list of 233 events from the validation set of the WEC corpus (Eirew et al., 2021). These events that are the target for coreference resolution in the original dataset. Similar to the Wiki-Ent collection, these event names have corresponding Wikipedia pages (we neither forcibly include nor exclude that page from search).

News Events. We select a list of 52 News Events that began in August 2023[2]. These events are chosen for their recency, i.e. they occurred after the knowledge cutoff of the LLM.

4.2 Methods Compared

We evaluate two models of different sizes from Google Cloud VertexAI[3], namely (the smaller) text-bison-001 and (the larger) text-unicorn-001 (Anil et al., 2023) with zero-shot instructions following the templates as described in §3. Additionally, we report the effectiveness of grounded generation with Plan-based Retrieval on the open-weight Mistral-7B-Instruct model in § 4.5.

We compare the following retrieval methods:

No Retrieval: Baseline approach described in § 3.1 where the language model prompted to use it's parametric knowledge to directly generate the output without any external context.

	Approach	AIS			Rouge Prec.			Length
		Strict	Macro	Micro	R1	R2	RL	# Tokens
Wiki-Event	No Retrieval	0.00	15.30	15.74	69.45	30.16	66.63	88.63
	One-Retrieval	74.54	90.46	90.35	96.23	84.94	95.09	79.26
	One-Retrieval (2x snippets)	75.39	90.99	91.10	97.65	87.01	97.04	78.63
	Plan-based Retrieval (Var.A)	**89.99**	**96.18**	**95.14**	**99.59**	**94.52**	**99.44**	82.76
	Plan-based Retrieval (Var.B)	72.96	92.02	91.85	99.43	90.30	99.19	113.92
Wiki-Ent	No Retrieval	0.00	13.80	16.27	55.67	16.95	53.14	97.29
	One-Retrieval	60.91	83.09	83.11	91.52	74.17	90.12	87.62
	One-Retrieval (2x snippets)	67.03	87.44	86.67	95.79	80.45	94.92	86.86
	Plan-based Retrieval (Var.A)	**84.77**	**95.16**	**94.23**	**99.59**	**92.99**	**99.44**	93.03
	Plan-based Retrieval (Var.B)	63.37	90.40	90.70	99.28	89.45	99.04	134.47
Researcher	No Retrieval	0.00	6.74	6.73	57.10	17.85	54.84	80.58
	One-Retrieval	62.89	83.68	84.70	91.55	73.90	90.53	74.24
	One-Retrieval (2x snippets)	69.81	86.34	87.57	95.90	80.03	95.27	75.01
	Plan-based Retrieval (Var.A)	**79.56**	**91.63**	**92.16**	**97.87**	**83.99**	**97.48**	77.71
	Plan-based Retrieval (Var.B)	64.78	88.31	88.93	97.72	83.40	97.23	98.00
News Events	No Retrieval	0.00	5.46	5.14	68.81	24.41	66.00	89.44
	One-Retrieval	64.10	84.76	87.77	95.95	80.93	94.61	83.51
	One-Retrieval (2x snippets)	61.54	86.72	87.62	97.63	80.70	96.50	79.83
	Plan-based Retrieval (Var.A)	**80.77**	**94.47**	**93.72**	**99.76**	**93.47**	**99.60**	92.15
	Plan-based Retrieval (Var.B)	66.67	90.88	91.49	99.61	92.53	99.46	124.67

Table 1: **Comparison of Generation Approaches using text-bison-001 model.** The plan-based retrieval models that are at the center of our analysis yield more attributable text than One- and No- Retrieval methods. Of the two variants, Var.A produces more attributable sentences and Var.B produces longer texts.

One-Retrieval: This is the approach described in § 3.2 which uses one round of web search and concatenates the results in-context to the model. This formulation provides the language model external knowledge for conditional text generation.

One-Retrieval (2x snippets): This approach uses a single round of retrieval, but uses double the number of search results compared to One-Retrieval. Note this is also double the number of search results used by the plan-based methods in their first round of search as well.

Plan-based Retrieval: The proposed approach that uses planning (§ 3.3) to retrieve additional web search results. There are two variants, which differ in how the additional plan-based retrieved information is used to write the final response: i) Variant A simply puts the web search results in-context, and ii) Variant B performs the aforementioned QA-based outline. We further describe and compare several ablations of Plan-based Retrieval in §4.4.

4.3 Metrics

Our goal is to measure the rate at which models generate hallucinated text about the given entities and events. We will focus on the case where hallucination detection is about having a source document that supports the generated text. (We acknowledge that this differs from the case of generating objectively true statements.) To measure whether or not a generated sentence is attributable to one of the

Figure 2: **Example Generation**. One of the hallucinations in the One-Retrieval model is the focus of one of the questions provided in the question-based plan.

sources used by the model to produce the generated text, we compare the following metrics:

Attribution. We evaluate whether a model generation is hallucinated by checking whether it can be attributed to the source documents. Specifically, we use an automated metric that approximates the Attributable to Identified Sources (AIS) metric of Rashkin et al. (2023). The AutoAIS model (Honovich et al., 2022)[4] is trained on several natural language inference (NLI) datasets and predicts

---

[4]https://huggingface.co/google/t5_xxl_true_nli_mixture

	Approach	AIS			Rouge Prec.			Length
		Strict	Macro	Micro	R1	R2	RL	# Tokens
Wiki-Event	No Retrieval	1.00	25.80	23.70	70.39	33.34	67.82	135.91
Wiki-Event	One-Retrieval	68.67	89.39	88.47	95.58	85.77	94.71	98.76
Wiki-Event	One-Retrieval (2x snippets)	67.53	88.32	85.96	96.74	86.65	95.99	100.90
Wiki-Event	Plan-based Retrieval (Var.A)	**88.27**	**96.33**	**96.25**	**99.66**	**94.45**	**99.52**	98.61
Wiki-Event	Plan-based Retrieval (Var.B)	82.40	94.50	94.13	99.36	92.82	99.19	115.45
Wiki-Ent	No Retrieval	0.00	17.79	20.31	56.11	19.00	53.53	115.45
Wiki-Ent	One-Retrieval	63.79	86.70	87.37	92.28	74.72	90.87	87.30
Wiki-Ent	One-Retrieval (2x snippets)	58.65	85.67	83.96	94.46	79.43	93.42	110.23
Wiki-Ent	Plan-based Retrieval (Var.A)	**76.54**	**94.31**	**94.25**	**99.41**	**91.81**	**99.26**	127.22
Wiki-Ent	Plan-based Retrieval (Var.B)	69.96	93.58	93.02	99.28	90.96	99.06	164.23
Researcher	No Retrieval	0.00	7.90	7.82	55.38	15.96	52.73	99.20
Researcher	One-Retrieval	62.26	86.38	87.36	91.95	75.05	90.99	78.82
Researcher	One-Retrieval (2x snippets)	63.21	85.09	86.04	95.91	82.08	95.26	93.65
Researcher	Plan-based Retrieval (Var.A)	**85.22**	**94.31**	**94.68**	**98.33**	**87.74**	**98.07**	87.73
Researcher	Plan-based Retrieval (Var.B)	81.76	93.29	93.81	98.12	86.86	97.84	104.62
News Events	No Retrieval	1.92	12.69	9.89	68.91	25.33	66.36	120.89
News Events	One-Retrieval	67.31	90.27	88.21	95.83	83.48	94.52	96.42
News Events	One-Retrieval (2x snippets)	63.46	88.42	85.79	96.37	82.81	95.46	104.11
News Events	Plan-based Retrieval (Var.A)	**87.18**	**96.86**	**96.68**	**99.73**	**94.05**	**99.62**	100.91
News Events	Plan-based Retrieval (Var.B)	82.69	95.85	96.01	99.72	93.62	99.55	135.72

Table 2: **Comparison of Generation Approaches using text-unicorn-001 model.** We observe that Plan-based Retrieval improves upon One-Retrieval and No Retrieval methods, even when One-Retrieval retrieves more results. Plan-based Retrieval Var.B produces much longer texts, which are more attributable compared to One-Retrieval. Var.A produces slightly more attributed texts and at shorter length.

	AIS			Rouge P.		Len.
	Strict	Macro	Micro	R2	RL	# Tok
One-Retr.	60.91	83.09	83.11	74.17	90.12	87.62
Plan-based (Var.B)	63.37	90.40	90.70	89.45	99.04	134.47
w/o 2nd search	61.32	85.46	85.13	80.05	94.11	98.72

Table 3: **Importance of Gathering Additional Information from Second Search.** We compare Plan-based Retrieval (Var.B) with text-bison-001 in two settings, where we use the typical (secondary retrieval step) and using only the original search results as the source for the documents. We measure performance on the Wiki-Ent dataset. We find that attribution is indeed improved by gathering more facts and information.

whether a given context passage supports a claim. We segment the language models' output into sentences and evaluate whether a sentence can be attributed to any of the context passages (retrieved snippets and QA model answers). We report 3 variants of the metric based on aggregation. **Strict AIS** measures what fraction of the model outputs have all sentences correctly attributed. **Macro AIS** reports what fraction of the output sentences per query can be attributed on average. **Micro AIS** reports what fraction of all output sentences across all prompts are attributable to the context. For No Retrieval, we check attribution to web search snippets

obtained with the original query.

Rouge Prec. We also measure the overlap between the generated text and the evidence sources using ROUGE precision metrics for ROUGE-1 (R1), ROUGE-2 (R2) and ROUGE-Lsum (RL) with the retrieved documents and answers as the reference.

Length Since different approaches use different prompts and different amounts of context, they naturally tend to differ in terms of the length of the final LM output. We report the number of words in the output from different approaches. Measuring length is important to consider when observing the strict AIS metric since writing more means there are a larger number of required sentences to be correctly attributed (or in other words, more chances for mistakes).

Human Evaluation The focus of this work is to improve attributable generation. However, with a human evaluation study (see Appendix A.1.1), we demonstrate that the generation quality does not degrade in fluency while being more informative. We additionally justify the use of AutoAIS evaluation with a small-scale validation study (see Table 11).

4.4 Empirical Results

We first consider a comparison of the approaches from § 4.2, in terms of the metrics described in

| Dataset | Approach | AIS | | | Rouge Prec. | | | Length |
		Strict	Macro	Micro	R1	R2	RL	# Tokens
Researcher	Plan-based Retrieval (Var.B)	81.76	93.29	93.81	98.12	86.86	97.84	104.62
	w/o unanswerable	79.25	92.49	93.01	98.24	86.77	97.88	108.28
News Events	Plan-based Retrieval (Var.B)	82.69	95.85	96.01	99.72	93.62	99.55	135.72
	w/o unanswerable	76.92	94.59	93.54	99.55	93.02	99.43	147.03

Table 4: **Importance of Marking Questions as Unanswerable.** We measure the importance of marking questions that did not have an answer from the QA step as unanswerable (first row) as opposed to dropping the question from the context in the generation prompt (second row). We report these metrics on the Researcher and News Events Dataset with the text-unicorn-001 model. Highlighting unaswerable questions leads to increase in the AIS metric.

§ 4.3 (AIS, ROUGE, length) on each of the four datasets from § 4.1. Table 1 presents the results in this setting using the text-bison-001 model and Table 2 presents results using text-unicorn-001. For each metric, we report the average across three different runs with nucleus sampling.

Observation 1. Using Parametric Knowledge Alone Yields Fewer Attributed Sentences. No Retrieval never generates summaries that are completely attributed (Strict AIS score is always 0). However, the non-zero Macro AIS scores suggest that the language models retain knowledge from large portions of their training corpus. In particular, we notice that the Macro AIS scores are higher for entities that had Wikipedia pages prior to the model knowledge cutoff (Wiki-Ent and Wiki-Event) than both less public (Researcher) and newer entities (News Events). Note that the more recently released text-unicorn-001 has higher Macro AIS score in this purely parametric setting, perhaps indicating more parametric knowledge.

Observation 2. Adding Retrieved Evidence Improves Attribution Compared to using Parametric Knowledge Alone. In comparison to No Retrieval, the retrieval-augmented approach, One-Retrieval, has a much higher attribution rate. The high ROUGE scores indicate that the model is copying and stitching together facts from the search snippets. This is the expected behavior given that the model is instructed to only use information present in its context. The text-unicorn-001 appears to be better at following this instruction than text-bison-001. We see that there is headroom for reducing hallucinations in the model outputs. One possibility to improve attribution is to simply double the number of search snippets (One-Retrieval (2x snippets)) provided to the model. However, we see that the gain from this approach is inconsistent and it only improves attribution in 3 out of 8 settings. Thus, appending more search results based on the initial query is not enough to improve attribution.

Observation 3. Using Planning Along with Retrieval Can Improve Attribution. Plan-based Retrieval achieves better or comparable Strict AIS score as compared to One-Retrieval in 6 out of 8 settings. The impact of planning first before writing can be seen on the generation length; when using plans, the final outputs are longer and cover more aspects of the biography. The outputs are around 20 words longer across all settings.

Observation 4. Gathering Information to Answer Questions in a Second Round of Search Improves Attribution. Recall that web search happens two times in the Plan-based Retrieval model. First, with the initial entity or event name and then second, after questions are generated independently for each paragraph. This second web search is used to gather more documents to answer the questions (§3.3, Figure 1). In the experiment presented in Table 3 we pose the question of whether this additional, per-question based retrieval leads to performance improvements over answering the questions using only the documents in the first search. Indeed we see this second search to be beneficial, we find that it produces improvements of about 5 points of Macro and Micro AIS and two points of strict AIS. Furthermore, the approach with the second search writes longer texts (134 tokens compared to 98 tokens).

Observation 5. Indicating Unanswerable Questions in the Plan Improves Attribution. Some of the questions that are generated by the model will not necessarily be answerable even when considering the second round of search. We hypothesize that these unanswerable questions could lead to hallucinations from the model, especially if text is generated to provide an answer to them. In Table 4, we compare two approaches for handling these unanswerable questions.[5] The first is to label unanswerable questions and provide

---

[5]We determine if a question is unanswerable by confidence threshold on the QA model's response.

	AIS			Rouge P.		Len
	Strict	Macro	Micro	R2	RL	# Tok
Plan-based (Var.A)	87.18	96.86	96.68	94.05	99.62	100.91
w/o plan	68.59	86.38	84.15	85.45	97.51	92.68
One-Retrieval	67.31	90.27	88.21	83.48	94.52	96.42

Table 5: **Importance of Using Outline for Question Generation (on News Events).** We compare different methods of retrieval of information while using the same final generation prompt with text-unicorn-001. We see that using the chain-of-thought style paragraph outline helps to produce more grounded responses than an approach that simply generates questions from the initial search results.

Evidence	AIS			Rouge Prec.		
	Strict	Macro	Micro	R1	R2	RL
One-Retrieval						
Normal	67.31	90.27	88.21	95.83	83.48	94.52
Expanded	69.23	91.78	90.45	98.71	88.48	98.22
Plan-based Retrieval (Var.B)						
Expanded	82.69	95.85	96.01	99.72	93.62	99.55

Table 6: **Effect of Evidence Set on AIS Score (on News Events).** We observe that the set of evidences used for AIS evaluation does not dramatically change the evaluation metrics for the One-Retrieval model. Results reported on the News Events dataset with text-unicorn-001. *Normal* evidence refers to the One-Retrieval models retrieved results. *Expanded* refers to the union of the retrieved results of One-Retrieval and Plan-based Retrieval (Var.B) models.

the model with an instruction to not write about that particular question (precisely, instead of an answer, we follow the question with the response `Not enough information. Skip this question.`). This is the default setting and corresponds to the results presented in other tables. The second is to remove the unanswerable questions and to not include them in the in-context prompt. We see that explicitly highlighting unanswerable questions leads to more grounded generation.

Observation 6. Effect of the Final Generation Prompt (Var.A vs. Var.B). In Tables 1 & 2, we compare the effect of prompt format given the same evidence in the model context. In particular, we compare Plan-based Retrieval (Var.B) (which uses the blueprint and query plan in the final generation prompt) with Plan-based Retrieval (Var.A) which uses the same prompt as One-Retrieval. Both methods use the same search results obtained by plan-based retrieval. We see that while Var.A improves in terms of the attribution metrics, the blueprint and query plan allow Var.B to generate longer outputs with a comparably high per-query Macro AIS score. Thus, this prompt allows for a trade-off between generation length and attribution. However, we note that this trend may depend on the underlying LLM model family (see § 4.5).

Observation 7. Generating an Outline Before Generating Questions Improves Attribution. In Table 5, we study how the query-writing method affects the attribution while holding the generation prompt constant. In particular, we ablate Plan-based Retrieval (Var.A) by skipping the outline generation step. The approach labeled '(w/o plan)' directly generates search queries based on the initial search results without first generating a paragraph outline. All 3 approaches use the same prompt during generation. We see that it is not the specific gen-

eration prompt but the outline-based retrieved information procured by Plan-Based Retrieval that leads to improvements in attribution. The significantly lower attribution rate of the '(w/o plan)' ablation compared to Plan-Based Retrieval shows that chain-of-thought style blueprint generation improves the utility of subsequently generated queries.

Observation 8. Document set used for AutoAIS evaluation of the One-Retrieval Minimally Affects Results. One might hypothesize that the improvement in AutoAIS score seen by Plan-based Retrieval is due to its having more evidence documents for attribution. We compare the AutoAIS score for the One-Retrieval model with two different sets of evidence documents. In Table 6, we see that the AutoAIS increases (from 67.31 strict AIS to 69.23 strict AIS), but is significantly lower than the 82.69 strict AIS results from Plan-based Retrieval.

4.5 Generalization to Open-source Models

To demonstrate that the studied approach of plan-based generation extends to model families not discussed in the main experiments of the paper, we conduct additional experiments with `Mistral-7B-Instruct-v0.3` (Jiang et al., 2023a), an open-weight model released by Mistral AI. We replicate the Wiki-Ent setting and prompt the model to generate biographies for entities on Wikipedia. We use the instruction-tuned `Mistral-7B-Instruct-v0.3` model with the same prompts as listed in A.2 with minor changes (chat-style formatting) suggested in the model card. Note that the absolute metric values are not comparable to the results in the main body

Approach	AIS			Rouge Prec.			Length
	Strict	Macro	Micro	R1	R2	RL	# Tokens
One-Retrieval	11.7	70.5	71.1	80.7	47.8	77.6	168.0
One-Retrieval (2x snippets)	20.3	76.6	76.6	87.9	57.3	84.9	169.6
Plan-based Retrieval (Var.A)	16.0	73.5	74.0	91.9	61.1	89.6	**177.6**
Plan-based Retrieval (Var.B)	**25.0**	76.7	**77.1**	92.8	64.2	90.8	161.8

Table 7: **Comparison of Generation Approaches using `Mistral-7B-Instruct-v0.3`.** We see that Plan-based Retrieval (Var.B) leads to the generation of more attributable text than the One-Retrieval and One-Retrieval (2x snippets) baselines methods. In this setting, Plan-based Retrieval (Var.A) improves over One-Retrieval but not One-Retrieval (2x snippets). These results demonstrate that plan-based retrieval with model-specific tuning is a useful approach for grounded, long-form generation.

of the paper since behaviour of search engines is non-reproducible (Chen et al., 2023b).

Plan Quality. From the sample plans in Tables 13,14, we see that the model is capable of generating outlines for the output and writing search queries. Thus, this is not a capability of the specific PALM-2 models studied in the main body of the paper. Moreover, if models are unable to write plans, the PALM-2 models studied can be used to generate synthetic training data to fine-tune plan generation models.

Attribution Quality. Our results (see Table 7) demonstrate that Plan-based Retrieval (Var.A) and Plan-based Retrieval (Var.B) allows the model to write biographies at a higher attribution rate than One-Retrieval. Thus, the benefits of planning can be extended to other language models so long as they possess sufficient instruction following capabilities. The One-Retrieval (2x snippets) baseline however performs better than the Plan-based Retrieval (Var.A) approach (even when both Plan-based Retrieval (Var.A) and Plan-based Retrieval (Var.B) use the same queries and additional retrieved documents). This highlights the potential sensitivity of `Mistral-7B-Instruct-v0.3` to the formatting of the search results. Thus, the same plan-based variants may not all perform consistently across models. Still, with appropriate model-specific tuning, plan-based retrieval can improve over a single search round and is an important method to include in the system-design toolbox.

5 Conclusion

This paper explores the use of planning to improve retrieval for grounded long-form generation. We investigate how planning can guide retrieval allowing LLMs to generate more attributable responses and provide detailed analysis of the effect of the granularity of the plans and the content and quantity of retrieved information. We show empirically

how the planning-based approaches outperform, in terms of automated groundedness metrics, standard retrieval-augmented generation approaches. Our empirical investigations span topics in the top-end and long-tail of the parametric knowledge of LLMs.

6 Limitations

Understanding the precise causes for hallucinations in language models is beyond the scope of our work. Instead, we focus on analyzing the role retrieval has in improving attribution. While we attempt to cover both operational ranges (queries within and outside of model parametric knowledge), future work would need further verification of how retrieval augmented models perform on other domains and tasks, with other language models, and with other retrieval systems. Concurrent work by Shao et al. (2024) demonstrates the utility of outline-based search and generation for long-form expository writing of new Wikipedia pages.

Our goal was not to compare models/retrievers, but rather analyze when and how retrieval can reduce hallucinations. Thus, we focus on evaluating the effect of retrieval scheme on one model class in different domains. We report a smaller scale evaluation of `Mistral-7B-Instruct-v0.3` (Jiang et al., 2023a) with Plan-based Retrieval in § 4.5. We see that the popular open-weight instruction-tuned model is capable of generating paragraph outlines and subsequent search queries. We demonstrate that Plan-based Retrieval effectively improves the rate of grounded generation.

Plan-based Retrieval makes multiple calls to the LLM: (1) for outline generation, (2) for question generation, (3) for final response generation. While this may slow down the generation process, we were not proposing a finalized or productionized system. Our results provide an avenue for future research on distilling the plan writing capabilities

into smaller LLMs to improve the scalability of our analyzed method. Moreover, this may be an acceptable trade-off if we can achieve more grounded long-form responses. Several past approaches have relied on multiple LLM calls for improving output quality along some metric of choice: retrieval on demand during generation (Jiang et al., 2023b; Trivedi et al., 2023), post-hoc revision of model generations (Gao et al., 2023a), multi-sample consistency for reasoning (Yao et al., 2023), inter alia.

Our attribution evaluation tool is a model-based one, Auto-AIS. These models are not perfect and themselves a focus of active research. Thus, the reported numbers, while effective for model rankings should not be taken at absolute face-value. See Min et al. (2023) and Bohnet et al. (2023) for further discussion on correlation between automated and human AIS evaluation. We conduct a small validation study and find high agreement between the predicted AutoAIS labels and human AIS judgement (see Table 11).

7 Broader Impact

There are broad and far reaching capabilities of language models. Hallucinations can lead to a number of undesirable outcomes when using language models, from confusing user experiences, misunderstandings, to misinformation, and more. Our work should not be seen as a final solution to the hallucination problem, and should not be seen to fully quantify the extent to which any one model hallucinates or does not hallucinate. Our work is a research exploration, not a final solution.

8 Acknowledgments

References

Pritom Saha Akash, Kashob Kumar Roy, Lucian Popa, and Kevin Chen-Chuan Chang. 2023. Long-form question answering: An iterative planning-retrieval-generation approach.

Reinald Kim Amplayo, Kellie Webster, Michael Collins, Dipanjan Das, and Shashi Narayan. 2023. Query refinement prompts for closed-book long-form QA. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7997–8012, Toronto, Canada. Association for Computational Linguistics.

Rohan Anil, Andrew M. Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, Eric Chu, Jonathan H. Clark, Laurent El Shafey, Yanping Huang, Kathy Meier-Hellstern, Gaurav Mishra, Erica Moreira, Mark Omernick, Kevin Robinson, Sebastian Ruder, Yi Tay, Kefan Xiao, Yuanzhong Xu, Yujing Zhang, Gustavo Hernandez Abrego, Junwhan Ahn, Jacob Austin, Paul Barham, Jan Botha, James Bradbury, Siddhartha Brahma, Kevin Brooks, Michele Catasta, Yong Cheng, Colin Cherry, Christopher A. Choquette-Choo, Aakanksha Chowdhery, Clément Crepy, Shachi Dave, Mostafa Dehghani, Sunipa Dev, Jacob Devlin, Mark Díaz, Nan Du, Ethan Dyer, Vlad Feinberg, Fangxiaoyu Feng, Vlad Fienber, Markus Freitag, Xavier Garcia, Sebastian Gehrmann, Lucas Gonzalez, Guy Gur-Ari, Steven Hand, Hadi Hashemi, Le Hou, Joshua Howland, Andrea Hu, Jeffrey Hui, Jeremy Hurwitz, Michael Isard, Abe Ittycheriah, Matthew Jagielski, Wenhao Jia, Kathleen Kenealy, Maxim Krikun, Sneha Kudugunta, Chang Lan, Katherine Lee, Benjamin Lee, Eric Li, Music Li, Wei Li, YaGuang Li, Jian Li, Hyeontaek Lim, Hanzhao Lin, Zhongtao Liu, Frederick Liu, Marcello Maggioni, Aroma Mahendru, Joshua Maynez, Vedant Misra, Maysam Moussalem, Zachary Nado, John Nham, Eric Ni, Andrew Nystrom, Alicia Parrish, Marie Pellat, Martin Polacek, Alex Polozov, Reiner Pope, Siyuan Qiao, Emily Reif, Bryan Richter, Parker Riley, Alex Castro Ros, Aurko Roy, Brennan Saeta, Rajkumar Samuel, Renee Shelby, Ambrose Slone, Daniel Smilkov, David R. So, Daniel Sohn, Simon Tokumine, Dasha Valter, Vijay Vasudevan, Kiran Vodrahalli, Xuezhi Wang, Pidong Wang, Zirui Wang, Tao Wang, John Wieting, Yuhuai Wu, Kelvin Xu, Yunhan Xu, Linting Xue, Pengcheng Yin, Jiahui Yu, Qiao Zhang, Steven Zheng, Ce Zheng, Weikang Zhou, Denny Zhou, Slav Petrov, and Yonghui Wu. 2023. Palm 2 technical report.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2024. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In *The Twelfth International Conference on Learning Representations*.

Bernd Bohnet, Vinh Q. Tran, Pat Verga, Roee Aharoni, Daniel Andor, Livio Baldini Soares, Massimiliano Ciaramita, Jacob Eisenstein, Kuzman Ganchev, Jonathan Herzig, Kai Hui, Tom Kwiatkowski, Ji Ma, Jianmo Ni, Lierni Sestorain Saralegui, Tal Schuster, William W. Cohen, Michael Collins, Dipanjan Das, Donald Metzler, Slav Petrov, and Kellie Webster. 2023. Attributed question answering: Evaluation and modeling for attributed large language models.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal,

Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Meng Cao, Yue Dong, and Jackie Cheung. 2022. Hallucinated but factual! inspecting the factuality of hallucinations in abstractive summarization. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3340–3354.

Ziqiang Cao, Furu Wei, Wenjie Li, and Sujian Li. 2018. Faithful to the original: Fact aware neural abstractive summarization. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1).

Anthony Chen, Panupong Pasupat, Sameer Singh, Hongrae Lee, and Kelvin Guu. 2023a. Purr: Efficiently editing language model hallucinations by denoising language model corruptions.

Jifan Chen, Grace Kim, Aniruddh Sriram, Greg Durrett, and Eunsol Choi. 2023b. Complex claim verification with evidence retrieved in the wild.

Sanghyuk Choi, Jeong in Hwang, Hyungjong Noh, and Yeonsoo Lee. 2021. May the force be with your copy mechanism: Enhanced supervised-copy method for natural language generation.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sashank Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2024. Palm: scaling language modeling with pathways. *J. Mach. Learn. Res.*, 24(1).

Souvik Das, Sougata Saha, and Rohini Srihari. 2022. Diving deep into modes of fact hallucinations in

dialogue systems. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 684–699, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. 2024. Chain-of-verification reduces hallucination in large language models. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 3563–3578, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Alon Eirew, Arie Cattan, and Ido Dagan. 2021. WEC: Deriving a large-scale cross-document event coreference dataset from Wikipedia. *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2498–2510.

Luyu Gao, Zhuyun Dai, Panupong Pasupat, Anthony Chen, Arun Tejasvi Chaganty, Yicheng Fan, Vincent Zhao, Ni Lao, Hongrae Lee, Da-Cheng Juan, and Kelvin Guu. 2023a. RARR: Researching and revising what language models say, using language models. *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16477–16508.

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. 2023b. Enabling large language models to generate text with citations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6465–6488, Singapore. Association for Computational Linguistics.

Zorik Gekhman, Jonathan Herzig, Roee Aharoni, Chen Elkind, and Idan Szpektor. 2023. TrueTeacher: Learning factual consistency evaluation with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2053–2070, Singapore. Association for Computational Linguistics.

Tanya Goyal and Greg Durrett. 2020. Evaluating factuality in generation with dependency-level entailment. *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3592–3603.

Or Honovich, Roee Aharoni, Jonathan Herzig, Hagai Taitelbaum, Doron Kukliansy, Vered Cohen, Thomas Scialom, Idan Szpektor, Avinatan Hassidim, and Yossi Matias. 2022. TRUE: Re-evaluating factual consistency evaluation. In *Proceedings of the*

Second DialDoc Workshop on Document-grounded Dialogue and Conversational Question Answering, pages 161–175, Dublin, Ireland. Association for Computational Linguistics.

Or Honovich, Leshem Choshen, Roee Aharoni, Ella Neeman, Idan Szpektor, and Omri Abend. 2021. Q^2: Evaluating factual consistency in knowledge-grounded dialogues via question generation and question answering. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7856–7870.

Fantine Huot, Joshua Maynez, Chris Alberti, Reinald Kim Amplayo, Priyanka Agrawal, Constanza Fierro, Shashi Narayan, and Mirella Lapata. 2024. *µPLAN: Summarizing using a content plan as cross-lingual bridge.* In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2146–2163, St. Julian's, Malta. 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, Dubrovnik, Croatia. Association for Computational Linguistics.

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

Ziwei Ji, Tiezheng Yu, Yan Xu, Nayeon Lee, Etsuko Ishii, and Pascale Fung. 2023b. Towards mitigating LLM hallucination via self reflection. *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1827–1843.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023a. Mistral 7b.

Zhengbao Jiang, Frank Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023b. Active retrieval augmented generation. *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7969–7992.

Nayeon Lee, Wei Ping, Peng Xu, Mostofa Patwary, Pascale Fung, Mohammad Shoeybi, and Bryan Catanzaro. 2024. Factuality enhanced language models for open-ended text generation. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Daliang Li, Ankit Singh Rawat, Manzil Zaheer, Xin Wang, Michal Lukasik, Andreas Veit, Felix Yu, and Sanjiv Kumar. 2023a. Large language models with controllable working memory. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1774–1793, Toronto, Canada. Association for Computational Linguistics.

Junyi Li, Xiaoxue Cheng, Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2023b. HaluEval: A large-scale hallucination evaluation benchmark for large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6449–6464, Singapore. Association for Computational Linguistics.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2022. What makes good in-context examples for GPT-3? In *Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 100–114, Dublin, Ireland and Online. Association for Computational Linguistics.

Nelson Liu, Tianyi Zhang, and Percy Liang. 2023. Evaluating verifiability in generative search engines. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7001–7025, Singapore. Association for Computational Linguistics.

Shayne Longpre, Kartik Perisetla, Anthony Chen, Nikhil Ramesh, Chris DuBois, and Sameer Singh. 2021. Entity-based knowledge conflicts in question answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7052–7063, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ximing Lu, Sean Welleck, Peter West, Liwei Jiang, Jungo Kasai, Daniel Khashabi, Ronan Le Bras, Lianhui Qin, Youngjae Yu, Rowan Zellers, Noah A. Smith, and Yejin Choi. 2022. NeuroLogic a*esque decoding: Constrained text generation with lookahead heuristics. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 780–799, Seattle, United States. Association for Computational Linguistics.

Ximing Lu, Peter West, Rowan Zellers, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2021. NeuroLogic decoding: (un)supervised neural text generation with predicate logic constraints. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4288–4299, Online. Association for Computational Linguistics.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9802–9822.

Potsawee Manakul, Adian Liusie, and Mark Gales. 2023. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9004–9017, Singapore. Association for Computational Linguistics.

Sewon Min, Suchin Gururangan, Eric Wallace, Weijia Shi, Hannaneh Hajishirzi, Noah A. Smith, and Luke Zettlemoyer. 2024. SILO language models: Isolating legal risk in a nonparametric datastore. In *The Twelfth International Conference on Learning Representations*.

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.

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.

Ella Neeman, Roee Aharoni, Or Honovich, Leshem Choshen, Idan Szpektor, and Omri Abend. 2023. DisentQA: Disentangling parametric and contextual knowledge with counterfactual question answering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10056–10070, Toronto, Canada. Association for Computational Linguistics.

Baolin Peng, Michel Galley, Pengcheng He, Hao Cheng, Yujia Xie, Yu Hu, Qiuyuan Huang, Lars Liden, Zhou Yu, Weizhu Chen, and Jianfeng Gao. 2023. Check your facts and try again: Improving large language models with external knowledge and automated feedback.

Fabio Petroni, Patrick Lewis, Aleksandra Piktus, Tim Rocktäschel, Yuxiang Wu, Alexander H. Miller, and Sebastian Riedel. 2020. How context affects language models' factual predictions. In *Automated Knowledge Base Construction*.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2463–2473, Hong Kong, China. Association for Computational Linguistics.

Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah Smith, and Mike Lewis. 2023. Measuring and narrowing the compositionality gap in language models. *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5687–5711.

Hannah Rashkin, Vitaly Nikolaev, Matthew Lamm, Lora Aroyo, Michael Collins, Dipanjan Das, Slav Petrov, Gaurav Singh Tomar, Iulia Turc, and David Reitter. 2023. Measuring attribution in natural language generation models. *Computational Linguistics*, 49(4):777–840.

Vipula Rawte, Swagata Chakraborty, Agnibh Pathak, Anubhav Sarkar, S.M Towhidul Islam Tonmoy, Aman Chadha, Amit Sheth, and Amitava Das. 2023. The troubling emergence of hallucination in large language models - an extensive definition, quantification, and prescriptive remediations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2541–2573, Singapore. Association for Computational Linguistics.

Paul Roit, Johan Ferret, Lior Shani, Roee Aharoni, Geoffrey Cideron, Robert Dadashi, Matthieu Geist, Sertan Girgin, Leonard Hussenot, Orgad Keller, Nikola Momchev, Sabela Ramos Garea, Piotr Stanczyk, Nino Vieillard, Olivier Bachem, Gal Elidan, Avinatan Hassidim, Olivier Pietquin, and Idan Szpektor. 2023. Factually consistent summarization via reinforcement learning with textual entailment feedback. *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6252–6272.

Mobashir Sadat, Zhengyu Zhou, Lukas Lange, Jun Araki, Arsalan Gundroo, Bingqing Wang, Rakesh Menon, Md Parvez, and Zhe Feng. 2023. DelucionQA: Detecting hallucinations in domain-specific question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 822–835, Singapore. Association for Computational Linguistics.

Yijia Shao, Yucheng Jiang, Theodore Kanell, Peter Xu, Omar Khattab, and Monica Lam. 2024. 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.

Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, YaGuang Li, Hongrae Lee, Huaixiu Steven Zheng, Amin Ghafouri, Marcelo Menegali, Yanping Huang,

Maxim Krikun, Dmitry Lepikhin, James Qin, Dehao Chen, Yuanzhong Xu, Zhifeng Chen, Adam Roberts, Maarten Bosma, Vincent Zhao, Yanqi Zhou, Chung-Ching Chang, Igor Krivokon, Will Rusch, Marc Pickett, Pranesh Srinivasan, Laichee Man, Kathleen Meier-Hellstern, Meredith Ringel Morris, Tulsee Doshi, Renelito Delos Santos, Toju Duke, Johnny Soraker, Ben Zevenbergen, Vinodkumar Prabhakaran, Mark Diaz, Ben Hutchinson, Kristen Olson, Alejandra Molina, Erin Hoffman-John, Josh Lee, Lora Aroyo, Ravi Rajakumar, Alena Butryna, Matthew Lamm, Viktoriya Kuzmina, Joe Fenton, Aaron Cohen, Rachel Bernstein, Ray Kurzweil, Blaise Aguera-Arcas, Claire Cui, Marian Croak, Ed Chi, and Quoc Le. 2022. Lamda: Language models for dialog applications.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2023. Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions. *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10014–10037.

Tu Vu, Mohit Iyyer, Xuezhi Wang, Noah Constant, Jerry Wei, Jason Wei, Chris Tar, Yun-Hsuan Sung, Denny Zhou, Quoc Le, and Thang Luong. 2024. FreshLLMs: Refreshing large language models with search engine augmentation. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 13697–13720, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

David Wan, Mengwen Liu, Kathleen McKeown, Markus Dreyer, and Mohit Bansal. 2023. Faithfulness-aware decoding strategies for abstractive summarization. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2864–2880, Dubrovnik, Croatia. Association for Computational Linguistics.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V Le. 2022. Finetuned language models are zero-shot learners. *International Conference on Learning Representations*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2024. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Sam Wiseman, Stuart Shieber, and Alexander Rush. 2017. Challenges in data-to-document generation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2253–2263, Copenhagen, Denmark. Association for Computational Linguistics.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik R

Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. *Thirty-seventh Conference on Neural Information Processing Systems*.

Ori Yoran, Tomer Wolfson, Ori Ram, and Jonathan Berant. 2024. Making retrieval-augmented language models robust to irrelevant context. In *The Twelfth International Conference on Learning Representations*.

Polina Zablotskaia, Misha Khalman, Rishabh Joshi, Livio Baldini Soares, Shoshana Jakobovits, Joshua Maynez, and Shashi Narayan. 2023. Calibrating likelihoods towards consistency in summarization models.

Shen Zheng, Jie Huang, and Kevin Chen-Chuan Chang. 2023. Why does chatgpt fall short in providing truthful answers?

Wangchunshu Zhou, Yuchen Eleanor Jiang, Ethan Wilcox, Ryan Cotterell, and Mrinmaya Sachan. 2023. Controlled text generation with natural language instructions. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Chen Zhu, Ankit Singh Rawat, Manzil Zaheer, Srinadh Bhojanapalli, Daliang Li, Felix Yu, and Sanjiv Kumar. 2020. Modifying memories in transformer models.

A Appendix

A.1 Qualitative Evaluation of System Output

In order to test whether plan-based retrieval and generation leads to degradation in other qualitative aspects of the output, we conduct additional evaluations beyond attribution quality.

A.1.1 Human Evaluation

Dataset	One Retrieval	No Preference	Plan-based Retrieval
text-unicorn-001			
Researcher	0%	100%	0%
Wiki-Ent	4%	92%	4%
Wiki-Event	4%	92%	4%
News Events	8%	84%	8%
Avg	4	92	4
text-unicorn-001			
Researcher	8%	71%	21%
Wiki-Ent	8%	88%	4%
Wiki-Event	4%	92%	4%
News Events	4%	92%	4%
Avg	6	86	8

Table 8: **Head-to-head Fluency Comparison.** Plan-based Retrieval and generation does not degrade the fluency of the output as compared to standard retrieval. In the most common scenario, the outputs are not distinguishable in terms of fluency.

Dataset	One Retrieval	No Preference	Plan-based Retrieval
text-unicorn-001			
Researcher	24%	28%	48%
Wiki-Ent	20%	44%	36%
Wiki-Event	17%	8%	75%
Curr-Event	8%	32%	60%
Avg	17	28	55
text-unicorn-001			
Researcher	16%	24%	60%
Wiki-Ent	36%	4%	60%
Wiki-Event	20%	48%	32%
Curr-Event	0%	36%	64%
Avg	18	28	54

Table 9: **Head-to-head Informativeness Comparison.** Plan-based Retrieval leads to model generations that are more informative about 55% more times than standard retrieval. The improvements are clearer when the LLM is queried about long-tail entities and recent events.

We had 3 annotators (2 authors and 1 colleague

	OR	-	PBR
OR	0	8%	0
-	8%	80%	4%
PBR	0	0	0

(a) Fluency

	OR	-	PBR
OR	28%	0	0
-	4%	4%	12%
PBR	4%	0	48%

(b) Informativeness

Table 10: **Inter-annotator agreement:** We report the confusion matrix of disagreements of two annotators. The low percentage of off-diagonal elements (especially the low percentage of extreme disagreements) is an indicator of high agreement. 'OR' means a preference for One-Retrieval, 'PBR' means a preference for Plan-based Retrieval and '-' indicates no preference.

familiar with the scope of the experiments) annotate 25 responses from each dataset to provide a comparison between the two best-performing methods: our proposed Plan-based Retrieval approach and the One-Retrieval baseline. The annotators are presented a pair of generations for the same entity/event prompt, and asked to compare the generations on two qualities: fluency and informativeness. Note that informativeness is independent of the truthfulness/correctness of the model response; we asked them to indicate whichever response provided more distinct, relevant facts. The source of each generation is hidden from the annotators.

> **Instruction for fluency:** For every pair of model generations, label which generation is more fluent or label "no preference" if the fluency is not noticeably different. Consider the following when judging fluency:
>
> - Does the generation present information in a coherent manner?
> - Does it switch back and forth between aspects? As an example, does it present biographical events in shuffled order?
> - Does the generation present appropriate information to the query?
> - Does the generation repeat information?
> - Do not base your judgement on the "ground truth" correctness of stated facts.

> **Instruction for informativeness:** For every pair of model generations, label which generation is provides more relevant information or label "no preference" if the informativeness is not noticeably different. Consider the following when judging informativeness:
>
> - Which generation provides more information "nuggets" while remaining rele-

> vant tot he query?
> - You may disregard small differences in the number of list-like facts. For example, when writing about an actor, if one generation lists 3 movies while the other lists 5 movies, do not consider this a significant difference.
> - Do give importance to facts that provide a more comprehensive description of the query topic. For example, if system A generates: "X was taught by Y." and model B generates: "X was taught by Y but would surpass Y in 1991.", then we would consider system B as more informative.
> - Do not base your judgement on the "ground truth" correctness of stated facts.

Regarding fluency (Table 8), we find that Plan-based Retrieval with `text-unicorn-001` yields generations that are as or more fluent than the baseline One-Retrieval response 92-100% of times. It does so while producing 10-20% more factual responses. In terms of informativeness (Table 9), responses from our Plan-based Retrieval were more informative 55% of the time compared to 18% win-rate for the baseline One-Retrieval. One-Retrieval performs well on queries in the head of the distribution (Wiki-Ent and Wiki-Event); Plan-based Retrieval shows its effectiveness on the long-tail query sets. We observe similar trends in responses from the `text-bison-001` model.

In order to report inter-annotator agreement, we collect additional judgements on generations by `ext-unicorn-001` on queries from the Wiki-Ent set. We report the judgements of the two annotators as confusion matrices in Tables 10a and 10b. 80% of the examples are on the diagonal indicating that the annotators agree very frequently. Strong disagreements (i.e. annotators preferring different systems) occurs only 4% of the time (see Table 10b). The most common disagreement is 'no preference' over any one system; this is expected given the subjectivity in evaluation.

We conduct a pilot validation study to support the use of the AutoAIS metric for evaluation. We randomly sampled set of 50 sentence-document pairs from real generations by the systems. The authors debate and annotate whether the generated sentence is supported by the paired document. We ensure that the validation set contains balanced number of claim-document pairs from each of the predicted labels from the AutoAIS model (sup-

		AutoAIS Prediction	
		Not Supp.	Supp.
Human Label	Not Supp.	41%	0%
	Supp.	6%	53%

Table 11: **Validating AutoAIS metric.** On the randomly sampled set of claim-document pairs, we note high agreement with the predicted AutoAIS label. We observe some instances of supported claims that are labeled as unattributable by the metric.

ported / not supported). We hide the predicted label during annotation. The results of the study are presented in Table 11. We find few instances of disagreement with the predicted AutoAIS label. Most supported claim-document pairs are highly extractive snippets from the document. Similarly, unsupported sentences often contain entities that do not appear in the paired document. When the annotators and the metric disagree, these are instances that require reasoning and table data manipulation. We point to the high agreement as evidence that the AutoAIS metric (while imperfect) provides reliable signal of model improvement.

A.1.2 N-gram Repetitiveness

Plan-based Retrieval uses the paragraph plan to generate search queries. Since queries for each paragraph are generated independently, there may be repeated queries in the final plan. To demonstrate that the final generations from Plan-based Retrieval are not repetitive, we compute the fraction of unique N-grams (based on word boundaries) out of all N-grams in the generation. If the generation was repetitive (increasing length by reiterating the same facts), then this metric would be low. From Table 12, we see that the {1,2,3}-gram uniqueness of texts from Plan-based Retrieval is comparable to the same metrics with the No Retrieval and One-Retrieval baselines. In fact, we see that the fraction of unique trigrams in the generation with Plan-based Retrieval is not distinguishable from the One-Retrieval setting and is generally higher than the fraction of unique N-grams in the No Retrieval settings (i.e., the generation incorporates newly retrieved information).

A.2 Prompt Formats

1. Direct Generation Prompt

```
Write a fluent, clear paragraph about
{entity}.
```

2. Search-Based Generation Prompt

Dataset	Approach	1-gram Unique.	2-gram Unique.	3-gram Unique.
text-unicorn-001				
Researcher	No Retrieval	71.90	95.13	98.81
	One-Retrieval	73.26	95.63	98.74
	Plan-based Retrieval(Var.A)	75.23	96.13	98.89
Wiki-Ent	No Retrieval	72.79	95.25	98.95
	One-Retrieval	75.57	96.35	99.18
	Plan-based Retrieval(Var.A)	77.93	96.74	99.29
Wiki-Event	No Retrieval	60.75	87.72	94.65
	One-Retrieval	69.33	93.78	98.12
	Plan-based Retrieval(Var.A)	67.89	93.57	98.19
News Events	No Retrieval	60.95	87.54	94.62
	One-Retrieval	73.30	95.14	98.29
	Plan-based Retrieval(Var.A)	66.53	92.13	97.37
text-bison-001				
Researcher	No Retrieval	73.34	95.70	99.02
	One-Retrieval	76.02	96.57	99.17
	Plan-based Retrieval(Var.A)	73.11	95.76	98.69
Wiki-Ent	No Retrieval	71.18	94.54	98.70
	One-Retrieval	78.91	97.28	99.39
	Plan-based Retrieval(Var.A)	74.20	95.97	99.03
Wiki-Event	No Retrieval	68.82	93.79	98.46
	One-Retrieval	73.65	95.41	98.67
	Plan-based Retrieval(Var.A)	71.08	94.25	98.04
News Events	No Retrieval	66.80	92.08	97.43
	One-Retrieval	75.25	95.96	98.90
	Plan-based Retrieval(Var.A)	70.87	94.86	98.74

Table 12: **Comparison of percentage of unique n-grams in model output.** We see no noticeable indication of repetition in the outputs of Plan-based Retrieval. Further human evaluation (Tab. 8) confirms that the output of Plan-based Retrieval is fluent.

```
Search Results:
Snippet Title: {result 1 title}
Snippet Text: {result 1 text}
Snippet Title: {result 2 title}
Snippet Text: {result 2 text}
...
Snippet Title: {result K title}
Snippet Text: {result K text}
Write a fluent, clear paragraph about
{entity} using only facts in the given
text.
```

3. Outline Prompt

```
Given the above search results, write
a list of instructions for how to
provide an answer to write a bio about
{entity} in the format:
Paragraph  1:   Instructions  for
paragraph 1
Paragraph  2:   Instructions  for
paragraph 2
...
```

Paragraph 1

- Begin by introducing Ramesses IV, the third Pharaoh of the Twentieth Dynasty of the New Kingdom of Ancient Egypt, with a mention of his original name, Heqamaatre-Setepenamun.
- Provide the dates of his reign, which were from 1155 BCE to 1149 BCE.
- Mention that he was the second son of Ramesses III, and likely named after Ramesses, the second son of Ramesses II.
- Emphasize that he was appointed crown prince by the 22nd year of his father's reign, following the deaths of his four older brothers.

Paragraph 2

- Describe the events leading to his ascension to the throne, such as the assassination of his father, Ramesses III, in 1156 BC, which made Ramesses IV the king at a middle-aged age.
- Mention that his reign was relatively short, lasting only six years.

Paragraph 3

- Discuss Ramesses IV's burial details, noting that he was originally buried in tomb KV2 in the Valley of the Kings, West Thebes, but was later reburied in the tomb of Amenhotep II (KV35).
- Highlight that his mummy was clumsily rewrapped and labeled by later priests.

...

Table 13: Partial plan generated zero-shot by Mistral-7B-Instruct-v0.3 for query: Ramesses IV

```
Paragraph  N:  Instructions  for
paragraph N
The outline should allow for each
paragraph to be written independently
in  parallel.  The  collection  of
paragraphs should form a bio for
{entity}. For each paragraph, write
a descriptive set of instructions
for  the  content  that  should  be
included  and  summarize  the  things
that should not be included because
they are written in other paragraphs.
All  facts,  dates,  and  years  must
be supported in the given search
results.
```

4. Question Prompt

```
Search Results:
Snippet Title: {result 1 title}
Snippet Text: {result 1 text}
Snippet Title: {result 2 title}
```

Paragraph 1

- Instructions: Begin by introducing Ronaldo, focusing on his full name, nationality, and the sport he is famous for. Mention his nickname, if it is not already clear from his name. Provide a brief background on his hometown and birth year.
- Do not include information about Cristiano Ronaldo, another famous athlete with a similar name.

Paragraph 2

- Instructions: Highlight Ronaldo's achievements and accolades early in his career, such as winning the FIFA World Cup with Brazil in 2002 and being named the tournament's top scorer (Golden Shoe award). Mention any individual awards he received during this period, such as the FIFA Player of the Year awards in 1996–97 and 2002.
- Do not repeat information about specific teams he played for at this stage, as it will be covered in a separate paragraph.

Paragraph 3

- Instructions: Describe Ronaldo's successful stint with European soccer teams, focusing on key clubs and the impact he made during his time there. Mention any notable accomplishments or awards he received during this period.
- Do not include personal details about his life outside of soccer at this stage.

...

Table 14: Partial plan generated zero-shot by Mistral-7B-Instruct-v0.3 for query: Ronaldo (Brazilian footballer)

```
Snippet Text: {result 2 text}
...
Snippet Title: {result K title}
Snippet Text: {result K text}

Given the above search results,
what are the questions you would
want answered to write the following
paragraph {paragraph description}
about {entity}? Write just the
questions separated by a new
line. Each question should be
understandable independently.
```

5. Outline-based Generation Prompt

```
Consider        the        following
question-answer pairs:
{Question 1} {Answer 1}
...
{Question N} {Answer N}
Write a fluent, clear paragraph
about {entity} using only facts in
```

the above.

A.3 Extended Related Work

Responsiveness to Context. When provided with factual knowledge in-context, we hope that LLMs use the facts and do not hallucinate. Previous studies have considered this for small models in the presence of producing LM facts (Petroni et al., 2019). Neeman et al. (2023) study how predictions change when facts in context contradict parametric knowledge. Longpre et al. (2021); Li et al. (2023a) performs similar analysis in the adversarial setting. Yoran et al. (2024) show that retrieving irrelevant information can hurt LLM performance. Concurrent with our work, Vu et al. (2024) demonstrate that including auxiliary data from the search engine and ordering the search results such that the most relevant results are presented at the end improves LLM grounding. We conduct a similar analysis of what and how to present search results to LLMs for long-form generation.

Understanding Hallucinations. The manner in which language models hallucinate is investigated both in dataset construction and empirical comparisons (Cao et al., 2022; Das et al., 2022; Li et al., 2023b; Rawte et al., 2023; Zheng et al., 2023). For instance, Li et al. (2023b) investigates hallucinate rates as a function of model as well as task and breaks down hallucination types into error categories (e.g., "comprehension", "factual", "specificity", and "inference"). Similarly, Sadat et al. (2023) presents analyze hallucinations in specific domains, in particular, car manuals.

Grounded Generation. Prior work has demonstrated that generating text that is grounded in given sources (be it tabular data, or natural language) can be challenging (Wiseman et al., 2017; Cao et al., 2018, inter alia). Methods have used constrained decoding (Lu et al., 2021, 2022) to produce text that is copied from given sources. Other methods have used architectural changes, which use attention-based re-weighting to make grounding more likely (Choi et al., 2021). AIS evaluation models have also been used during training (Zablotskaia et al., 2023) and decoding (Roit et al., 2023; Wan et al., 2023) to increase faithfulness of model generations.

Attribution by Self-Verification. As discussed before, it is well known that language models store a large amount of knowledge in their parameters (Petroni et al., 2019). Additionally, instruction-tuned LLMs have demonstrated the ability to perform natural language inference with zero-shot prompting (Wei et al., 2022). Dhuliawala et al.

(2024) use this ability for muilt-answer QA to verify generated responses with the same LLM. Ji et al. (2023b) apply a similar strategy for medical domain generative-QA and show that iterative verification and patching (fixing factually incorrect statements) leads to more factually correct responses.

In-Context Learning. The method of providing task-specific examples in the context of language models to solve a tasks, i.e., in-context learning (Brown et al., 2020), is similar to our work in that it requires being responsive to the context provided to the language model. There exists parallels between our work and prior methods that retrieve in-context examples (Liu et al., 2022).