When Context Leads but Parametric Memory Follows in Large Language Models

Yufei Tao, Adam Hiatt, Erik Haake, Antonie J. Jetter, Ameeta Agrawal
Portland State University, USA
{yutao, ahiatt, ehaake, ajetter, ameeta}@pdx.edu

Abstract

Large language models (LLMs) have demonstrated remarkable progress in leveraging diverse knowledge sources. This study investigates how nine widely used LLMs allocate knowledge between local context and global parameters when answering open-ended questions in knowledge-consistent scenarios. We introduce a novel dataset, `WikiAtomic`[1], and systematically vary context sizes to analyze how LLMs prioritize and utilize the provided information and their parametric knowledge in knowledge-consistent scenarios. Additionally, we also study their tendency to hallucinate under varying context sizes. Our findings reveal consistent patterns across models, including a consistent reliance on both contextual (around 70%) and parametric (around 30%) knowledge, and a decrease in hallucinations with increasing context. These insights highlight the importance of more effective context organization and developing models that use input more deterministically for robust performance.

1 Introduction

Large language models (LLMs) have significantly advanced the capabilities of natural language processing. When generating responses, LLMs can use the contextual information provided in a prompt along with or instead of the parametric knowledge embedded during pretraining (Petroni et al., 2019; Brown et al., 2020; Heinzerling and Inui, 2021).

In order to generate accurate and coherent responses, LLMs need to effectively combine their parametric knowledge with provided contextual information, and understanding the balance between these two sources of information is crucial (Neeman et al., 2022; Li et al., 2024). Most prior work has explored this question through the lens of counterfactual data where the parametric and contextual knowledge are in conflict (Longpre et al., 2021; Xu et al., 2024a). Some studies suggest that the models prefer contextual knowledge, while others suggest they prioritize parametric knowledge (Krishna et al., 2021; Zhou et al., 2023a).

In many real-world scenarios such as question answering or summarization, however, contextual knowledge may *complement* rather than conflict with parametric knowledge. Understanding how models integrate different sources of knowledge in knowledge-consistent scenarios is critical.

Moreover, while parametric knowledge enables LLMs to generate coherent text, it also contributes to the risk of hallucinations where responses are coherent and confident but factually incorrect or irrelevant (Ji et al., 2023; Tian et al., 2023; Wang et al., 2023; Luo et al., 2024). As such, our work also explores the hallucination tendency in knowledge-consistent scenarios.

Real-world applications often present LLMs with varying amounts of context. Building robust models requires understanding how effectively they leverage this information. This research investigates how the volume of context influences knowledge preference and hallucination tendencies in LLMs within a question answering (QA) framework. We study nine widely used LLMs to determine how different models navigate varying context sizes in knowledge-consistent settings, aiming to answer the main question: How do LLMs prioritize contextual and parametric knowledge when generating responses? Additionally, we investigate the likelihood of generating hallucinations with increasing context, which parts of contexts are used, how similar are various types of knowledge, and further analyses exploring potentially unseen knowledge.

Figure 1 provides an overview of our investigation pipeline. We create a new dataset,

[1] The dataset is available at https://github.com/PortNLP/WikiAtomic.

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 4034–4058
November 12-16, 2024

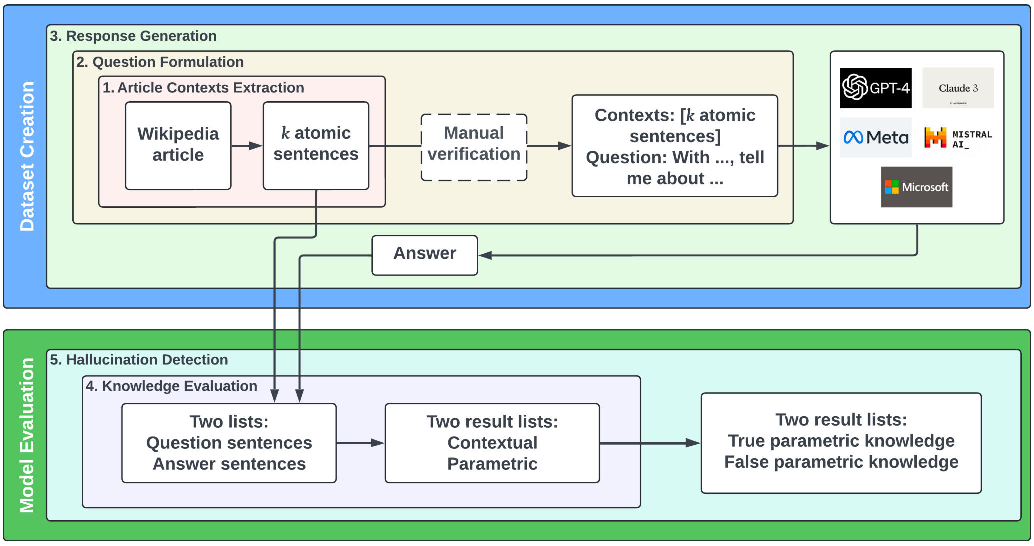


Figure 1: Overview of the dataset creation and model evaluation pipeline

`WikiAtomic`, of atomic sentences serving as our context. We systematically adjust the amount of factual context provided to the models and prompt them in an open-ended QA task to study the responses generated by various models. Then, we measure the amount of contextual knowledge models recall, the additional parametric knowledge they rely on, and the extent of hallucinations found in their responses.

Our results suggest that all models behave very similarly, integrating up to 30% parametric knowledge in their responses. For smaller contexts, models recall information from all parts of the context. However, for longer contexts, they predominantly focus on the first half, potentially missing key information from other parts of the input. The sequence of information in the responses largely mirrors the context and the additional parametric knowledge included is moderately similar to the contextual knowledge. Lastly, the hallucination score decreases as more context is added. These findings highlight the importance of effective context organization and models that utilize input more predictably.

Our key contributions include:

- A novel dataset, `WikiAtomic`, comprised of prompt/response pairs in which provided context is incrementally increased.
- Comprehensive analyses showing how models utilize contextual and parametric and in what proportion.
- We discuss the relationship between hallucination tendency and the amount of context.

2 Task and Terminology

We structure our investigation as an open-ended question answering task to gain deeper insights into how LLMs utilize and integrate contextual knowledge with their parametric knowledge. By incrementally increasing the amount of context, we analyze the models' behavior, focusing on their ability to prioritize different types of information and their tendency toward factual hallucination in knowledge-consistent scenarios.

We next introduce some key terminology used throughout the paper.

Topic: This is the title of the Wikipedia article. The `WikiAtomic` dataset includes 200 distinct articles, and questions are asked about these topics.

Atomic Sentence: This is a sentence that contains a single piece of information that cannot be broken down into simpler components without losing its meaning (Liu et al., 2023).

Context: For each topic, a context consists of k atomic sentences provided in a prompt. In our experiments, $2 \leq k \leq 50$, where contexts consist of increments of 2 sentences from 2 to 30 (e.g., 2, 4, 6, etc.), followed by increments of 5 sentences until reaching a total of 50 sentences. For each topic, this results in contexts of 20 different sizes, creating a total of 4,000 topic-context instances.

Response: Given the context, the model generates a response that includes some degree of both contextual and parametric knowledge. This response is then atomized. A sentence directly derived from the provided context is considered as **contextual (local) knowledge**. In contrast, a sentence that is *not* entailed from the context is considered as **parametric (global) knowledge**. Following prior work (Neeman et al., 2022), contextual knowledge comes from external sources provided during inference, while parametric knowledge is knowledge encoded (or "memorized") in the model parameters. As such, if a sentence introduces information not present in the context, it is considered a parametric knowledge sentence.

3 `WikiAtomic` Dataset

We created a novel dataset – `WikiAtomic` – consisting of 200 articles from Wikipedia. We selected Wikipedia as the basis for creating our knowledge-consistent dataset because its extensive data has historically been integral to the training of numerous LLMs, explicitly in models like GPT-2, GPT-3, BERT, T5, and BLOOM, and implicitly in large-scale aggregate corpora such as The Pile (Gao et al., 2020) and Common Crawl (Luccioni and Viviano, 2021). This extensive use allows us to reasonably assume that information from Wikipedia, particularly from older articles, is present in the pretraining data of several LLMs we study.

Extracting Wikipedia Articles We selected 200 high-quality articles from Wikipedia[2], each over 1000 words, covering diverse topics from science and technology to history, culture, and prominent figures. After manually removing low-quality texts, we ensured a diverse collection for our open-ended question answering task.

Converting to Atomic Sentences To precisely control the number of contexts in our question, all articles are decomposed into a set of atomic sentences (Liu et al., 2023). Sentences containing multiple pieces of information can complicate evaluations, particularly those involving entailment (Kim et al., 2024). By breaking down sentences into atomic sentences, we ensure that each unit presents a single, unambiguous piece of information which enhances the accuracy of entailment assessments. Following Min et al. (2023), we use GPT-4o to extract atomic sentences by providing the definition of atomic facts and instructing the model to perform multiple passes on an article to break each sentence down to individual atomic sentences (the prompt is included in Appendix A). Consequently, k atomic sentences are extracted for each article (in our experiments, $k = 50$) for a total of $200 * 50 = 10000$ atomic sentences in `WikiAtomic`. Example 1 shows an original sentence and corresponding atomic sentences.

[2]https://huggingface.com/datasets/wikipedia

Example 1: Atomic Sentence

Original Sentence:
`Mariah Carey, born on March 27, 1969, in Huntington, New York.`

Atomic Sentences:
`1. Mariah Carey was born on March 27, 1969.`
`2. Mariah Carey was born in Huntington, New York.`

We manually verify a subset of 1000 atomic sentences (20 articles, each with 50 atomic sentences) to ensure that the sentences (i) were correctly atomized and, (ii) originate from the corresponding Wikipedia article. Only 12 out of 1000 sentences were found to be insufficiently atomic or had their meaning changed. GPT-4o tended to struggle the most with sentences that have complex structures, such as those with multiple clauses or extensive use of commas. This led to errors in atomic sentence extraction where the sentences either contained too much information to be considered atomic or the meaning got altered in the process. Here is an example of each failure case:

(a). Not atomic enough example:

Original sentence: `In 1774, the British passed the Intolerable Acts to punish the colonists in Boston for the Boston Tea Party.`
GPT-4o's atomic sentence: `'In 1774, the British passed the Intolerable Acts to punish the colonists in Boston.', ...`
Correct atomic sentences: `'In 1774, the British passed the Intolerable Acts.', 'The Intolerable Acts were intended to punish the colonists in Boston.'` ...

(b). Changed meaning example:

Original sentence: `... However, Denmark, on the losing side of the Napoleon wars, lost Norway to Sweden, on the winning side.`
GPT 4o's atomic sentences: `... 'Norway was on the losing side of the Napoleonic Wars.', 'Sweden was on the winning side of the Napoleonic Wars.'`

Overall, GPT-4o achieved 98.8% accuracy in atomic sentence extraction. The sentences that did not meet our criteria for atomicity were revised by referring back to their original Wikipedia articles.

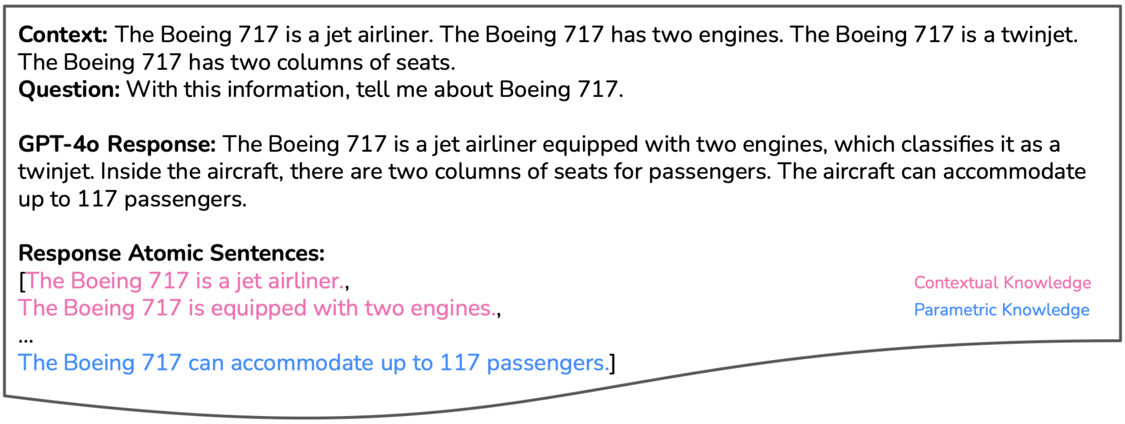


Figure 2: An example of Context, Question, Model Response (GPT-4o) and the list of Atomic Response mapped to contextual knowledge and parametric knowledge

4 Experiments and Evaluation

This section outlines the experiment setup, evaluation metrics, and the models studied.

4.1 Experiments

Figure 2 illustrates a sample instance of our question answering task that includes the input (a list of atomic sentences as context along with a question prompt) and the output which is a model's response further atomized.

Question Formulation We prompt the models to answer a question given the contexts using a *semi-restrict* format: "`With this information, tell me about {Topic}`"[3]. The effects of varying the question format are further explored in Section 5.5.

Response Generation The responses generated by the LLMs are converted into atomic sentences using the same method described earlier in Section 3[4].

Ultimately, we obtain two lists for each question-answer pair: atomic contexts (atomic sentences from the context) and atomic responses (atomic sentences from the response), allowing us to directly compare them and minimize discrepancies.

4.2 Evaluation

Here, we describe the metrics for detecting contextual and parametric knowledge, as well as model hallucination, focusing on evaluating the faithfulness and factual accuracy of responses.

Contextual vs. parametric knowledge detection

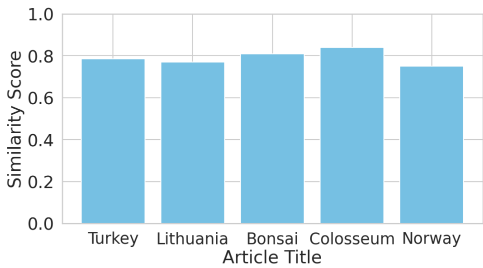


Figure 3: Knowledge-consistency between parametric knowledge and input context of `WikiAtomic` topics, computed using SBERT (Reimers and Gurevych, 2019)

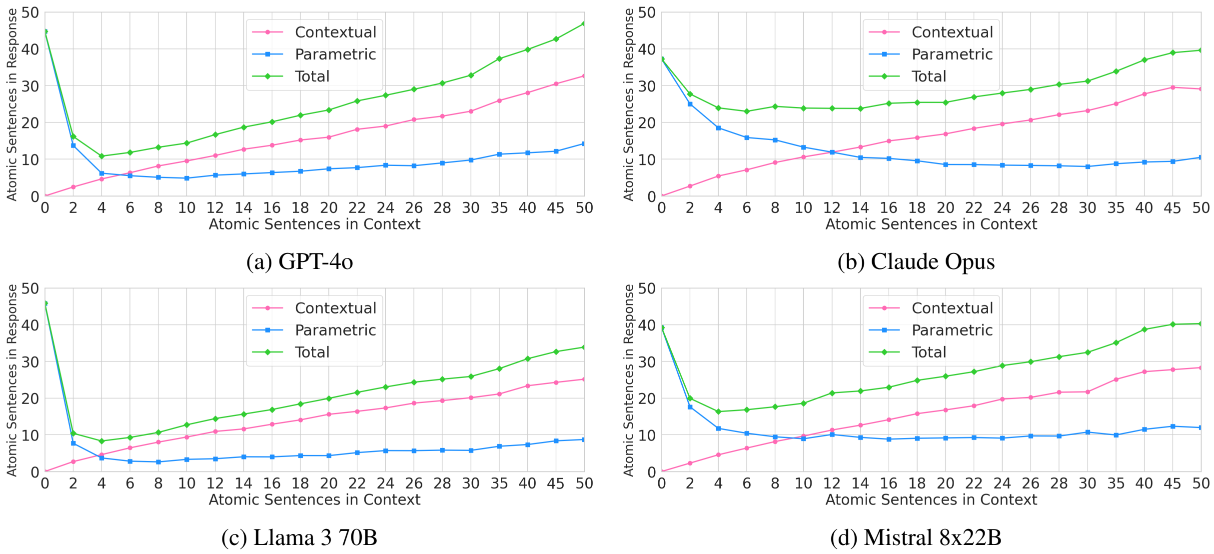
When contextual and parametric knowledge largely align with subtle differences, it becomes challenging to distinguish between the two. When no context is provided, the model's response serves as a baseline of global parametric knowledge. We calculate the overlap between the *response* (at $k = 0$) and the longest *input context* ($k = 50$) to estimate knowledge consistency. As shown in Figure 3, parametric knowledge across five topics shows high consistency with `WikiAtomic`. This shows that the model can produce similar responses whether it is provided with detailed context or not, supporting our experiments in a non-conflict setting. This differentiates our work from previous studies that relied on counterfactual datasets.

Using the *atomic contexts* as a reference, we categorize the *atomic responses* as either contextual or parametric knowledge. We use the Natural Language Inference INFUSE framework[5] (Zhang et al., 2024) to assess the faithfulness of responses. INFUSE calculates entailment scores for each sentence in what is considered the summary (responses

[3]When there is no context ($k = 0$), the question is simply "`Tell me about` {Topic}".

[4]We used the same converting to atomic sentences extraction pipeline which was validated earlier.

[5]`https://github.com/HJZnlp/Infuse`



(a) GPT-4o (b) Claude Opus (c) Llama 3 70B (d) Mistral 8x22B

Figure 4: Contextual (local), parametric (global), and total sentences in responses for (a) GPT-4o, (b) Claude Opus, (c) Llama 3 70B, and (d) Mistral 8x22B. On the x-axis, $k = 0$ serves as the baseline when no context is provided.

from the model), with scores ranging from 0 to 1. A score of 0 indicates that the sentence is not entailed from the context, whereas a score of 1 signifies fully entailed from the context. We empirically set the threshold at 0.5 based on an examination of preliminary results, thus helping us identify sentences in the response as either contextual or parametric[6]. This process involved three annotators who independently reviewed a subset of these ambiguously scored sentences. They assessed whether each of these sentences were contextual or parametric based on the definition and whether the scores align with their decisions. The threshold of 0.5 was chosen based on a consensus from this process, as it most effectively distinguished between the two categories. Preliminary experiments with other metrics such as ROUGE, METEOR, and partial match yielded similar patterns (results included in Appendix B).

Hallucination detection For sentences classified as parametric knowledge, we further assess whether they are factually accurate or hallucinated as LLMs are known to hallucinate (Huang et al., 2023; Xu et al., 2024b; Bai et al., 2024). We use the FActScore framework[7] (Min et al., 2023) that uses an external knowledge source to verify each sentence with scores ranging from 0 to 100; a higher score indicates fewer hallucinations and more factually accurate responses.

4.3 Models

We study nine models ranging from small models to state-of-the-art LLMs, including both open-source and closed-source options (for implementation details, see Appendix C). These models include: **GPT-4o** (`gpt-4o-2024-05-13`), **Claude 3 Opus** (`claude-3-opus-20240229`), **Sonnet** (`claude-3-sonnet-20240229`) and **Haiku** (`claude-3-haiku-20240307`)), **Llama 3**[8] **70B** (`Meta-Llama-3-70B-Instruct`) and **8B** (`Meta-Llama-3-8B-Instruct`), **Mixtral** 8x22b (`Mixtral-8x22B-Instruct-v0.1`), **Mistral 7B** (`Mistral-7B-Instruct-v0.2`) (Jiang et al., 2023), and **Phi-3**[9] (`Phi-3-mini-4k-instruct-gguf`) (Abdin et al., 2024).

5 Results and Analysis

Now, we present and discuss the results of our experiments.

5.1 In knowledge-consistent setting, how do models prioritize sources of knowledge?

Figure 4 shows how four models prioritize sources of knowledge (similar results were obtained from other models, plots included in Appendix D). The plots display changes in the average number of local, global, and total atomic sentences in responses as context increases. Somewhat surprisingly, we found that all models (except Phi-3 in certain cases)

[6]Most sentences were clearly either entailed or not, with only about 10% falling in the middle ambiguous range. We discuss the results of an ablation study later in Section 5.5.

[7]https://github.com/shmsw25/FActScore

[8]https://ai.meta.com/blog/meta-llama-3/

[9]https://huggingface.co/microsoft/Phi-3-mini-4k-instruct-gguf

Model	$k = 10$	$k = 20$	$k = 30$	$k = 40$	$k = 50$
GPT-4o	0.69 / 0.31	0.68 / 0.32	0.69 / 0.31	0.68 / 0.32	0.67 / 0.33
Claude 3 Opus	0.48 / 0.52	0.67 / 0.33	0.73 / 0.27	0.74 / 0.26	0.72 / 0.28
Claude 3 Sonnet	0.50 / 0.50	0.64 / 0.36	0.68 / 0.32	0.67 / 0.33	0.66 / 0.34
Claude 3 Haiku	0.69 / 0.31	0.75 / 0.25	0.75 / 0.25	0.71 / 0.29	0.72 / 0.28
Llama 3 70B	0.73 / 0.27	0.75 / 0.25	0.74 / 0.26	0.74 / 0.26	0.72 / 0.28
Llama 3 8B	0.62 / 0.38	0.63 / 0.37	0.65 / 0.35	0.65 / 0.35	0.67 / 0.33
Mixtral 8x22B	0.58 / 0.42	0.65 / 0.35	0.66 / 0.34	0.69 / 0.31	0.69 / 0.31
Mistral 7B	0.48 / 0.52	0.61 / 0.39	0.66 / 0.34	0.67 / 0.33	0.69 / 0.31
Phi-3	0.24 / 0.76	0.37 / 0.63	0.42 / 0.58	0.46 / 0.54	0.49 / 0.51
All Models	0.56 / 0.44	0.64 / 0.36	0.67 / 0.33	0.67 / 0.33	0.67 / 0.33

Table 1: Ratios of contextual/parametric knowledge in responses

show consistently similar patterns. This uniformity among the models suggests a shared underlying mechanism in processing and responding to contextual information.

With no context provided ($k = 0$), models tend to be most verbose as expected from an initially open prompt. They also peak in parametric knowledge, which drops drastically with the first four contexts. Interestingly, total response lengths generally match context lengths. From $k = 2$ onward, the total sentences and local contextual knowledge steadily increase as context increases. More contexts were utilized but none of these models utilized 100% of them in their responses. Global facts show a consistent trend: after the initial open-ended prompt, the global knowledge drops sharply and remains low, though never to zero, with some models slowly increasing or decreasing but always including some global knowledge in their responses.

Table 1 presents the proportions of contextual and parametric knowledge in responses. Larger contexts generally increase the model's reliance on contextual knowledge (about 70%) while reducing dependence on parametric knowledge (about 30%). However, in smaller contexts, models show different preferences, with some prioritizing contextual knowledge (GPT-4o, Claude Haiku, Llama 3 70B, Llama 3 8B, and Mixtral 8x22B), and others, parametric knowledge (Claude Opus, Mistral 7B, Phi-3). Moreover, the average proportion remains similar for k = 30, 40, and 50 suggesting that the ratio of contextual/parametric knowledge is maintained at these different context lengths.

5.2 Which parts of context are used?

The results of the previous experiment intrigued us, leading us to the next question: which parts of the provided context do LLMs use in their responses? Using the INFUSE framework, we reversed the input position to give each atomic sentence in the context an entailment score, considering sentences with scores greater than 0.5 to be included in the model's response. For $k > 4$, inputs were split into quartiles for clearer analysis.

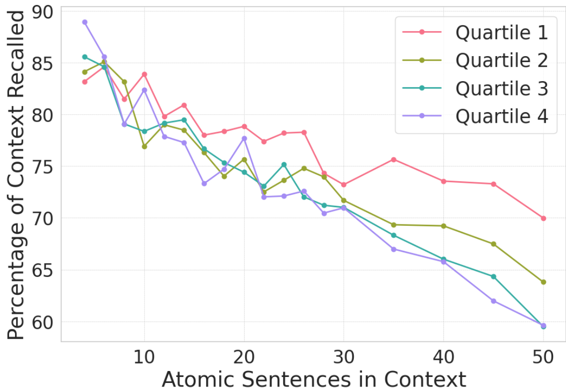


Figure 5: Percentage of each quartile of context recalled in response (GPT-4o)

Figure 5 (additional results in Appendix E) shows that for smaller contexts ($k < 10$), the model treats all portions of the context equally. As context increases, the model predominantly focuses on the first quartile. For $k \geq 25$, the preference gap between quartile 1 and all others further increases, with quartile 2 becoming the next most preferred section. We observe selective attention where certain parts of the data receive more focus than others (Liu et al., 2024), and find that models struggle more with recalling the bottom half of the data rather than the middle, further confirming the importance of initial contexts.

To determine where provided contexts are used in the responses, we divide contexts and responses into four quartiles and compare each section using cosine similarity of SBERT embeddings (Reimers and Gurevych, 2019), essentially mapping the context quartiles to response quartiles. From Figure 6, we observe that the first context quartile maps most closely to the first, and to a lesser extent, the adjacent response quartile (additional results in Ap-

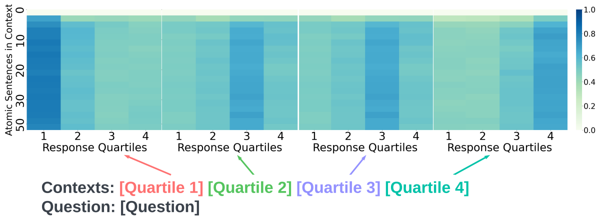


Figure 6: Mapping context quartiles to response quartiles (GPT-4o)

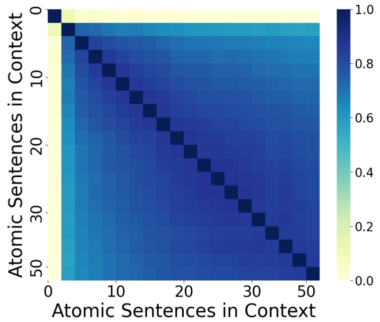
pendix F). This pattern continues with each subsequent context quartile showing the highest similarity to its corresponding response quartile, indicating that the model prefers to match the positions of data. This pattern likely mimics the natural sequence of text when describing a topic, aligning with how information is typically structured.

5.3 How similar are various types of knowledge?

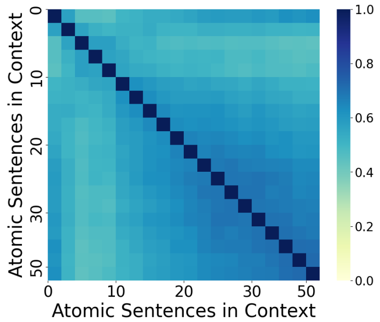
Our analysis so far suggests that LLMs consistently add parametric knowledge regardless of the amount of context provided in the question. This prompted us to investigate the relationships between the different types of knowledge. The graphs were generated as follows: For each Wikipedia topic containing 20 questions (up to 50 contexts), we obtained 20 responses. Each response contains sentences marked as contextual or parametric knowledge.

Local vs. Local We perform a pairwise comparison of **contextual knowledge** in responses across various context sizes (Figure 7a, full set of results in Appendix G). We observe that the local context remains moderately similar in smaller contexts but becomes increasingly similar with larger contexts. This pattern suggests that models tend to focus on certain types of context, often the earlier parts.

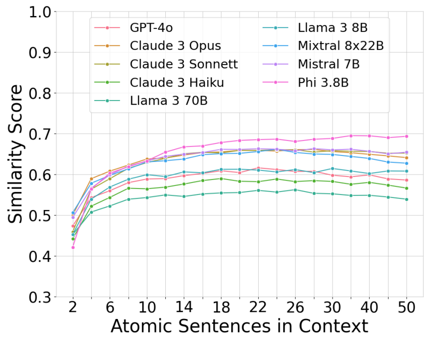
Global vs. Global Let us now turn to Figure 7b (full set of results in Appendix H) which shows pairwise similarity between **parametric knowledge** in each response. The darker plots near the diagonal line indicate that the model's parametric knowledge remains similar when context sizes are close, such as context size 8's parametric knowledge being most similar to that of sizes 6 or 10. Initially, with minimal context, models provide a larger variety of information. As context size increases, some models show increased similarity across broader ranges, suggesting responses align more closely with the overall theme of the context. This indicates that models tend to repeatedly add certain pieces of information from a small pool of



(a) Local vs. local (GPT-4o)



(b) Global vs. global (GPT-4o)



(c) Local vs. global (All models)

Figure 7: Similarity between contextual knowledge of responses obtained with varying input lengths (a), similarity between parametric knowledge of responses obtained with varying input lengths (b), and similarity between contextual and parametric knowledge (c).

knowledge, though it is not as uniformly homogeneous as local knowledge.

Local vs. Global Our analysis naturally leads us to our next question: what type of global parametric knowledge is incorporated with local contextual knowledge? Our assumption is that low similarity indicates complementary parametric knowledge, while high similarity suggests it is supplemental. Figure 7c (detailed graphs in Appendix I) shows consistent trends across all models. For smaller contexts, the similarity between **contextual knowledge and parametric knowledge** is much lower, suggesting that models incorporate complementary information not directly provided by the context. As context increases, the similarity rises to around 0.6 indicating moderate similarity, with global knowledge sharing some concepts with the provided context, but not merely echoing it.

5.4 In knowledge-consistent setting, (how much) do models hallucinate?

In Figure 8, we present the FActScore hallucination scores across all models. For smaller contexts, models have higher hallucination rate, which improves with additional context and converges with as little as 10 sentences in context. Larger models (GPT-4o, Claude Opus, Claude Sonnet, Claude Haiku, Llama 3 70B, and Mistral 8x22B) consistently show higher FActScores, indicating they are less prone to hallucinations and benefit significantly from additional context. Among smaller models, Mistral 7B and Llama 3 8B perform well, showing significant improvement with increased context, while Phi-3 tends to hallucinate more, although it also improves with context. Larger models generally stabilize at higher FActScores with fewer fluctuations, indicating a stronger ability to leverage additional context to minimize hallucinations. Smaller models show a more gradual improvement, with some like Phi-3, displaying more variability and a lower overall FActScore, suggesting they are more context-sensitive but can still improve with more information (for detailed results of false parametric knowledge for each model, see Appendix J).

5.5 Further Analyses

(Potentially) Unseen Knowledge In knowledge-consistent scenarios, models rely on a mix of contextual and parametric knowledge. But in unseen knowledge scenarios, where the context contains new information that the models have potentially not seen, how do models respond? To explore this, we create a recent example around the pro-Palestinian university protests of April 2024 and prompt the models for information on this topic.

Without context, most models stated they lacked information on the topic and provided a general background on protests related to Palestine (see Figure 9, with additional results in Appendix K). However, some models, Claude Haiku, Llama 8B, and Mixtral 22x8B, all confidently respond to the prompt as if they have seen this knowledge in their pre-training data. We hypothesize that this may be due to similar events that have occurred in the past that the model has seen in its pretraining data. With just two contexts, all models referenced the context but some included facts not in the context.

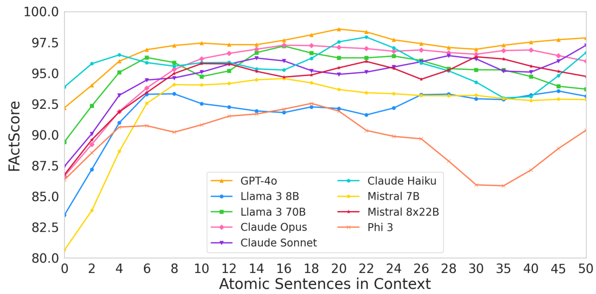


Figure 8: FActScore across all LLMs (higher score indicates lower hallucination)

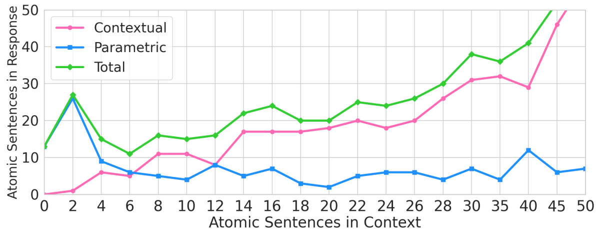


Figure 9: Unseen knowledge results (GPT-4o)

With 50 contexts, they effectively incorporated the contextual knowledge without much additional information. Our findings suggest that care must be taken when querying LLMs on new information as without context or with minimal context, models may still confidently provide seemingly accurate information not in their parametric knowledge.

Prompt Sensitivity Our naturalistic *semi-restrict* question prompt ("With this information, tell me about [topic].") encourages the models to consider the provided contexts while still allowing them to use their parametric knowledge. Considering that LLMs are sensitive to prompts (Lu et al., 2021), we tested two alternative phrasings: a *no-restrict* prompt which allows the models complete freedom to draw on their parametric and contextual knowledge as they see fit ("Tell me about [topic]") and a *strict* prompt which tries to restrict the models to use only the contexts provided ("Using the provided context only, tell me about [topic].").

To conduct this ablation study, we randomly selected 20 topics from our dataset, resulting in a total of 400 instances of varying context sizes, and obtained responses different prompts. While the *no-restrict* plots look similar to the *semi-restrict* prompts presented earlier, the *strict* plots show that the models are entirely focused on contextual knowledge with very little parametric knowledge (supporting graphs in Appendix L). These results confirm model sensitivity to prompts and highlight that, with simple prompt adjustments, models can be guided to leverage different ratios of contextual and parametric knowledge.

Disregard Ambiguous Sentences When we used INFUSE to generate entailment scores, a small number of sentences had scores close to 0.5, indicating lower confidence in categorizing them as entailed or not. We decided to exclude these sentences to assess whether the patterns from Section 5.1 remained valid. After ignoring sentences with scores between 0.3 and 0.7 (see Appendix M), the patterns persisted, confirming the robustness of our approach and the reliability of the INFUSE framework for this task.

6 Discussion

Understanding how LLMs utilize contextual and parametric knowledge is important for real-world knowledge-consistent use cases. We summarize our key findings and their implications:

- Models process context very similarly suggesting a standardized approach to context processing across all selected models.
- Context is never fully utilized and some parametric knowledge is always included highlighting the need for developing models that utilize input contexts more deterministically.
- Earlier information is prioritized in longer contexts, with responses following the order of presented information highlighting the importance of organizing context effectively.
- Responses tend to include similar contextual knowledge but supplemental parametric knowledge.
- Hallucinations decrease as context increases implying that more context helps models generate more accurate responses, reducing reliance on potentially incorrect parametric knowledge.

7 Related Work

Contextual Grounding A persistent challenge for LLMs lies in reconciling contradictory or superseded information between the provided context and their internal knowledge base (Li et al., 2022; Zhou et al., 2023a), with most recent work exploring this interplay between local and global knowledge in knowledge-conflict scenarios using counterfactual datasets (Si et al., 2022; Qian et al., 2023; Feng et al., 2024; Yang et al., 2024). The preference for contextual or parametric knowledge in counterfactual settings is not straightforward. Larger models seem better at adapting to counterfactual context, while smaller ones often prioritize their learned knowledge (Si et al., 2022). When faced with counterfactual prompts, LLMs adjust their responses to align with the given context, even if it contradicts their pretrained parametric knowledge (Li et al., 2022; Feng et al., 2024). In other cases, however, steering LLMs away from generating outputs aligned with their vast but potentially inaccurate pretrained knowledge, even when contradicted by the context, has been challenging (Zhou et al., 2023b). Our work contributes to contextual grounding studies by prompting LLMs in a *knowledge-consistent* scenario and is the first to systematically analyze how the amount of atomic information in a prompt affects context utilization.

Hallucination There has been extensive work on hallucination detection with methods falling into two general categories: internal parameter based methods (Chen et al., 2024; Su et al., 2024; Duan et al., 2024) and external response based methods (Manakul et al., 2023; Min et al., 2023; Yu et al., 2024; Manakul et al., 2023; Sun et al., 2024). Similar to our work, Hu et al. (2024) analyze the tendency of LLMs to generate factual hallucinations in the presence of varying local context in question answering. Unlike our approach, they study the effects of entirely switching out local context, whereas our approach incrementally increases the amount of context while comparing hallucination tendency between these responses.

8 Conclusion

Understanding how LLMs handle different context sizes is crucial for developing robust models. Our evaluation of nine widely used LLMs with our `WikiAtomic` dataset showed that all models process context similarly, balancing contextual and parametric knowledge, while adjusting response lengths consistently. They never use all provided contexts, always include some parametric knowledge, and prioritize information sequentially. As context increases, while contextual knowledge remains similar, parametric knowledge becomes more aligned with the context, and hallucination rates decrease. These insights highlight the importance of organizing context effectively and developing models that utilize input more deterministically.

Limitations

Our work, while yielding some interesting findings, is not without limitations. Our method of extracting atomic sentences from Wikipedia articles often split the source sentences into multiple, occasionally creating sentences that begin with an indirect reference to a subject. Because of the natural flow of presenting information in such a format, we did not randomize the order. Further work could look at how the order of contexts used by models compared when the provided contexts are shuffled.

We utilized the INFUSE method to classify model response contexts into ‘contextual’ and ‘parametric’. We empirically set the threshold for determining the categorization based on manual verification. Further work could look at more sophisticated methods to set this threshold for even better results. A limitation of the metric we adopt for hallucination detection is that it relies on a single knowledge source Wikipedia as its knowledge source. Information that is factually accurate, but not present in the knowledge source could be incorrectly classified as a hallucination.

Acknowledgments

We thank the anonymous reviewers as well as the members of PortNLP lab for their insightful comments that helped improve this paper. This research was supported by the National Science Foundation grant SAI-P 2228783.

References

Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Qin Cai, Martin Cai, Caio César Teodoro Mendes, Weizhu Chen, Vishrav Chaudhary, Dong Chen, Dongdong Chen, Yen-Chun Chen, Yi-Ling Chen, Parul Chopra, Xiyang Dai, Allie Del Giorno, Gustavo de Rosa, Matthew Dixon, Ronen Eldan, Victor Fragoso, Dan Iter, Mei Gao, Min Gao, Jianfeng Gao, Amit Garg, Abhishek Goswami, Suriya Gunasekar, Emman Haider, Junheng Hao, Russell J. Hewett, Jamie Huynh, Mojan Javaheripi, Xin Jin, Piero Kauffmann, Nikos Karampatziakis, Dongwoo Kim, Mahoud Khademi, Lev Kurilenko, James R. Lee, Yin Tat Lee, Yuanzhi Li, Yunsheng Li, Chen Liang, Lars Liden, Ce Liu, Mengchen Liu, Weishung Liu, Eric Lin, Zeqi Lin, Chong Luo, Piyush Madan, Matt Mazzola, Arindam Mitra, Hardik Modi, Anh Nguyen, Brandon Norick, Barun Patra, Daniel Perez-Becker, Thomas Portet, Reid Pryzant, Heyang Qin, Marko Radmilac, Corby Rosset, Sambudha Roy, Olatunji Ruwase, Olli Saarikivi, Amin Saied, Adil Salim, Michael Santacroce, Shital Shah, Ning Shang, Hiteshi Sharma, Swadheen Shukla, Xia Song, Masahiro Tanaka, Andrea Tupini, Xin Wang, Lijuan Wang, Chunyu Wang, Yu Wang, Rachel Ward, Guanhua Wang, Philipp Witte, Haiping Wu, Michael Wyatt, Bin Xiao, Can Xu, Jiahang Xu, Weijian Xu, Sonali Yadav, Fan Yang, Jianwei Yang, Ziyi Yang, Yifan Yang, Donghan Yu, Lu Yuan, Chengruidong Zhang, Cyril Zhang, Jianwen Zhang, Li Lyna Zhang, Yi Zhang, Yue Zhang, Yunan Zhang, and Xiren Zhou. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *Preprint*, arXiv:2404.14219.

Zechen Bai, Pichao Wang, Tianjun Xiao, Tong He, Zongbo Han, Zheng Zhang, and Mike Zheng Shou. 2024. Hallucination of multimodal large language models: A survey. *Preprint*, arXiv:2404.18930.

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.

Chao Chen, Kai Liu, Ze Chen, Yi Gu, Yue Wu, Mingyuan Tao, Zhihang Fu, and Jieping Ye. 2024. Inside: Llms' internal states retain the power of hallucination detection. *Preprint*, arXiv:2402.03744.

Hanyu Duan, Yi Yang, and Kar Yan Tam. 2024. Do llms know about hallucination? an empirical investigation of llm's hidden states. *Preprint*, arXiv:2402.09733.

Shangbin Feng, Weijia Shi, Yuyang Bai, Vidhisha Balachandran, Tianxing He, and Yulia Tsvetkov. 2024. Knowledge card: Filling LLMs' knowledge gaps with plug-in specialized language models. In *The Twelfth International Conference on Learning Representations*.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. 2020. The pile: An 800gb dataset of diverse text for language modeling. *Preprint*, arXiv:2101.00027.

Benjamin Heinzerling and Kentaro Inui. 2021. Language models as knowledge bases: On entity representations, storage capacity, and paraphrased queries. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1772–1791, Online. Association for Computational Linguistics.

Xinshuo Hu, Baotian Hu, Dongfang Li, Xiaoguang Li, and Lifeng Shang. 2024. Does the generator mind its contexts? an analysis of generative model faithfulness under context transfer. *Preprint*, arXiv:2402.14488.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2023. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *Preprint*, arXiv:2311.05232.

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 Comput. Surv.*, 55(12).

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. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Yekyung Kim, Yapei Chang, Marzena Karpinska, Aparna Garimella, Varun Manjunatha, Kyle Lo, Tanya Goyal, and Mohit Iyyer. 2024. Fables: Evaluating faithfulness and content selection in book-length summarization. *Preprint*, arXiv:2404.01261.

Kalpesh Krishna, Aurko Roy, and Mohit Iyyer. 2021. Hurdles to progress in long-form question answering. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4940–4957, Online. Association for Computational Linguistics.

Daliang Li, Ankit Singh Rawat, Manzil Zaheer, Xin Wang, Michal Lukasik, Andreas Veit, Felix Yu, and Sanjiv Kumar. 2022. Large language models with controllable working memory. *arXiv preprint arXiv:2211.05110*.

Yongqi Li, Mayi Xu, Xin Miao, Shen Zhou, and Tieyun Qian. 2024. Prompting large language models for counterfactual generation: An empirical study. *Preprint*, arXiv:2305.14791.

Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173.

Yixin Liu, Alex Fabbri, Pengfei Liu, Yilun Zhao, Linyong Nan, Ruilin Han, Simeng Han, Shafiq Joty, Chien-Sheng Wu, Caiming Xiong, and Dragomir Radev. 2023. Revisiting the gold standard: Grounding summarization evaluation with robust human evaluation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4140–4170, Toronto, Canada. 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.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 2021. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. *arXiv preprint arXiv:2104.08786*.

Alexandra Sasha Luccioni and Joseph D. Viviano. 2021. What's in the box? a preliminary analysis of undesirable content in the common crawl corpus. *Preprint*, arXiv:2105.02732.

Junliang Luo, Tianyu Li, Di Wu, Michael Jenkin, Steve Liu, and Gregory Dudek. 2024. Hallucination detection and hallucination mitigation: An investigation.

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, 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.

Ella Neeman, Roee Aharoni, Or Honovich, Leshem Choshen, Idan Szpektor, and Omri Abend. 2022. Disentqa: Disentangling parametric and contextual knowledge with counterfactual question answering. *Preprint*, arXiv:2211.05655.

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.

Cheng Qian, Xinran Zhao, and Sherry Tongshuang Wu. 2023. "merge conflicts!" exploring the impacts of external distractors to parametric knowledge graphs. *Preprint*, arXiv:2309.08594.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. 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 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Chenglei Si, Zhe Gan, Zhengyuan Yang, Shuohang Wang, Jianfeng Wang, Jordan Boyd-Graber, and Lijuan Wang. 2022. Prompting gpt-3 to be reliable. *arXiv preprint arXiv:2210.09150*.

Weihang Su, Changyue Wang, Qingyao Ai, Yiran Hu, Zhijing Wu, Yujia Zhou, and Yiqun Liu. 2024. Unsupervised real-time hallucination detection based on the internal states of large language models. *CoRR*, abs/2403.06448.

Yuhong Sun, Zhangyue Yin, Qipeng Guo, Jiawen Wu, Xipeng Qiu, and Hui Zhao. 2024. Benchmarking hallucination in large language models based on unanswerable math word problem. *Preprint*, arXiv:2403.03558.

Katherine Tian, Eric Mitchell, Huaxiu Yao, Christopher D. Manning, and Chelsea Finn. 2023. Fine-tuning language models for factuality. *ArXiv*, abs/2311.08401.

Cunxiang Wang, Xiaoze Liu, Yuanhao Yue, Xiangru Tang, Tianhang Zhang, Cheng Jiayang, Yunzhi Yao, Wenyang Gao, Xuming Hu, Zehan Qi, Yidong Wang, Linyi Yang, Jindong Wang, Xing Xie, Zheng Zhang, and Yue Zhang. 2023. Survey on factuality in large language models: Knowledge, retrieval and domain-specificity.

Rongwu Xu, Zehan Qi, Cunxiang Wang, Hongru Wang, Yue Zhang, and Wei Xu. 2024a. Knowledge conflicts for llms: A survey. *Preprint*, arXiv:2403.08319.

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

Linyao Yang, Hongyang Chen, Zhao Li, Xiao Ding, and Xindong Wu. 2024. Give us the facts: Enhancing large language models with knowledge graphs for fact-aware language modeling. *IEEE Transactions on Knowledge and Data Engineering*, PP:1–20.

Jifan Yu, Xiaozhi Wang, Shangqing Tu, Shulin Cao, Daniel Zhang-Li, Xin Lv, Hao Peng, Zijun Yao, Xiaohan Zhang, Hanming Li, Chunyang Li, Zheyuan Zhang, Yushi Bai, Yantao Liu, Amy Xin, Kaifeng Yun, Linlu GONG, Nianyi Lin, Jianhui Chen, Zhili Wu, Yunjia Qi, Weikai Li, Yong Guan, Kaisheng Zeng, Ji Qi, Hailong Jin, Jinxin Liu, Yu Gu, Yuan Yao, Ning Ding, Lei Hou, Zhiyuan Liu, Xu Bin, Jie Tang, and Juanzi Li. 2024. KoLA: Carefully benchmarking world knowledge of large language models. In *The Twelfth International Conference on Learning Representations*.

Huajian Zhang, Yumo Xu, and Laura Perez-Beltrachini. 2024. Fine-grained natural language inference based faithfulness evaluation for diverse summarisation tasks. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1701–1722, St. Julian's, Malta. Association for Computational Linguistics.

Wenxuan Zhou, Sheng Zhang, Hoifung Poon, and Muhao Chen. 2023a. Context-faithful prompting for large language models. *Preprint*, arXiv:2303.11315.

Wenxuan Zhou, Sheng Zhang, Hoifung Poon, and Muhao Chen. 2023b. Context-faithful prompting for large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 14544–14556, Singapore. Association for Computational Linguistics.

Example Contexts

Contexts: `Mariah Carey was born on March 27, 1969. She was born in Huntington, New York. She is a highly celebrated American artist. She is known for her work as a singer. She is known for her work as a songwriter...(continue)`

A Wikipedia Article Contexts Extraction Prompt

Figure A.1 shows the exact prompt we used to extract atomic facts from Wikipedia articles. We have tested multiple versions of prompts to break down sentences into atomic sentences. We found if we only asking the model directly to break down sentences to atomic sentences, or only adding the definition of atomic sentence and asking it to break down sentences. The model performed very poorly, the main problem with these approaches was this version of atomic sentences weren't atomic enough. For example:

`April is the fourth month of the year in both the Julian and Gregorian calendars.`

This sentence could be treated as a single information. But this definitly could be further broke down to smaller pieces:

`1. April is the fourth month of the year in the Julian calendar`
`2. April is the fourth month of the year in the Gregorian calendar.`

Figure A.1's prompt was the most effective version in our experiments, it could handle the above example very well and at the same time add the appropriate subject to each atomic sentences. We used this version of prompt also to break down responses from each model into atomic sentences (Figure A.2).

B Alternative Metrics for Contextual/Parametric Knowledge Evaluation

Figure B.3 shows preliminary results using ROUGE-L, METEOR, and a partial match approach to evaluate GPT-4o model responses. In the figure, local facts and global facts represent contextual knowledge and parametric knowledge, respectively.

C Implementation Detail

Here we go through implementation details, model parameters. We used default parameters across all models where temperature and top_p set to 1, presence_penalty and frequency_penalty set to 0. For atomic sentence extraction both for Wikipedia articles and models responses, we set GPT4-o's the max_tokens to 2048 because the output is a JSON object that contains 50 atomic sentences. We think this number is sufficient for this size of output. When generating responses from different models once we have all the questions set up, we set the max_token to 512 to mimic the real world application setting.

API calls were made to OpenAI and Anthropic to obtain responses from GPT-4o, Claude 3 Opus, Sonnet, and Haiku, completing in 4 hours. For large open-source models Mixtral 8x22B and Llama 3 70B, we used a third-party service, which took 2 hours. The INFUSE evaluation of responses from 9 LLMs was conducted on 9 A100 GPUs over 10 hours. Inference for smaller models (Llama 3 8B and Phi 3) was performed on an M1 Ultra, taking 5 hours.

D Model-specific contextual/parametric evaluation

Figure D.4 shows the rest of models' contextual/parametric evaluation results.

E Context Focus Position Analysis

Figure E.5 shows the percentage of contexts were present in responses grouped by quartiles for each models.

F Context Response Mapping Analysis

Figure F.6 shows results of each model's quartile contexts in response areas.

G Local vs. Local Similarity

Figure G.7 shows contextual knowledge similarity across rest of 9 models other than the Claude Sonnet mentioned in the paper.

H Parametric vs. Parametric Knowledge in Response

Figure H.8 shows parametric knowledge similarity across each set of example for rest of the models not mentioned in main paper.

I Contextual vs. Parametric Knowledge in Response

Figure I.9 shows similarity score for contextual and parametric knowledge in each response across different model.

J FactScore False Parametric Knowledge

Figure J.10 show number of false parametric knowledge for all models.

K New knowledge

Figure K.11, K.12 show how each behave when asking question about a new information.

L Further Analyses

This section shows the difference in the model's use of local contexts when using a strict vs unrestricted prompt (Figure L.13, L.14 and L.15). The strict prompt instructs the model to only use the provided context when answering the question. The unrestricted prompt provides the context but doesn't instruct the model to use it in any way.

M Contextual vs. Parametric while disregarding ambiguous sentences

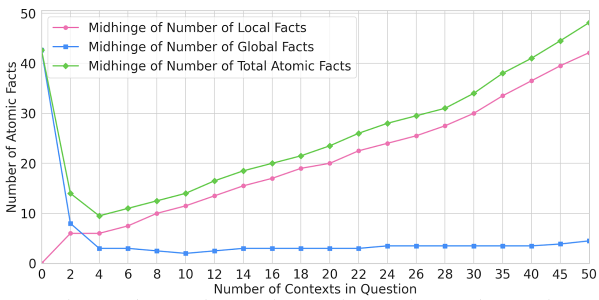
This section shows the alternative graphs for number of contextual and parametric knowledge in each model (Figure M.16) while disregarding sentences received INFUSE score between 0.3 and 0.7.

```
(No list, not bullet points, everything should be on the same line)
Definition of Atomic: An atomic sentence is a type of declarative sentence which is
either true or false, also referred to as a proposition, statement, or truth-bearer. It
cannot be broken down into simpler sentences without losing its meaning.

You will do multiple passes for first 60 sentences, using appropriate NLP method is
needed, for each sentence:
The first pass: remove comma in the sentence, rewrite the sentence into multiple smaller
sentences if needed.
The second pass: remove 'and' and 'or' in the sentene, rewrite the sentence into
multiple smaller sentences if needed.
The third pass: replace indirect references with direct references(topic word) to
maintain clarity and focus on the text’s main topic.
The fourth pass: separate temporal information (dates, times) from the main action into
distinct sentences.
The final pass: make sure each sentence contains extactly only one information. Nothing
more than one information, even the inforamtion that are dependend on each other.

The goal of these passes is to break down each long sentences into very small atomic
sentences that contains one single inseperable information.

Output Format:
A JSON object with the following elements:
atomic_sentences: A list of 60 atomic sentences.
count: The number of sentences in the atomic_sentences.
Not in JSON output but you need to think:
The process of each pass, are you sure you removed all commas, are you sure you removed
all 'and' and 'or', are you sure you replaces all indrect reference to actual topic
word.
```

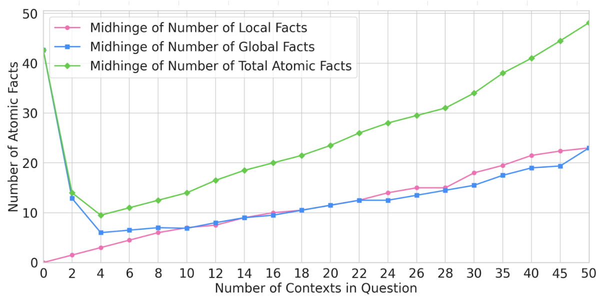
Figure A.1: Prompt to extract atomic facts from article

```
(No list, not bullet points, everything should be on the same line)
Definition of Atomic: An atomic sentence is a type of declarative sentence which is
either true or false, also referred to as a proposition, statement, or truth-bearer. It
cannot be broken down into simpler sentences without losing its meaning.

You will do multiple passes for following text, using appropriate NLP method is needed,
for each sentence:
The first pass: remove comma in the sentence, rewrite the sentence into multiple smaller
sentences if needed.
The second pass: remove 'and' and 'or' in the sentene, rewrite the sentence into
multiple smaller sentences if needed.
The fourth pass: separate temporal information (dates, times) from the main action into
distinct sentences.
The final pass: make sure each sentence contains extactly only one information.
The goal of these passes is to break down each long sentences into very small atomic
sentences that contains one single inseperable information.

Output Format:
A JSON object with the following elements:
atomic_sentences: A list of all atomic sentences.
count: The number of sentences in the atomic_sentences.
Not in JSON output but you need to think:
The process of each pass, are you sure you removed all commas, are you sure you removed
all 'and' and 'or', are you sure you replaces all indrect reference to actual topic
word.
```

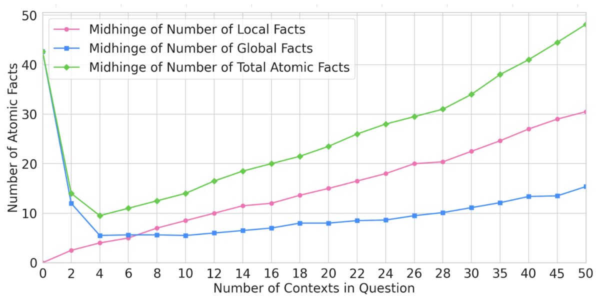
Figure A.2: Response Atomization Prompt



(a) Partial match approach results on GPT-4o

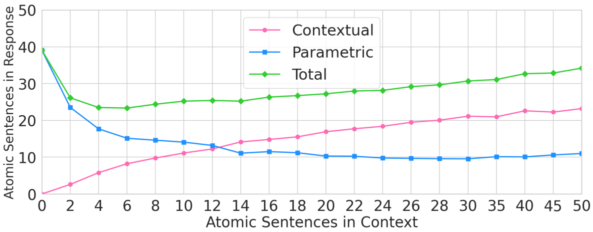


(b) ROUGE-L approach results on GPT-4o

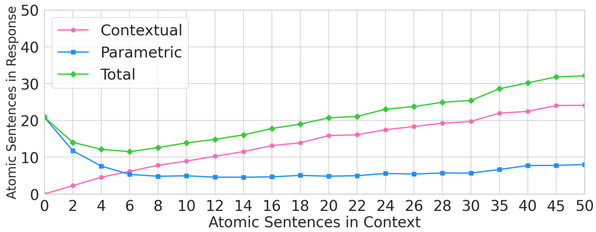


(c) METEOR approach results on GPT-4o

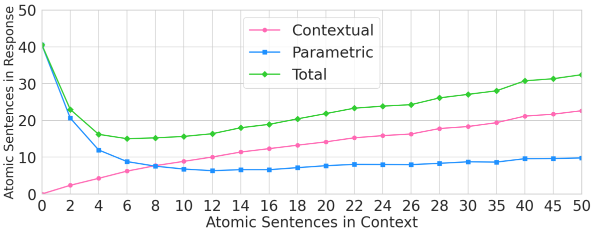
Figure B.3: Alternative methods to evaluate contextual/parametric knowledge



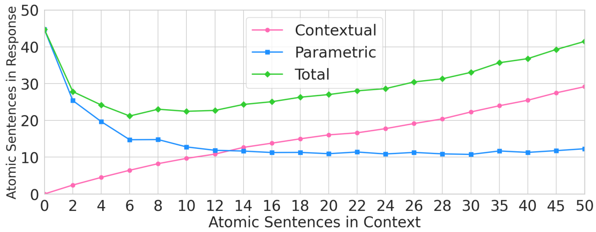
(a) Claude Sonnet



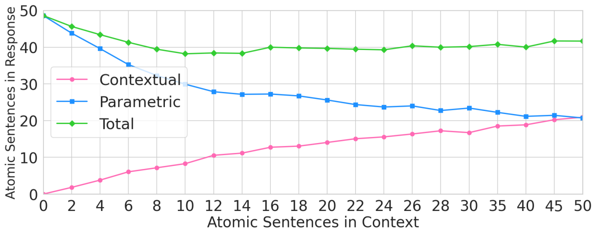
(b) Claude Haiku



(c) Llama 3 8B

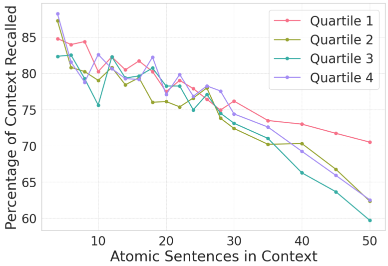


(d) Mistral 7B

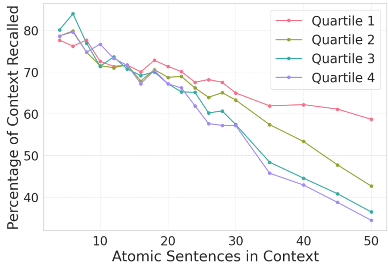


(e) Phi-3 3

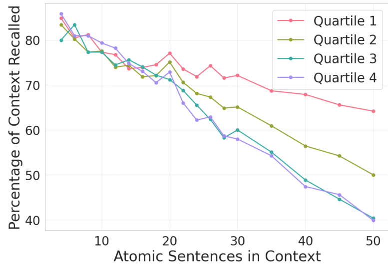
Figure D.4: Rest of models' combined local, global, and total facts figures



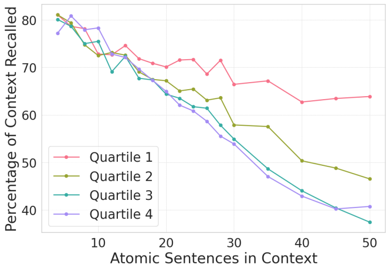
(a) Claude Opus contexts positions



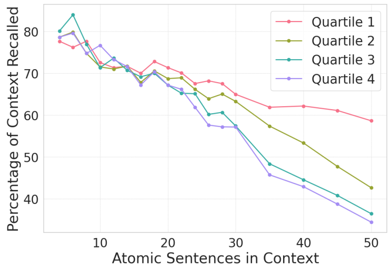
(b) Claude Sonnet contexts positions



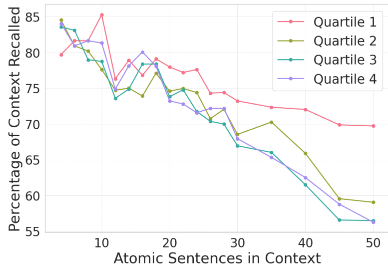
(c) Claude Haiku contexts positions



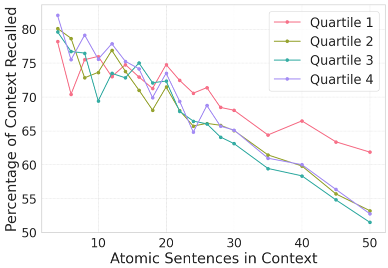
(d) Llama 3 70B contexts positions



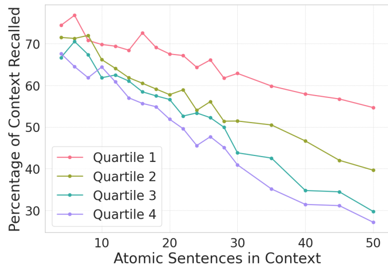
(e) Llama 3 8B contexts positions



(f) Mixtral 8x22B contexts positions

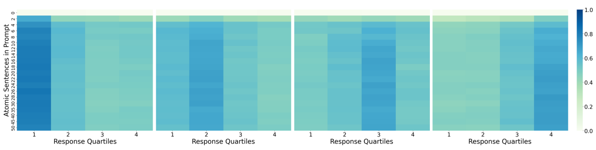


(g) Mixtral 7B contexts positions

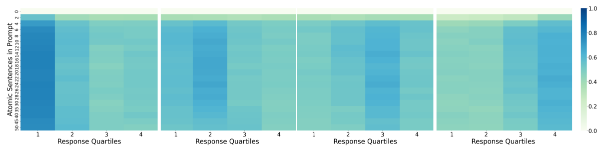


(h) Phi 3 contexts positions

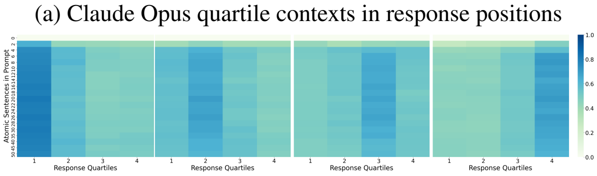
Figure E.5: Global context positions for various models



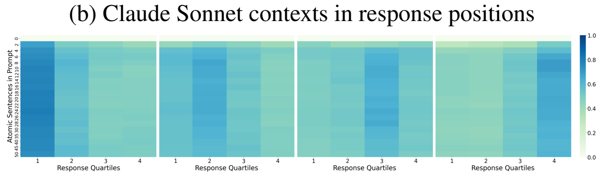
(a) Claude Opus quartile contexts in response positions



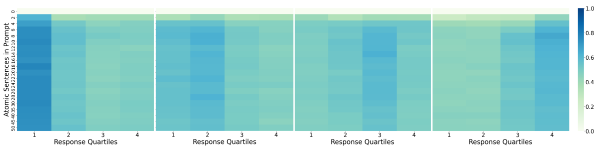
(b) Claude Sonnet contexts in response positions



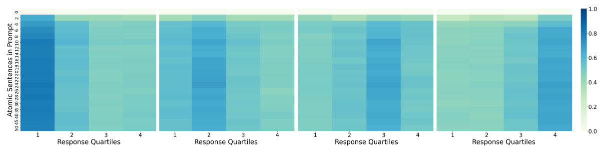
(c) Claude Haiku contexts in response positions



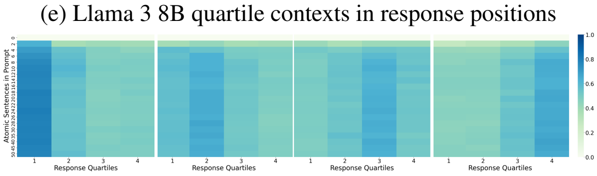
(d) Llama 3 70B contexts in response positions



(e) Llama 3 8B quartile contexts in response positions



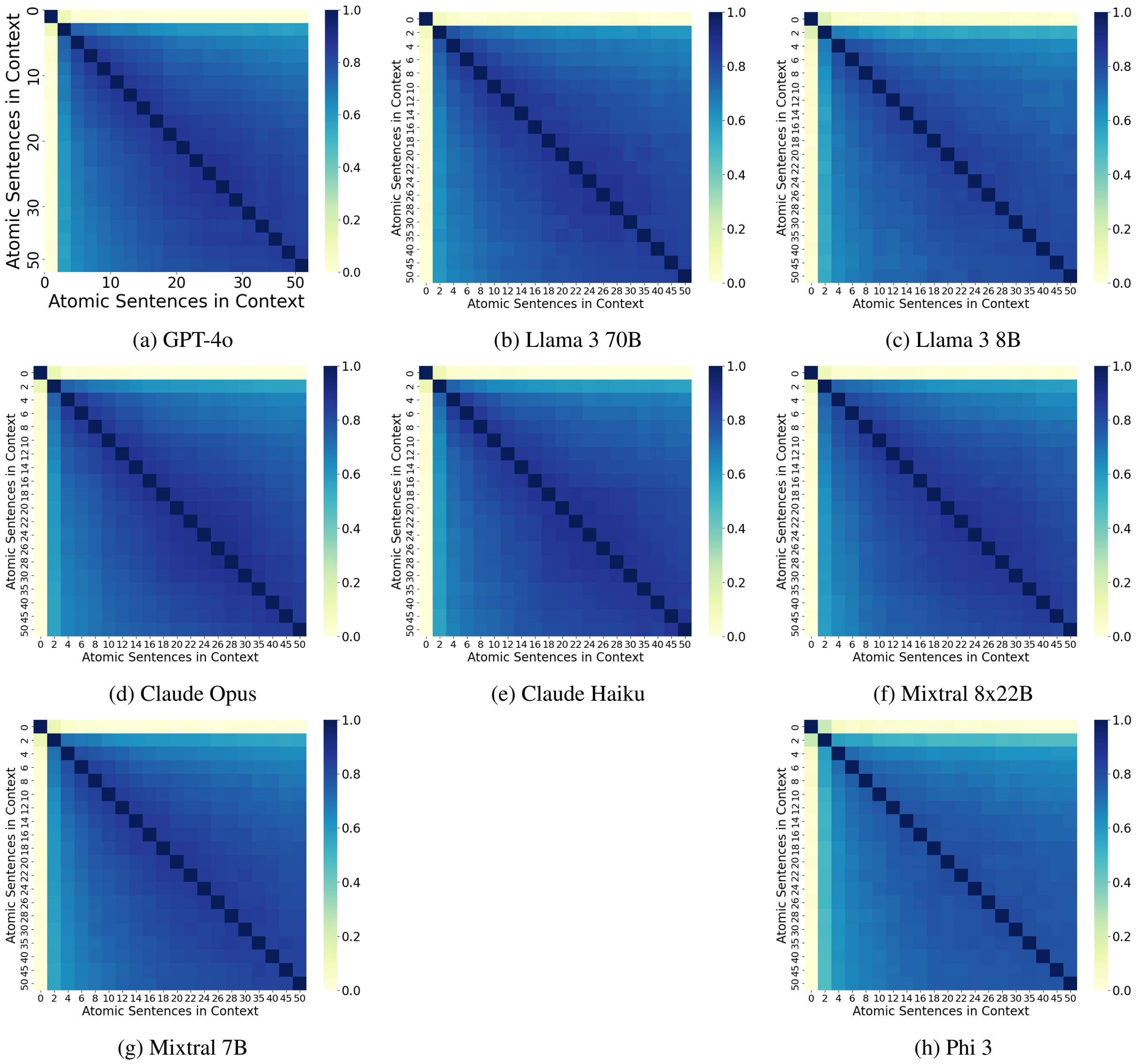
(f) Mixtral 8x22B contexts in response positions



(g) Mistral 7B quartile contexts in response positions

(h) Phi quartile contexts in response positions

Figure F.6: Quartile contexts in response positions for various models

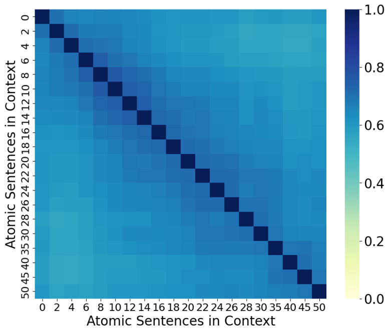


(a) GPT-4o (b) Llama 3 70B (c) Llama 3 8B

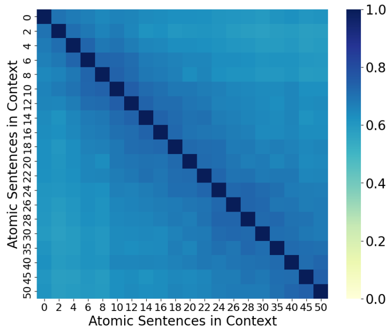
(d) Claude Opus (e) Claude Haiku (f) Mixtral 8x22B

(g) Mixtral 7B (h) Phi 3

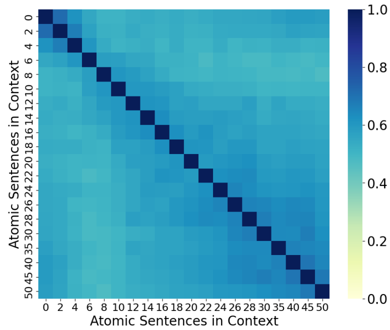
Figure G.7: Similarity between contextual knowledge within a set of example for each model



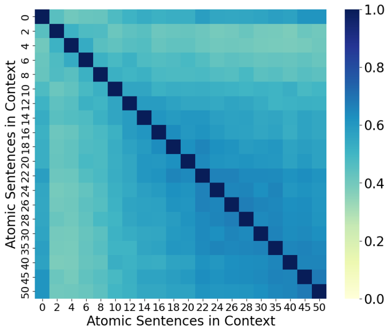
(a) Opus global vs global



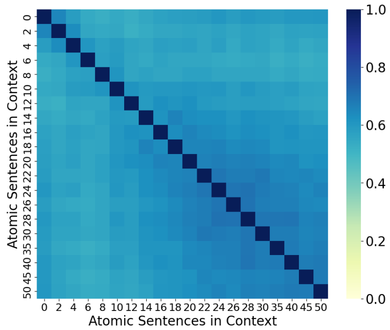
(b) Sonnet global vs global



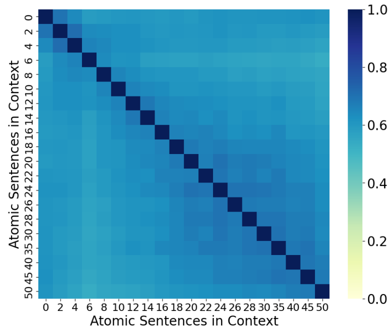
(c) Haiku global vs global



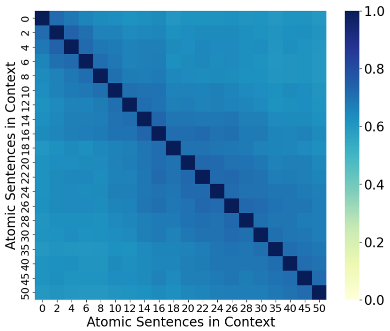
(d) Llama 70B global vs global



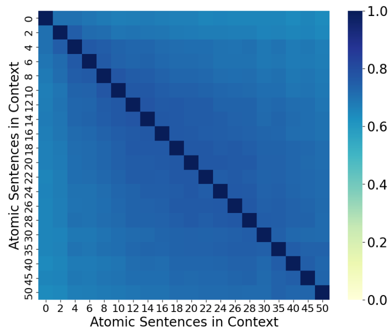
(e) Llama 7B global vs global



(f) Mistral 8x22B global vs global

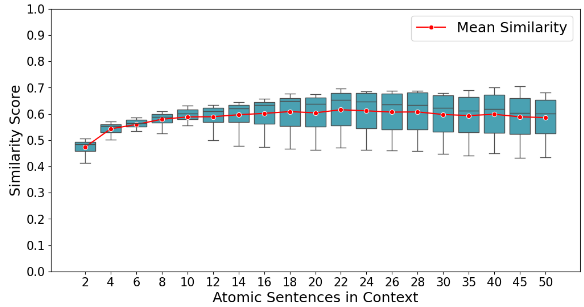


(g) Mistral 7B global vs global

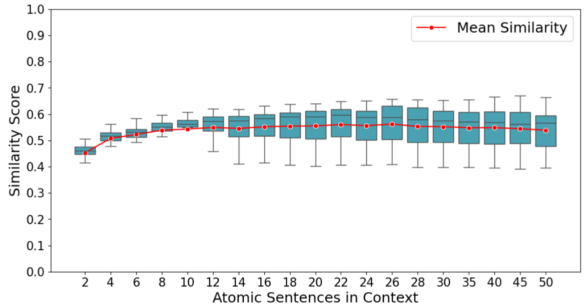


(h) Phi global vs global

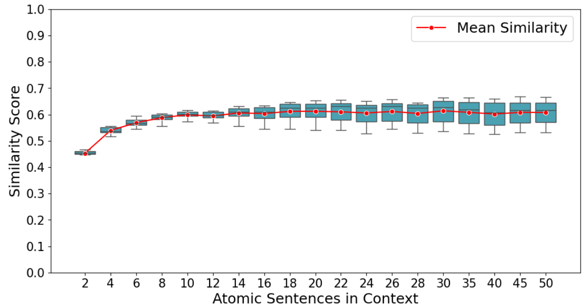
Figure H.8: Global vs Global Similarity Analysis for Various Models



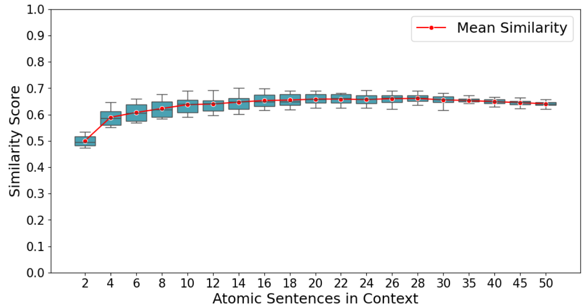
(a) GPT-4o



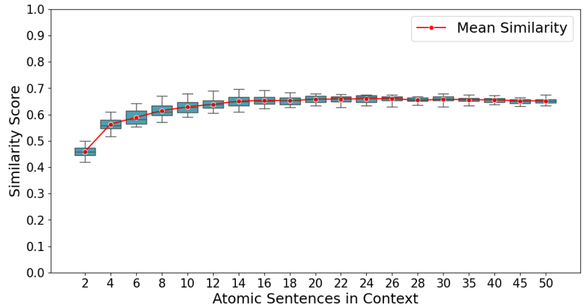
(b) Llama 3 70B



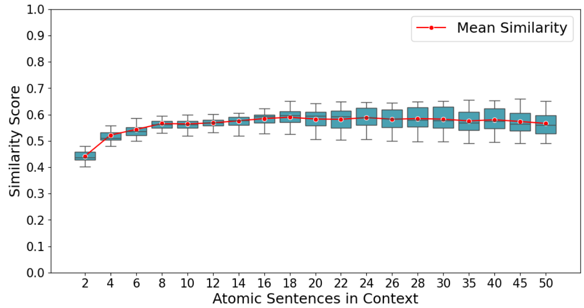
(c) Llama 3 8B



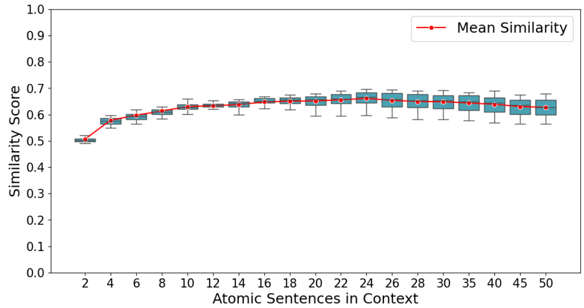
(d) Claude Opus



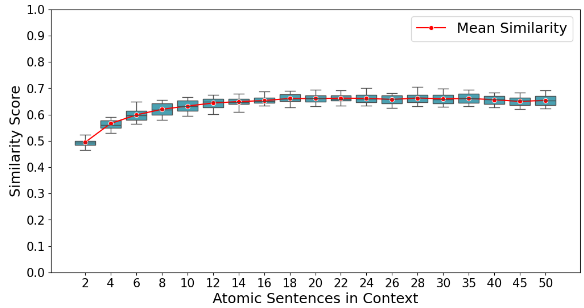
(e) Claude Sonnet



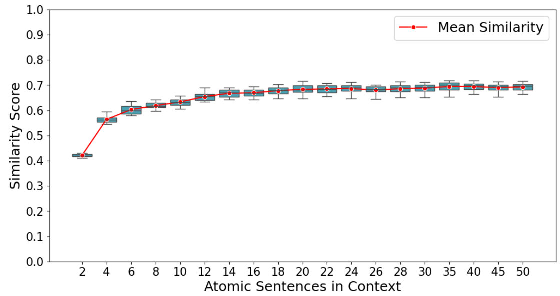
(f) Claude Haiku



(g) Mixtral 8x22B

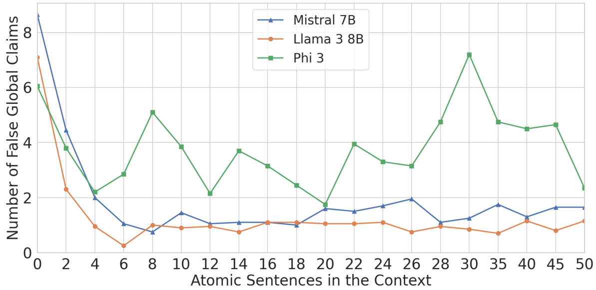


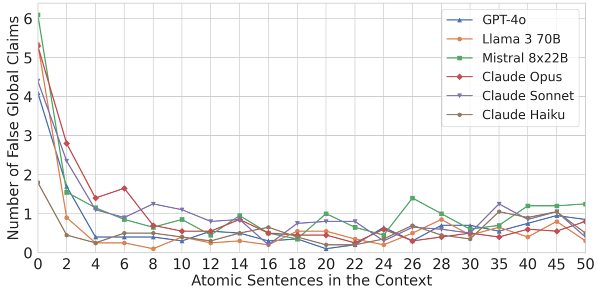
(h) Mixtral 7B



(i) Phi 3

Figure I.9: Similarity between contextual knowledge and parametric knowledge within a set of example for each model

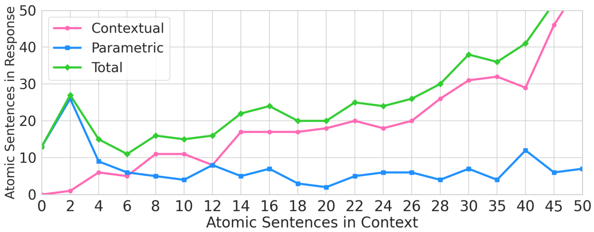




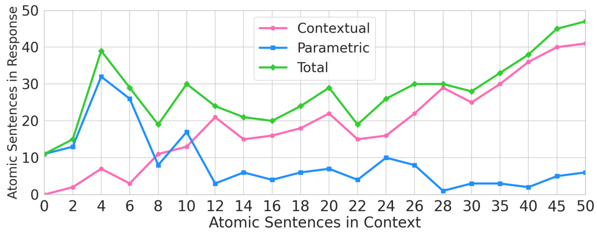
(a) Number of false parametric knowledge for Mistal 7B, Llama 3 8B and Phi 3

(b) Number of false parametric knowledge for GPT-4o, Llama 3 70B, Mistral 8x22B and Claude Opus, Sonnet and Haiku

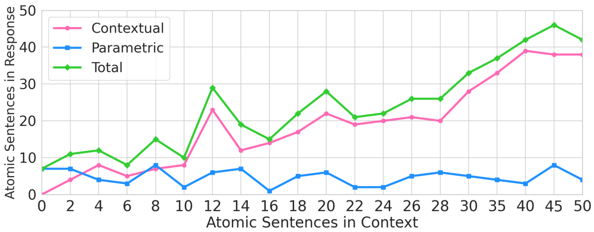
Figure J.10: Number of false parametric knowledge across all LLMs



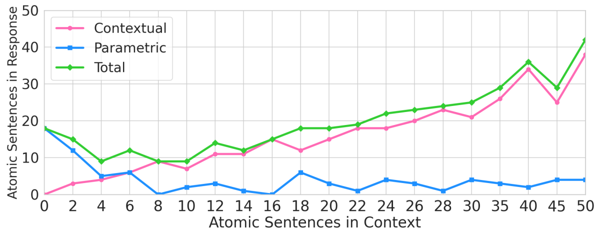
(a) New knowledge (GPT-4o)



(b) New knowledge (Opus)

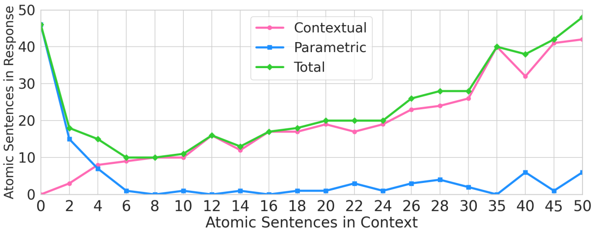


(c) New knowledge (Sonnet)

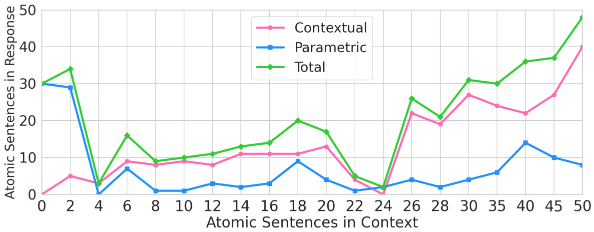


(d) New knowledge (Haiku)

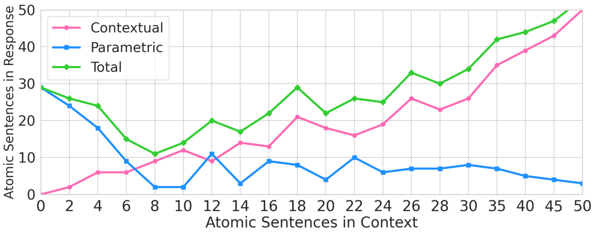
Figure K.11: New knowledge (All Claude models)



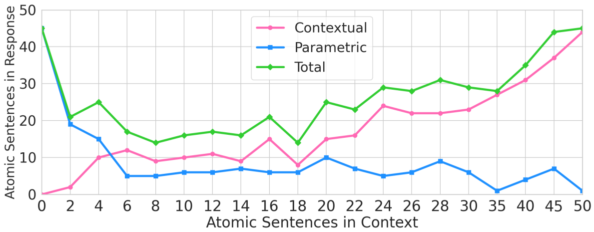
(a) New knowledge (Llama 3 70B)



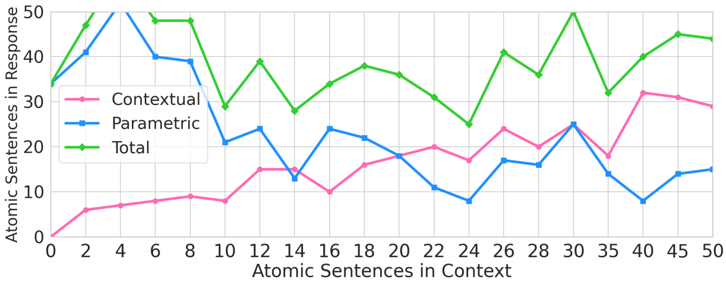
(b) New knowledge (Llama 3 8B)



(c) New knowledge (Mixtral 8x22B)

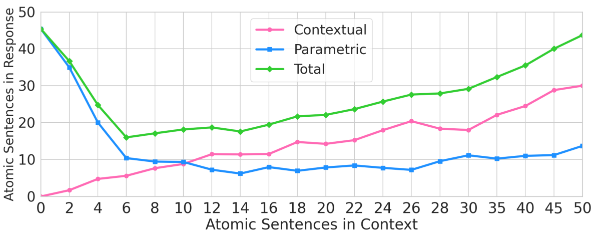


(d) New knowledge (Mistral 7B)

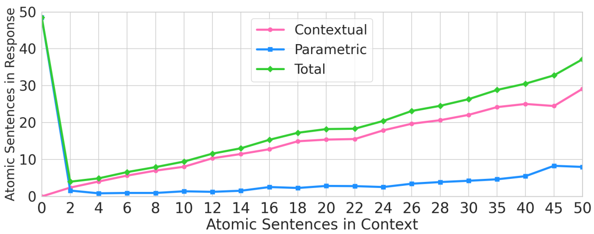


(e) New knowledge (Phi 3)

Figure K.12: New knowledge (Phi 3, Mistral AI and Llama 3 models)

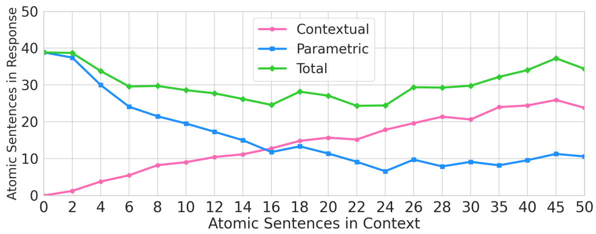


(a) No restriction in question (GPT-4o)

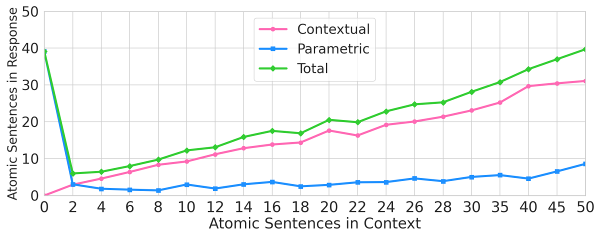


(b) Strict restriction in question (GPT-4o)

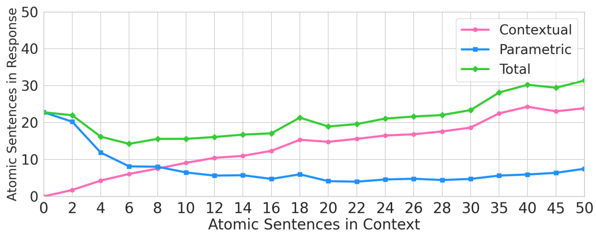
Figure L.13: Comparing two methods of prompting the model to answer, one instructing to strictly adhere to provided context and the other not imposing any restriction.



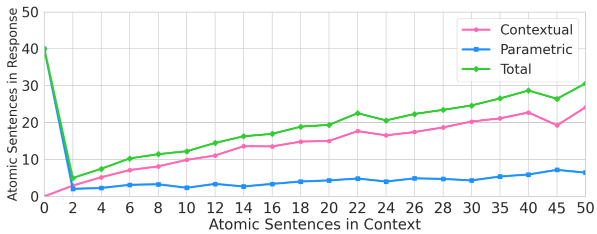
(a) No restriction in question (Opus)



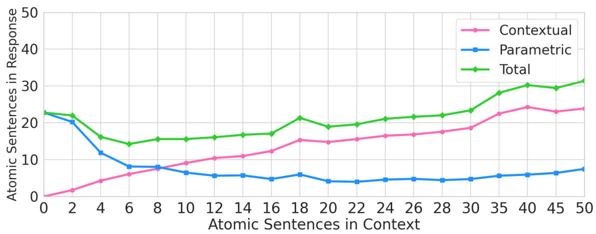
(b) Strict restriction in question (Opus)



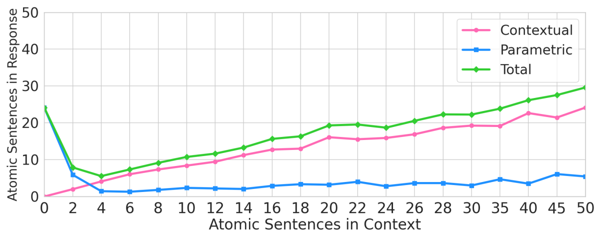
(c) No restriction in question (Sonnet)



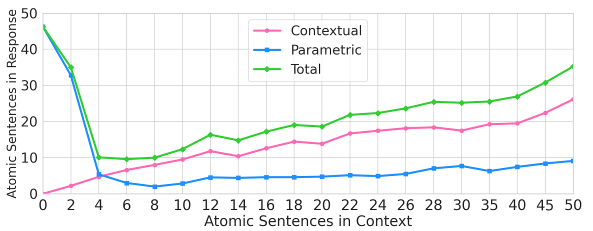
(d) Strict restriction in question (Sonnet)



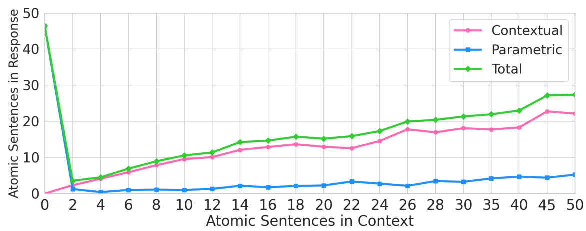
(e) No restriction in question (Haiku)



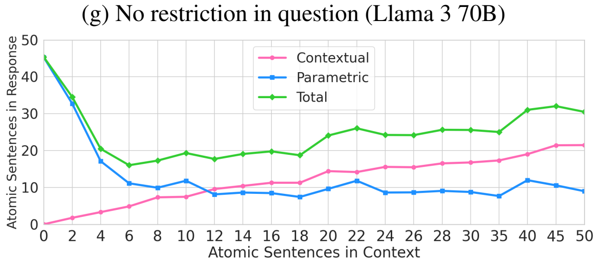
(f) Strict restriction in question (Haiku)



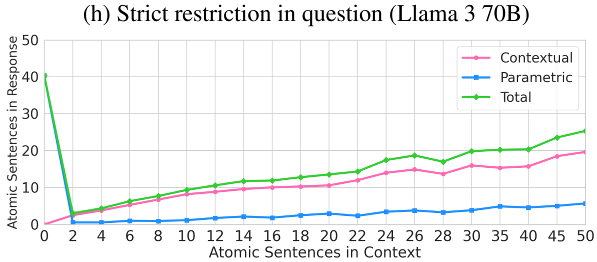
(g) No restriction in question (Llama 3 70B)



(h) Strict restriction in question (Llama 3 70B)

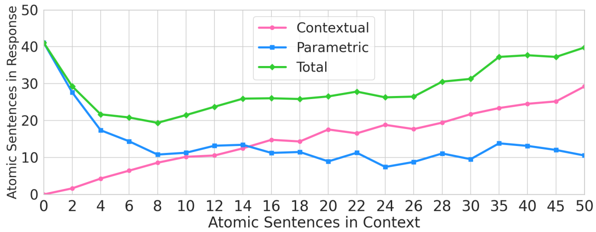


(i) No restriction in question (Llama 3 8B)

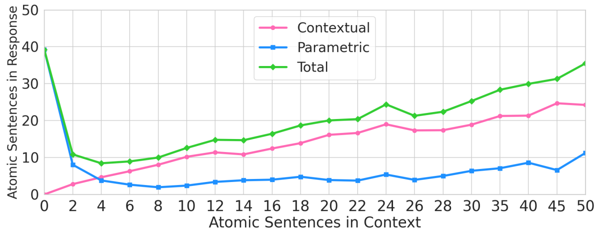


(j) Strict restriction in question (Llama 3 8B)

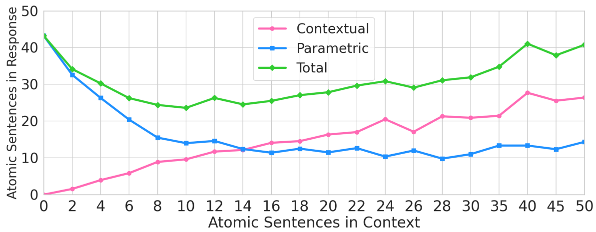
Figure L.14: Comparing two methods of prompting the model to answer, one instructing to strictly adhere to provided context and the other not imposing any restriction.



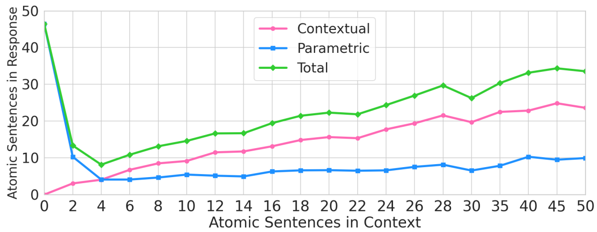
(a) No restriction in question (Mixtral 8x22B)



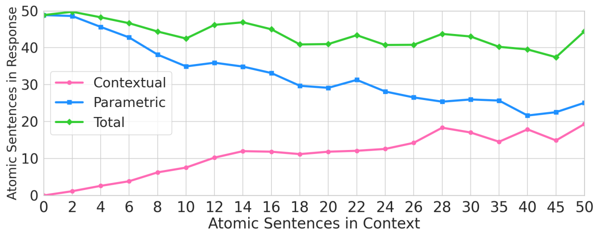
(b) Strict restriction in question (Mixtral 8x22B)



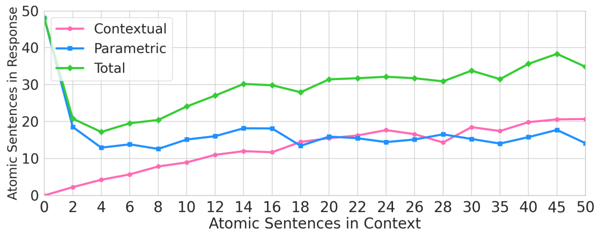
(c) No restriction in question (Mistral 7B)



(d) Strict restriction in question (Mistral 7B)

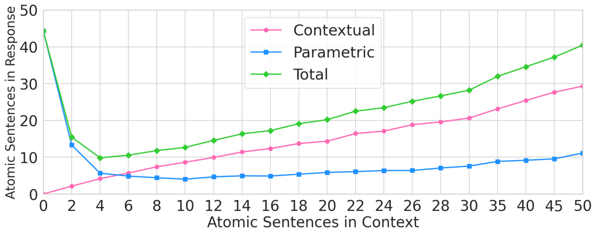


(e) No restriction in question (Phi 3)

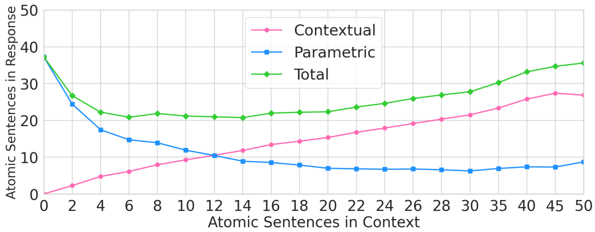


(f) Strict restriction in question (Phi 3)

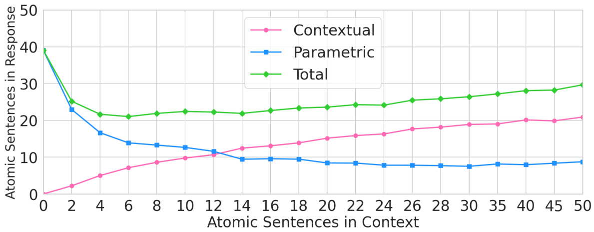
Figure L.15: Comparing two methods of prompting the model to answer, one instructing to strictly adhere to provided context and the other not imposing any restriction.



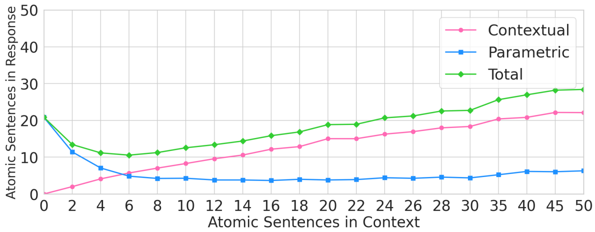
(a) GPT-4o



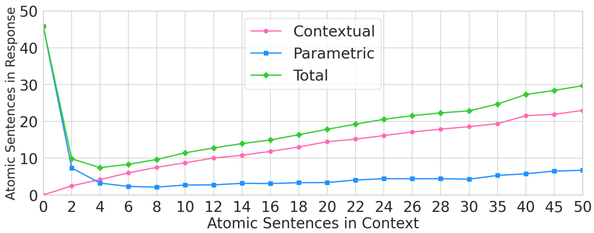
(b) Claude Opus



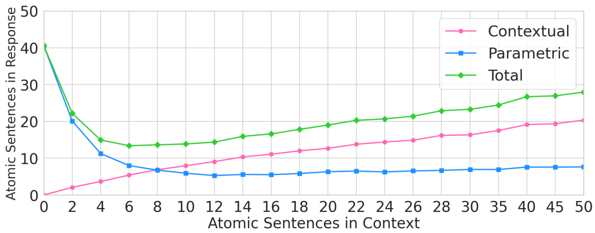
(c) Claude Sonnett



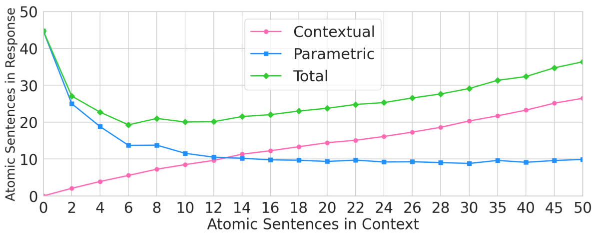
(d) Claude Haiku



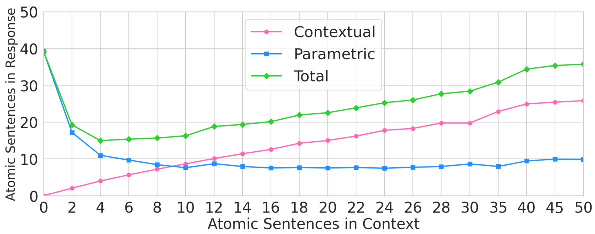
(e) Llama 3 70B



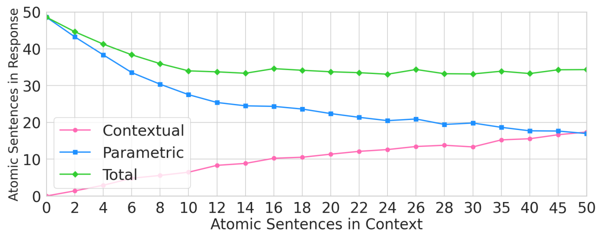
(f) Llama 3 8B



(g) Mistral 7B



(h) Mixtral 8x22B



(i) Phi 3

Figure M.16: Contextual (local), parametric (global), and total sentences in responses for each model while disregarding sentences received INFUSE score between 0.3 and 0.7.