What am I missing here?: Evaluating Large Language Models for Masked Sentence Prediction

Charlie Wyatt* **Aditya Joshi** **Flora Salim**
University of New South Wales, Sydney, Australia

Abstract

Transformer-based models primarily rely on Next Token Prediction (NTP), which predicts the next token in a sequence based on the preceding context. However, NTP's focus on single-token prediction often limits a model's ability to plan ahead or maintain long-range coherence, raising questions about how well LLMs can predict longer contexts, such as full sentences within structured documents. While NTP encourages local fluency, it provides no explicit incentive to ensure global coherence across sentence boundaries—an essential skill for reconstructive or discursive tasks. To investigate this, we evaluate three commercial LLMs (GPT-4o, Claude 3.5 Sonnet, and Gemini 2.0 Flash) on Masked Sentence Prediction (MSP) — the task of infilling a randomly removed sentence — from three domains: ROCStories (narrative), Recipe1M (procedural), and Wikipedia (expository). We assess both **fidelity** (similarity to the original sentence) and **cohesiveness** (fit within the surrounding context). Our key finding reveals that commercial LLMs, despite their superlative performance in other tasks, are poor at predicting masked sentences in low-structured domains, highlighting a gap in current model capabilities.

1 Introduction

Large Language Models (LLMs) have rapidly advanced natural language processing, achieving strong performance across a wide range of benchmarks (Anthropic, 2024; OpenAI et al., 2024; DeepMind, 2024). These models are trained with Next Token Prediction (NTP), a token-level objective that rewards fluent continuation. However, NTP struggles with tasks that require long-term planning, coherence across extended contexts, or discourse-level structure (Bachmann and Nagarajan, 2024; Maharana et al., 2024).

This trade-off raises a fundamental question: To what extent do LLMs understand and model sentence-level structure within broader document context? Can they generate missing content that is not only fluent but also faithful and contextually grounded?

We argue that evaluating LLMs solely on token-level fluency masks deeper limitations in their ability to reason over and reconstruct global context. This is particularly relevant for applications like summarization, document editing, or repair, where sentence-level understanding is essential.

By measuring the ability to reason over long dependencies, MSP provides a lens to evaluate LLMs' understanding of global document structure. We provide diagnostic insights into both when and why these models fail at sentence-level reconstruction, and empirical evidence of how document structure affects task difficulty. These findings have direct implications for practitioners selecting appropriate reconstruction tasks: our results suggest that procedural texts with clear sequential structure are suitable to automatic reconstruction, while narrative and expository domains may require more sophisticated approaches.

To explore these limitations, we study how LLMs handle sentence-level uncertainty using the task of **Masked Sentence Prediction (MSP)** across fidelity and cohesiveness.

We apply MSP to three distinct domains: narrative (ROCStories), procedural (Recipe1M), and expository (Wikipedia). Across all three, we evaluate generations from GPT-4o, Claude 3.5 Sonnet, and Gemini 2.0 Flash. We find that fidelity metrics measuring n-gram overlap are often low, whereas semantic comparative measures such as BLEURT and SBERT cosine similarity vary significantly across domain and model. Notably, fidelity improves in more structured domains like Recipe1M, while perceived cohesion suffers.

Such generations reveal a limitation in

* charles.wyatt@student.unsw.edu.au

Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 273–283

document-level understanding—one that poses risks in contexts requiring factual accuracy and precise reconstruction, like legal, historical, or journalistic documents.

2 Methodology

We define MSP as the task of infilling a missing sentence s_i within a document D. For a document $D = \{s_0, \ldots, s_m\}$, a model M is asked to generate a sentence $s' = M(D - s_i)$.

Figure 1 illustrates our MSP approach. First, we perform **sentence segmentation** through the `en_core_web_sm` model from spaCy (Honnibal and Montani, 2017). Next, we apply **masking** by replacing a single sentence with the special token `<|mask_id|>`, following the masking strategies described in Section 2.1. Finally, for **sentence generation**, we pass the masked document to the model with the following prompt:

> *[Document with masked sentence] Fill in the masked sentence of the text. Do not say anything else. Do not use quotation marks. It should be a complete sentence with punctuation.*

2.1 Experimental Variables

We vary two key conditions to probe model behavior:

1. Text Domain: We evaluate across three domains—narrative (ROCSTORIES), procedural (RECIPE1M), and expository (WIKIPEDIA).

2. Masking Strategy: We manipulate context by changing (a) *mask position*, masking the first, last, or a middle sentence, and (b) *mask density*, masking multiple contiguous sentences to assess performance under larger information gaps.

3 Experimental Setup

3.1 Datasets

We evaluate across three publicly available corpora:

- **Narrative:** ROCSTORIES (Mostafazadeh et al., 2016), a set of 5-sentence commonsense narratives.
- **Procedural:** RECIPE1M (Marin et al., 2019), the cooking–instruction portion of Recipe1M
- **Expository:** WIKIPEDIA-2022-ENGLISH, encyclopedic articles

We randomly sample 400 test documents per dataset to balance statistical power with API constraints.

3.2 Models

We evaluate the current flagship LLMs from three major vendors:

i. GPT-4O (OpenAI)
ii. CLAUDE 3.5 SONNET (Anthropic)
iii. GEMINI 2.0 FLASH (Google)

All models are accessed via public APIs with default decoding settings (e.g., temperature = 1.0).

3.3 Evaluation Metrics

Our evaluation focuses on two behavioral dimensions: **fidelity**, the similarity between the generated and original sentence, and **cohesion**, how well the generated sentence fits with the surrounding context.

3.3.1 Fidelity Metrics (Automatic)

We use standard similarity metrics to quantify fidelity:

- **Semantic: BLEURT** (Sellam et al., 2020) and **SBERT cosine similarity** (Reimers and Gurevych, 2019) for semantic similarity.
- **Lexical: ROUGE-1** and **BLEU**, for n-gram overlap.

3.3.2 Cohesiveness Evaluation (Human)

To assess cohesion, we conduct a blind human preference test. An annotator was shown both the original and generated sentences (in randomized order) within the document context and asked to indicate which sentence they prefer, if any.

4 Results

4.1 Fidelity

While BLEURT and SBERT cosine similarity scores close to zero indicate a poor semantic match, these metrics are most informative comparatively. From Table 1, we can observe that structured domains like RECIPE1M consistently yield higher fidelity than the more open-ended ROCSTORIES and WIKIPEDIA datasets. Moreover, the much more challenging task of lexically reconstructing the original sentence - captured by ROUGE-1 and BLEU - shows clear model limitations. ROUGE-1 measures unigram (word-level) overlap between the generated and reference sentences, while BLEU measures n-gram precision for longer contiguous sequences. In our results, BLEU scores hover around $\approx$0.05 and ROUGE-1 around $\approx$0.20, which are both well below the thresholds ($\approx$0.2 BLEU or $\approx$0.4 ROUGE-1) typically associated with acceptable lexical overlap.

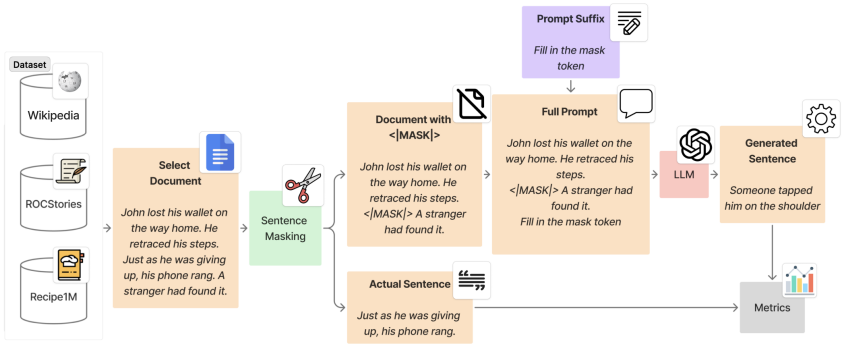

Figure 1: Our experimental pipeline to evaluate masked sentence prediction.

Table 1: Fidelity scores for Masked Sentence Prediction. Full results are in Appendix A.

Model	Dataset	BLEURT	SBERT	ROUGE-1	BLEU
GPT-4o	ROCSTORIES	0.3839	0.4652	0.2069	0.0225
Claude 3.5 Sonnet	ROCSTORIES	**0.4181**	**0.5110**	0.2445	0.0272
Gemini 2.0 Flash	ROCSTORIES	0.3747	0.4644	**0.2455**	**0.0412**
GPT-4o	RECIPE1M	0.4987	0.5816	0.2969	0.0913
Claude 3.5 Sonnet	RECIPE1M	**0.5259**	**0.6082**	**0.3551**	0.0980
Gemini 2.0 Flash	RECIPE1M	0.4930	0.5675	0.3261	**0.1096**
GPT-4o	WIKIPEDIA	0.3248	0.4302	0.1856	0.0257
Claude 3.5 Sonnet	WIKIPEDIA	**0.3416**	**0.4396**	**0.2094**	**0.0471**
Gemini 2.0 Flash	WIKIPEDIA	0.3135	0.4064	0.1852	0.0257

These low values indicate that the models rarely reproduce the original wording, even if the generated sentence is semantically or contextually appropriate.

As shown in Figure 2, Claude 3.5 Sonnet achieves higher average BLEURT scores and lower variance compared to Gemini. Consistency is important for applications like missing value reconstruction in damaged or historical documents, where variations in generated text can cause inaccuracies and compromise content integrity. The wide BLEURT distribution of Gemini shows significant variation between high and low quality generations, potentially limiting its usability.

We observe that fidelity also strongly correlates with **domain structure**. Structured domains like RECIPE1M consistently yield higher fidelity than the more open-ended ROCSTORIES and WIKIPEDIA.

4.1.1 Qualitative Error Patterns

Fidelity failures often undermine logical consistency, factual accuracy, or tone. We identified several common failure modes, illustrated with examples in Table 6 (Appendix B):

- In **ROCStories and Wikipedia**, models sometimes generate vague or tonally inconsistent sentences (e.g., inserting overly formal language in casual stories).
- In **Recipe1M**, failures often stem from logical or sequencing issues (e.g., placing 'Serve immediately' before 'Mix ingredients').

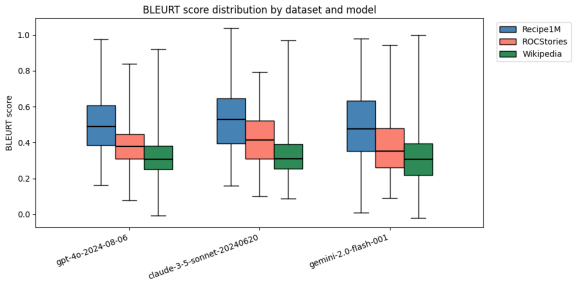

Figure 2: Distribution of BLEURT scores by model and domain.

These failure modes underscore the importance of separating reconstructive fidelity from contextual cohesiveness: a sentence can be contextually fluent but still pragmatically or factually invalid in structured domains.

4.1.2 Mask Position Analysis

The position of the masked sentence also influences fidelity. As shown in Figure 3, models perform best when the masked sentence appears in the **middle** of the text, where both preceding and following context help guide generation.

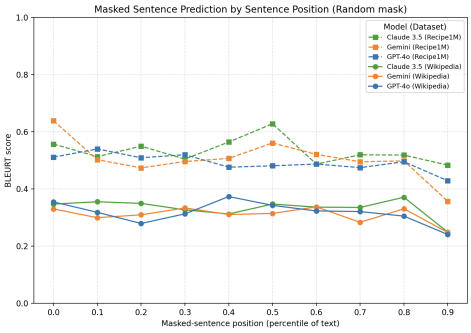

Figure 3: Fidelity by sentence position (BLEURT).

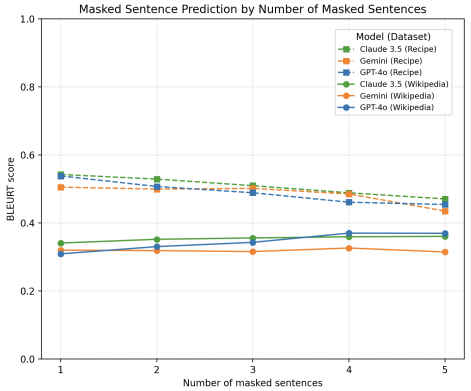

Figure 4: BLEURT by number of masked sentences.

Masking the **final** sentence yields the lowest fidelity. This improvement is likely due to the model having access to both preceding and succeeding context, which provides richer information for sentence reconstruction.

4.1.3 Multi-Sentence Masking

To examine how models handle greater uncertainty, we masked multiple contiguous sentences. Figure 4 shows domain-specific trends:

- In RECIPE1M, fidelity declines steadily as more steps are masked— representing the increasing difficulty of the task and the sensitivity of this structured domain to masking.
- In WIKIPEDIA, fidelity is relatively consistent.

4.2 Cohesion

Table 2 summarizes human preferences between the original and generated sentences. Despite low fidelity in ngram overlap and highly variable semantic similiarity, generations were often scored with an 'Equal Preference'. ROCSTORIES and WIKIPEDIA showed particularly high rates of equal preference (over 60%), reflecting the models' ability to generate plausible substitutes in open-ended domains. By contrast, in RECIPE1M, annotators preferred the original sentence more often—highlighting the stricter structural demands of procedural text. Interestingly, this suggests an inverse relationship between fidelity and cohesion: while structured domains like RECIPE1M make it easier for models to reproduce the original sentence (higher fidelity), they also make errors more conspicuous to human evaluators. In contrast, the looser expectations in narrative and expository domains allow models to maintain surface-level cohesion even when the generated sentence diverges semantically from the original.

Table 2: Human preference for generated vs. actual sentences ($n = 50$ per condition).

Dataset	Preference	GPT-4o	Claude	Gemini
Stories	Equal Preference	33	35	32
	Prefer Generated	8	7	8
	Prefer Actual	9	8	10
Recipes	Equal Preference	28	33	30
	Prefer Generated	4	8	5
	Prefer Actual	18	9	15
Wikipedia	Equal Preference	37	35	38
	Prefer Generated	3	4	3
	Prefer Actual	10	11	9

5 Related Work

MSP is commonly used during pre-training to enhance downstream performance. BART (Lewis et al., 2019) and T5 (Raffel et al., 2023) are notable examples, by masking small, contiguous spans of text. These objectives focus on token-level or span-level corruption rather than full-sentence prediction. In contrast, our study treats MSP as a standalone task and evaluates commercial LLMs without fine-tuning.

Beyond pre-training, TIGS (Liu et al., 2019) predicts text by optimizing missing words as parameterized vectors, focusing on shorter text gaps rather than full sentences. INSET (Huang et al., 2020) bridges gaps with intermediate sentences using a structured three-step generation process. The Masked Sentence Model (MSM) (Zhang et al., 2023) improves cross-lingual dense retrieval by modeling sequential sentence relations using a hierarchical contrastive loss.

While prior work often involves fine-tuning models, we fill a critical gap by assessing commercial LLMs out of the box, highlighting their limitations in predicting masked sentences.

6 Conclusion & Future Work

We evaluated three commercial LLMs—GPT-4o, Claude 3.5 Sonnet, and Gemini 2.0 Flash—on the task of MSP across narrative, procedural, and expository domains. Our results highlight key limitations in current LLMs' ability to model sentence-level coherence and long-range dependencies.

Models perform better in procedurally structured domains like Recipe1M, where the flow is predictable and local context often suffices for accurate reconstruction. In contrast, performance drops in narrative and expository texts, where global coherence and subtle discourse cues are required. This suggests that NTP—which optimizes for short-term fluency—does not adequately support tasks requiring holistic understanding over entire passages.

Future work should explore architectures or training strategies that explicitly capture both local and global context, such as hierarchical attention mechanisms or planning-based generation. Additionally, fine-tuning on MSP-style objectives, incorporating richer prompting strategies (e.g., chain-of-thought), and expanding evaluation to include diverse domains and human judgments could yield deeper insights into model reliability and coherence in real-world settings.

We hope this work serves as a foundation for future research into sentence-level understanding in LLMs. Key directions include: (1) establishing human baselines to assess inherent task difficulty, (2) using post-training-cutoff data or open models to disentangle memorization from generation capability, and (3) exploring whether fine-tuning on MSP-style objectives can improve global coherence.

Limitations

Our study has several limitations that provide opportunities for future work.

First, our analysis is based on 400 samples from each of the three datasets. While sufficient to reveal clear behavioral patterns and provide a meaningful diagnostic signal, larger-scale experiments would be beneficial to ensure stronger statistical significance.

Second, while our scope covers narrative, procedural, and expository texts, a broader evaluation across more genres would further generalize our findings about the interplay between domain structure and generation strategy.

A further limitation is our use of closed-source commercial models. These evaluation datasets were most likely included in the models' vast training corpora. This data contamination could mean that in some instances, the models are performing a form of memorization rather than true, zero-shot generation. Future work could directly address this by replicating our study using a fully open model like OLMo (Groeneveld et al., 2024), whose open training data allows for the definitive exclusion of evaluation sets. This would provide a cleaner signal on the models' inherent generative capabilities.

An important consideration is whether MSP is inherently difficult even for humans. Our study does not include a human baseline for reconstructing masked sentences, making it difficult to establish whether the observed 'low' fidelity represents a model limitation or an inherently challenging task. Future work should include human performance benchmarks to better separate task complexity from model-specific limitations.

Finally, our human evaluation was conducted by a single annotator who is also an author of this paper. While the evaluations were performed in a blind setting to minimize bias, this setup may still limit the generalizability and objectivity of the results. A larger-scale evaluation with multiple independent annotators would strengthen future conclusions.

Ethical Considerations

Since we use publicly available LLMs and datasets there are no known ethical considerations that have been left out.

Data and Model Usage. All artifacts used in this study were accessed in accordance with their licenses and terms of service. Our datasets include: **ROCStories** (Mostafazadeh et al., 2016), distributed under the CC BY 4.0 license; **Recipe1M** (Marin et al., 2019), used under terms permitting non-commercial research; and articles from **Wikipedia**, used in accordance with the Creative Commons Attribution-ShareAlike (CC BY-SA) license. All models (GPT-4o, Claude 3.5 Sonnet, and Gemini 2.0 Flash) were accessed via their official public APIs and used in compliance with their respective terms of service. The `spaCy` library (Honnibal and Montani, 2017),

used for sentence segmentation, is open-source under the MIT license.

Potential Risks and Societal Impact. Our central finding—that LLMs prioritize plausible generation over high-fidelity reconstruction—carries potential societal implications. This behavior is a double-edged sword. On one hand, it can be highly beneficial for creative applications, brainstorming, and generating diverse linguistic paraphrases.

On the other hand, it presents a clear risk. The models' ability to confidently generate plausible but factually incorrect sentences could be misused to create convincing misinformation, to alter the meaning of records, or to automate the production of "authentic-looking" but false content. While our work deals with low-stakes domains like stories and recipes, the underlying behavior we identify is domain-general. We believe it is crucial for the research community and practitioners to be aware of this dual-use nature when deploying such models in high-stakes applications where factual precision is paramount.

Use of AI in Research Preparation. We utilized AI-assisted tools during the development of this work. Specifically, ChatGPT was used for drafting, grammatical editing, and proofreading the manuscript. Additionally, GitHub Copilot was employed to assist with coding tasks and debugging. We ensured that all outputs generated by these tools were carefully reviewed and validated by the authors to maintain accuracy and correctness.

Acknowledgments

Some of the illustrations in this paper incorporate icons sourced from Flaticon.com. We gratefully acknowledge the individual artists who designed and shared these assets. The research was funded by UNSW Sydney, via an HDR scholarship awarded to Charlie Wyatt. Aditya Joshi would like to dedicate this paper to the memory of Prof. Pushpak Bhattacharyya, IIT Bombay, whose guidance and mentorship produced several generations of researchers, and who instilled in his students the philosophy of NLP grounded in linguistic intuition.

References

Anthropic. 2024. Claude 3.5 sonnet model card addendum. Accessed: 2025-05-19.

Gregor Bachmann and Vaishnavh Nagarajan. 2024. The pitfalls of next-token prediction. *Preprint*, arXiv:2403.06963.

Google DeepMind. 2024. Introducing gemini 2.0: our new ai model for the agentic era. Accessed: 2025-05-19.

Dirk Groeneveld, Iz Beltagy, Pete Walsh, Akshita Bhagia, Rodney Kinney, Oyvind Tafjord, Ananya Harsh Jha, Hamish Ivison, Ian Magnusson, Yizhong Wang, Shane Arora, David Atkinson, Russell Authur, Khyathi Raghavi Chandu, Arman Cohan, Jennifer Dumas, Yanai Elazar, Yuling Gu, Jack Hessel, Tushar Khot, William Merrill, Jacob Morrison, Niklas Muennighoff, Aakanksha Naik, Crystal Nam, Matthew E. Peters, Valentina Pyatkin, Abhilasha Ravichander, Dustin Schwenk, Saurabh Shah, Will Smith, Emma Strubell, Nishant Subramani, Mitchell Wortsman, Pradeep Dasigi, Nathan Lambert, Kyle Richardson, Luke Zettlemoyer, Jesse Dodge, Kyle Lo, Luca Soldaini, Noah A. Smith, and Hannaneh Hajishirzi. 2024. Olmo: Accelerating the science of language models. *Preprint*, arXiv:2402.00838.

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

Yichen Huang, Yizhe Zhang, Oussama Elachqar, and Yu Cheng. 2020. INSET: Sentence infilling with INter-SEntential transformer. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2502–2515, Online. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *Preprint*, arXiv:1910.13461.

Dayiheng Liu, Jie Fu, Pengfei Liu, and Jiancheng Lv. 2019. Tigs: An inference algorithm for text infilling with gradient search. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, page 4146–4156. Association for Computational Linguistics.

Adyasha Maharana, Dong-Ho Lee, Sergey Tulyakov, Mohit Bansal, Francesco Barbieri, and Yuwei Fang. 2024. Evaluating very long-term conversational memory of llm agents. *Preprint*, arXiv:2402.17753.

Javier Marin, Aritro Biswas, Ferda Ofli, Nicholas Hynes, Amaia Salvador, Yusuf Aytar, Ingmar Weber, and Antonio Torralba. 2019. Recipe1m+: A dataset for learning cross-modal embeddings for cooking recipes and food images. *Preprint*, arXiv:1810.06553.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. 2016. A corpus and evaluation framework for deeper understanding of commonsense stories. *arXiv preprint arXiv:1604.01696*.

OpenAI, Aaron Hurst, and et al. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2023. Exploring the limits of transfer learning with a unified text-to-text transformer. *Preprint*, arXiv:1910.10683.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *Preprint*, arXiv:1908.10084.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Shunyu Zhang, Yaobo Liang, Ming Gong, Daxin Jiang, and Nan Duan. 2023. Modeling sequential sentence relation to improve cross-lingual dense retrieval. *Preprint*, arXiv:2302.01626.

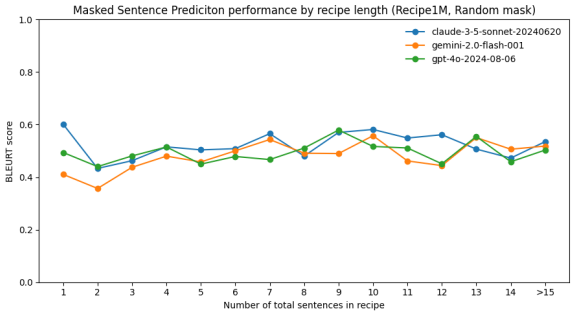

Figure 5: MSP performance (BLEURT) by the number of sentences in the RECIPE1M dataset.

A Detailed Experimental Results

This section provides detailed numerical results supplementing the analysis in the main paper. **Table 3** expands on the main results table by breaking down performance by the position of the masked sentence. **Tables 4 and 5** provide further analysis on the RECIPE1M dataset, showing how fidelity scores vary with the total number of sentences and the relative masked position, respectively.

B Qualitative Generation Examples

This section provides the qualitative examples referenced in Section 4.1.1 of the main text. Table 7 illustrates the fidelity-plausibility trade-off, while Table 6 categorizes common failure modes.

Table 3: Full performance breakdown by model, dataset, and masked sentence position. For ROCStories and Recipe1M, "Random" refers to the average score across all positions, while "First" and "Last" refer to masking only the first or last sentence. For Wikipedia, only random middle sentences were masked.

Model	Dataset	Mask Position	BLEURT	SBERT	ROUGE-1	BLEU
GPT-4o	ROCStories	Random	0.3839	0.4652	0.2069	0.0225
		First	0.3754	0.4305	0.1807	0.0160
		Last	0.3754	0.4305	0.1807	0.0160
Claude 3.5 Sonnet	ROCStories	Random	0.4181	0.5110	0.2445	0.0272
		First	0.4606	0.5803	0.2892	0.0488
		Last	0.4056	0.4836	0.2076	0.0215
Gemini 2.0 Flash	ROCStories	Random	0.3747	0.4644	0.2455	0.0412
		First	0.4195	0.5468	0.2899	0.0719
		Last	0.3550	0.4144	0.1931	0.0247
GPT-4o	Recipe1M	Random	0.4987	0.5816	0.2969	0.0913
		First	0.5195	0.6224	0.3340	0.1013
		Last	0.4848	0.5027	0.2439	0.0554
Claude 3.5 Sonnet	Recipe1M	Random	0.5259	0.6082	0.3551	0.0980
		First	0.5335	0.6459	0.3811	0.1110
		Last	0.4858	0.5057	0.2721	0.0705
Gemini 2.0 Flash	Recipe1M	Random	0.4930	0.5675	0.3261	0.1096
		First	0.4841	0.5897	0.3316	0.1167
		Last	0.4433	0.4597	0.2255	0.0743
GPT-4o	Wikipedia	Random	0.3248	0.4302	0.1856	0.0257
Claude 3.5 Sonnet	Wikipedia	Random	0.3416	0.4396	0.2094	0.0471
Gemini 2.0 Flash	Wikipedia	Random	0.3135	0.4064	0.1852	0.0257

Table 4: BLEURT scores on the RECIPE1M dataset, broken down by the total number of sentences in the recipe.

Model	Total Number of Sentences in Recipe														
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	≥15
GPT-4o	0.4939	0.4402	0.4808	0.5149	0.4496	0.4787	0.4670	0.5108	0.5794	0.5169	0.5111	0.4508	0.5551	0.4586	0.5032
Claude 3.5 Sonnet	0.6016	0.4333	0.4626	0.5154	0.5039	0.5086	0.5655	0.4804	0.5704	0.5815	0.5490	0.5615	0.5068	0.4726	0.5353
Gemini 2.0 Flash	0.4105	0.3570	0.4375	0.4800	0.4580	0.5002	0.5433	0.4911	0.4899	0.5578	0.4618	0.4440	0.5507	0.5063	0.5184

Table 5: BLEURT scores on the WIKIPEDIA and RECIPE1M datasets, broken down by masked position (0.0–0.9). All experiments used a randomly masked sentence.

Dataset	Model	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
Wikipedia	GPT-4o	0.3546	0.3183	0.2794	0.3129	0.3734	0.3427	0.3232	0.3210	0.3050	0.2408
	Claude 3.5 Sonnet	0.3471	0.3553	0.3495	0.3271	0.3126	0.3472	0.3365	0.3354	0.3709	0.2492
	Gemini 2.0 Flash	0.3302	0.2996	0.3096	0.3337	0.3106	0.3145	0.3373	0.2834	0.3306	0.2481
Recipe1M	GPT-4o	0.5111	0.5398	0.5091	0.5198	0.4762	0.4809	0.4872	0.4745	0.4953	0.4287
	Claude 3.5 Sonnet	0.5569	0.5130	0.5496	0.5044	0.5641	0.6280	0.4871	0.5194	0.5185	0.4832
	Gemini 2.0 Flash	0.6384	0.5024	0.4737	0.4961	0.5068	0.5609	0.5210	0.4951	0.4981	0.3568

Table 6: **Examples of common failure modes in plausible generation.** These sentences are often locally coherent but fail to respect the global context, leading to logical inconsistencies, unresolved narratives, or formatting errors.

Model	Generated Sentence	Actual Sentence	Failure Type	Dataset
Gemini 2.0 Flash	Keep it in the fridge.	Serve and enjoy.	Logical Inconsistency	Recipe
GPT-4o	In a large bowl, mix together the flour, baking powder, and salt.	1.In a large bowl, mix together flour, salt, baking powder, cinnamon, nutmeg, ground cloves, ginger and Flax Seeds; set aside.	Formatting Inconsistency	Recipe
Claude 3.5 Sonnet	Frustrated, Ulrich decided to rely on the video's visual demonstrations and hoped he could figure out the repair process without understanding the narration.	Ulrich was able to fix the scanner just by the video alone.	Verbosity	Stories
Gemini 2.0 Flash	Don was shocked, but the CEO insisted it was necessary for the company's future.	Don refused.	Unresolved	Stories
Gemini 2.0 Flash	I had built them a wonderful cat condo.	I had noticed that my cats seemed kind of bored.	No Introduction	Stories
GPT-4o	He asked his wife to pour vinegar over the sting to neutralize the venom.	He scraped the sting with a credit card to help stop the stinging.	More Formal	Stories

Table 7: **Qualitative examples of the Fidelity-Plausibility Trade-off.** High-scoring examples show high fidelity (reconstruction). Low-scoring examples often exhibit high plausibility but low fidelity, where the model generates a coherent but novel sentence (e.g., explaining a dry pond with a "drought" instead of the author's "surprise").

Model	Score Type	Context (with mask)	Actual Sentence	Generated Sentence
GPT-4o	High Fidelity	Tom had never been on an airplane. <|mask_id|> His first time was a bumpy ride. They hit a lot of turbulence. Tom was frightened the whole time.	Tom had never been on an airplane.	Tom had never been on an airplane before.
	Low Fidelity	I used to run around a pond. It was there one fall. The next fall it had vanished. There was nothing but dirt. <|mask_id|>	I was surprised.	A long drought had caused the water to evaporate completely.
Claude 3.5 Sonnet	High Fidelity	<|mask_id|> When it came in he added all his favorite apps. He bought a new case. Then as he was out one day, he dropped it face down. Josh shattered his phone within hours of getting it.	Josh was so excited about his new phone.	Josh was excited to receive his brand new smartphone.
	Low Fidelity	Tom went to a new church. It was full and a lot of fun. He started going more often. So did a lot of other people. <|mask_id|>	Eventually Tom wasn't able to fit anymore.	The church grew rapidly as word spread about its welcoming atmosphere and engaging services.
Gemini 2.0 Flash	High Fidelity	<|mask_id|> He decides to snack on some peanuts. He eats a handful of peanuts. He is glad that he ate some peanuts. He feels much less hungry afterwards.	John is feeling a little hungry.	John is feeling a little hungry.
	Low Fidelity	Before led bike lights, I used battery powered lights. The batteries wore out too quickly. <|mask_id|> I tried a generator, but it never produced much light. I found an incandescent flashing light which used less energy.	I drove forty five minutes in the dark.	Replacing them was expensive and inconvenient.