Self-Consistency Falls Short! The Adverse Effects of Positional Bias on Long-Context Problems

Adam Byerly and **Daniel Khashabi**
Johns Hopkins University, USA
abyerly2@jhu.edu, danielk@cs.jhu.edu

Abstract

Self-consistency (SC) improves the performance of large language models (LLMs) across various tasks and domains that involve short content. However, does this support its effectiveness for long-context problems? We challenge the assumption that SC's benefits generalize to long-context settings, where LLMs often struggle with position bias—the systematic over-reliance on specific context regions—which hinders their ability to utilize information effectively from all parts of their context. Through comprehensive experimentation with varying state-of-the-art models, tasks, and SC formulations, we find that SC not only fails to improve but actively degrades performance on long-context tasks. This degradation is driven by persistent position bias, which worsens with longer context lengths and smaller model sizes but remains invariant to prompt format or task type. Unlike short-context tasks, where SC diversifies reasoning paths, long-context SC amplifies positional errors. These comprehensive results provide valuable insight into the limitations of current LLMs in long-context understanding and highlight the need for more sophisticated approaches.

1 Introduction

Large Language Models (LLMs) have shown remarkable versatility in performing various tasks through prompting (Brown et al., 2020, inter alia). However, these models exhibit various forms of brittleness across tasks (Mishra et al., 2022; Frieder et al., 2024; Mirzadeh et al., 2025; Wu et al., 2024), including catastrophic failures on simple problems easily solvable by humans (Nezhurina et al., 2024). To improve reliability, *self-consistency* (SC) (Wang et al., 2023a) has emerged as a powerful strategy that aggregates multiple sampled responses to mitigate failures. SC has been highly effective in *short* tasks (e.g., ≤ 100 tokens), but its ability to handle *long* contexts (e.g., $\geq 10K$ tokens) remains underexplored. As real-world applications increasingly demand processing lengthy inputs—such as legal analysis, medical diagnostics, or scientific literature review—understanding SC's scalability is crucial for developing robust solutions. This raises a critical question: Does SC's benefit scale with context length, or do longer contexts introduce challenges that inherently alter its effectiveness?

Fundamentally, SC assumes that errors in individual samples are independent. However, in long-context tasks, systematic position biases induce *correlated errors*, violating SC's core assumption. While prior work (Wang et al., 2023b; Zheng et al., 2023; Liu et al., 2024) has documented position bias as a standalone challenge in long-context reasoning, our study provides a diagnostic analysis of how these biases interact with self-consistency, compounding errors through correlated sampling rather than resolving them. We hypothesize that because SC aggregates multiple responses from the same biased model, it reinforces these correlated errors, ultimately amplifying position bias (Figure 1).

To test this hypothesis and explore the broader implications of SC in long-context scenarios, we conduct extensive experiments across state-of-the-art models (GPT-4o [OpenAI, 2024], LLaMA-3.3-70B [Grattafiori et al., 2024], and Qwen-2.5-72B [Yang et al., 2024]) and their smaller variants, evaluating nine diverse long-context tasks spanning summarization, question answering, multi-hop reasoning, and controlled position experiments. Our comprehensive evaluation combines traditional metrics with LLM-based assessment and explores multiple SC implementations, including: majority voting (Wang et al., 2023a) for tasks with discrete answer choices, universal SC (Chen et al., 2024) for open-ended generation tasks, and soft SC (Wang et al., 2024) as an alternative open-ended generation aggregation strategy.

Contributions. We provide the first systematic study of self-consistency in long-context settings,

Transactions of the Association for Computational Linguistics, vol. 14, pp. 292–317, 2026. https://doi.org/10.1162/TACL.a.625
Action Editor: Asli Celikyilmaz. Submission batch: 3/2025; Revision batch: 6/2025; Published 4/2026.

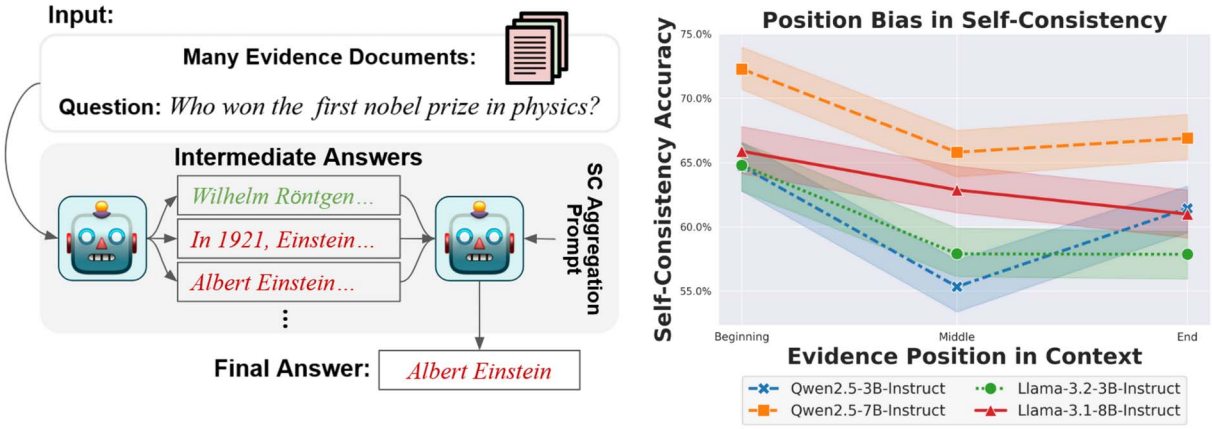


Figure 1: Schematic of self-consistency in long-context, ‘‘needle-in-a-haystack’’ scenarios. Input consists of a query and multiple evidence documents, one of which contains the correct answer, with a model generating diverse intermediate answers via stochastic sampling (non-zero temperature). However, aggregation yields incorrect answers due to position bias, highlighting a key challenge in long-context reasoning. Sampling from a model with inherent position bias **amplifies rather than mitigates errors**, as all samples inherit the same structural biases, violating SC's core assumption of error independence.

revealing three key findings from across 651 carefully designed experiments:

(i) **Failure of SC in long-contexts:** SC consistently degrades performance across models and tasks. This degradation persists across synthetic and real-world tasks, challenging the assumption that SC can universally improve LLM reliability.

(ii) **Robustness to metric and implementation:** SC's long-context limits are confirmed across multiple metrics, aggregation methods, and SC parameterizations.

(iii) **Mechanistic role of position bias:** Through controlled experiments, we show that SC amplifies position bias in both retrieval and reasoning steps. Varying prompt structure and SC configuration reveals that SC's failures stem from *correlated positional errors*.

These insights challenge the assumption that SC universally improves performance and provides guidance for developing more robust approaches to long-context tasks. Our findings underscore the need to rethink aggregation strategies in the face of position bias, paving the way for future research into context-aware methods, such as position-aware voting or debiased sampling, that enhance reliability in long-context tasks.

2 Related Work

Self-Consistency. Building on chain-of-thought prompting (Wei et al., 2022), self-consistency (SC) (Wang et al., 2023a) leverages the intuition that diverse reasoning spaths converge toward correct answers when aggregated through methods like majority voting, likelihood scoring (Wang et al., 2024), LLM Judges (Chen et al., 2024; Lin et al., 2024), or game theoretic consensus (Jacob et al., 2024). However, these studies predominantly focus on short tasks (e.g., math problems, brief QA), where stochastic sampling produces *independent errors*. This assumption breaks down in long settings where *systemic errors* emerge due to position bias. While Chen et al. (2024) evaluates SC on long-context summarization, their reliance on ROUGE metrics, which measure surface-level alignment rather than holistic quality, leaves SC's impact on long-context understanding insufficiently addressed. Our work bridges this gap by comprehensively evaluating SC implementations across multiple long-context tasks, employing automated metrics (ROUGE, F1, accuracy) and LLM judges, with bootstrapped 95% confidence intervals to rigorously quantify statistical significance.

Position Bias. Position bias, the preference for certain context regions, is a fundamental challenge for LLMs in long-context processing. Prior work highlights three manifestations: *primacy* (overweighting early context) (Wang et al., 2023b),

Dataset	Split/Source	#Eval	Task Type	Metric	Avg. # Words
GovReport	LongBench (Bai et al., 2024)	200	Summ	ROUGE / Judge	8,734
QMSum	LongBench (Bai et al., 2024)	200	QB-Summ	ROUGE / Judge	10,614
SQuALITY	Test Split (Wang et al., 2022)	1040	QB-Summ	ROUGE / Judge	5,208
Qasper	LongBench (Bai et al., 2024)	200	QA	F1	3,619
NarrativeQA	LongBench (Bai et al., 2024)	200	QA	F1	18,409
MuSiQue	LongBench (Bai et al., 2024)	200	Multi-Hop QA	F1	11,214
QuALITY	Dev Split (Pang et al., 2022)	2086	MCQA	Accuracy	5,183
NQ-Open (QA)	Liu et al. (2024)	2655	MDQA	Accuracy	–
NQ-Open (TR)	Liu et al. (2024)	2655	Retrieval	Accuracy	–

Table 1: **Overview of tasks and datasets used in our experiments.** The QuALITY dev split was used as the annotations for the test split are not publicly released. QB-Summ indicates query-based summarization, MCQA indicates multiple choice question answering, and MDQA indicates multi-document question answering. The ''#Eval'' column refers to the exact number of examples in each dataset split we evaluate. For summarization tasks, we report both ROUGE and a GPT-4o-based Judge score; for QA tasks, we report F1 or accuracy depending on the output format. The average number of words in NQ-Open examples varies depending upon the total number of documents used.

recency (preferring recent information) (Zheng et al., 2023), and a *U-shaped* pattern where middle-context information is neglected (Liu et al., 2024). These biases persist across architectures, unaffected by instruction tuning (Liu et al., 2024) or context scaling (Lee et al., 2024), suggesting that they are intrinsic to transformer-based attention mechanisms. While recent efforts propose mitigations through context compression (Jiang et al., 2024), attention calibration (Hsieh et al., 2024b), or fine-tuning (Xiong et al., 2025), such approaches incur computational overhead and merely alleviate symptoms rather than address the root cause. Critically, these approaches target individual model outputs but do not consider how aggregation methods, such as SC, might amplify positional errors–*a gap our work uncovers*. By rigorously quantifying this interaction, we expose SC's incompatibility with long-context processing and underscore the need for bias-aware aggregation.

Long-Context Evaluation. Long-context evaluation has evolved from early synthetic tasks to more comprehensive benchmarks. Tay et al. (2021) introduced the Long-Range Arena (LRA) to compare efficient Transformers, albeit via relatively constrained tasks. More recent efforts, such as RULER (Hsieh et al., 2024a), extend beyond ''needle-in-a-haystack'' (Kamradt, 2023) retrieval by incorporating multi-hop reasoning and aggregation-style tasks. However, Yen et al. (2024) raise concerns that synthetic benchmarks may not reliably predict downstream performance, underscoring the need for more representative long-form evaluations. Meanwhile, benchmarks like SCROLLS (Shaham et al., 2022), its zero-shot variant ZeroSCROLLS (Shaham et al., 2023), and HELMET (Yen et al., 2024) collate tasks featuring more realistic, naturally long text. Our evaluation combines *controlled position probing* and *naturally long tasks*, ensuring that our findings generalize to synthetic and real-world scenarios where systemic position biases persist, regardless of context origin.

3 Methodology

To systematically evaluate SC in long-context settings, we design experiments spanning diverse models, task types, and aggregation strategies. Our methodology combines comprehensive benchmarking across real-world tasks with controlled experiments investigating position bias.

3.1 Tasks and Datasets

For our dataset-level evaluations (§4), we utilize seven datasets with distinct challenges in long-context processing, from dense information synthesis to targeted retrieval and reasoning (Table 1). Summarization tasks include GovReport (Huang et al., 2021), QMSum (Zhong et al., 2021), and SQuALITY (Wang et al., 2022), which test dense information synthesis. We use Qasper (Dasigi et al., 2021), NarrativeQA (Kočiskỳ et al., 2018), and QuALITY (Pang et al., 2022) for question answering, where models must understand details in long narratives. We also evaluate

MuSiQue (Trivedi et al., 2022) for multi-hop reasoning across documents. Datasets are sourced from LongBench (Bai et al., 2024), a standardized benchmark for long-context evaluation—except SQuALITY (Wang et al., 2022) and QuALITY (Pang et al., 2022), where we use the original test and dev splits, as they were omitted from LongBench.

For our position bias investigation (§5), we utilize NQ-Open (Kwiatkowski et al., 2019; Lee et al., 2019), comprising 2,655 real user queries paired with paragraph-length answers from Wikipedia. We design two tasks: (1) Question Answering (QA): Given a query and a collection of documents (with one gold document), models must generate the correct answer. This task examines the model's ability to locate and use information from lengthy contexts. (2) Text Retrieval (TR): Models are asked to identify which document contains the correct answer. By isolating the retrieval step, this task enables us to disentangle the effects of self-consistency on information localization from those of answer generation. These tasks have been used in prior works (Liu et al., 2024; Lee et al., 2024) as two canonical long-text tasks, providing a solid foundation for comparing our results with existing literature. Task prompt templates are illustrated in §C.5.1.

3.2 Self-Consistency Implementations

To assess self-consistency across diverse task formats, we differentiate our SC implementation based on the nature of the task output. For tasks requiring open-ended generation (GovReport, QMSum, SQUALITY, Qasper, NarrativeQA, MuSiQue, and NQ-Open QA/TR), we primarily employ Universal Self-Consistency (USC), which utilizes an LLM judge for response aggregation. For tasks with discrete, categorical answer options, specifically the multiple-choice question answering dataset QUALITY, we use traditional Majority Voting SC. We also evaluate Soft-SC across a subset of tasks as an alternative aggregation mechanism. By comparing these aggregation strategies across diverse tasks, we ensure that our conclusions are not specific to any single SC method. All implementations maintain consistent sampling parameters (eight samples, temperature 1.0) unless explicitly varied for experimental purposes.

SC (For QuALITY). We implement traditional majority voting SC as described in Wang et al. (2023a) for QuALITY's multiple-choice format. This implementation selects the most frequently generated answer across eight samples, breaking ties by selecting the first most frequent answer.

USC (For Open-Ended Tasks). Our primary implementation follows the USC baseline of Chen et al. (2024), generating eight samples per input at temperature 1.0. These parameters balance sampling diversity against computational cost while maintaining output quality. For aggregation, we employ GPT-4o as a judge (§C.3) to select the highest-quality response from the generated candidates, using carefully designed evaluation criteria specific to each task type.

Soft-SC. We also implement Soft-SC following Wang et al. (2024). This variant replaces discrete voting with a continuous scoring mechanism based on model likelihood. For each generated response, we compute the mean token likelihood across all response tokens. The final output is the response with the highest mean token likelihood, providing a more nuanced selection mechanism incorporating model uncertainty. By replacing discontinuous majority voting with likelihood-based aggregation, Soft-SC tests whether model confidence (rather than frequency) can mitigate correlated errors.

In short, we use USC for open-ended tasks, majority-vote SC for the QuALITY multiple-choice dataset, and soft-SC as a robustness ablation against aggregation mechanism.

3.3 Model Selection and Configuration

We evaluate eight instruction-tuned models spanning proprietary and open-source families, ranging from 3B to 72B parameters. All models are deployed in a standardized environment to eliminate implementation-specific confounders. Proprietary models representing state-of-the-art commercial systems include OpenAI's GPT-4o and GPT-4o-mini (OpenAI, 2024). For open-source architectures, we test Meta's LLaMA-3-Instruct series (3.2-3B, 3.1-8B, 3.3-70B; Grattafiori et al., 2024) and Alibaba's Qwen-2.5 series (3B, 7B, 72B; Yang et al., 2024). We excluded base models from our analysis as their documented underperformance on generative tasks would confound our investigation of SC's impact. Deployment was via the vLLM framework (Kwon et al., 2023), which

provides efficient serving with consistent runtime characteristics across all models. Serving configurations and hardware environments (NVIDIA A100 80GB and A6000 48GB) are identical for all experiments.

3.4 Evaluation Framework

We employ task-specific performance metrics (ROUGE, F1, accuracy) alongside LLM-based evaluation for summarization to ensure a rigorous assessment of SC's impact. We further quantify performance shifts using bootstrapped 95% confidence intervals, allowing us to determine whether SC effects are statistically significant.

Multi-Metric Assessment. For summarization, we extend Shaham et al.'s (2023) evaluation protocol, reporting both the geometric mean of ROUGE-1, ROUGE-2, and ROUGE-L (Lin, 2004) and GPT-4o-as-a-judge scores (0–100 scale (§C.4)). While ROUGE quantifies surface-level alignment, it fails to capture essential quality dimensions, such as factual consistency and narrative flow (Fabbri et al., 2021). Our GPT-4o judge evaluates summaries on consistency (alignment with references), relevance (focus on key points), fluency (grammatical coherence), and informativeness (coverage of critical details). SQuALITY additionally uses QA-specific criteria, wherein the judge evaluates summaries on correctness (factual accuracy), relevance (focus on key points), and fluency (grammatical coherence).

For QA tasks (Qasper, NarrativeQA, MuSiQue), we compute unigram F1 scores between outputs and annotated answers to accommodate paraphrasing better than exact match. QuALITY uses accuracy for its closed-set multiple-choice format. For the position bias experiments, we also use accuracy as our primary metric: For QA, accuracy measures the exact match between the generated answer and gold answer, and for TR, accuracy measures whether models correctly identify the source document containing the answer.

Performance Difference and Statistical Rigor. To quantify the impact of self-consistency, we measure the direct performance difference:

$$\text{Difference} = \text{Perf. w/ SC} - \text{Perf. w/o SC} \quad (1)$$

Statistical significance of the performance differences was assessed via bootstrapped 95% confidence intervals using 10K resamples. We use bootstrapped confidence intervals (CIs) to avoid challenges related to the unknowns of the underlying distribution of differences. A CI containing 0 indicates no significant change, while an entirely positive or negative CI indicates a significant performance gain or degradation, respectively.

Summary. By employing a multi-metric approach, we ensure that our findings are robust across different evaluation criteria, thereby mitigating the impact of any single metric's limitations. Our analysis presents baseline performance using greedy decoding, compares it to SC performance, and demonstrates the significance of observed differences through bootstrapped 95% confidence intervals. This evaluation framework ensures that our findings are statistically robust while accounting for variance in model outputs and task difficulty.

4 How Effective is Self-Consistency for Long-Context Problems?

Our experiments reveal systematic challenges for SC in long-context tasks. In the section that follows, we first analyze performance degradation across tasks (§4.1), then examine model-scale trends (§4.2), and finally explore how robust these effects are to variations in SC implementations (§4.3).

4.1 Performance Degradation Across Tasks

Table 2 shows the performance difference when applying self-consistency across numerous tasks and models, along with the specific aggregation method used for each task (USC for open-ended generation, Majority Voting for QuALITY). **Applying majority voting SC or USC leads to virtually no significant gains across many long-context task-model pairs.** Among 56 dataset-model pairs, only three illustrate statistically significant improvement from applying SC or USC.

ROUGE metrics demonstrate significant declines in 79% of model-dataset pairs (19/24), with smaller models experiencing the most severe degradation (e.g., Qwen-3B: -5.0 ROUGE). While LLM Judge scores show more modest effects, they still indicate degradation in 21% of cases (5/24), with no instances of significant improvement. This consistent pattern

Dataset	Model	ROUGE		Judge	
		Difference	95% CI	Difference	95% CI
GovReport (USC)	GPT-4o	$-2.2_{(23.6\leftarrow25.8)}$	[−2.6, −1.8]	$-0.8_{(80.0\leftarrow80.8)}$	[−2.3, 0.6]
	GPT-4o-Mini	$-1.7_{(23.8\leftarrow25.5)}$	[−2.0, −1.4]	$0.5_{(79.1\leftarrow78.6)}$	[−1.4, 2.2]
	LLaMA-3.3-70B	$0.1_{(29.5\leftarrow29.4)}$	[−0.4, 0.5]	$-0.6_{(81.6\leftarrow82.2)}$	[−2.1, 0.8]
	LLaMA-3.1-8B	$-1.1_{(29.4\leftarrow30.5)}$	[−1.5, −0.6]	$-0.1_{(80.1\leftarrow80.2)}$	[−1.9, 1.5]
	LLaMA-3.2-3B	$-2.2_{(27.0\leftarrow29.2)}$	[−2.8, −1.6]	$-1.0_{(76.8\leftarrow77.8)}$	[−3.3, 0.9]
	Qwen-2.5-72B	$-0.7_{(28.4\leftarrow29.1)}$	[−1.1, −0.3]	$0.3_{(80.7\leftarrow80.4)}$	[−1.4, 1.9]
	Qwen-2.5-7B	$-2.7_{(25.8\leftarrow28.7)}$	[−3.3, −2.4]	$-0.9_{(79.8\leftarrow80.7)}$	[−2.1, 0.4]
	Qwen-2.5-3B	$-5.0_{(23.9\leftarrow28.9)}$	[−5.6, −4.5]	$-2.3_{(76.4\leftarrow78.7)}$	[−4.4, −0.3]
QMSum (USC)	GPT-4o	$-1.0_{(17.8\leftarrow18.8)}$	[−1.5, −0.5]	$-0.2_{(58.0\leftarrow58.2)}$	[−2.6, 2.1]
	GPT-4o-Mini	$-0.3_{(18.0\leftarrow18.3)}$	[−0.9, 0.3]	$-0.7_{(57.9\leftarrow58.6)}$	[−3.1, 1.5]
	LLaMA-3.3-70B	$-0.4_{(18.9\leftarrow19.3)}$	[−1.1, 0.3]	$0.5_{(57.6\leftarrow57.1)}$	[−2.0, 3.1]
	LLaMA-3.1-8B	$-1.3_{(17.4\leftarrow18.7)}$	[−2.0, −0.5]	$0.4_{(53.6\leftarrow53.2)}$	[−1.9, 2.8]
	LLaMA-3.2-3B	$-1.3_{(15.3\leftarrow16.6)}$	[−2.1, −0.6]	$0.7_{(48.8\leftarrow48.1)}$	[−2.0, 3.3]
	Qwen-2.5-72B	$-0.6_{(17.8\leftarrow18.4)}$	[−1.1, 0.1]	$-0.4_{(56.8\leftarrow57.2)}$	[−2.7, 2.0]
	Qwen-2.5-7B	$-0.8_{(17.4\leftarrow18.6)}$	[−1.8, −0.7]	$-0.4_{(56.4\leftarrow56.8)}$	[−2.7, 2.0]
	Qwen-2.5-3B	$-2.3_{(14.4\leftarrow16.7)}$	[−3.0, −1.7]	$-1.5_{(50.8\leftarrow52.3)}$	[−4.0, 1.2]
SQuALITY (USC)	GPT-4o	$-1.5_{(17.2\leftarrow18.7)}$	[−1.7, −1.3]	$-1.1_{(80.9\leftarrow82.0)}$	[−1.9, −0.4]
	GPT-4o-Mini	$-1.0_{(16.4\leftarrow17.4)}$	[−1.1, −0.8]	$-0.4_{(79.5\leftarrow79.9)}$	[−1.2, 0.4]
	LLaMA-3.3-70B	$-0.1_{(19.6\leftarrow19.7)}$	[−0.2, 0.1]	$0.5_{(80.1\leftarrow79.6)}$	[−0.3, 1.2]
	LLaMA-3.1-8B	$-1.2_{(18.3\leftarrow19.5)}$	[−1.4, −1.0]	$0.1_{(77.4\leftarrow77.3)}$	[−0.7, 0.9]
	LLaMA-3.2-3B	$-1.7_{(16.7\leftarrow18.4)}$	[−2.0, −1.5]	$-0.3_{(75.6\leftarrow75.9)}$	[−1.1, 0.5]
	Qwen-2.5-72B	$-0.6_{(18.9\leftarrow19.5)}$	[−0.7, −0.4]	$-1.6_{(79.9\leftarrow81.6)}$	[−2.4, −0.9]
	Qwen-2.5-7B	$-1.3_{(17.6\leftarrow18.9)}$	[−1.4, −1.1]	$-1.3_{(77.2\leftarrow78.5)}$	[−2.1, −0.6]
	Qwen-2.5-3B	$-2.6_{(16.1\leftarrow18.7)}$	[−2.8, −2.4]	$-1.5_{(75.5\leftarrow77.0)}$	[−2.3, −0.8]

Dataset	Model	Difference	95% CI
Qasper (USC)	GPT-4o	$3.3_{(51.5\leftarrow48.2)}$	[1.0, 5.9]
	GPT-4o-Mini	$0.1_{(48.0\leftarrow47.9)}$	[−1.9, 2.0]
	LLaMA-3.3-70B	$-2.5_{(49.4\leftarrow51.9)}$	[−5.0, −0.4]
	LLaMA-3.1-8B	$-15.1_{(32.0\leftarrow47.1)}$	[−19.6, −11.0]
	LLaMA-3.2-3B	$-1.3_{(38.9\leftarrow40.2)}$	[−4.9, 2.3]
	Qwen-2.5-72B	$-0.8_{(49.7\leftarrow50.5)}$	[−3.0, 1.0]
	Qwen-2.5-7B	$-7.3_{(41.1\leftarrow48.4)}$	[−11.9, −3.4]
	Qwen-2.5-3B	$-10.8_{(31.3\leftarrow42.1)}$	[−15.9, −6.5]
Narrative QA (USC)	GPT-4o	$0.3_{(30.1\leftarrow29.8)}$	[−1.6, 1.6]
	GPT-4o-Mini	$0.6_{(28.2\leftarrow27.6)}$	[−0.4, 2.0]
	LLaMA-3.3-70B	$-0.7_{(15.8\leftarrow16.5)}$	[−2.2, 0.8]
	LLaMA-3.1-8B	$-2.3_{(10.6\leftarrow12.9)}$	[−4.7, −0.6]
	LLaMA-3.2-3B	$-0.6_{(8.2\leftarrow8.8)}$	[−2.6, 1.2]
	Qwen-2.5-72B	$0.4_{(15.5\leftarrow15.1)}$	[−0.6, 1.7]
	Qwen-2.5-7B	$-0.2_{(11.7\leftarrow11.9)}$	[−2.3, 2.1]
	Qwen-2.5-3B	$-1.7_{(9.5\leftarrow11.2)}$	[−4.6, 0.2]

Dataset	Model	Difference	95% CI
MuSiQue (USC)	GPT-4o	$0.3_{(38.5\leftarrow38.2)}$	[−2.6, 3.4]
	GPT-4o-Mini	$2.8_{(36.1\leftarrow33.3)}$	[0.2, 5.8]
	LLaMA-3.3-70B	$2.5_{(44.7\leftarrow42.2)}$	[−0.5, 5.6]
	LLaMA-3.1-8B	$4.6_{(25.5\leftarrow20.9)}$	[1.7, 8.4]
	LLaMA-3.2-3B	$-3.0_{(17.9\leftarrow20.9)}$	[−6.6, 0.1]
	Qwen-2.5-72B	$-0.2_{(50.3\leftarrow50.5)}$	[−2.6, 1.8]
	Qwen-2.5-7B	$0.9_{(29.6\leftarrow28.7)}$	[−2.3, 4.1]
	Qwen-2.5-3B	$0.3_{(12.8\leftarrow12.5)}$	[−3.0, 3.5]
QuALITY (Maj. Vote)	GPT-4o	$-1.0_{(90.4\leftarrow91.4)}$	[−1.7, −0.3]
	GPT-4o-Mini	$-0.3_{(79.7\leftarrow80.0)}$	[−1.1, 0.4]
	LLaMA-3.3-70B	$-0.3_{(87.0\leftarrow87.3)}$	[−0.8, 0.1]
	LLaMA-3.1-8B	$-6.0_{(63.7\leftarrow69.7)}$	[−7.8, −4.3]
	LLaMA-3.2-3B	$-0.7_{(52.4\leftarrow53.1)}$	[−2.4, 1.0]
	Qwen-2.5-72B	$-0.7_{(86.8\leftarrow87.5)}$	[−1.3, −0.1]
	Qwen-2.5-7B	$0.0_{(74.1\leftarrow74.1)}$	[−0.6, 0.7]
	Qwen-2.5-3B	$-0.3_{(63.6\leftarrow63.9)}$	[−0.9, −0.4]

Table 2: SC and USC performance difference across tasks and models. Universal Self-Consistency (USC) is used for all tasks, except QuALITY, which uses majority voting, for aggregation. *Red* indicate statistically significant degradation (95% CI); *green* highlights rare improvements. **Application of SC or USC leads to virtually no significant gains across many long-context task-model pairs (53 of 56 pairs show no statistically significant improvement).** Larger models (LLaMA-70B, Qwen-72B) show milder degradation but no consistent gains, underscoring self-consistency's fundamental limitations.

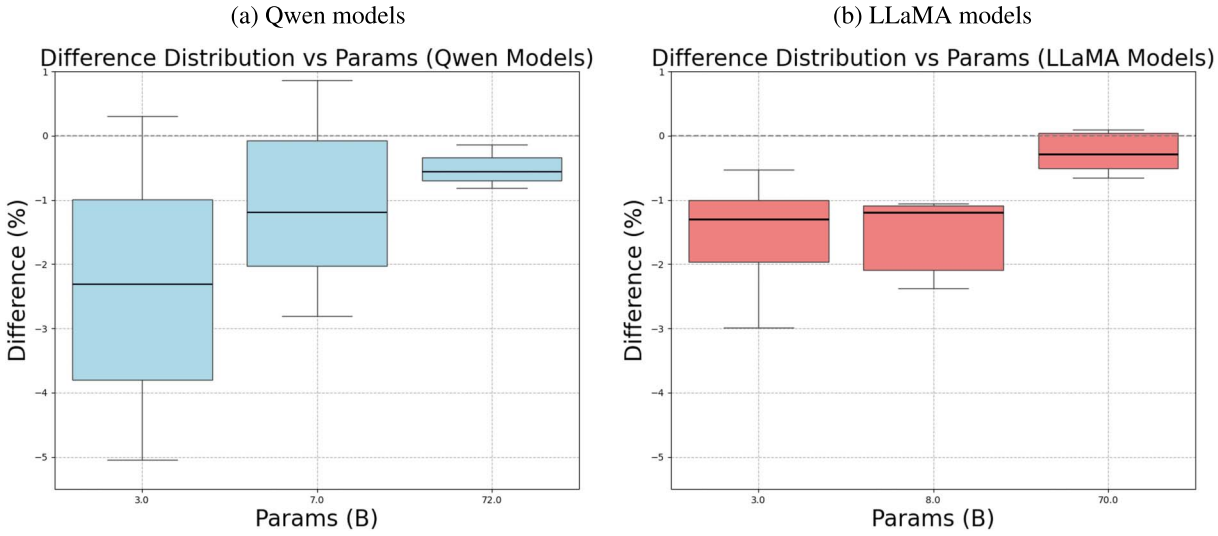


Figure 2: Average performance difference distribution for Qwen (left) and LLaMA (right) models across all tasks (excluding NQ-Open) in Table 2. The y-axis shows the difference in performance between SC and baseline approaches (negative values indicate degradation). Box plots show quartiles with whiskers extending to min/max values. Both model families demonstrate reduced performance degradation as model size increases, **but even the largest models still fail to break-even**.

across surface-level alignment and holistic judgment metrics suggests that SC fundamentally compromises summarization quality.

While results show slightly more variation in the QA domain, they still predominantly challenge SC's effectiveness. Of 32 model-dataset pairs, only three exhibited statistically significant improvements: GPT-4o on Qasper ($3.3_{(51.5\leftarrow 48.2)}$), GPT-4o-mini on MuSiQue ($2.8_{(36.1\leftarrow 33.3)}$), and LLaMA-3.1-8B on MuSiQue ($4.6_{(25.5\leftarrow 20.9)}$). However, these improvements are isolated cases rather than indicative of a broader trend. Notably, LLaMA-3.1-8B's performance varies dramatically across datasets, showing improvement on MuSiQue but suffering a severe $-15.1_{(32.0\leftarrow 47.1)}$ degradation on Qasper, which we attribute to poor instruction following despite adequate performance on other datasets. Eight pairs show significant degradation, particularly among smaller models.

Robustness to Choice of Metric. Despite differences in magnitude, the consistency between ROUGE and LLM Judge metrics in indicating performance degradation suggests that our findings reflect genuine limitations of SC rather than artifacts of specific evaluation methods. The more moderate effects observed in LLM Judge scores indicate that human-like evaluation demonstrates greater robustness to surface-level variations in outputs while still capturing the overall negative impact of SC on model performance.

4.2 Model-Scale Dependency

Figure 2 shows scaling between model size and performance degradation under SC and USC, with greater declines in smaller models. We compute Pearson and Spearman correlations between model size and performance difference. We find weak-to-moderate monotonicity, with stronger effects under ROUGE (Spearman's $\rho = 0.50$, $p < 0.01$) than LLM-Judge (Spearman's $\rho = 0.31$, $p < 0.05$). However, this relationship is nonlinear, as the Pearson correlation is *not* statistically significant under either ROUGE nor the LLM Judge. While larger models exhibit more stability than smaller ones, this ''resistance'' to degradation rarely translates into actual improvements. This pattern suggests that **while increased model scale might mitigate SC's adverse effects, it fails to unlock SC's promised benefits in long-context scenarios.**

4.3 Robustness to SC Aggregation Method

Having established the generally negative impact of USC (on generative tasks) and Majority Voting SC (on QUALITY), we further investigated whether an alternative aggregation strategy could alter these findings. To ensure our findings are not

Model	Difference	95% CI
GovReport (ROUGE)		
GPT-4o	$-1.8_{(24.0 \leftarrow 25.8)}$	[−2.2, −1.4]
GPT-4o-mini	$-1.3_{(24.2 \leftarrow 25.5)}$	[−1.6, −1.0]
GovReport (Judge)		
GPT-4o	$-1.3_{(79.5 \leftarrow 80.8)}$	[−2.6, 0.1]
GPT-4o-mini	$-0.4_{(78.2 \leftarrow 78.6)}$	[−2.0, 1.2]
Qasper		
GPT-4o	$-0.1_{(48.1 \leftarrow 48.2)}$	[−3.5, 2.8]
GPT-4o-mini	$0.1_{(48.0 \leftarrow 47.9)}$	[−1.6, 1.3]
MuSiQue		
GPT-4o	$-1.9_{(36.3 \leftarrow 38.2)}$	[−4.9, 0.5]
GPT-4o-mini	$1.5_{(34.8 \leftarrow 33.3)}$	[−0.7, 4.0]

Table 3: Soft self-consistency results on select datasets. Alternative SC implementation fails to show improvement, showing significant degradation (red) on GovReport (ROUGE). Results suggest **SC's limitations persist across implementation strategies.** Performance patterns mirror standard SC across metrics and model sizes.

artifacts of our primary SC implementations, we evaluated Soft SC on a subset of tasks (Table 3).

SC's limitation persists across implementation strategies. This approach fails to improve over baseline, showing significant degradation on GovReport (ROUGE) and no significant gains across other tasks. The consistent degradation observed with both standard SC/USC and Soft SC implementations suggests that the limitations are fundamental to the SC approach rather than artifacts of specific implementation choices. This consistent pattern of failure across models, tasks, and implementations strengthens our central hypothesis: that position bias undermines SC's core assumption of error independence by inducing strongly correlated errors. This suggests a systematic, mechanistic cause, which we isolate in the following section through a series of controlled experiments designed to explain the observed performance degradations. Together with the SC and USC results in Table 2, these findings demonstrate that SC's limitations persist regardless of aggregation strategy.

5 Understanding the Phenomenon: Position Bias and Self-Consistency

To probe the failures cataloged in §4, we ask if *positional bias* creates correlated errors that SC merely amplifies. We test (i) context-length ablations (§5.2), (ii) paired QA–retrieval evaluations (§5.2), and (iii) robustness checks (§5.3), each confirming that the degradation stems from position-induced bias, not SC.

5.1 Controlled Exploration of Position Bias

We evaluate SC's impact through controlled experiments on QA and TR tasks (§3). The QA task tests information synthesis, while TR isolates information retrieval, allowing us to separate position effects on access from those on reasoning. Our experimental design follows that of Liu et al. (2024), using their variant of NQ-Open with varying context sizes (10, 20, and 30 documents) and gold document positions (beginning, middle, or end) to assess position sensitivity. We employ three complementary metrics to quantify performance: QA accuracy measures answer correctness, TR accuracy evaluates document identification success, and the performance delta (Equation 1), which reveals SC's relative impact across positions. To ensure our findings are robust to implementation choices, we evaluate multiple configurations by varying sample counts (4, 8, 16) and temperatures (0.2, 1.0, 1.8).

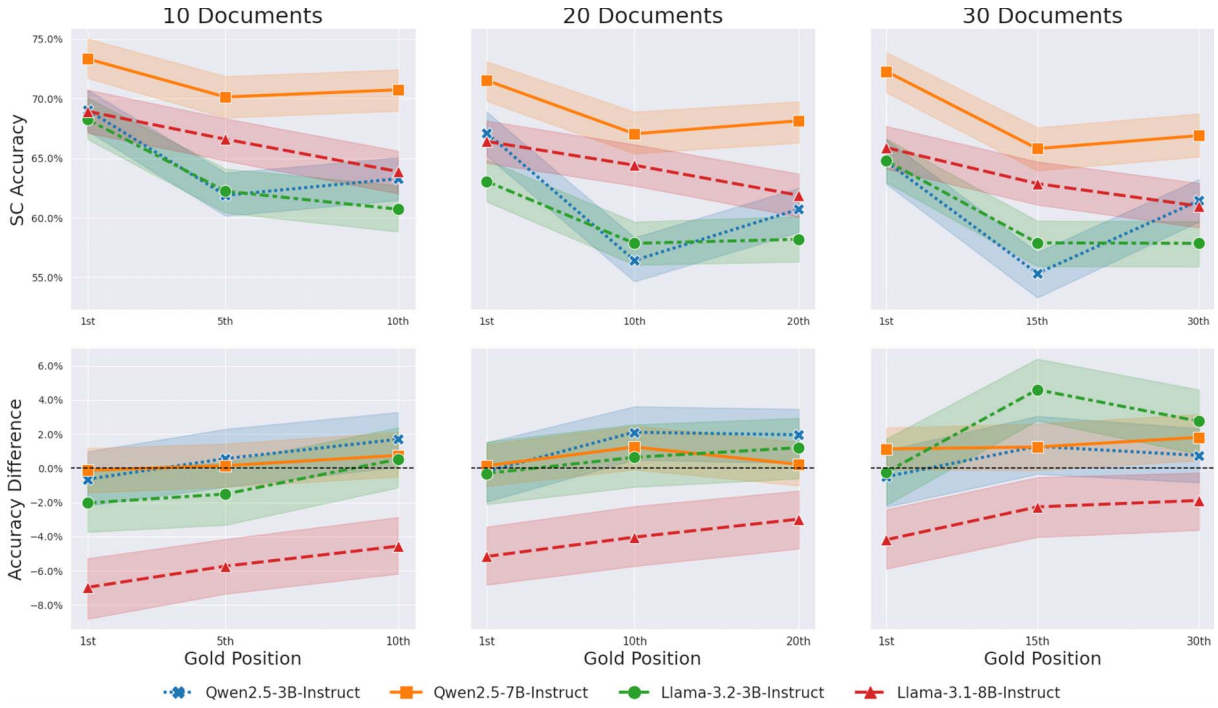
Extensive literature (Perez et al., 2021; Lu et al., 2022; Mishra et al., 2022, *inter alia*) has demonstrated that LLM performance is susceptible to the language and structure of their prompts. To examine the effects of prompt formatting, we test three distinct prompt formulations: documents-then-question (Doc-Q), question-then-documents (Q-Doc), and a query-aware approach with the question bracketing the documents (Q-Doc-Q).

5.2 Evidence of Correlated Errors

Our analysis reveals that SC's impact varies systematically with document position, fundamentally challenging previous assumptions about its effectiveness. The data demonstrate distinct patterns across task types and positions, providing strong evidence for position-induced correlated errors.

Question Answering Tasks. In QA tasks, SC performance exhibits a distinctive U-shaped curve (Figure 3a), with accuracy peaking when relevant information appears at context boundaries but dropping significantly for middle positions; this pattern persists across all models and context lengths. The position sensitivity becomes more

(a) Question Answering



(b) Text Retrieval

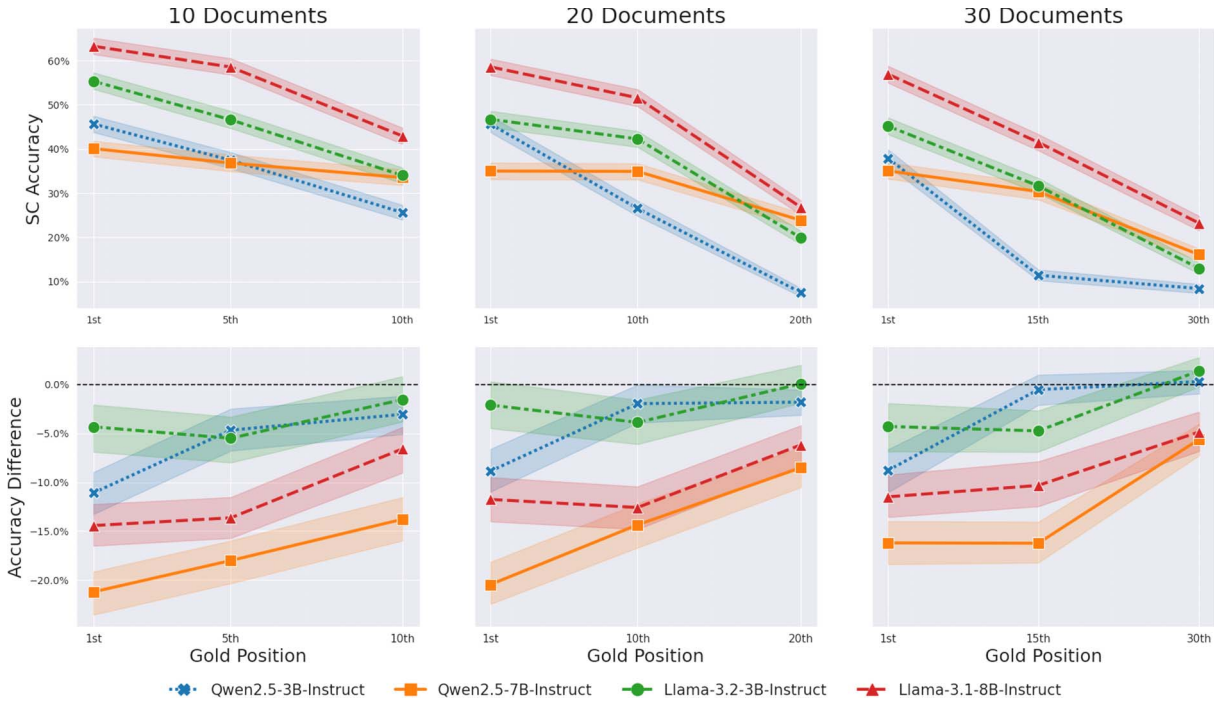


Figure 3: Self-consistency accuracy across models for NQ-Open, showing positional bias. (a) QA accuracy is highest at the beginning or end of context, with LLaMA-3.1 degrading under SC across positions; a U-shaped pattern persists across context lengths and model sizes. The accuracy difference reveals an upward trend as the gold document position moves later in context–**SC provides relatively less harm for later positions but never achieves the performance levels of baseline models on early positions.** (b) TR accuracy also peaks at the start, with severe drops as context length and position of relevant information increase. The corresponding difference plots demonstrate consistent negative impact across positions, with particularly severe degradation (20–25%) for early positions in longer contexts.

pronounced as context length increases—for instance, in 30-document contexts, Qwen-2.5-7B shows a $15\% \pm 2\%$ accuracy drop for middle positions compared to boundary positions.

When examining the performance delta, the relationship between position and SC effectiveness becomes more apparent. This reveals a consistent upward trend as the gold document moves later in the context, suggesting that SC's harmful effect is relatively lessened for information appearing at the end of the context. However, this requires careful interpretation: While the performance delta improves for later positions, absolute performance under SC never matches the baseline on early-position documents, demonstrating that **SC may partially compensate for position-related degradation but cannot overcome the underlying position bias**.

Text Retrieval Tasks. TR tasks reveal a more severe, qualitatively different manifestation of position bias (Figure 3b). Unlike QA's U-shaped curve, TR accuracy shows a monotonic decline as the gold document moves later in the context. The performance degradation becomes particularly severe with longer contexts—in 30-document scenarios, accuracy drops by $40\% \pm 3\%$ when the gold document appears in the final third of the context. SC consistently deteriorates performance across all positions, with larger models like Qwen-2.5-7B and LLaMA-3.1-8B showing significant 20-25% performance reductions. To illustrate this failure mode, when the gold document appears early in a 30-document context, Qwen-2.5-7B generates multiple samples that consistently focus on later, irrelevant documents, amplifying its position bias through repeated sampling. The consistency of this pattern provides strong evidence that **SC not only fails to mitigate position bias but can actively exacerbate it**, particularly in tasks requiring precise information retrieval.

Analysis of Intermediate Generations. To trace the origin of these aggregation-level failures, we analyzed the distribution of correct answers within the intermediate generations before SC is applied (Figure 4). This analysis reveals that the biases observed in the final aggregated output are deeply rooted in the initial sampling step. For QA tasks, the proportion of cases where all or most of the eight samples are correct follows the same U-shaped pattern seen in the final output, dropping when the gold answer is in the middle of the context. The effect is even more pronounced in TR tasks, which exhibit a stark monotonic decline. For instance, in 30-document contexts with the answer at the end, nearly 60% of the time, none of the eight intermediate samples correctly identified the source document. This provides direct evidence that the **errors are strongly correlated across samples**, stemming from the model's inherent positional bias and violating the core assumption of independence that underpins self-consistency.

5.3 Robustness Analysis

We conduct extensive robustness checks across prompt structure and sampling parameters, known to sway LLM behavior (Perez et al., 2021; Lu et al., 2022; Mishra et al., 2022). Our results demonstrate that implementation choices can affect absolute performance but cannot resolve the underlying challenges of position bias.

Prompt Engineering Effects. Even carefully engineered prompts cannot overcome position-dependent performance degradation. Among the three tested formats, the query-aware approach (Q-Doc-Q) shows a modest advantage (+2–3%) for early positions in QA tasks. However, this benefit diminishes with longer contexts and middle/end positions (Figure 5). TR tasks demonstrate even more striking limitations: While format choice can influence overall accuracy by up to 20%, all formats suffer severe degradation for later positions in long contexts. For example, the documents-then-question format (Doc-Q) drops to approximately $10\% \pm 2\%$ accuracy in 30-document contexts when the target document appears in the final third, regardless of prompt engineering efforts.

Parameter Sensitivity. Varying SC implementation parameters reveal similar limitations (Figure 6). Increasing sample count yields diminishing returns, where doubling generations from 8 to 16 yields less than 1% improvement in absolute accuracy while doubling the computational overhead in the QA task. More concerning, larger sample counts actively harm TR performance, likely because they amplify the model's tendency to fixate on positionally favored but incorrect documents. Temperature variations primarily affect overall performance levels rather than positional

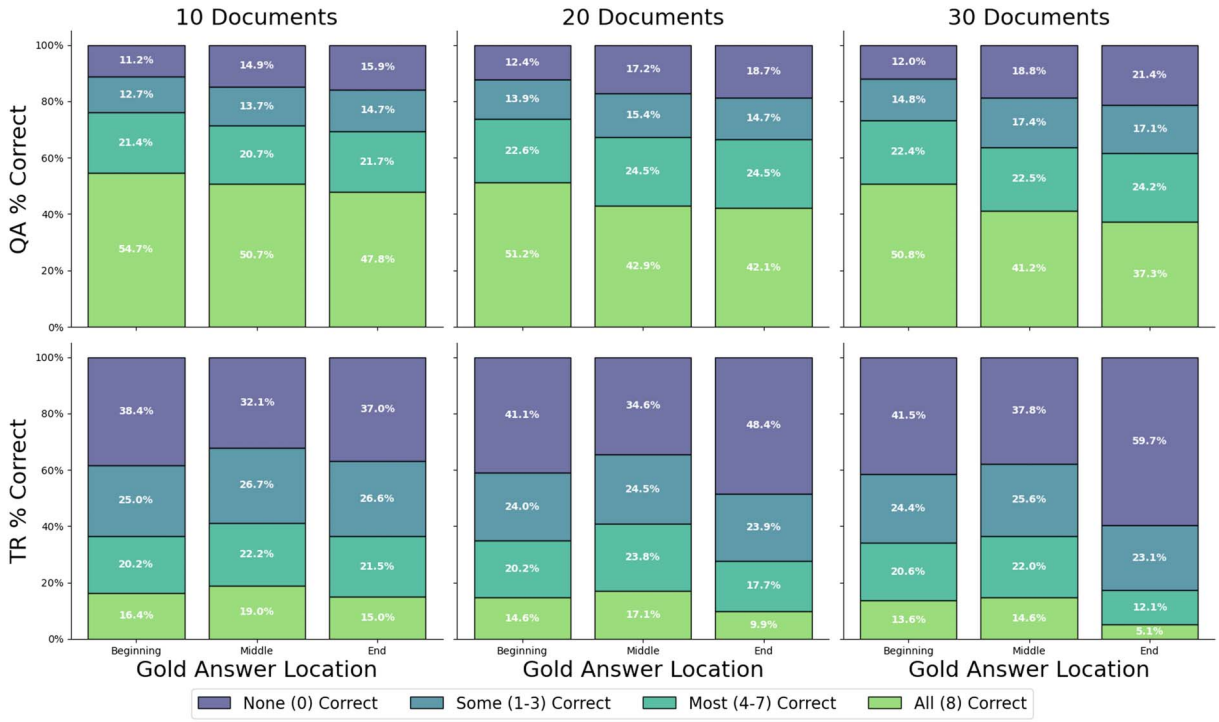


Figure 4: Distribution of correctness across eight intermediate generations, demonstrating positional bias before self-consistency aggregation. The stacked bars show the percentage of cases where none (0), some (1–3), most (4–7), or all (8) of the intermediate samples were correct. (Top Row) For Question Answering (QA), the prevalence of highly correct sample sets (green bars) follows a U-shaped curve, degrading when information is in the middle. (Bottom Row) For Text Retrieval (TR), correctness declines monotonically, with a dramatic increase in cases where zero samples are correct as the gold document is placed later in longer contexts. This visualization confirms that **the errors amplified by SC are systemic and correlated**, originating from the model's fundamental bias rather than the aggregation process.

patterns, with extreme temperatures leading to universal degradation.

6 Discussion

Our findings reveal a fundamental limitation in applying self-consistency (SC) to long-context tasks. Whether using sophisticated judge-based aggregation for open-ended generation or traditional majority voting for multiple-choice questions, the result is the same: Performance either degrades or fails to improve. This pattern persists across both automated metrics and human-like evaluations, indicating a systemic issue. We argue that this failure stems from positional bias, which induces correlated errors that SC incorrectly amplifies.

6.1 Effects of Position Bias on Long-Context

SC was designed to filter uncorrelated errors by leveraging diverse reasoning paths, but our experiments show this assumption fails in long contexts. Positional bias induces strongly correlated errors across samples, and these biases (e.g., over-relying on early context) persist despite varying sample counts or temperature. While one might argue that these failures are not specific to SC but are merely a symptom of models struggling with the difficulty of long-context reasoning, our controlled experiments refute this. The fact that model performance is systematically dependent on the position of the correct information, not just its presence in a long context, demonstrates that the issue is biased information access, which SC incorrectly amplifies.

Building on this insight, we find that SC reinforces these positional blind spots, effectively amplifying rather than mitigating systemic errors. This amplification is a direct result of the aggregation process operating on a set of samples already skewed by bias, where, as shown in our analysis of intermediate generations (Figure 4), a majority of reasoning paths are often predisposed to the same positional error. This mechanism manifests most dramatically in our text retrieval experiments, where SC led to 20–25% performance reductions in models like Qwen-2.5-7B and LLaMA-3.1-8B.

We frequently observed that a majority of samples incorrectly selected documents from early positions when the correct document was located in the middle or at the end, resulting in a unanimous but incorrect consensus. For instance, in Appendix B, we illustrate an example in which seven of the eight generations incorrectly identify information from the first document as being correct when, in actuality, the gold document was at the end. These findings suggest that traditional aggregation methods may be fundamentally unsuited for long-context tasks, motivating exploration into position-aware aggregation methods that explicitly account for and counteract positional effects.

6.2 Practical Implications and Future Directions

The unsuitability of standard SC for long-context tasks has significant practical implications. In critical fields such as legal analysis or medical diagnostics, overlooking key information buried in a lengthy document can have severe consequences. Practitioners cannot assume that this common inference-time technique will improve reliability and should instead explore alternatives.

Several directions emerge from our findings. First, position-aware aggregation methods could explicitly account for document position when combining multiple samples, perhaps by weighting responses based on their attention patterns across the context. Second, contrastive sampling strategies might reduce the correlation between errors by explicitly encouraging diversity in the context regions to which models attend across samples. Third, attention recalibration techniques, such as those proposed by Hsieh et al. (2024b), could be adapted specifically for sampling-based methods to mitigate positional effects before aggregation. Finally, retrieval-augmented approaches could be combined with SC, allowing models to process small context windows where traditional SC remains effective.

These approaches represent testable hypotheses for follow-up work addressing the challenge of correlated positional errors. Future work may unlock more reliable performance in long-context scenarios while maintaining the benefits of aggregation by developing techniques that directly counter position bias rather than simply applying SC.

7 Conclusion

Our study provides the first in-depth analysis of self-consistency in long-context scenarios, challenging its presumed universality. Across 651 experiments and eight models, SC consistently *failed to improve performance*—and often degraded it—due to positional biases. Even state-of-the-art models like GPT-4o exhibited these failures, with this degradation persisting across sampling configurations, prompt formats, and evaluation metrics, underscoring SC's fundamental incompatibility with long-context tasks. These findings suggest that inference-time techniques like SC may not generalize to long contexts and that addressing these challenges, particularly the effects of positional bias, will require deeper architectural innovations in attention and aggregation.

For practitioners, our work highlights the risks of porting short-context methods to long-context applications. Future research should investigate hybrid approaches, such as combining SC with positional debiasing, rather than relying on standard SC. By rigorously establishing SC's limitations, this work provides the foundational analysis needed to drive targeted innovations in context-aware inference methods. Rather than attempting to characterize and solve this complex challenge simultaneously, we provide a comprehensive empirical analysis necessary to drive targeted innovations in context-aware inference methods.

Limitations

While our study reveals fundamental challenges for SC in long-context settings, several limitations warrant discussion. Our evaluation covers contexts up to 30 documents (5–10K tokens), leaving open questions about extremely long contexts (100K+ tokens) where different dynamics might emerge. However, recent work (Modarressi et al., 2025) suggests that the effective context window of many models is far shorter than their advertised limits. This finding reinforces our focus as a critical analysis area where models are functional yet exhibit the systematic biases we investigate.

Our focus on textual tasks (summarization, QA, multi-hop reasoning) leaves questions about SC's utility in multi-modal long-context scenarios open. Additionally, while NQ-Open's controlled setup effectively isolates position bias, real-world

contexts may exhibit more complex positional interactions than those we captured.

We deliberately focus on problem characterization rather than solution development, which enables more effective future mitigation strategies by establishing a mechanistic understanding of why SC fails. This comprehensive groundwork (651 experiments across multiple models, tasks, and implementations) provides a robust foundation for developing targeted approaches for sampling-based methods for long-context reasoning.

Acknowledgments

We would like to thank Asli Celikyilmaz and Minlie Huang, who served as our TACL action editors, and the anonymous reviewers for their comments and feedback, as well as Zhouxiang Fang, Hannah Gonzalez, Emily Guan, Dongwei Jiang, Tianjian Li, Jiefu Ou, Angad Sandhu, Andrew Wang, and Jack Zhang for their constructive discussions which have helped to improve this work. This work was carried out in part at the Advanced Research Computing at Hopkins (ARCH) core facility, which is supported by the National Science Foundation (NSF) grant number OAC1920103.

References

Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. 2024. LongBench: A bilingual, multitask benchmark for long context understanding. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3119–3137. https://doi.org/10.18653/v1/2024.acl-long.172

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.

Xinyun Chen, Renat Aksitov, Uri Alon, Jie Ren, Kefan Xiao, Pengcheng Yin, Sushant Prakash, Charles Sutton, Xuezhi Wang, and Denny Zhou. 2024. Universal self-consistency for large language models. In *ICML 2024 Workshop on In-Context Learning*.

Pradeep Dasigi, Kyle Lo, Iz Beltagy, Arman Cohan, Noah A. Smith, and Matt Gardner. 2021. A dataset of information-seeking questions and answers anchored in research papers. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4599–4610. https://doi.org/10.18653/v1/2021.naacl-main.365

Alexander R. Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. Summeval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409. https://doi.org/10.1162/tacl_a_00373

Simon Frieder, Luca Pinchetti, Ryan-Rhys Griffiths, Tommaso Salvatori, Thomas Lukasiewicz, Philipp Petersen, and Julius Berner. 2024. Mathematical capabilities of chatgpt. In *Advances in Neural Information Processing Systems*, volume 36, pages 27699–27744.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis,

Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco

Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan Mc-Phie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U., Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A., Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783v3*.

Cheng-Ping Hsieh, Simeng Sun, Samuel Kriman, Shantanu Acharya, Dima Rekesh, Fei Jia, and Boris Ginsburg. 2024a. RULER: What's the real context size of your long-context language models? In *First Conference on Language Modeling*.

Cheng-Yu Hsieh, Yung-Sung Chuang, Chun-Liang Li, Zifeng Wang, Long Le, Abhishek Kumar, James Glass, Alexander Ratner, Chen-Yu Lee, Ranjay Krishna, and Tomas Pfister. 2024b. Found in the middle: Calibrating positional attention bias improves long context utilization. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 14982–14995. https://doi.org/10.18653/v1/2024.findings-acl.890

Luyang Huang, Shuyang Cao, Nikolaus Parulian, Heng Ji, and Lu Wang. 2021. Efficient attentions for long document summarization.

In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1419–1436. https://doi.org/10.18653/v1/2021.naacl-main.112

Athul Paul Jacob, Yikang Shen, Gabriele Farina, and Jacob Andreas. 2024. The consensus game: Language model generation via equilibrium search. In the *Twelfth International Conference on Learning Representations*.

Huiqiang Jiang, Qianhui Wu, Xufang Luo, Dongsheng Li, Chin-Yew Lin, Yuqing Yang, and Lili Qiu. 2024. Longllmlingua: Accelerating and enhancing llms in long context scenarios via prompt compression. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 1658–1677. https://doi.org/10.18653/v1/2024.acl-long.91

Greg Kamradt. 2023. Needle in a haystack - pressure testing llms. *GitHub*.

Tomáš Kočiskỳ, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gábor Melis, and Edward Grefenstette. 2018. The NarrativeQA reading comprehension challenge. *Transactions of the Association for Computational Linguistics*, 6:317–328. https://doi.org/10.1162/tacl_a_00023

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural Questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466. https://doi.org/10.1162/tacl_a_00276

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with PagedAttention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*. https://doi.org/10.1145/3600006.3613165

Jinhyuk Lee, Anthony Chen, Zhuyun Dai, Dheeru Dua, Devendra Singh Sachan, Michael Boratko, Yi Luan, Sébastien M. R. Arnold, Vincent Perot, Siddharth Dalmia, Hexiang Hu, Xudong Lin, Panupong Pasupat, Aida Amini, Jeremy R. Cole, Sebastian Riedel, Iftekhar Naim, Ming-Wei Chang, and Kelvin Guu. 2024. Can long-context language models subsume retrieval, rag, sql, and more? *arXiv preprint arXiv:2406.13121v1*.

Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. 2019. Latent retrieval for weakly supervised open domain question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6086–6096. https://doi.org/10.18653/v1/P19-1612

Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *ACL Workshop on Text Summarization Branches Out*. https://doi.org/10.18653/v1/2024.findings-acl.230

Lei Lin, Jiayi Fu, Pengli Liu, Qingyang Li, Yan Gong, Junchen Wan, Fuzheng Zhang, Zhongyuan Wang, Di Zhang, and Kun Gai. 2024. Just ask one more time! Self-agreement improves reasoning of language models in (almost) all scenarios. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 3829–3852.

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. https://doi.org/10.1162/tacl_a_00638

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 2022. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In *Annual Meeting of the Association for Computational Linguistics (ACL)*. https://doi.org/10.18653/v1/2022.acl-long.556

Seyed Iman Mirzadeh, Keivan Alizadeh, Hooman Shahrokhi, Oncel Tuzel, Samy Bengio, and Mehrdad Farajtabar. 2025. GSM-symbolic: Understanding the limitations of mathematical reasoning in large language models. In the *Thirteenth International Conference on Learning Representations*.

Swaroop Mishra, Daniel Khashabi, Chitta Baral, Yejin Choi, and Hannaneh Hajishirzi. 2022. Reframing instructional prompts to gptk's language. In *Annual Meeting of the Association for Computational Linguistics (ACL) - Findings*. https://doi.org/10.18653/v1/2022.findings-acl.50

Ali Modarressi, Hanieh Deilamsalehy, Franck Dernoncourt, Trung Bui, Ryan A. Rossi, Seunghyun Yoon, and Hinrich Schuetze. 2025. NoLiMa: Long-context evaluation beyond literal matching. In *Forty-second International Conference on Machine Learning*.

Marianna Nezhurina, Lucia Cipolina-Kun, Mehdi Cherti, and Jenia Jitsev. 2024. Alice in wonderland: Simple tasks showing complete reasoning breakdown in state-of-the-art large language models. *arXiv preprint arXiv:2406.02061v4*.

OpenAI. 2024. Hello GPT-4o.

Richard Yuanzhe Pang, Alicia Parrish, Nitish Joshi, Nikita Nangia, Jason Phang, Angelica Chen, Vishakh Padmakumar, Johnny Ma, Jana Thompson, He He, and Samuel Bowman. 2022. QuALITY: Question answering with long input texts, yes! In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5336–5358. https://doi.org/10.18653/v1/2022.naacl-main.391

Ethan Perez, Douwe Kiela, and Kyunghyun Cho. 2021. True few-shot learning with language models. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Uri Shaham, Maor Ivgi, Avia Efrat, Jonathan Berant, and Omer Levy. 2023. ZeroSCROLLS: A zero-shot benchmark for long text understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7977–7989. https://doi.org/10.18653/v1/2023.findings-emnlp.536

Uri Shaham, Elad Segal, Maor Ivgi, Avia Efrat, Ori Yoran, Adi Haviv, Ankit Gupta, Wenhan Xiong, Mor Geva, Jonathan Berant, and Omer Levy. 2022. SCROLLS: Standardized comparison over long language sequences. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 12007–12021. https://doi.org/10.18653/v1/2022.emnlp-main.823

Yi Tay, Mostafa Dehghani, Samira Abnar, Yikang Shen, Dara Bahri, Philip Pham, Jinfeng Rao, Liu Yang, Sebastian Ruder, and Donald Metzler. 2021. Long range arena: A benchmark for efficient transformers. In *International Conference on Learning Representations*.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2022. MuSiQue: Multihop questions via single-hop question composition. *Transactions of the Association for Computational Linguistics*, 10:539–554. https://doi.org/10.1162/tacl_a_00475

Alex Wang, Richard Yuanzhe Pang, Angelica Chen, Jason Phang, and Samuel Bowman. 2022. SQuALITY: Building a long-document summarization dataset the hard way. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1139–1156. https://doi.org/10.18653/v1/2022.emnlp-main.75

Han Wang, Archiki Prasad, Elias Stengel-Eskin, and Mohit Bansal. 2024. Soft self-consistency improves language models agents. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 287–301. https://doi.org/10.18653/v1/2024.acl-short.28

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023a. Self-consistency improves chain of thought reasoning in language models. In the *Eleventh International Conference on Learning Representations*.

Yiwei Wang, Yujun Cai, Muhao Chen, Yuxuan Liang, and Bryan Hooi. 2023b. Primacy effect of ChatGPT. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 108–115. https://doi.org/10.18653/v1/2023.emnlp-main.8

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-Thought prompting elicits reasoning

in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837.

Zhaofeng Wu, Linlu Qiu, Alexis Ross, Ekin Akyürek, Boyuan Chen, Bailin Wang, Najoung Kim, Jacob Andreas, and Yoon Kim. 2024. Reasoning or reciting? Exploring the capabilities and limitations of language models through counterfactual tasks. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1819–1862. https://doi.org/10.18653/v1/2024.naacl-long.102

Zheyang Xiong, Vasilis Papageorgiou, Kangwook Lee, and Dimitris Papailiopoulos. 2025. From artificial needles to real haystacks: Improving retrieval capabilities in LLMs by finetuning on synthetic data. In the *Thirteenth International Conference on Learning Representations*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115v2*.

Howard Yen, Tianyu Gao, Minmin Hou, Ke Ding, Daniel Fleischer, Peter Izasak, Moshe Wasserblat, and Danqi Chen. 2024. HELMET: How to evaluate long-context language models effectively and thoroughly. *arXiv preprint arXiv:2410.02694v2*.

Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. 2023. Large language models are not robust multiple choice selectors. In the *Twelfth International Conference on Learning Representations*.

Ming Zhong, Da Yin, Tao Yu, Ahmad Zaidi, Mutethia Mutuma, Rahul Jha, Ahmed Hassan Awadallah, Asli Celikyilmaz, Yang Liu, Xipeng Qiu, and Dragomir Radev. 2021. QMSum: A new benchmark for query-based multi-domain meeting summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5905–5921. https://doi.org/10.18653/v1/2021.naacl-main.472

APPENDIX

A Robustness Results

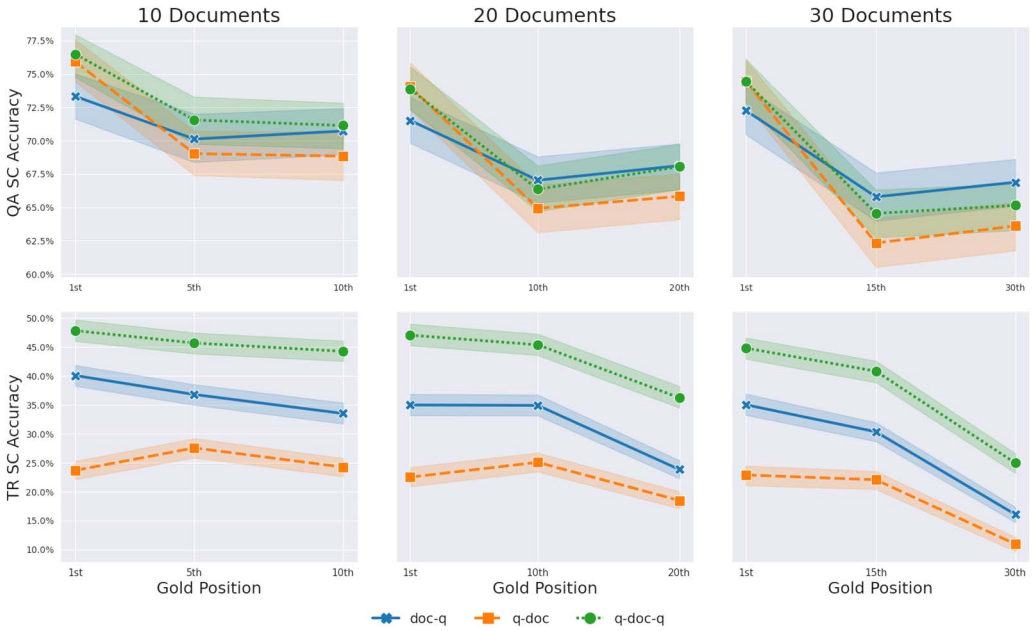
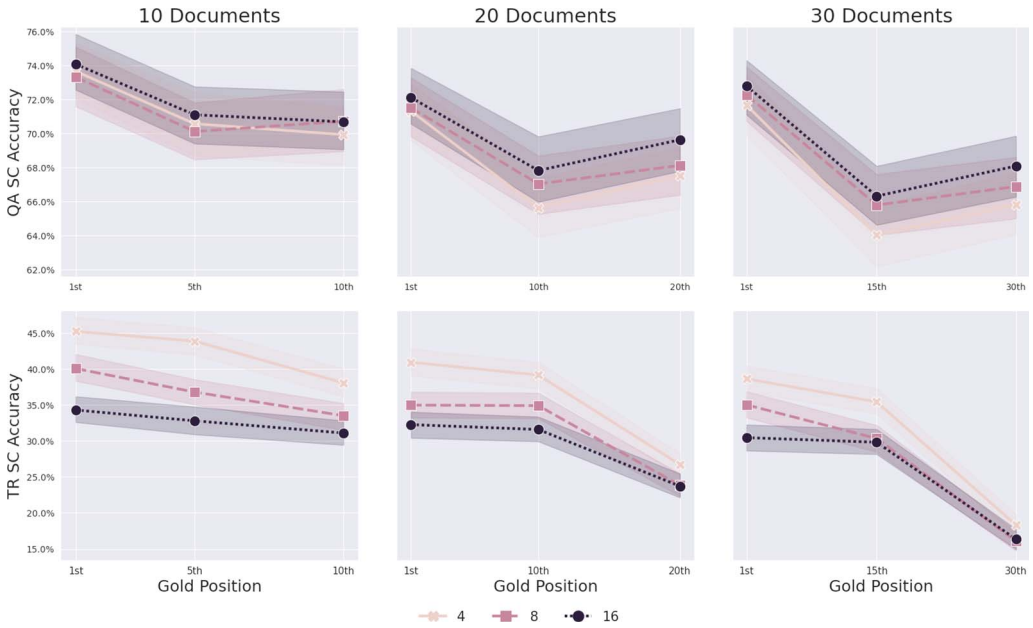


Figure 5: Effect of prompt format on QA and TR self-consistency accuracy for the Qwen-2.5-7B model. Different prompt formats show minimal impact on mitigating positional bias. While Q-Doc-Q slightly improves overall QA performance (top row), TR performance (bottom row) is more sensitive to format choice, with up to 20% performance *degradation* between formats. The consistent degradation pattern across gold positions indicates that position bias persists regardless of query-document ordering.

(a) # of Generations



(b) Temperature

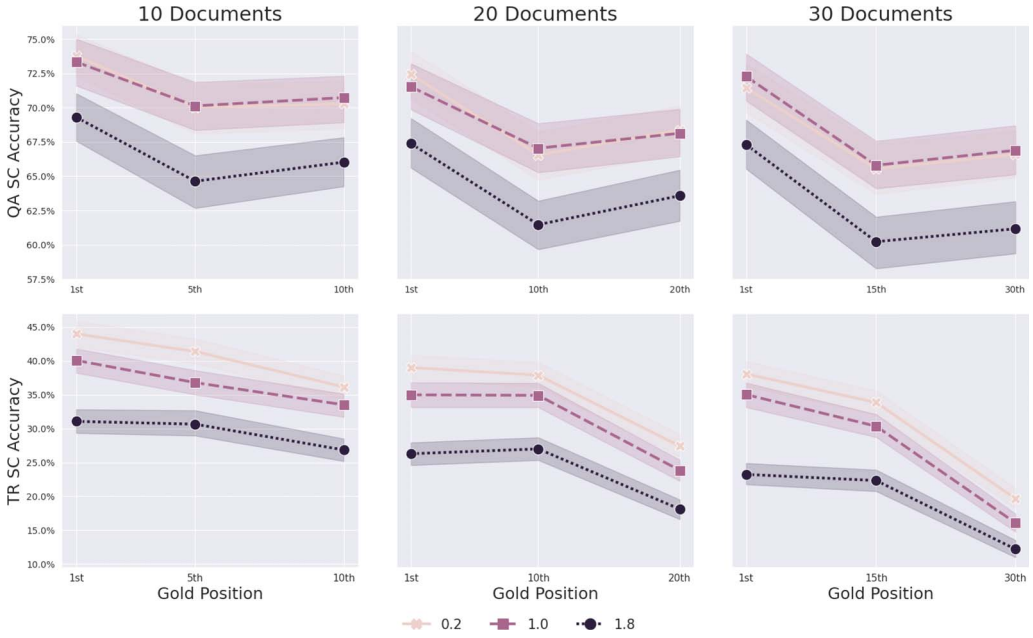


Figure 6: Effect of SC parameter variations on QA and TR accuracy for the Qwen-2.5-7B model. Increasing generations (a) slightly boosts QA accuracy but negatively impacts TR. Higher temperatures (b) degrade performance, especially for TR tasks, persisting positional degradation as gold information shifts deeper into the context.

B Error Analysis of Self-Consistency Failures

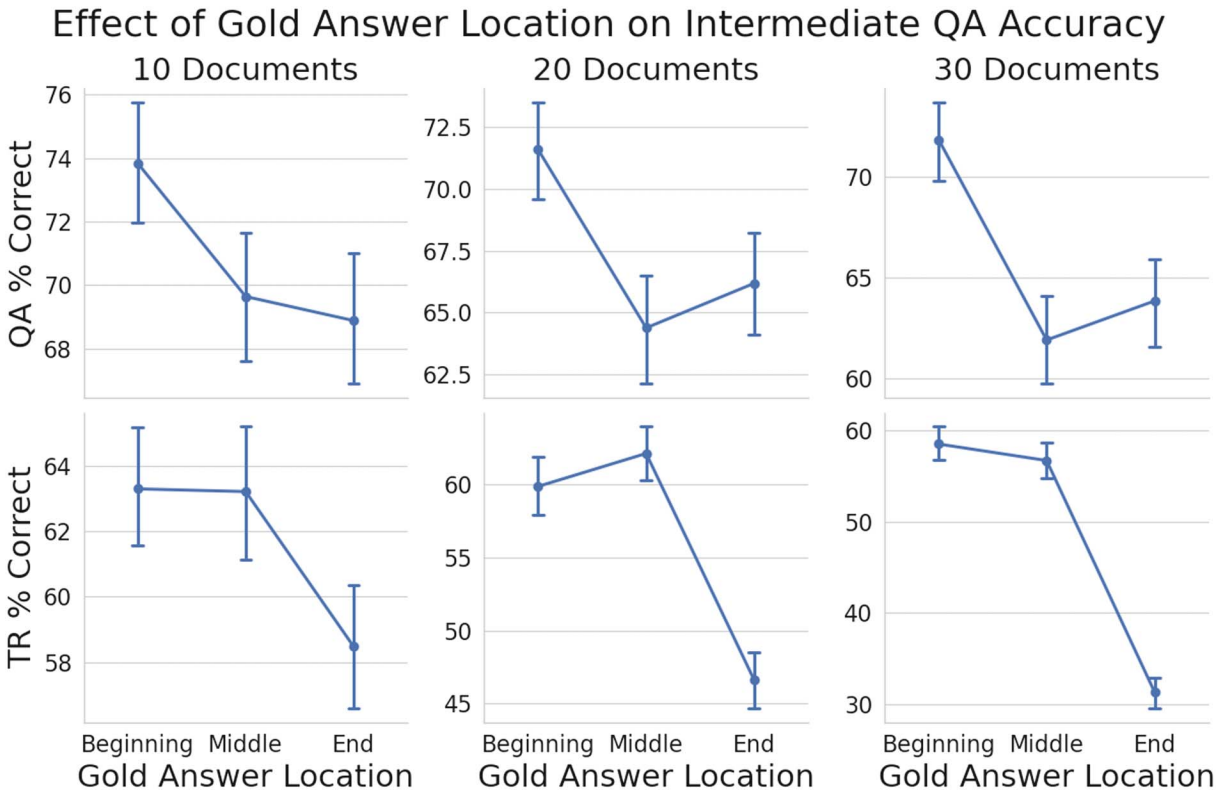


Figure 7: Error analysis of intermediate generations for the Qwen-2.5-7B model. The accuracy of the eight intermediate samples is plotted before aggregation, revealing the origin of the positional bias observed in Figure 3. For Question Answering (QA) tasks (top row), accuracy follows a U-shaped curve, degrading in the middle of the context. For Text Retrieval (TR) tasks (bottom row), accuracy shows a monotonic decline as the gold document is placed later in the context.

Illustrative Example: Correlated Errors in Question Answering
Prompt:
Write a high-quality answer for the given question using only the provided search results (some of which might be irrelevant).

Document [1](Title: Drafting of the Universal Declaration of Human Rights) ...John Peters Humphrey was newly appointed as Director of the Division of Human Rights within the United Nations Secretariat...
. . .
Document [20](Title: Cyrus Cylinder) The cylinder [of Cyrus] has also been referred to as...the first declaration of universal human rights...

Question: who wrote the first declaration of human rights
Answer:

Expected Answer:
`Cyrus`

Intermediate Generations (8 Samples at Temperature 1.0):
The model generates eight different reasoning paths. However, due to a systemic position bias, the majority of these paths incorrectly fixate on documents from the beginning of the context, violating SC's core assumption of independent errors.

1. `The first direct quote suggesting the author of the first Declaration of human rights is John Peters Humphrey can be found in multiple documents.`
2. `The first draft of the Universal Declaration of Human Rights, which is often regarded as the first comprehensive declaration of human rights on a global scale, was written by John Peters Humphrey.`
3. `Based on the provided information, specifically from document [8] and document [6], the first draft of what became the Universal Declaration of Human Rights was prepared by John Peters Humphrey.`
4. `The first draft of the Universal Declaration of Human Rights was written by John Peters Humphrey, a Canadian legal scholar, jurist, and human rights advocate.`
5. `Based on the provided documents, John Peters Humphrey is credited with writing the first draft of the Universal Declaration of Human Rights.`
6. `Based on the provided search results, the first draft of the Universal Declaration of Human Rights, often considered a foundational document in human rights, was written by John Peters Humphrey.`
7. `The first explicit declaration of human rights, the Declaration of the Rights of Man and of the Citizen, was set by France's National Constituent Assembly in 1789.`
8. `The first declaration of human rights, the Universal Declaration of Human Rights, was primarily drafted by John Peters Humphrey.`

Final Aggregated Answer:
`John Peters Humphrey`

Analysis:
This example demonstrates the mechanistic failure of self-consistency when faced with strong positional bias. The correct answer, **Cyrus**, is explicitly mentioned in a document near the *end* of the context (Document [20]). However, the model exhibits a powerful primacy bias, causing it to fixate on **John Peters Humphrey**, a name found in one of the *first* documents. This bias induces strongly correlated errors across the generated samples, with seven of the eight confidently proposing the same incorrect name. The majority voting process, unable to distinguish this systemic flaw from random noise, latches onto the incorrect consensus and amplifies it.

C Prompt Templates

C.1 Summarization Task Prompts

GovReport

You are given a report by a government agency. Write a one-page summary of the report.

Report:
{REPORT}

Summary:

QMSum

You are given a meeting transcript and a query containing a question or instruction. Answer the query in one or more sentences.

Transcript:
{TRANSCRIPT}

Query:
{QUERY}

Answer:

SQuALITY

You are given a story and a question. Answer the question in a paragraph.

Story:
{STORY}

Question:
{QUESTION}

Answer:

C.2 Question Answering Task Prompts

Qasper

You are given a scientific article and a question. Answer the question as concisely as you can, using a single phrase or sentence if possible. If the question cannot be answered based on the information in the article, write ‘‘unanswerable’’. If the question is a yes/no question, answer ‘‘yes’’, ‘‘no’’, or ‘‘unanswerable’’. Do not provide any explanation.

Article:
{ARTICLE}

Question:
{QUESTION}

Answer:

NarrativeQA

You are given a story, which can be either a novel or a movie script, and a question. Answer the question as concisely as you can, using a single phrase if possible. Do not provide any explanation.

Story:
{STORY}

Question:
{QUESTION}

Answer:

QuALITY

You are provided a story and a multiple-choice question with 4 possible answers (marked by A, B, C, D). Choose the best answer by writing its corresponding letter (either A, B, C, or D). Do not provide any explanation.

Story:
{STORY}

Question and Possible Answers:
{QUESTION_AND_OPTIONS}

Answer:

MuSiQue

You are given several paragraphs from Wikipedia and a question. Answer the question as concisely as you can, using a single phrase if possible. If the question cannot be answered based on the information in the paragraphs, write ‘‘unanswerable’’. Do not provide any explanation.

Paragraphs:
{PARAGRAPHS}

Question:
{QUESTION}

Answer:

C.3 USC Prompt

USC

I have generated the following responses to the question: {question}

{RESPONSES}

Evaluate these responses. Select the most consistent response based on majority consensus. Start your answer with ‘‘The most consistent response is Response X’’ (without quotes).

C.4 LLM Judge Prompts

LLM Judge (Summarization)

You are an expert evaluator assessing the quality of a generated summary compared to a reference summary. Consider the following factors in your evaluation:
- Consistency: Does the generated summary convey the same key information as the reference summary without contradictions?
- Relevance: Is the generated summary focused on the main points of the reference summary?
- Fluency: Is the generated summary grammatically correct and free of repetition or incoherence?
- Informativeness: Does the generated summary adequately cover the most important details in the reference?

Provide your evaluation as a single score on a scale from 0 to 100:
- 0 means the generated summary is completely unacceptable (e.g., incoherent, irrelevant, or contradictory).
- 100 means the generated summary is perfect (e.g., consistent, relevant, fluent, and informative).

Reference Summary:
{REFERENCE}

Generated Summary:
{PREDICTION}

Score (0–100):

LLM Judge (QA)

You are an expert evaluator assessing the quality of an answer generated in response to a question. Consider the following factors in your evaluation:
- Correctness: Does the generated answer accurately address the question based on the ground truth?
- Relevance: Is the generated answer focused and does it address the question without unnecessary information?
- Fluency: Is the generated answer grammatically correct and free of repetition or incoherence?

Provide your evaluation as a single score on a scale from 0 to 100:
- 0 means the generated answer is completely unacceptable (e.g., incorrect, irrelevant, or incoherent).
- 100 means the generated answer is perfect (e.g., correct, relevant, and fluent).

Question:
{QUESTION}

Ground Truth Answer:
{REFERENCE}

Generated Answer:
{PREDICTION}

Score (0–100):

C.5 Position Bias Analysis Prompts

C.5.1 Task Prompts

QA Task:

{DOCUMENTS}

Question: {QUESTION}
Answer:

TR Task:

{DOCUMENTS}

Which documents are needed to answer the following query: {QUESTION}
Answer:

C.5.2 Prompt Formats

DOC-Q:

{DOCUMENTS}

Question: {QUESTION}
Answer:

Q-DOC:

Question: {QUESTION}

{DOCUMENTS}

Answer:

Q-DOC-Q:

Question: {QUESTION}

{DOCUMENTS}

Question: {QUESTION}
Answer: