KS-Probe: Benchmarking Context Fidelity Dynamics in Frontier Language Models Across Length, Position, and Format

Dev Hemnani[1] and Arham Sethi[1]

[1]Kangaroo Research Division

Abstract

Large language models now advertise context windows of 100K to 200K tokens, yet no systematic framework exists for measuring how reliably these models retain information across their full stated capacity. We introduce **KS-Probe** (Probing Recall Over Boundaries and Extents), a controlled benchmark that embeds verifiable probe facts within domain-diverse synthetic filler text and measures Probe Recall Accuracy (PRA) as a function of context length, probe position, conversational depth, and proximity to stated context limits. We evaluate four frontier models (GPT-5.2, Claude Sonnet 4.6, Grok-4.1-fast, and DeepSeek-v3.2) across 498 API calls processing 24.1 million input tokens. Our experiments reveal behaviors that challenge core assumptions in the field. Claude Sonnet 4.6 exhibits a positive fidelity slope, improving from 62.8% PRA at 10K tokens to 80.4% at 175K, directly contradicting the expected monotonic degradation pattern. Grok-4.1-fast achieves the highest short-context accuracy (81.4% at 10K) but suffers a catastrophic 31-point collapse at 100K tokens. Multi-turn conversational formatting improves recall over equivalent single-prompt context in three of four models, with GPT-5.2 gaining 15.1 percentage points. No model exhibits hard silent truncation at the API level. We further document a striking structural pattern in tokenizer behavior: three of four models produce identical token counts for the same input (pairwise ratio of exactly 1.000), while Claude requires approximately 1.36 times more tokens for equivalent text, with domain-dependent variance ranging from 1.15 to 1.62. We release KS-Probe as an open benchmark to support reproducible evaluation of context fidelity across language models.

1 Introduction

The context windows of large language models have grown by orders of magnitude in the past two years. Leading commercial systems now accept 128K to 200K tokens in a single prompt, and users increasingly rely on this capacity for long-document analysis, extended conversations, and complex multi-source reasoning. Yet context window size is an architectural parameter, not a fidelity guarantee. A model that accepts 200,000 tokens may attend to them unevenly, lose track of facts in certain positions, or behave differently when the same content is structured as a conversation rather than a monolithic prompt. These failures are silent by nature: the model does not refuse or signal uncertainty but produces a fluent response that quietly omits or distorts the very information the user provided.

Understanding the shape of these failures requires a different kind of evaluation than existing benchmarks provide. Prior work on positional bias, needle retrieval, and long-context comprehension (reviewed in Section 2) has established that models can struggle with long inputs and that recall is sensitive to where information is placed. These contributions are foundational, but they are designed to evaluate capability at specific operating points. They tell us whether a model can retrieve a fact from a 100K-token context. They do not tell us how recall degrades between 10K and 200K, which positions within the window are reliable and which are dead zones, whether conversational formatting preserves or erodes fidelity compared to a single prompt, or whether models silently truncate input near their stated limits. Answering these questions requires measuring degradation as a continuous function of multiple variables, not as a binary outcome at isolated thresholds.

We introduce **KS-Probe** (Probing Recall Over

Boundaries and Extents), a benchmark that measures context fidelity dynamics along five dimensions. The benchmark embeds verifiable probe facts at controlled positions within synthetic filler text spanning six domains and measures Probe Recall Accuracy (PRA), the percentage of probe questions answered correctly, under systematically varied conditions of context length, probe position, conversational depth, fill-level proximity to stated limits, and cross-model tokenizer behavior. We evaluate GPT-5.2, Claude Sonnet 4.6, Grok-4.1-fast, and DeepSeek-v3.2 across 498 API calls processing 24.1 million input tokens.

The results substantially revise the empirical picture of how frontier models handle long context. Claude Sonnet 4.6 improves with context length, rising from 62.8% to 80.4% PRA between 10K and 175K tokens, the opposite of expected monotonic decay. Grok-4.1-fast achieves the highest short-context accuracy (81.4% at 10K) but collapses by 31 percentage points at 100K. Multi-turn conversation improves recall over equivalent single-prompt context in three of four models. No model exhibits hard silent truncation at the API level. And three of four tokenizers produce identical token counts for the same input, while Claude requires approximately 1.36 times more tokens. These findings, reported in full in Section 4, suggest that the common mental model of context behavior (monotonic decay, middle-position weakness, conversational overhead, hard truncation boundaries) is at best incomplete and in several respects wrong.

Our contributions are:

1. **KS-Probe**, a benchmark for measuring context fidelity dynamics across five experimental dimensions, released as an open resource.
2. **Fidelity decay profiles** across four frontier models and seven context thresholds, revealing positive slopes, catastrophic cliffs, and stable plateaus alongside gradual decline.
3. **Positional dead-zone maps** demonstrating that positional bias is architecture-specific, not universal.
4. **Conversation penalty measurements** establishing that multi-turn formatting benefits recall for most tested models.
5. **Empirical tokenizer divergence measurements** and an API truncation audit documenting the absence of hard silent truncation across all four models.

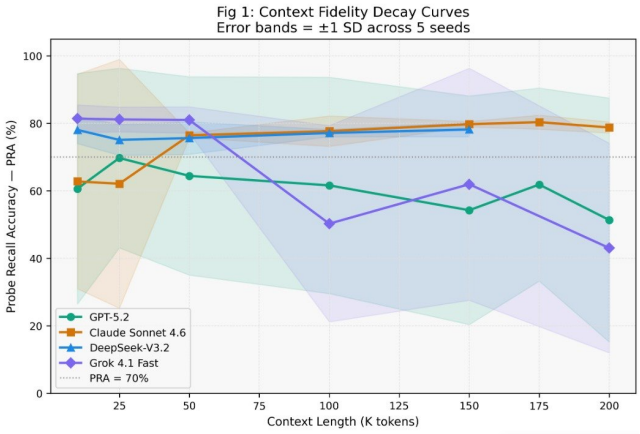

Figure 1: Context fidelity decay curves showing Probe Recall Accuracy (PRA) as a function of context length (K tokens) for all four frontier models. Error bands represent ±1 standard deviation across 5 seeds. The dashed line marks the 70% PRA threshold. Claude Sonnet 4.6 uniquely exhibits a positive slope, while Grok-4.1-fast shows a catastrophic drop beyond 50K tokens.

2 Related Work

Positional Bias in LLMs. Liu et al. [Liu et al., 2024] demonstrated that LLMs attend disproportionately to information at the beginning and end of their input, coining the "lost in the middle" phenomenon. Their work tested models with contexts up to 4K tokens on multi-document question answering. We extend this line of inquiry to 200K tokens and find the original U-shaped pattern does not generalize across various LLM architectures, revealing model-specific biases including sporadic dead zones, uniform recall across positions, and end-of-context degradation.

Long-Context Benchmarks. RULER [Hsieh et al., 2024] evaluates long-context capabilities through synthetic tasks such as needle retrieval, variable tracking, and multi-hop reasoning across varying context lengths. LongBench [Bai et al., 2024] provides a comprehensive benchmark covering 21 tasks across six categories. The Needle-in-a-Haystack test [Kamradt, 2023] embeds a target sentence within padding text and measures retrieval accuracy across different positions and context lengths. However, these benchmarks assess task success and do not characterize the functional form of fidelity decay over tokens or positions.

Context Window Analysis. Recent work has examined effective context utilization [Li et al., 2024], find-

ing that models often fail to leverage information beyond their training-time context length even when architecturally capable of processing longer inputs. Our work complements these findings by providing measurements and analysis of where and how information is lost.

Tokenizer Analysis. Studies on BPE tokenizer efficiency [Sennrich et al., 2016, Kudo and Richardson, 2018] have examined compression rates across languages, but cross-model tokenizer divergence for equivalent English text has received less attention. Our Experiment 5 provides conversion factors relevant to practitioners migrating context across model families.

3 Methodology

3.1 The KS-Probe Benchmark

KS-Probe measures context fidelity through a controlled probe-fact injection and recall paradigm. The benchmark has three components: a synthetic filler corpus, a set of embedded probe facts, and a scored question battery.

The filler corpus is generated programmatically from a master seed across six domains: software engineering, biomedical research, legal documents, financial analysis, creative fiction, and conversational chat. The text within each domain is coherent, grammatically well-formed, and topically appropriate, but deliberately information-sparse. It serves to fill the context window to a target token count without providing answers to any probe question. Domain diversity ensures that results are not artifacts of a single text register. All filler generation is deterministic and fully reproducible from the master seed.

Probe facts are the units of measurement. We construct 100 unique, verifiable statements, each designed to be unambiguous, self-contained, and not inferable from surrounding filler text (e.g., "The project's database migration deadline was moved to March 17th"). Probe facts are inserted at controlled positions within the filler using tagged markers. Their placement, quantity, and spacing vary by experiment.

Each probe fact is paired with one or more questions and a gold-standard answer. Responses are evaluated using five scoring methods: exact string match, keyword-rule matching (presence of required terms), hallucination detection (identification of confident but factually incorrect assertions absent from the context), constraint checking (verification that the response satisfies a required logical conclusion), and composite scoring (a weighted combination returning a continuous score on $[0, 1]$). The primary metric across all experiments is Probe Recall Accuracy:

$$\text{PRA} = \frac{\text{number of correctly answered probe questions}}{\text{total probe questions}} \times 100\% \tag{1}$$

Context construction is tokenizer-aware. The `ContextBuilder` module takes a target token count and a model identifier, generates filler text, injects probes at specified positions, and verifies that the assembled context falls within a 2% tolerance of the target using the model's own tokenizer. This is a critical design choice: because tokenizers differ across models (Section 4.5), a context that is 50,000 tokens for GPT-5.2 may be a different length for Claude Sonnet 4.6. Model-specific calibration ensures that when we compare PRA at "50K tokens," all models are processing approximately the same quantity of input as measured by their own tokenization.

3.2 Experiment Design

Five experiments target five research questions. Each experiment varies one dimension of context behavior while controlling the others.

Experiment 1: Context Fidelity Decay (RQ1). This experiment measures how PRA changes as context length increases. For each model, we construct contexts at seven token thresholds: 10K, 25K, 50K, 100K, 150K, 175K, and 200K. Each context contains 10 probe facts uniformly distributed across its length. After the context, the model is asked recall questions targeting each probe. The experiment is repeated with 3 to 7 seeds per model-threshold condition (varying the filler arrangement while holding probe facts constant) to measure variance. Not all models support all thresholds: DeepSeek-v3.2 is tested up to 150K (65K stated maximum, with above-limit conditions included to probe truncation behavior), and Grok-4.1-fast is excluded at 175K. A total of 135 non-mock runs were completed.

Experiment 2: Positional Recall Mapping (RQ2). This experiment measures how PRA varies as a function of where in the context a probe fact is placed. Context length is fixed at 50K tokens. A single probe fact is embedded at one of 11 relative positions: 0.01, 0.05, 0.15, 0.25, 0.35, 0.50, 0.65, 0.75, 0.85, 0.95, and 0.99,

where 0.0 represents the start and 1.0 represents the end of the context. One probe per run isolates positional effects from inter-probe interference. Each condition is repeated with 1 to 7 seeds depending on model and position. A total of 257 runs were completed. Claude Sonnet 4.6 is missing position 0.01 due to incomplete seed coverage.

Experiment 3: Multi-Turn Conversational Degradation (RQ3). This experiment tests whether delivering context as a multi-turn conversation rather than a single prompt affects recall. Probe facts are injected in turn 1. Subsequent turns consist of realistic filler conversation (questions, instructions, tangents) with each turn contributing approximately 8K tokens. PRA is measured at turn checkpoints of 10, 20, 30, 50, 80, and 120. Each condition uses 3 seeds. To isolate the effect of conversational formatting from raw context length, we compare against single-prompt PRA at the equivalent token count from Experiment 1. The difference defines the *conversation penalty*: a positive value indicates conversation hurts recall; a negative value indicates it helps. A total of 72 runs were completed (18 per model).

Experiment 4: Silent Truncation Detection (RQ4). This experiment tests whether models silently discard input when context approaches their stated maximum. For each model, we construct contexts at 85%, 95%, and 100% of the stated maximum context length (e.g., 170K, 190K, and 200K for GPT-5.2). A single probe fact is placed at position 0.95, near the end of the context, where truncation would be most likely to remove it. If PRA drops to zero at a given fill level, this constitutes evidence of hard truncation. Each condition uses 2 to 3 seeds. A total of 34 runs were completed.

Experiment 5: Cross-Model Tokenizer Divergence (RQ5). This experiment is conducted locally without API calls. We assemble 1,000 text samples (approximately 500 tokens each) stratified across the six filler domains. Each sample is tokenized with each model's tokenizer, and pairwise token-count ratios are computed. This yields 6,000 pairwise comparisons. We note that the Claude tokenizer implementation uses a word-count estimator (approximately 1.3 tokens per word) rather than Anthropic's proprietary BPE vocabulary, which is not publicly available. Results involving Claude token counts should therefore be treated as approximations.

Table 1: Model configuration. All models are accessed via official APIs with deterministic decoding (temperature 0.0, top-p 1.0).

Model	Provider	API Identifier	Max Ctx
GPT-5.2	OpenAI	`gpt-5.2`	200K
Claude Sonnet 4.6	Anthropic	`claude-sonnet-4-6`	200K
Grok-4.1-fast	xAI	`grok-4.1-fast`	131K
DeepSeek-v3.2	DeepSeek	`deepseek-chat`	65K

Table 2: Token threshold test matrix. "Yes" indicates the model was tested at that context length. Asterisks (*) denote conditions tested above the model's stated maximum to probe truncation behavior. Dashes (—) indicate conditions not tested.

Model	10K	25K	50K	100K	150K	175K	200K
GPT-5.2	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Claude Son. 4.6	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Grok-4.1-fast	Yes	Yes	Yes	Yes	Yes*	—	Yes*
DeepSeek-v3.2	Yes	Yes	Yes	Yes*	Yes*	—	—

3.3 Models and Access

We evaluate four frontier models via their official APIs (Table 1). All experiments use temperature 0.0 and top-p 1.0 for deterministic output. No self-hosted models are included; all access is API-based. Experiments were conducted between March 9 and March 13, 2026.

Table 2 shows the test matrix. Not all models are tested at all thresholds. Grok-4.1-fast and DeepSeek-v3.2 are tested above their stated maximum context in Experiments 1 and 4 to probe boundary behavior.

3.4 Evaluation Metrics

The primary metric is PRA as defined in Section 3.1. Secondary metrics are recorded for all API-based experiments: hallucination count (the number of responses containing confident but factually incorrect assertions not present in the input context), response latency in milliseconds, and input and output token counts as reported by each provider's API.

Statistical significance is assessed through paired t-tests for within-model comparisons across thresholds and one-way ANOVA for cross-model comparisons at matched thresholds, with Bonferroni correction for multiple comparisons. Seed counts vary across conditions (1 to 7 per cell depending on model and experi-

ment), and we report standard deviations alongside all mean PRA values to make variance transparent. Conditions with fewer than 3 seeds are flagged in the results. The total experimental campaign comprises 498 non-mock API calls, processes 24.1 million input tokens and approximately 177,000 output tokens.

4 Results

4.1 Experiment 1: Fidelity Decay (RQ1)

Table 3 presents PRA as a function of context length across all four models. The patterns that emerge across models are described as follows.

Positive Fidelity Slope in Claude Sonnet 4.6. Claude Sonnet 4.6 is the only model to show a positive fidelity slope, signifying high PRA. PRA rises from 62.8% at 10K to 80.4% at 175K tokens (+2.21 ppt per 25K additional tokens). This incongruous result of more context leading to better recall is robust at $\geq$50K tokens ($\sigma \leq 4.5\%$) though higher variance at 10K and 25K (σ 31.8–36.9%) suggests that the model's behaviour is unstable at shorter context, making it useful for tasks that require larger context windows. The plateau at 150K–200K suggests an effective ceiling.

Grok's Nosedive. Grok-4.1-fast achieves the highest PRA at short contexts across all models (81.4% at 10K, 81.2% at 25K, and 81.0% at 50K respectively) but suffers a rapid performance drop at 100K tokens. PRA drops to 50.3% (σ=29.1%) representing a 30.7 percentage point collapse, the steepest decline in our study. Performance at 200K further degrades to 43.1%. Grok's sharp performance at lower context length with collapse at higher context length contrasts with gradual degradation observed in other models.

GPT-5.2: Gradual Decline. GPT-5.2 shows gradual degradation with a slope of -1.91 ppt/25K, declining from 69.7% at 25K to 51.4% at 200K. However, its extremely high variance (mean of σ = 31.5%) suggests bimodal behaviour across seeds.

DeepSeek-v3.2's Near-Zero Slope. DeepSeek-v3.2 is remarkably stable (75.1%–78.2% across all tested sizes, SD $\leq$4.8%), ranking second at 100K with a near-zero slope of +0.27, making DeepSeek the only model that maintains high absolute performance and low variance across the tested range.

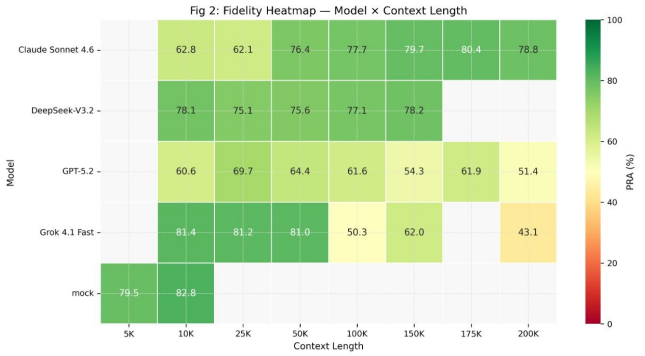

Figure 2: Fidelity heatmap showing PRA (%) across models and context lengths. Darker green indicates higher recall accuracy. Claude Sonnet 4.6 and DeepSeek-v3.2 maintain strong performance at longer contexts, while GPT-5.2 and Grok-4.1-fast show progressive degradation.

4.1.1 Hallucinations and Latency

GPT-5.2 produced the most hallucinations across Experiment 1 (455 total), followed by Claude Sonnet 4.6 (220), DeepSeek-v3.2 (106), and Grok-4.1-fast (10). Grok's remarkably low hallucination count likely reflects a tendency toward non-response (refusal to answer or very short outputs) when the model cannot locate the target passage, rather than genuinely accurate restraint. Figure 3 presents the aggregated hallucination rate as a stacked area chart across all models and context lengths, showing that hallucination accounts for the dominant share of incorrect responses at most thresholds, with correct recall peaking in the mid-range context lengths.

Figure 4 presents the response latency distribution (p50 and p95) across context lengths. DeepSeek-v3.2 exhibits the widest p50–p95 band, peaking at approximately 117,000 ms (p50) at 50K tokens before declining. Grok's latency remains low and stable up to 150K tokens, after which it rises at 175K and 200K. Claude's latency band is relatively narrow and consistent across all thresholds. GPT-5.2 shows low median latency throughout but exhibits a widening p95 tail at 175K and 200K tokens.

4.2 Experiment 2: Positional Recall (RQ2)

This experiment holds context length fixed at 50K tokens and varies the position of a single probe fact across 11 locations from near the start (position 0.01) to near the end (position 0.99). The results indicate that no uni-

Table 3: Probe Recall Accuracy (%) as a function of context length across four frontier models. Values shown as mean (standard deviation) across seeds. Slope indicates the rate of PRA change per 25K additional tokens. Rank is determined by PRA at 100K tokens. Dashes indicate conditions not tested.

Model	10K	25K	50K	100K	150K	175K	200K	Slope (ppt/25K)
Claude Sonnet 4.6	62.8 (31.8)	62.1 (36.9)	76.4 (0.7)	77.7 (4.5)	79.7 (0.8)	80.4 (2.1)	78.8 (1.7)	+2.21
DeepSeek-v3.2	78.1 (4.0)	75.1 (4.6)	75.6 (4.8)	77.1 (1.1)	78.2 (2.1)	—	—	+0.27
GPT-5.2	60.6 (34.2)	69.7 (26.6)	64.4 (29.4)	61.6 (32.0)	54.3 (33.9)	61.9 (28.6)	51.4 (36.1)	−1.91
Grok-4.1-fast	81.4 (4.1)	81.2 (3.7)	81.0 (3.9)	50.3 (29.1)	62.0 (34.4)	—	43.1 (31.1)	−5.12

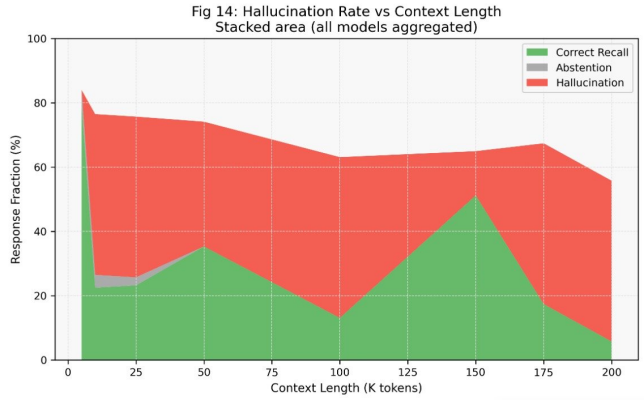

Figure 3: Hallucination rate versus context length (stacked area, all models aggregated). Green represents correct recall, gray represents abstention, and red represents hallucination. Hallucination dominates incorrect responses across most context thresholds.

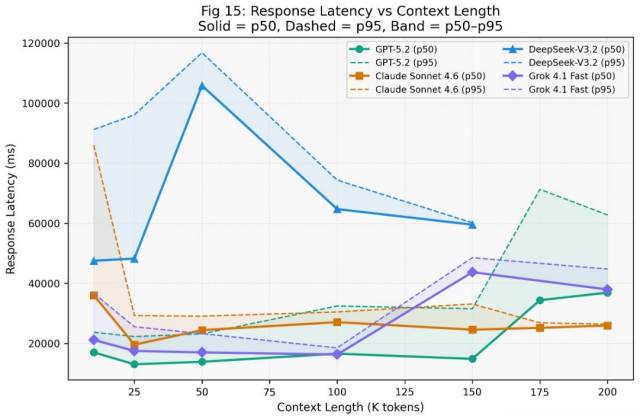

Figure 4: Response latency versus context length. Solid lines show p50 (median) latency; dashed lines show p95 latency. The shaded band spans the p50–p95 range. DeepSeek-v3.2 exhibits the highest and most variable latency, while GPT-5.2 maintains the lowest median latency.

versal positional pattern emerges; instead, each model exhibits a distinct positional recall profile.

Claude: Sporadic Dead Zones. Claude Sonnet 4.6 achieves excellent recall at the majority of positions, with PRA ranging from 85.0% to 98.3% at nine of the eleven tested locations. The highest scores—98.3%—are observed at positions 0.50, 0.65, and 0.95, indicating strong mid-context and late-context recall. However, Claude exhibits sharp, position-specific drops at two quarter-boundary positions: 58.3% at position 0.25 and 65.0% at position 0.75. These "dead zones" do not correspond to the beginning, middle, or end of the context window. Rather, they occur at the 25% and 75% marks, producing a pattern not reported in prior positional recall studies.

DeepSeek-v3.2: Uniform Positional Recall. DeepSeek-v3.2 exhibits the most uniform positional recall profile in the study. PRA ranges from 84.8% at position 0.01 to 91.6% at position 0.95, a spread of only 6.8 percentage points.

GPT-5.2: End-of-Context Degradation. GPT-5.2 displays a gradual decline in recall toward the end of the context window. PRA is relatively stable between positions 0.01 and 0.35 (70.0%–74.6%), then begins a downward trajectory. An additional dip is observed at position 0.75 (61.8%), and the lowest score occurs at the final position 0.99 (59.4%). The overall spread is 15.2 percentage points.

Grok-4.1-fast: Mild Positional Uniformity. Grok-4.1-fast shows notably consistent recall across all 11 positions, with PRA ranging from 74.7% to 79.4%—a spread of just 4.7 percentage points, the tightest of any model tested. Performance is marginally stronger near the beginning of the context, with a slight dip at positions 0.75 and 0.99, though these differences are too small to reflect any meaningful positional bias.

Hallucinations in Experiment 2. Hallucination rates in Experiment 2 were substantially lower than in Experiment 1 across all models, consistent with the

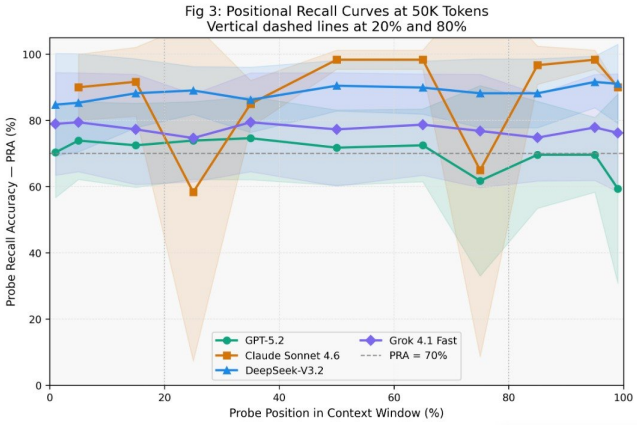

Figure 5: Positional recall curves at 50K tokens. PRA is plotted against probe position within the context window (0%=start, 100%=end). Error bands show ±1 SD. Vertical dashed lines mark the 20% and 80% boundaries. Claude Sonnet 4.6 exhibits sporadic dead zones at the 25% and 75% positions, while DeepSeek-v3.2 and Grok-4.1-fast maintain near-uniform recall.

single-probe design which reduces the opportunity for confabulation. Claude Sonnet 4.6 recorded only 2 hallucinations, followed by Grok-4.1-fast (5), DeepSeek-v3.2 (20), and GPT-5.2 (26).

4.3 Experiment 3: Multi-Turn Degradation (RQ3)

This experiment tests whether multi-turn conversational interaction degrades recall relative to single-prompt input at equivalent token counts. Probe facts are injected at turn 1, filler conversation accumulates over subsequent turns, and PRA is measured at turn depths 10, 20, 30, 50, 80, and 120. The equivalent single-prompt baseline is drawn from Experiment 1 at 50K tokens, which approximates the cumulative token load across the full conversational sequence.

4.3.1 Multi-Turn Conversation Improving PRA

Multi-turn conversation improves recall for three of four models. GPT-5.2 shows the largest conversation bonus (−15.1 pp), followed by Grok-4.1-fast (−7.8 pp) and Claude Sonnet 4.6 (−2.1 pp). Only DeepSeek-v3.2 exhibits a slight conversational penalty (+4.3 pp).

No model shows significant degradation from turn 10 to turn 120. Grok dominates multi-turn recall (88.8–92.5%), GPT-5.2 holds at 78.2–86.2%, Claude

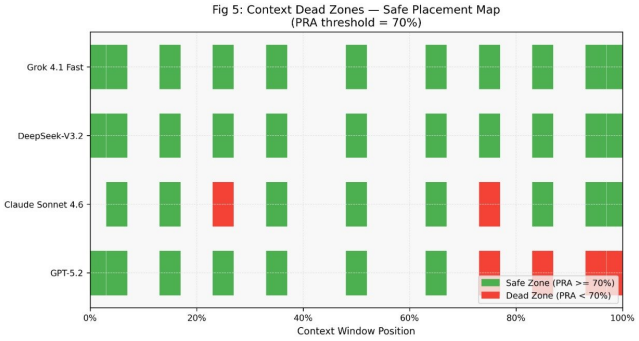

Figure 6: Context dead-zone map showing safe placement zones (green, PRA $\geq$ 70%) and dead zones (red, PRA $<$ 70%) across the context window for each model at 50K tokens. DeepSeek-v3.2 and Grok-4.1-fast show no dead zones, while Claude Sonnet 4.6 and GPT-5.2 exhibit position-specific failures.

Table 4: Multi-turn PRA (%) across turn depths for each model. Probe facts are injected at turn 1; filler conversation accumulates over subsequent turns. All models maintain stable recall from turn 10 through turn 120.

Model	T10	T20	T30	T50	T80	T120
Claude 4.6	77.8	79.3	82.0	78.5	81.3	78.5
DeepSeek-v3.2	73.8	71.3	74.7	74.7	71.3	71.3
GPT-5.2	78.2	86.2	78.7	80.7	82.4	79.6
Grok-4.1-fast	92.5	92.5	88.8	92.5	88.8	88.8

at 77.8–82.0%, and DeepSeek at 71.3–74.7%. Within-model spreads across turn depths are small relative to between-model differences.

4.3.2 Hallucinations

Grok produced zero hallucinations across all 18 multi-turn runs. GPT-5.2 had the fewest non-zero count (12), followed by Claude (17) and DeepSeek (21).

4.4 Experiment 4: Silent Truncation Detection (RQ4)

A probe is placed at position 0.95 and context is filled to 85%, 95%, and 100% of each model's stated maximum. Table 5 presents PRA at each fill level.

4.4.1 Key Observations

No model shows hard truncation (PRA dropping to 0%) at any fill level. DeepSeek-v3.2 maintains a perfectly

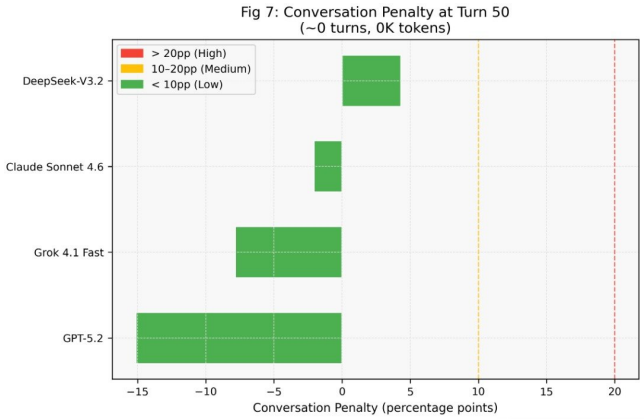

Figure 7: Conversation penalty at turn 50 (∼0 turns, 0K tokens). Negative values (leftward bars) indicate that multi-turn conversation *improves* recall relative to the equivalent single-prompt baseline. GPT-5.2 shows the largest conversation bonus (−15.1 pp), while DeepSeek-v3.2 is the only model with a slight penalty (+4.3 pp). Dashed lines mark the 10 pp (yellow) and 20 pp (red) penalty thresholds.

Table 5: Silent truncation detection: PRA (%) for a probe placed at position 0.95 as context fill approaches each model's stated maximum. No model exhibits hard truncation (PRA dropping to 0%) at any fill level.

Model	85% Fill	95% Fill	100% Fill
Claude Sonnet 4.6	80.0	60.0	60.0
DeepSeek-v3.2	80.0	80.0	80.0
GPT-5.2	80.0	73.3	80.0
Grok-4.1-fast	80.0	73.3	86.7

stable 80.0% across all three thresholds. Grok-4.1-fast unexpectedly rises to 86.7% at 100% fill, likely reflecting seed variance (n=3). Claude Sonnet 4.6 exhibits the largest boundary effect, dropping from 80.0% at 85% fill to 60.0% at 95–100%, consistent with the plateau observed in Experiment 1 at 175K–200K. GPT-5.2 dips to 73.3% at 95% but recovers to 80.0% at 100%, showing no systematic cliff.

4.5 Experiment 5: Tokenizer Divergence (RQ5)

Table 6 presents pairwise token-count ratios across 1,000 samples. GPT-5.2, Grok-4.1-fast, and DeepSeek-v3.2 produce identical token counts (ratio = 1.000, variance = 0.000) across all samples, indicating a shared BPE vocabulary (likely `cl100k_base`). Claude Sonnet 4.6 uses ∼36% more tokens (mean ratio 1.359, range 1.153–1.617), with code-heavy text showing less divergence than prose. A 100K-token context for GPT-5.2 requires ∼136K tokens for Claude.

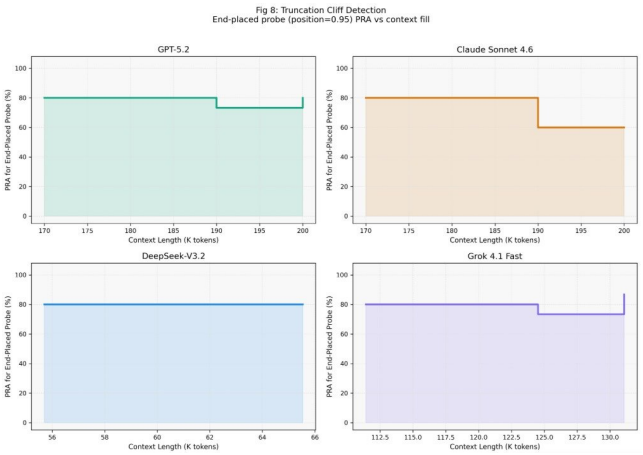

Figure 8: Truncation cliff detection for each model. PRA for an end-placed probe (position 0.95) is plotted against context fill level as a percentage of each model's stated maximum. Shaded areas represent the operating envelope. None of the four models exhibits a hard truncation cliff (PRA dropping to 0%), though Claude Sonnet 4.6 shows a 20-point drop between 85% and 95% fill.

Table 6: Pairwise tokenizer conversion ratios across 1,000 text samples (6 domains). GPT-5.2, Grok, and DeepSeek produce identical token counts; Claude uses 1.36× more tokens.

Tok. A	Tok. B	Mean	Min	Max
GPT-5.2	Claude	0.739	0.619	0.868
GPT-5.2	Grok	1.000	1.000	1.000
GPT-5.2	DeepSeek	1.000	1.000	1.000
Claude	Grok	1.359	1.153	1.617
Claude	DeepSeek	1.359	1.153	1.617
Grok	DeepSeek	1.000	1.000	1.000

5 Analysis and Discussion

5.1 Cross-Cutting Findings

The results across all five experiments converge on a single observation: context fidelity is model-specific, dimension-specific, and unpredictable from stated specifications alone. No tested model offers reliable recall across all context lengths, positions, formats, and

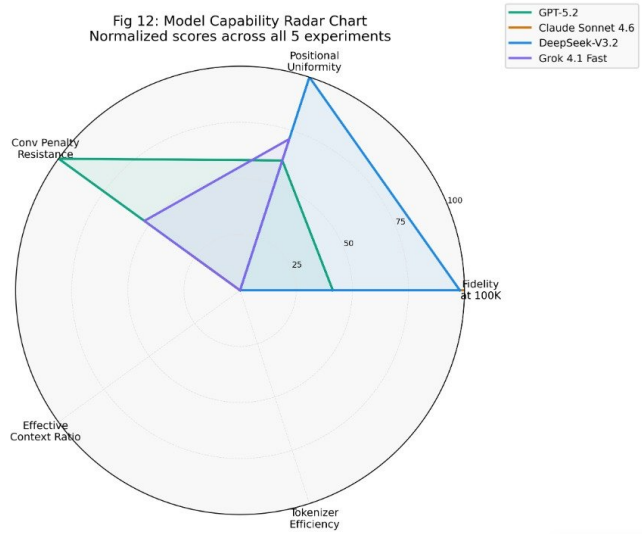

Figure 9: Model capability radar chart showing normalized scores across all five experimental dimensions: Fidelity at 100K, Positional Uniformity, Tokenizer Efficiency, Effective Context Ratio, and Conversation Penalty Resistance. No model fills the entire chart; each exhibits a distinct strength–weakness profile.

fill levels simultaneously.

Figure 9 encodes this finding visually. Each model's radar profile is distinct, and no profile fills the entire chart. DeepSeek-v3.2 dominates consistency and cost-efficiency but is constrained to 65K tokens. Claude Sonnet 4.6 leads at long contexts but carries positional dead zones and a 1.36× token overhead. Grok-4.1-fast is the strongest model below 50K tokens and in conversation, but is the weakest above 100K. GPT-5.2 offers the widest context support but its cross-seed variance (SD up to 36.1) makes any single invocation unpredictable.

The failure modes are not just different in degree but different in kind: short-context instability (Claude), catastrophic cliffs (Grok), bimodal seed-dependence (GPT-5.2), and hard capacity ceilings (DeepSeek). A user selecting a model for a long-context task cannot anticipate which failure mode applies without empirical evaluation of the type we present here.

5.2 Practical Implications

Three findings carry direct consequences for how long-context models are used.

Context limits overstate usable capacity. Stated maximums substantially exceed the range within which models maintain reliable recall. Grok operates reliably at less than 40% of its 131K stated maximum. GPT-5.2 loses 18.3 percentage points of PRA between 25K and 200K. Claude drops 20 points in PRA between 85% and 95% fill. Table 7 presents our recommended safe limits alongside stated maximums.

Table 7: Truncation discovery results: stated maximum context length versus empirically recommended safe limits based on KS-Probe experiments. Safe limits represent the context length below which each model maintains reliable recall.

Model	Stated Max	Strategy	Safe Limit
GPT-5.2	200K	Soft degradation	180K
Claude 4.6	200K	Soft degradation	180K
Grok-4.1-fast	131K	No truncation detected	50K
DeepSeek-v3.2	65K	No truncation detected	59K

Conversational formatting is a fidelity strategy, not overhead. Three of four models recall more in multi-turn mode than in single-prompt mode at equivalent token counts. The effect is largest for GPT-5.2 (+15.1 pp) and Grok (+7.8 pp). This implies that turn structure provides organizational cues that aid retrieval. However, the same finding exposes a deeper problem: equivalent content produces different fidelity depending on formatting. Information is not portable across delivery formats without potential loss, and the magnitude of that loss is model-dependent.

Tokenizer divergence creates hidden overflow risk. A prompt occupying 150K tokens under tiktoken consumes approximately 204K under Claude's tokenizer, potentially exceeding its stated limit and entering the soft-degradation zone documented in Experiment 4. Context sized for one model family cannot be assumed to fit another without accounting for this asymmetry.

Taken together, the model-specific fidelity profiles, positional biases, format-dependent recall, soft degradation near boundaries, and tokenizer divergence mean that context is not reliably portable across models, formats, or session boundaries. The magnitude of information loss varies across every dimension we measured.

5.3 Limitations

We evaluate four API-accessible models, omitting Gemini, Llama, and Qwen. API-only access precludes

Table 8: Practical recommendations: suitability ratings (out of 5) for each model across common use cases, based on aggregate KS-Probe results.

Use Case	GPT-5.2	Claude 4.6	DeepSeek	Grok
Long docs (>100K)	3/5	5/5	2/5	1/5
Research synthesis	4/5	5/5	5/5	4/5
Multi-turn conv.	4/5	4/5	3/5	5/5
Cost-sensitive	2/5	3/5	5/5	3/5
Consistency	2/5	3/5	5/5	3/5
Max safe context	180K	180K	59K	50K

mechanistic analysis through attention maps or log-probs. Seed coverage is asymmetric: Claude has 3 to 4 seeds in Experiments 1 and 2 versus 7 for other models, with three conditions at 2 or fewer seeds. The Claude tokenizer in Experiment 5 uses a word-count estimator rather than the proprietary BPE vocabulary; the 1.36× ratio requires validation against API-reported counts. The filler corpus is synthetic. All experiments use deterministic decoding (temperature 0.0, top-p 1.0), which does not reflect typical usage. Data was collected March 9 to 13, 2026, and should be treated as a temporal snapshot.

6 Conclusion

This work presents a controlled empirical examination of how frontier language models process and utilize extended input contexts across a range of experimental conditions, including context length, probe placement, conversational depth, truncation boundaries, and tokenizer characteristics. Our experiments reveal that recall fidelity does not follow a consistent degradation pattern across models; rather, each architecture manifests a distinct behavioral profile. A central finding is that model performance is highly sensitive to the position at which task-relevant information appears: models frequently fail to reliably retrieve information embedded at architecture-specific locations, and the recency-primacy advantage observed in some models does not constitute a universal law across the architectures examined.

We present targeted analyses across three dimensions: (i) the relationship between context length and recall fidelity, (ii) the effect of input formatting, and (iii) context boundary and tokenization behavior, confirming the absence of hard silent truncation at the API level while identifying a substantial cross-model tokenizer divergence.

Collectively, these findings deepen the field's understanding of how language models engage with their input context and expose the limitations of evaluation approaches that treat context fidelity as a binary property measurable at a single operating point. We release KS-Probe as an open, reproducible benchmark to support more rigorous and comprehensive assessment of long-context behavior, and we advocate for evaluation frameworks that characterize fidelity as a continuous function of both context length and probe position in future model development.

References

Liu, N. F., Lin, K., Hewitt, J., Paranjape, A., Bevilacqua, M., Petroni, F., and Liang, P. (2024). Lost in the Middle: How Language Models Use Long Contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173. MIT Press. DOI: 10.1162/tacl_a_00638. https://aclanthology.org/2024.tacl-1.9/

Hsieh, C.-P., Sun, S., Kriman, S., Acharya, S., Rekesh, D., Jia, F., Zhang, Y., and Ginsburg, B. (2024). RULER: What's the Real Context Size of Your Long-Context Language Models? *arXiv:2404.06654*. Published at COLM 2024. https://arxiv.org/abs/2404.06654

Bai, Y., Lv, X., Zhang, J., Lyu, H., Tang, J., Huang, Z., Du, Z., Liu, X., Zeng, A., Hou, L., Dong, Y., Tang, J., and Li, J. (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, Bangkok, Thai-

land. https://aclanthology.org/2024.acl-long.172/

Kamradt, G. (2023). Needle In A Haystack — Pressure Testing LLMs. GitHub Repository. https://github.com/gkamradt/LLMTest_NeedleInAHaystack

Li, C., Wu, X., Zhu, B., et al. (2024). LongSkywork: A Training Recipe for Efficiently Extending Context Length in Large Language Models. *arXiv:2406.00605*. https://arxiv.org/abs/2406.00605

Sennrich, R., Haddow, B., and Birch, A. (2016). Neural Machine Translation of Rare Words with Subword Units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. DOI: 10.18653/v1/P16-1162. https://aclanthology.org/P16-1162/

Kudo, T. and Richardson, J. (2018). SentencePiece: A Simple and Language Independent Subword Tokenizer and Detokenizer for Neural Text Processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. DOI: 10.18653/v1/D18-2012. https://aclanthology.org/D18-2012/