Breaking the Flow: How Topic Disruptions Degrade LLM Performance in Multi-Turn Conversation

Anonymous ACL submission

Abstract

Multi-turn dialogue is the dominant interaction paradigm for large language models (LLMs), yet current architectures treat conversational history as an undifferentiated context stream. In real interactions, users frequently shift topics, abandon threads, and later return to earlier discussions, patterns that may degrade model performance but remain largely unexplored. We introduce TOPICBREAK, a benchmark paired with a four-category taxonomy of conversational breakage, to systematically study how natural topic disruptions affect LLM responses across 8 model families. Our results show that topic disruption causes substantial degradation ($\Delta I = -0.64$), nearly double the effect of context length alone. The effect is amplified by dialogue depth and semantic distance, and varies across breakage modes, with silent abandonment producing the largest degradation. Reasoning-augmented models exhibit greater resilience, and topic interference correlates strongly with prompt sensitivity ($\rho = 0.72$). These findings establish conversational topic structure as a key factor in multi-turn reliability and suggest that future architectures should explicitly detect and manage topic boundaries in dialogue history.

1 Introduction

Consider a common scenario: *a user debugging a Python script with an LLM briefly asks, "what's a good restaurant for a date this weekend?", receives a recommendation, and then returns with "ok, back to the error on line 42."* Such non-linear interactions, including abrupt topic shifts and dropped conversational threads, are not edge cases but routine patterns in real-world human–AI dialogue (Zhao et al., 2024; Zheng et al., 2023a). Human interlocutors manage these transitions through **conversational grounding** (Clark and Brennan, 1991), using cues and repair markers to maintain mutual understanding. LLMs, by contrast, process the entire dialogue history as a continuous stream (Li et al., 2025) and lack mechanisms to isolate abandoned conversational threads.

Prior work has established that conversational history can interfere with LLM performance. Gupta et al. (2024) introduced a task-switch framework and proposed a reference-free sensitivity metric $\tau(\cdot)$ to quantify this vulnerability. Similarly, Laban et al. (2025) reported a nearly 40% drop in performance in multi-turn settings compared to single-turn prompts. A recent follow-up argues that this "Lost in Conversation" phenomenon may stem from an intent alignment gap between users and models (Liu et al., 2026). Complementary evidence comes from model architecture: Liu et al. (2024) showed that LLMs tend to underweight information in the middle of long contexts. This bias may become problematic when abandoned topics remain buried in the dialogue history. Research on prompt sensitivity (Sclar et al., 2023; Chatterjee et al., 2024) shows that even small changes in input formatting can alter model outputs. Conversation analysis has also shown that topic management follows recognizable patterns in human dialogue (Sacks et al., 1974; Schegloff, 2007).

Despite this evidence, three gaps remain. First, existing benchmarks study artificial task switches instead of the natural topic changes that occur in real conversations (Handa et al., 2025). Second, prior studies cannot tell whether performance drops are caused by topic changes or simply by longer conversation history. Third, prior work has not examined whether different types of topic changes, such as returning to an earlier topic or abandoning a topic completely, affect models in different ways. Recent advances in topic shift detection now make this analysis possible (Lin et al., 2023; Wang et al., 2025).

In this paper, we study how abrupt topic changes in real conversations affect LLM behavior. ***Our experiments show that topic disruptions leave a measurable residual effect from earlier conversa-***

tion. Performance on a new topic becomes worse as the abandoned discussion grows longer. We also find that different types of topic changes, such as active pivots and silent abandonment, affect models in different ways, with large differences across model families. Our contributions are threefold:

1. We provide the first systematic study of topic disruptions in multi-turn LLM conversations using large-scale real dialogue data.

2. We introduce TOPICBREAK, a benchmark that enables controlled evaluation of topic changes while separating them from context-length effects.

3. We show that different types of topic changes produce different levels of interference, revealing new patterns of sensitivity in multi-turn LLM interaction.

2 Related Work

2.1 Multi-Turn Degradation and Task-Switching

A growing body of work shows that LLM performance declines as conversations extend. Gupta et al. (2024) introduced a task-switch framework demonstrating how earlier tasks can interfere with later instructions within the same dialogue. Laban et al. (2025) further showed that models often "get lost" in multi-turn interactions, with degradation driven more by rising unreliability than by loss of underlying capability. Recent surveys (Li et al., 2025), evaluation frameworks such as BotChat (Duan et al., 2024), and long-term memory benchmarks like LoCoMo (Maharana et al., 2024) further underscore the difficulty of maintaining coherence across extended interactions, Liu et al. (2026) attributes these failures to an alignment gap between evolving user intent and the model's interpretation. However, these studies focus on task switches or on performance within a single task. They do not study how topic shifts in conversation affect model behavior.

2.2 Context Utilization and Prompt Sensitivity

Several lines of evidence help explain why topic shifts may interfere with model responses. The "Lost in the Middle" effect (Liu et al., 2024) shows that LLMs pay less attention to information located in the middle of long contexts. Shi et al. (2023) demonstrated more directly that irrelevant context acts as a powerful distractor, dramatically reducing reasoning accuracy even when the added information is trivially unrelated to the task. LLMs are also highly sensitive to small changes in input formatting (Sclar et al., 2023), a property measured by metrics such as POSIX (Chatterjee et al., 2024). In addition, models often rely too strongly on patterns from earlier turns that no longer match the user's current intent (Herlihy et al., 2024). However, these studies examine perturbations within a single prompt or turn. In contrast, the interference we study is *cross-turn*, cumulative, and arises from the topical structure of the dialogue history itself.

2.3 Dialogue Topic Dynamics and Detection

Topic transitions have long been studied in conversation analysis, where Sacks et al. (1974) and Schegloff (2007) describe how speakers manage shifts between topics during dialogue (Clark and Brennan, 1991). In NLP, recent work has developed methods for detecting topic shifts. Lin et al. (2023) proposed multi-granularity prompting for boundary detection, and Wang et al. (2025) introduced a dual-process simulation to distinguish gradual drifts from abrupt pivots. Topic modeling tools such as BERTopic (Grootendorst, 2022) further enable the analysis of topic structure in dialogue.

These advances make it possible to study topic changes in real conversations. We build on these methods to construct TOPICBREAK, a benchmark based on topic disruptions extracted from large-scale conversation logs.

3 Methodology

3.1 Data Source and Topic Breakage Identification

We derive our experimental data from WildChat (Zhao et al., 2024) and LMSYS-Chat-1M (Zheng et al., 2023a), two large-scale datasets of human–LLM conversations. We filter for English-language, multi-turn dialogues containing at least six user turns to ensure sufficient conversational depth for analysis.

To identify potential topic breakage points, we employ a hybrid detection pipeline combining local semantic signals with global topic structure. First, we compute the cosine similarity between sentence embeddings (via `all-mpnet-base-v2`) of adjacent user turns U_t and U_{t+1} to capture local semantic discontinuity. Second, we apply BERTopic (Groo-

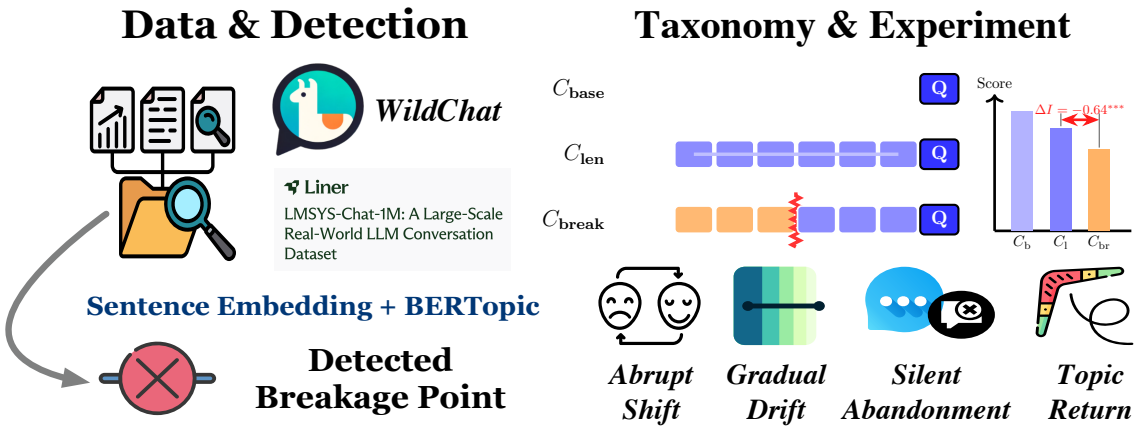

Figure 1: The overview of this paper.

tendorst, 2022) to the full corpus to obtain latent topic clusters, enabling detection of shifts that span thematic boundaries rather than surface-level lexical changes alone. A turn transition is flagged as a breakage candidate when the pairwise similarity falls below a threshold $\delta = 0.15$ *and* the two turns belong to distinct BERTopic clusters. To validate this automated pipeline, two authors independently annotated 200 flagged transitions, yielding a precision of 0.91 and recall of 0.84 for identifying genuine intent pivots. We then classify each disruption using the taxonomy defined in Section 3.2.

3.2 Taxonomy of Topic Breakage

We introduce a four-category taxonomy of topic breakage in real conversations. Each category is defined by the magnitude of semantic shift and the interactional pattern surrounding the transition:

Abrupt Shift: A sudden leap to a semantically unrelated domain (e.g., *from debugging Python to asking for a pasta recipe*), with no transitional cue or closing of the prior thread.

Gradual Drift: A progressive migration where the topic evolves through intermediate turns until the original intent is no longer active (e.g., *discussing a film, then its director, then the director's political views*).

Silent Abandonment: The user stops engaging with an ongoing thread and ignores model questions or drops the current line of inquiry, then initiates a new topic without explicit closure.

Topic Return: A non-linear pattern where the user pivots from Topic A to Topic B, then resumes Topic A in a later turn.

For each type, we extract representative seed examples from the real-world logs. In our annotated sample, *Abrupt Shifts* account for approximately 41% of detected breakages, followed by *Silent Abandonment* (28%), *Topic Return* (19%), and *Gradual Drift* (12%). Two annotators independently classified 300 instances; inter-annotator agreement reached Cohen's $\kappa = 0.82$, indicating high reliability. This taxonomy directly addresses our goal of examining whether different breakage modes produce qualitatively distinct interference patterns.

3.3 Controlled Evaluation Design

The core of our methodology is a controlled setup designed to disentangle the interference of topic breakage from the confound of context length. For every target query Q (*the "post-break" question*), we construct three experimental conditions:

- **Clean Baseline** (C_{base}): Q is presented in a fresh session with no prior history, measuring the model's peak performance on the query.
- **Same-Length Control** (C_{len}): Q is presented after a coherent, single-topic history of N turns, isolating the effect of increased context length alone.
- **Topic-Break Condition** (C_{break}): Q is presented after a history of the same N turns that contains one of the four breakage types from Section 3.2, isolating the specific interference cost of the disrupted flow.

Comparing C_{base} against C_{break} quantifies the overall degradation caused by topic disruption (*quantifies the overall effect of topic disruption*).

The contrast between C_{len} and C_{break}, which share the same token budget but differ in topical coherence, isolates the effect of topic disruption from that of context length. This directly addresses the second gap identified in our introduction. To investigate what variables modulate this interference, we manipulate three factors: (i) *Dialogue Depth* ($N \in \{2, 5, 10, 20\}$ turns), (ii) *Semantic Distance* between the pre- and post-break topics (low vs. high, measured by embedding similarity), and (iii) *Breakage Mode* (the four types from Section 3.2).

We evaluate 8 model families selected to span three dimensions of variation: open- vs. closed-weight (***Llama 3 8B/70B*** (Grattafiori et al., 2024), ***DeepSeek-V3*** (Guo et al., 2025) vs. ***GPT-4o*** (Hurst et al., 2024), ***Claude 3.5 Sonnet*** (Anthropic, 2024)), model scale (8B to frontier-class), and the presence of explicit reasoning capabilities (***Llama 4 Maverick*** (Adcock et al., 2026), ***DeepSeek-R1*** (Guo et al., 2025)). This selection allows us to examine whether architectural or training differences correlate with resilience to topic disruption.

3.4 Evaluation Metrics

Following the LLM-as-Judge paradigm (Zheng et al., 2023b), we use G-Eval (Liu et al., 2023) with GPT-4o as the evaluator. Each response is scored on a 1–5 scale across three dimensions:

- **Relevance:** Does the response address the new intent without referring to the abandoned topic?
- **Accuracy:** Does residual context introduce hallucinations or reasoning errors?
- **Completeness:** Is the response as thorough as its C_{base} counterpart?

To measure the effect of topic disruption while controlling for context length, we define the *Interference Delta* as $\Delta I = S(C_{\text{len}}) - S(C_{\text{break}})$, where $S(\cdot)$ denotes the average score across the three dimensions. Because C_{len} and C_{break} share the same token budget, ΔI isolates the interference caused by the topic break rather than longer conversation history.

To validate the automated evaluation, we conducted human verification on a stratified subset of 500 responses, obtaining a Spearman correlation of $\rho = 0.78$ between human judgments and G-Eval scores.

4 Results

4.1 Overall Impact of Topic Disruption

Our results show that topic disruption significantly degrades model performance beyond what can be explained by context length alone. Table 1 summarizes the mean scores across experimental conditions. The Interference Delta ($\Delta I = -0.64$) represents the net effect of topic breakage: while simple context expansion (C_{len}) causes moderate decay, a topic break nearly doubles the performance penalty ($p < 0.01$).

Table 1: Mean scores (1–5) and Interference Delta (ΔI) by evaluation dimension. Red cells highlight the most severe degradation.

Dimension	C_{base}	C_{len}	C_{break}	ΔI
Relevance	4.71	4.25	3.43	**-0.82**
Accuracy	4.55	4.12	3.55	-0.57
Completeness	4.60	4.17	3.64	-0.53
Overall	**4.62**	**4.18**	**3.54**	**-0.64**

The severe hit to Relevance ($\Delta I = -0.82$) suggests a failure in attentional compartmentalization: models frequently fail to purge abandoned constraints, producing responses that are technically coherent but misaligned with the new intent.

4.2 Modulating Variables

Residual Context Effect. As shown in Figure 2, the Interference Delta correlates positively with dialogue depth, increasing from $\Delta I = -0.21$ at 2 turns to $\Delta I = -0.89$ at 20 turns. This reveals a residual context effect: ***interference grows as the abandoned conversation becomes deeper.***

Semantic Distance. Counter-intuitively, high semantic distance (orthogonal topics) produced higher interference than low distance. We attribute this to a ***redirect cost***. Partially related topics allow models to reuse portions of the latent conversational representation, whereas distant shifts require a complete attentional reset. In this case, the model's reliance on stale priors acts as an anchor.

Breakage Mode. Different modes yield distinct interference signatures, validating the empirical utility of our taxonomy. Table 2 presents the full breakdown. Silent Abandonment ($\Delta I = -0.81$) and Abrupt Shift ($\Delta I = -0.74$) cause the most disruption, followed by Topic Return ($\Delta I = -0.52$).

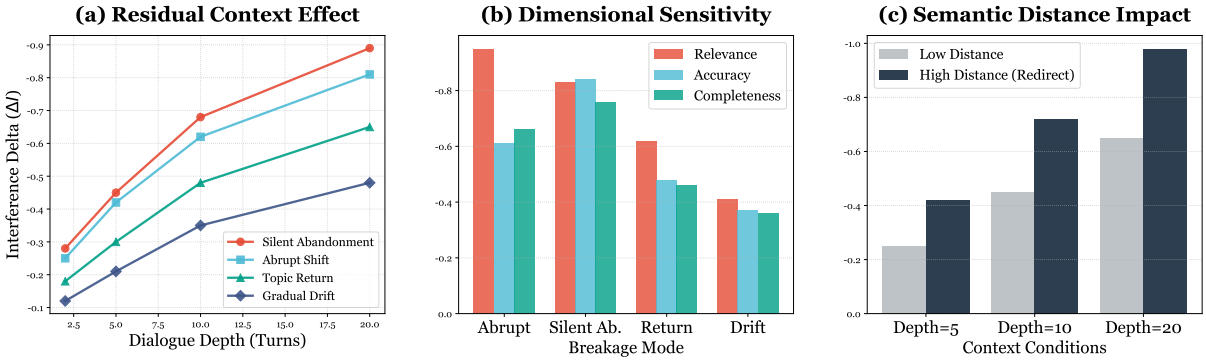

Figure 2: **Comprehensive analysis of conversational interference.** ***(a) The Residual Context Effect:*** Interference Delta (ΔI) as a function of dialogue depth, stratified by breakage mode. ***(b) Dimensional Sensitivity:*** Breakdown of ΔI across evaluation metrics. ***(c) Redirect Cost:*** Interaction between semantic distance and context depth.

Gradual Drift ($\Delta I = -0.38$) is the least disruptive, suggesting that transitional turns act as a semantic buffer that facilitates smoother attention reallocation. ***Notably, the interference profile differs across modes:*** Abrupt Shift disproportionately harms Relevance, while Silent Abandonment degrades Accuracy most severely. This pattern is consistent with models making ungrounded assumptions when a thread is left unresolved.

4.3 Cross-Model Variation

We observed substantial variation across model families. Table 3 presents the Interference Delta for each model, broken down by breakage mode.

Reasoning models (DeepSeek-R1, Llama 4 Maverick) were the most resilient, particularly under Silent Abandonment, where explicit thinking traces allow re-verification of current constraints before generation. Smaller standard models (Llama 3 8B) and those with high prompt-sensitivity indices were the most vulnerable. Figure 3 visualizes this variation as a heatmap, making the interaction between model capability and breakage mode immediately apparent.

We found a Spearman correlation of $\rho = 0.72$ between a model's POSIX score (Chatterjee et al., 2024) and its Interference Delta, suggesting that the mechanisms driving sensitivity to formatting perturbations also underlie vulnerability to topic disruption. Notably, while larger models (Llama 3 70B vs. 8B) show better absolute scores, scaling alone does not close the gap with reasoning-augmented architectures.

4.4 Interaction Effects: When Depth Meets Breakage Mode

The analyses above examine depth and breakage mode separately. Their interaction reveals an important pattern: breakage modes do not scale equally with dialogue depth. Table 4 reports ΔI for each mode at shallow (2 turns) and deep (20 turns) contexts.

Two patterns emerge. First, interference increases with depth across all modes, but the amplification differs. Silent Abandonment grows from $\Delta I = -0.29$ to -1.18 ($4.1\times$), whereas Gradual Drift increases more modestly from -0.12 to -0.41 ($3.4\times$). As conversations deepen, the gap between the most and least disruptive modes widens substantially. Second, the interaction suggests a practical threshold. When conversations remain short (under 5 turns), breakage mode has limited impact (ΔI ranges from -0.12 to -0.29). Beyond roughly 10 turns, however, the mode of topic transition becomes a dominant factor. In short exchanges, even abrupt shifts are tolerable; in longer sessions, the *manner* of transition becomes critical for maintaining response quality.

4.5 Semantic Distance and Breakage Mode Interaction

We further analyze how semantic distance interacts with breakage mode. Table 5 reports ΔI under low- and high-distance conditions for each mode.

The distance penalty varies substantially across modes. Silent Abandonment and Abrupt Shift show the largest increase in interference ($\Delta \approx -0.41$), nearly doubling the gap between low- and high-distance conditions. In contrast, Gradual Drift is relatively robust to semantic distance ($\Delta = -0.13$),

Table 2: Granular analysis of Interference Delta (ΔI) across breakage modes and evaluation dimensions. Values show mean ΔI (std). Red = most severe per row; green = least severe. Bottom rows report statistical significance and effect size.

Dimension	Abrupt Shift	Silent Abandonment	Topic Return	Gradual Drift	Peak Sensitivity
Relevance	**-0.95** (0.12)	-0.83 (0.09)	-0.62 (0.15)	-0.41 (0.08)	-1.24 (GPT-3.5)
Accuracy	-0.61 (0.14)	**-0.84** (0.11)	-0.48 (0.10)	-0.37 (0.07)	-1.12 (Llama-3-8B)
Completeness	-0.66 (0.10)	**-0.76** (0.13)	-0.46 (0.09)	-0.36 (0.06)	-0.98 (GPT-3.5)
Overall	-0.74 (0.11)	**-0.81** (0.12)	-0.52 (0.11)	-0.38 (0.07)	-1.15 (GPT-3.5)
P-value	< 0.001	< 0.001	< 0.01	< 0.05	—
Effect Size (d)	0.82 (Large)	0.88 (Large)	0.54 (Med.)	0.31 (Small)	—

Table 3: Interference Delta (ΔI) by model and breakage mode, sorted by overall resilience. Cell shading intensity reflects degradation severity: green = most resilient, red = most vulnerable.

Model	Abrupt Shift	Silent Ab.	Topic Return	Gradual Drift	Overall ΔI
DeepSeek-R1	-0.34	**-0.27**	-0.24	-0.18	**-0.29**
Llama 4 Maverick	-0.41	-0.33	-0.30	-0.22	-0.35
GPT-4o	-0.52	-0.58	-0.39	-0.28	-0.46
DeepSeek-V3	-0.55	-0.54	-0.42	-0.31	-0.48
Claude 3.5 Sonnet	-0.61	-0.63	-0.45	-0.33	-0.53
Llama 3 70B	-0.78	-0.82	-0.55	-0.40	-0.66
Llama 3 8B	-0.93	-1.01	-0.68	-0.49	-0.81
GPT-3.5-Turbo	-1.05	**-1.12**	-0.74	-0.52	**-0.94**

Table 4: Interference Delta (ΔI) at shallow vs. deep contexts, by breakage mode.

Breakage Mode	Depth=2	Depth=20	Amp.
Silent Abandonment	-0.29	**-1.18**	4.1×
Abrupt Shift	-0.26	-1.02	3.9×
Topic Return	-0.18	-0.71	3.9×
Gradual Drift	-0.12	-0.41	3.4×

suggesting that transitional turns allow the model to adapt its representation incrementally even when the new topic is distant. Topic Return falls between these extremes: when the intervening topic remains semantically close, recovery is easier, whereas distant interruptions lead to significant detail loss, consistent with the Return Gap errors described in Section 4.7.

Table 5: Interference Delta (ΔI) by semantic distance and breakage mode.

Breakage Mode	Low Dist.	High Dist.	Δ
Silent Abandonment	-0.62	**-1.03**	-0.41
Abrupt Shift	-0.54	-0.96	-0.42
Topic Return	-0.41	-0.64	-0.23
Gradual Drift	-0.31	-0.44	-0.13

4.6 Dimension-Level Degradation Across Depth

We analyze how the three evaluation dimensions degrade as dialogue depth increases. Figure 4 plots dimension-specific ΔI across depths under the C_{break} condition.

The three dimensions show distinct degradation patterns. Relevance declines immediately, from $\Delta I_{\text{Rel}} = -0.38$ at 2 turns to -1.12 at 20 turns, indicating that models tend to reference stale context whenever residual history is present. Accuracy remains relatively stable at shallow depths ($\Delta I_{\text{Acc}} = -0.14$ at 2 turns) but deteriorates sharply beyond 10 turns (-0.81 at 20 turns), suggesting that hallucinations and reasoning errors emerge only when irrelevant context accumulates. Completeness follows a more gradual trajectory, decreasing from -0.11 to -0.72. This pattern has practical implications: in short conversations the main risk is off-topic responses (Relevance), whereas in longer sessions factual reliability (Accuracy) becomes the dominant concern.

4.7 Error Patterns

We observe three recurring failure modes aligned with our taxonomy. ***Contextual Leakage***, common in Abrupt Shifts, occurs when models carry over safety or stylistic constraints from an abandoned topic into an unrelated response. ***State Paralysis***,

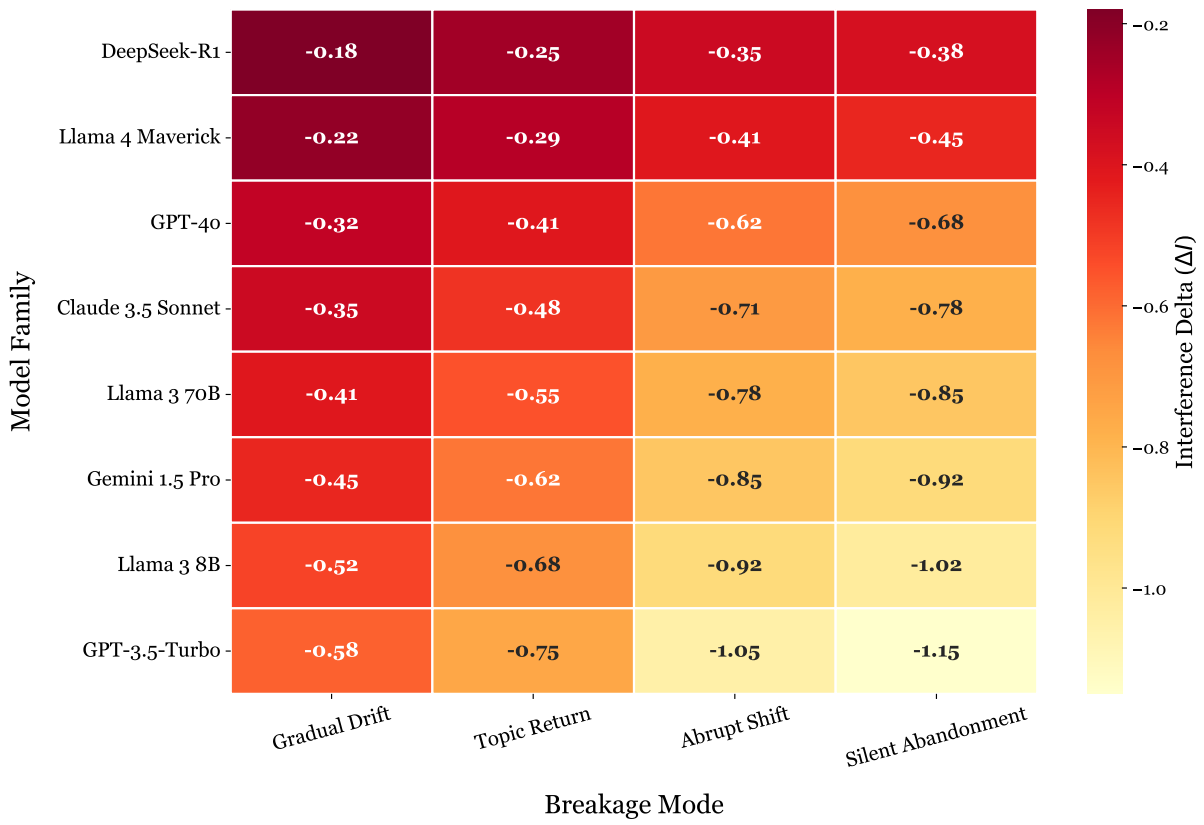

Figure 3: Heatmap of Interference Delta (ΔI) across 8 model families and 4 breakage modes. Darker cells indicate more severe degradation. Reasoning models (top rows) show uniformly lighter shading, while the gap widens most under Silent Abandonment.

also associated with Abrupt Shifts, arises when models enter clarification loops about the previous topic instead of addressing the new intent. The ***Return Gap***, specific to Topic Return, reflects models' difficulty retrieving details from a thread once it has been interrupted by an intermediate conversation.

Taken together, these findings show that topic breakage degrades LLM performance through mechanisms beyond simple context expansion. The effect is strongest when conversations are deep, topics are semantically distant, or threads are abandoned without closure, highlighting a limitation in how current architectures manage topical coherence across turns.

5 Discussion

5.1 Theoretical Mechanisms: The Inertia of Attention

Our findings on the *Residual Context Effect* and the strong sensitivity to *Silent Abandonment* ($\Delta I = -0.81$) highlight a limitation of the Transformer's attention mechanism. In human discourse, topic boundaries are managed through **conversational grounding** (Clark and Brennan, 1991), where speakers use explicit signals to close one thread before moving to another. LLMs lack such cues. As a result, the self-attention mechanism may exhibit **attentional inertia**, continuing to assign weight to representations from abandoned threads rather than discounting them when the topic changes. Testing whether this effect appears directly in attention distributions is an important direction for future mechanistic analysis.

The stronger interference observed for high semantic distance further supports this interpretation and is consistent with **distractor theory** (Shi et al., 2023). When a new query is unrelated to the previous dialogue, the model must redirect its latent representation entirely. In this situation, reliance on historical conversational priors (Herlihy et al., 2024) can anchor the model to stale context, producing the State Paralysis and Contextual Leakage patterns described in Section 4.4.

5.2 Refining Multi-Turn Reliability Research

Our results refine the current understanding of multi-turn degradation. While Gupta et al. (2024)

Table 6: Representative failure cases from the TOPICBREAK benchmark. Each pattern is mapped to a breakage mode from our taxonomy, illustrated with a concrete dialogue excerpt, and annotated with the most affected model.

Failure Pattern	Breakage Mode	Dialogue Excerpt	Most Affected
Contextual Leakage	*Abrupt Shift*	User discusses medication side effects for 8 turns, then asks: *"Plan a weekend trip to Kyoto."* Model responds with travel advice but appends: *"Please consult a healthcare provider before traveling with medication."*	GPT-4o($\Delta I = -1.24$)
State Paralysis	*Abrupt Shift*	User debugs a race condition for 12 turns, then asks: *"What's a good birthday gift for a 5-year-old?"* Model replies: *"Before we move on, shall I finish resolving the asyncio issue?"* and loops for 3 turns.	Llama 3 8B($\Delta I = -1.01$)
Return Gap	*Topic Return*	User defines `config = {lr: 3e-4}` in turn 3, pivots to meal planning for 6 turns, then resumes: *"Use the config I defined earlier."* Model hallucinates `lr: 1e-3` instead of retrieving the original value.	Llama 3 70B($\Delta I = -0.82$)

reported an average 15% performance drop in structured task-switching settings, we show that this effect is not uniform and depends strongly on the **thematic structure** of the conversation. For example, Gradual Drift ($\Delta I = -0.38$) causes substantially less interference than Silent Abandonment ($\Delta I = -0.81$), suggesting that transitional turns act as a semantic buffer that helps models reallocate attention. This extends the findings of Laban et al. (2025) by showing that multi-turn unreliability is not simply a function of sequence length but is driven by how topic transitions disrupt the model's internal state.

We also observe a strong correlation ($\rho = 0.72$) between topic interference and the POSIX index (Chatterjee et al., 2024). This result suggests that structural topic disruption and surface-level formatting perturbations may share a common cause: current architectures struggle to distinguish the user's current instruction from irrelevant context in the dialogue history.

5.3 Practical Implications

Our findings suggest practical guidelines for two audiences. For **users**, the Residual Context Effect implies that when a preceding dialogue is deep (>10 turns) and semantically unrelated to a new intent, starting a fresh session is often more effective than continuing the same conversation. When continuity is necessary, providing an explicit transitional cue (e.g., "Let's switch topics to...") can partially reduce interference. The relatively small ΔI observed for Gradual Drift indicates that even minimal bridging between topics can help.

For **model developers**, the strong resilience of reasoning models (DeepSeek-R1, $\Delta I = -0.29$) points to a potential architectural direction. Models may benefit from an explicit intent verification step that checks the current query against the dialogue state before generation, combined with topic boundary detection that discounts attention to abandoned segments.

Returning to our opening scenario, a user who shifts from debugging Python to asking for restaurant recommendations and later resumes the coding task may experience measurable degradation under current architectures. Better session management or topic-aware inference could mitigate this issue.

6 Conclusion

We study how abrupt topic changes affect large language models using the TOPICBREAK benchmark and a four-category taxonomy of conversational breakage. Our results show that topic disruption significantly degrades performance ($\Delta I = -0.64$), nearly doubling the effect of context length alone. This effect becomes stronger with deeper conversations, larger semantic shifts, and more disruptive breakage modes, with Silent Abandonment causing the largest drop. These findings suggest that multi-turn performance depends not only on context length but also on how topics change over time. Since this interference persists across model scales, scaling alone is not sufficient. Instead, the stronger performance of reasoning-augmented models points to a more promising direction: explicitly checking the user's current intent before generating responses.

Limitations

Our experiments focus on English-language, general-domain dialogues. Topic management may vary across languages and application domains, particularly in high-stakes settings such as medical or legal consultation. Our four-category taxonomy captures common patterns observed in real-world data but is not exhaustive. More fine-grained distinctions may further refine the analysis of topic disruption. We propose attentional inertia as a possible explanation for the observed effects, but do not directly validate this hypothesis through model internals. Future work could investigate this mechanism using interpretability methods.

Ethics Statement

This study uses publicly available conversation datasets (WildChat and LMSYS-Chat-1M) that were collected with user consent as described in their respective data release documentation (Zhao et al., 2024; Zheng et al., 2023a). We do not collect any new human data from users. The human evaluation component involved three graduate student annotators who participated voluntarily and were compensated at a rate above the local minimum wage. No personally identifiable information was used or generated at any stage of this research. Our work reveals that LLMs are vulnerable to performance degradation under naturalistic topic shifts. While we frame this as a robustness concern, we acknowledge that the detection pipeline and taxonomy we introduce could in principle be repurposed to exploit model weaknesses adversarially. We believe that transparently documenting these vulnerabilities serves the broader goal of building more robust conversational AI systems, and that the benefits of public disclosure outweigh the risks.

References

Aaron Adcock, Aayushi Srivastava, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pande, Abhinav Pandey, Abhinav Sharma, Abhishek Kadian, Abhishek Kumawat, Adam Kelsey, et al. 2026. The llama 4 herd: Architecture, training, evaluation, and deployment notes. *arXiv preprint arXiv:2601.11659*.

Anthropic. 2024. The claude 3 model family: Opus, sonnet, haiku.

Anwoy Chatterjee, HSVNS Kowndinya Renduchintala, Sumit Bhatia, and Tanmoy Chakraborty. 2024. Posix: A prompt sensitivity index for large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14550–14565.

Herbert H Clark and Susan E Brennan. 1991. Grounding in communication.

Haodong Duan, Jueqi Wei, Chonghua Wang, Hongwei Liu, Yixiao Fang, Songyang Zhang, Dahua Lin, and Kai Chen. 2024. Botchat: Evaluating llms' capabilities of having multi-turn dialogues. In *Findings of the association for computational linguistics: NAACL 2024*, pages 3184–3200.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Akash Gupta, Ivaxi Sheth, Vyas Raina, Mark Gales, and Mario Fritz. 2024. Llm task interference: An initial study on the impact of task-switch in conversational history. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 14633–14652.

Kunal Handa, Alex Tamkin, Miles McCain, Saffron Huang, Esin Durmus, Sarah Heck, Jared Mueller, Jerry Hong, Stuart Ritchie, Tim Belonax, et al. 2025. Which economic tasks are performed with ai? evidence from millions of claude conversations. *arXiv preprint arXiv:2503.04761*.

Christine Herlihy, Jennifer Neville, Tobias Schnabel, and Adith Swaminathan. 2024. On overcoming miscalibrated conversational priors in llm-based chatbots. *arXiv preprint arXiv:2406.01633*.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Philippe Laban, Hiroaki Hayashi, Yingbo Zhou, and Jennifer Neville. 2025. Llms get lost in multi-turn conversation. *arXiv preprint arXiv:2505.06120*.

Yubo Li, Xiaobin Shen, Xinyu Yao, Xueying Ding, Yidi Miao, Ramayya Krishnan, and Rema Padman. 2025. Beyond single-turn: A survey on multi-turn interactions with large language models. *arXiv preprint arXiv:2504.04717*.

Jiangyi Lin, Yaxin Fan, Xiaomin Chu, Peifeng Li, and Qiaoming Zhu. 2023. Multi-granularity prompts for topic shift detection in dialogue. In *International Conference on Intelligent Computing*, pages 511–522. Springer.

Geng Liu, Fei Zhu, Rong Feng, Changyi Ma, Shiqi Wang, and Gaofeng Meng. 2026. Intent mismatch causes llms to get lost in multi-turn conversation. *arXiv preprint arXiv:2602.07338*.

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.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: Nlg evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pages 2511–2522.

Adyasha Maharana, Dong-Ho Lee, Sergey Tulyakov, Mohit Bansal, Francesco Barbieri, and Yuwei Fang. 2024. Evaluating very long-term conversational memory of llm agents. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13851–13870.

Harvey Sacks, Emanuel A Schegloff, and Gail Jefferson. 1974. A simplest systematics for the organization of turn-taking for conversation. *language*, 50(4):696–735.

Emanuel A Schegloff. 2007. *Sequence organization in interaction: A primer in conversation analysis I*, volume 1. Cambridge university press.

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. 2023. Quantifying language models' sensitivity to spurious features in prompt design or: How i learned to start worrying about prompt formatting. *arXiv preprint arXiv:2310.11324*.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed H Chi, Nathanael Schärli, and Denny Zhou. 2023. Large language models can be easily distracted by irrelevant context. In *International Conference on Machine Learning*, pages 31210–31227. PMLR.

Huiyao Wang, Peifeng Li, Yaxin Fan, and Qiaoming Zhu. 2025. Simulating dual-process thinking in dialogue topic shift detection. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 2592–2602.

Wenting Zhao, Xiang Ren, Jack Hessel, Claire Cardie, Yejin Choi, and Yuntian Deng. 2024. Wildchat: 1m chatgpt interaction logs in the wild. *arXiv preprint arXiv:2405.01470*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Tianle Li, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zhuohan Li, Zi Lin, Eric P Xing, et al. 2023a. Lmsys-chat-1m: A large-scale real-world llm conversation dataset. *arXiv preprint arXiv:2309.11998*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023b. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623.

A Detection Pipeline Validation

To validate our hybrid topic breakage detection pipeline (Section 3.1), two authors independently annotated 200 flagged transitions sampled uniformly across dialogue lengths and source datasets. Each transition was labeled as either a genuine topic shift or a false positive (e.g., a follow-up question with low lexical overlap but the same underlying intent). Table 7 reports the results.

Table 7: Detection pipeline validation metrics. Green highlights strong performance.

Metric	Score
Precision	**0.91**
Recall	0.84
F1	0.87
Cohen's κ (inter-annotator)	0.85

The most common source of false negatives was gradual drifts spanning 3+ turns, where no single adjacent pair crossed the similarity threshold $\delta = 0.15$ but the cumulative shift was substantial. We address this by additionally flagging sequences where similarity drops below 0.3 over a sliding window of 3 turns.

B Taxonomy Annotation Guidelines

Annotators were provided with the following definitions and decision procedure for classifying detected breakage points into one of four categories:

1. **Abrupt Shift:** The user's turn U_{t+1} introduces a topic with no semantic or pragmatic connection to the preceding dialogue. No transitional cue is present. *Example: Turn 1–6 discuss a Python bug; Turn 7 asks "What's a good birthday gift for a 5-year-old?"*

2. **Gradual Drift:** The topic migrates through a chain of related subtopics across 2+ turns, such that the final topic is unrelated to the starting point but each adjacent pair is locally coherent. *Example: discussing a movie → its director → the director's political controversy.*

3. **Silent Abandonment:** The user stops engaging with an ongoing thread (e.g., ignores the model's follow-up question or drops a line of inquiry) and initiates a new topic without acknowledging the transition. *Example: Model asks "Shall I also add error handling?" and the user responds "How do I cook risotto?"*

4. **Topic Return:** The user pivots to a new topic for 1+ turns and then explicitly or implicitly resumes the earlier thread. *Example: Turns 1–5 on coding, Turns 6–8 on dinner plans, Turn 9: "Ok back to the code—what about line 42?"*

Decision priority: if a transition qualifies as both Abrupt Shift and Silent Abandonment (e.g., an abrupt pivot that also ignores a model question), annotators default to **Silent Abandonment**, as the absence of closure signal is the more diagnostically relevant feature. Inter-annotator agreement on the 300-instance annotation set was Cohen's $\kappa = 0.82$.

C Prompt Templates

C.1 G-Eval Scoring Prompt

The following template was used for LLM-as-Judge evaluation with GPT-4o:

```
You are evaluating the quality of an
AI assistant's response in a multi-turn
conversation. The user's most recent
query is: [QUERY]. The assistant's
response is: [RESPONSE].

Rate the response on a scale of 1-5 for
each of the following dimensions:

1. Relevance: Does the response address
the current query without referencing
information from earlier, unrelated
parts of the conversation?

2. Accuracy: Is the response factually
correct and free of hallucinations
that may have been induced by the
conversation history?

3. Completeness: Is the response as
thorough and detailed as one would
expect if the query were asked in a fresh
session?

Output your scores as JSON: {"relevance":
X, "accuracy": X, "completeness": X}
```

C.2 User Simulation Prompt

For constructing controlled C_{len} histories, we used the following prompt to generate coherent single-topic conversations of a specified length:

```
Simulate a natural multi-turn
conversation between a user and
an AI assistant about [TOPIC]. The
conversation should last exactly [N]
turns. The user should ask progressively
deeper questions about the same topic
without changing subjects. Output
the conversation as alternating
User/Assistant turns.
```

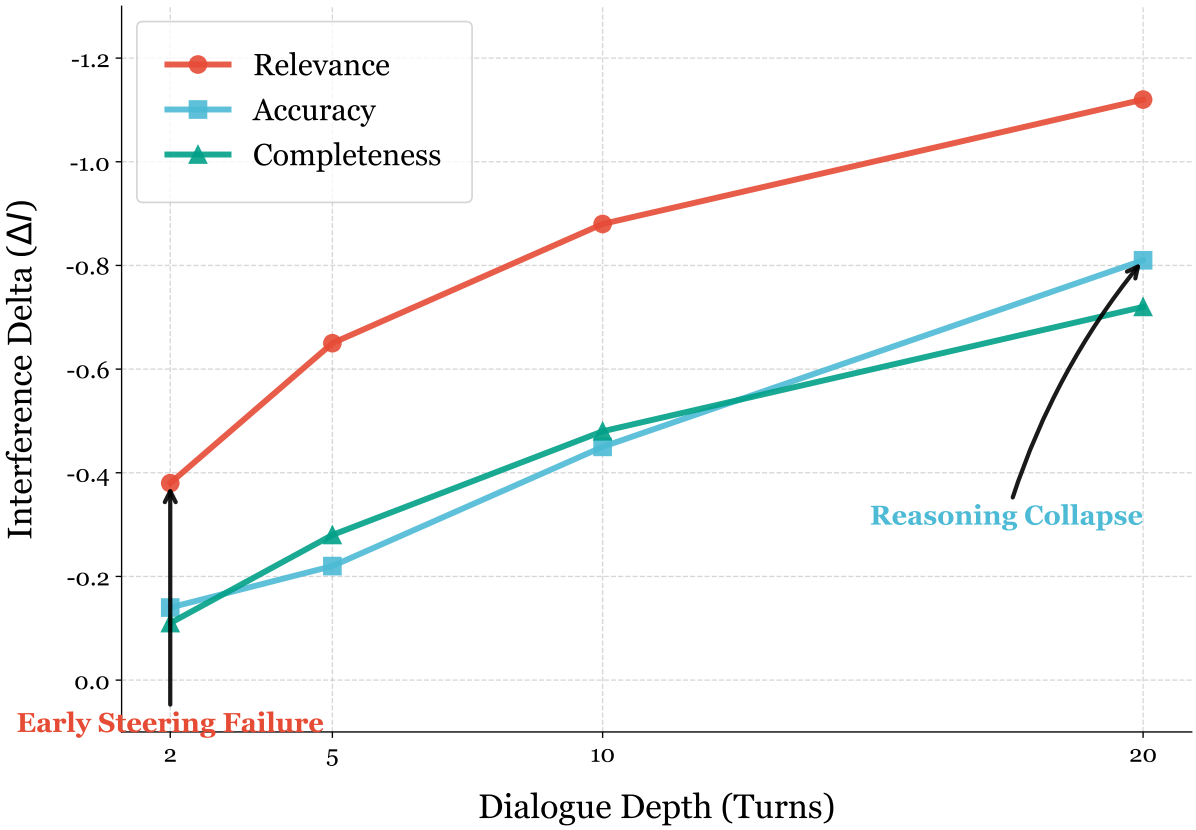

Figure 4: Dimension-specific Interference Delta across dialogue depths. Relevance degrades early and steeply; Accuracy remains stable at shallow depths but collapses beyond 10 turns.

D Full Model Configuration

Table 8 lists the exact model versions, access methods, and inference parameters used in our experiments. All models were evaluated with temperature 0 for reproducibility.

E Human Evaluation Details

To validate the LLM-as-Judge scores, three graduate students with NLP backgrounds independently rated a stratified sample of 500 responses (balanced across breakage modes and model families). Each response was scored on the same 1–5 scale for Relevance, Accuracy, and Completeness. Table 9 reports the agreement statistics.

Relevance achieved the highest agreement, consistent with its more observable nature (off-topic content is easy to spot). Accuracy had lower agreement due to cases requiring domain knowledge to verify factual claims. The overall Spearman $\rho = 0.78$ between human and automated scores supports the reliability of our G-Eval-based evaluation pipeline.

F Breakage Mode Distribution in Source Datasets

Table 10 reports the distribution of breakage modes across the two source datasets after applying our detection pipeline.

The distributions are broadly consistent across datasets, with Abrupt Shift being the most prevalent mode in both. The slightly higher proportion of Silent Abandonment in LMSYS may reflect its inclusion of more exploratory, benchmark-style interactions where users test multiple capabilities in a single session.

Table 8: Model configurations. Blue = reasoning models; gray = standard models.

Model	Version	Access	Max Tokens	Temp.
DeepSeek-R1	deepseek-r1-250120	API	4096	0.0
Llama 4 Maverick	llama-4-maverick-17b-128e	Together AI	4096	0.0
GPT-4o	gpt-4o-2024-11-20	OpenAI API	4096	0.0
DeepSeek-V3	deepseek-v3-250324	API	4096	0.0
Claude 3.5 Sonnet	claude-3-5-sonnet-20241022	Anthropic API	4096	0.0
Llama 3 70B	llama-3-70b-instruct	Together AI	4096	0.0
Llama 3 8B	llama-3-8b-instruct	Together AI	4096	0.0
GPT-3.5-Turbo	gpt-3.5-turbo-0125	OpenAI API	4096	0.0

Table 9: Human evaluation agreement and correlation with G-Eval. Green = strongest agreement.

Dimension	Human κ	Spearman ρ
Relevance	**0.79**	**0.81**
Accuracy	0.74	0.76
Completeness	0.72	0.74
Overall	**0.75**	**0.78**

Table 10: Breakage mode distribution (%) by source dataset. Red = most prevalent mode.

Mode	WildChat	LMSYS
Abrupt Shift	**43.2**	**38.5**
Silent Abandonment	26.1	30.4
Topic Return	18.7	19.8
Gradual Drift	12.0	11.3