BEYOND CONTINUITY: CHALLENGES OF CONTEXT SWITCHING IN MULTI-TURN DIALOGUE WITH LLMS

Aditya Sinha*
Netflix Inc.
Los Gatos, CA, USA
adityasinha@netflix.com

Harald Steck
Netflix Inc.
Los Gatos, CA, USA
hsteck@netflix.com

Vito Ostuni
Netflix Inc.
Los Gatos, CA, USA
vostuni@netflix.com

Matteo Rinaldi
Netflix Inc.
Los Gatos, CA, USA
matteorinaldi@netflix.com

ABSTRACT

Users interacting with Large Language Models (LLMs) in a multi-turn conversation routinely *refine* their requests or *pivot* to new topics. LLMs, however, often miss these topic shifts and carry over irrelevant context from previous turns, leading to inaccurate responses. In this paper, we stress-test the multi-turn understanding of LLMs and study the following two sub-tasks: (1) detecting whether the user *pivots* or *refines* in the current turn, and (2) shortlisting relevant context from previous turns. To this end, we construct synthetic benchmarks based on real-world datasets from varied domains, as to simulate context shifts of different levels of difficulty. We then evaluate the *zero-shot* performance of ten LLMs (open-weight, closed-source and reasoning), and demonstrate that only some reasoning and strongly instructed LLMs are accurate in detecting *pivots*; open-weight LLMs struggle with the task and frequently carry stale context even with explicit *cues*; and all models suffer from a position bias. Based on the results, we discuss key takeaways for improving long-term robustness in multi-turn capabilities for LLMs.

1 INTRODUCTION

In real-world dialogues with conversational assistants, users routinely change topics mid-dialogue (i.e., without explicitly starting a new session). In contrast, LLM-based conversational assistants typically expect coherence and continuity in interactions, which leads to the carrying of stale information across turns. Deployment of *off-the-shelf* models in chat based interfaces Casheekar et al. (2024) without any task-specific fine-tuning is a realistic scenario for small enterprises, yet the ability of these models to *reliably detect topic shifts* and *reset context* at the *current* user utterance in a multi-turn conversation is understudied in literature. To stress-test the multi-turn comprehension of the models, we frame this task as having two key aspects - (1) a classification problem predicting whether a user's intent is to *pivot* to a new topic at the current turn (relative to the previous turns), or to *refine* the current topic, and (2) a context selection task identifying prior turns that are relevant to the current turn. We evaluate these behaviours across several LLMs via prompt-based *zero-shot* detection in controlled settings to probe the inherent generalization capability of out-of-the-box LLMs in multi-turn dialogue and expose their failure modes. Overall, our contributions are as follows:

- We formalize the *pivot* vs. *refine* detection task together with appropriate *context resetting* at *pivots*, and construct synthetic conversations from real world datasets over diverse domains with precise control of pivot boundaries and their positions in a session.
- We introduce the *over-carry rate* metric to quantify context carry-over errors, complementing standard F1 measures, offering a direct readout of *reset* quality, as a useful evaluation metric for future work in assessing the quality of multi-turn understanding.

*corresponding author

- Since different real-world applications have varied latency/token constraints, we conduct an extensive *zero-shot* prompt-based evaluation of ten LLMs spanning reasoning/non-reasoning and open/closed-weight families, providing (to our knowledge) one of the first direct comparisons of out-of-the-box models on *pivot* detection and *context resetting* capabilities.
- We analyze the impact of reasoning and cue phrases in the pivot detection and context resetting tasks. We also uncover how conversation structure with different pivot positions influences model performance, and offer insights into the failure modes for each model. This could be helpful for improving the performance of LLMs (and specifically open-weight models) for enhancing their multi-turn understanding.

2 Related Work

Impressive progress has been made in conversational assistants' abilities to interact with human users in multi-turn conversations. Challenges like long-horizon decision-making or dealing with delayed rewards have been tackled by reinforcement learning (RL) in the context of multi-turn conversations with human users, e.g., Zhou et al. (2024; 2025); Jaques et al. (2020); Jang et al. (2022), as well as in the context of web-agents Wei et al. (2025); Chae et al. (2025); He et al. (2024). Maintaining relevant context in memory is another challenge in multi-turn conversations, and has been addressed by query-rewriting in Yang et al. (2024); Lai et al. (2025) and memory augmentations in Xu et al. (2021a); Packer et al. (2024); Wang et al. (2023); Zhong et al. (2024). Humans often tend to start a conversation with under-specified queries, which was identified as an important unsolved problem in Laban et al. (2025). Challenges and failure modes of LLMs for long-context management have been discussed in Liu et al. (2023); Lu et al. (2024). A common underlying assumption in all these works is that the conversation follows a coherent topic, i.e., the user does not suddenly pivot to a different/unrelated topic during the conversation.

In real-world conversations, however, users routinely change topics suddenly Soni et al. (2021). The task of topic shift detection in real-time dialogues was defined in Xie et al. (2021): *Does the utterance of the user in the current turn of the conversation change the topic compared to the preceding turns?* Early methods to solve this problem used topic segmentation Hearst (1997); Eisenstein & Barzilay (2008); Galley et al. (2003) and specialized classifiers Konigari et al. (2021), however they often struggled with incorporating subtle cues. Recent work uses hierarchical representations Lin et al. (2023) or task-specific methodology Xie et al. (2021); Hwang et al. (2024) with smaller models, such as T5. We discuss additional related works in Section A in the Appendix owing to space constraints.

This serves as the primary motivation in our effort to construct a systematic study for stress-testing several recent LLMs in their *inherent* ability to detect topic shifts in conversations via prompting alone, i.e. *zero-shot* on base models without any task-specific fine-tuning, and expose their failure modes in detecting context switches.

3 Problem Setting

In this paper, we focus on a simple yet fundamental scenario: we assume that users either intend to PIVOT from the topics discussed in the previous turn(s), or to REFINE them further. Such a binary setup is very common in practice, as evidenced from prior research in real-world conversations Soni et al. (2021). We then construct controlled "hard switches" by concatenating head segments of different conversations, which also provides us with the ground-truth turn where the pivot(s) take place, as well as with the ground-truth shortlist of relevant prior turns.

Formally, let a conversation with T turns be defined by the sequence $\{(r_t, x_t)\}_{t=1}^{T}$ with role $r_t \in \{\text{USER}, \text{ASSISTANT}\}$ and utterance x_t. For each USER turn t, we define:

$$\text{Ground-truth label} \quad y_t \in \{\text{PIVOT}, \text{REFINE}\} \tag{1}$$

$$\text{Ground-truth shortlist of indices} \quad G_t \subseteq \{1, \ldots, t-1\} \tag{2}$$

$$\text{Predicted label} \quad \hat{y}_t = f(x_{<t}, x_t) \in \{\text{PIVOT}, \text{REFINE}\} \tag{3}$$

$$\text{Predicted shortlist of indices} \quad \hat{G}_t = g(x_{<t}, x_t) \subseteq \{1, \ldots, t-1\} \tag{4}$$

where the shortlist of indices G_t contains all the previous turns $t' < t$ that are relevant for the current turn t, based on the relevance of their utterances $x_{<t}$ to the current utterance x_t. It is easy to see that

if the current turn is a pivot ($y_t = \text{PIVOT}$), then $G_t = \varnothing$. It is important to note that the predictions are functions (f and g) of the current turn t and the previous ones ($< t$), but not of future ones.

To evaluate the performance of a model, we report the classification metrics F1_{PIVOT} and $\text{F1}_{\text{REFINE}}$ computed from turn level predictions $\hat{y}_t$ averaged across all the conversations. While F1_{PIVOT} emphasizes a model's sensitivity to true topic shifts (often sparse but consequential), $\text{F1}_{\text{REFINE}}$ tracks stability on the majority class. Reporting both the metrics mitigates any issues arising from class-imbalances. Further, *context resetting* is evaluated from the predicted shortlist $\hat{G}_t$, where at all the *pivot* turns t in a conversation, the shortlist $G_t = \varnothing$ (see above). Hence, we define the *over-carry* rate of stale context at pivot-turns as $\text{OC} = \text{mean}_{\, t \in \mathcal{I}_{\text{PIVOT}}} \mathbf{1}\{\, |\hat{G}_t| > 0 \,\}$, where $\mathcal{I}_{\text{PIVOT}} = \{t : y_t = \text{PIVOT}\}$ is the set of pivot-turns in the conversation (see Appendix Section C.1). This metric directly probes whether a model can *reset* its retrieval policy at a boundary, a frequently observed failure in long-context models Liu et al. (2023). Hence, the lower the value of OC for a model, the better it is at *resetting context*, and the smaller is the "stickiness" of stale context.

4 EXPERIMENTS

In our experiments, we investigate the following three research questions: **RQ1:** How do the models perform on the tasks of PIVOT/REFINE detection and context resetting jointly? **RQ2**: How does the performance of the models for detection and context resetting change as the number of turns grow?, and finally **RQ3:** Do explicit context-switch signals and reasoning help in these tasks?

Models and Scope: We evaluate three different classes of LLMs: (1) open-weight models, (2) closed-source models, and (3) reasoning models – please see Table 1 for a detailed list. We *zero-shot* prompt the LLMs to generate a response at each turn t, comprised of a predicted label $\hat{y}_t \in \{\text{PIVOT}, \text{REFINE}\}$ and a predicted shortlist $\hat{G}_t$ of relevant previous turns. (See Appendix Section C.2 for more details on the models, setup and prompts.). We focus on the task and evaluation for *zero-shot* multi-turn understanding specifically, rather than downstream performance (See Section E).

Dataset Construction: We leverage two datasets from different domains – TopiOCQA Adlakha et al. (2022) with topic driven Q&A, and MSC Xu et al. (2021a) with *persona* based chat conversations – for constructing synthetic sessions by merging the head conversations from different conversations within the dataset. Specifically, we construct three settings – **V1** with hard concatenation of conversations, **V2** with a *cue-phrase* inserted at the pivot when concatenating conversations, making a pivot more natural and easier for the models to detect, and **V3** with a single pivot inserted at different positions (turn length) in the session. More details about the datasets, merging strategies and sample conversations can be found in Section B in the Appendix.

Results: Table 1 shows the results on the V1 and V2 settings for both datasets. Furthermore, we report results on the V3 setting of the TopiCOQA dataset in Figure 1, and discuss key findings below.

RQ1: Open-weight models are *sticky*, Closed-source models perform better. `gpt-4o` dominates the non-reasoning baselines, combining high F1_{PIVOT} and high OC, followed by `gpt-4o-mini` which is competitive on detection but *over-carries* dramatically, revealing a strong continuity bias. Amongst open weight models, two clear failure modes emerge evidently, where the first ones "always-carry" the context, such as `gemma3-12.2B` and `llama-3.1-8B` which show reasonable F1_{PIVOT} but very high OC as seen on TOPIOCQA and more so on MSC. They exhibit high OC even when they predict PIVOT, i.e. they acknowledge that the conversation is *pivoting* but still carry unrelated context. This indicates that their "classification-head" and "context selection" are potentially not well aligned, possibly requiring further instruction tuning. The second failure mode which is evident, is the *low/empty* carrying of the context, for instance `phi4-14.7B`, which shows very low OC, but also a lower F1_{PIVOT} indicating a conservative context resetting, but weak detection pattern. `mistral-7.2B` underperforms on both F1_{PIVOT} and OC in our settings.

RQ2: Position bias of context. Figure 1 shows a trend of degraded F1_{PIVOT} performance across models as the pivot position increases, indicating that *later* pivots in a session become much harder to detect, while OC fluctuates. This reveals a clear **position bias** and complements the unreliability

Table 1: Results on TopiOCQA and MSC for V1 (hard concat) and V2 (cue) across families of models. Mean ± std over repeated runs for $F1_{PIVOT}$, $F1_{REFINE}$, and pivot over-carry OC (lower is better). We **highlight** the best result in each category and underline the overall best result.

Model	Dataset	TopiOCQA			MSC		
		$F1_{PIVOT}$ (↑)	$F1_{REFINE}$ (↑)	OC (↓)	$F1_{PIVOT}$ (↑)	$F1_{REFINE}$ (↑)	OC (↓)
Open-weight (non-reasoning)							
gemma3-12.2B	V1	0.546 ± 0.001	0.887 ± 0.002	0.781 ± 0.003	**0.634 ± 0.001**	0.901 ± 0.002	0.957 ± 0.001
	V2	0.591 ± 0.003	0.908 ± 0.001	0.837 ± 0.004	**0.652 ± 0.002**	0.906 ± 0.001	0.929 ± 0.003
llama3.1-8B	V1	**0.622 ± 0.001**	**0.936 ± 0.004**	0.882 ± 0.001	0.548 ± 0.001	**0.937 ± 0.005**	0.945 ± 0.002
	V2	**0.699 ± 0.007**	**0.946 ± 0.002**	0.877 ± 0.001	0.774 ± 0.007	**0.960 ± 0.002**	0.927 ± 0.001
mistral-7.2B	V1	0.214 ± 0.001	0.180 ± 0.004	0.756 ± 0.006	0.295 ± 0.002	0.380 ± 0.003	0.991 ± 0.001
	V2	0.216 ± 0.003	0.191 ± 0.001	0.655 ± 0.009	0.299 ± 0.004	0.384 ± 0.001	0.994 ± 0.009
phi4-14.7B	V1	0.367 ± 0.003	0.731 ± 0.001	**0.002 ± 0.004**	0.486 ± 0.002	0.791 ± 0.003	**0.248 ± 0.001**
	V2	0.385 ± 0.003	0.756 ± 0.006	**0.003 ± 0.004**	0.503 ± 0.002	0.808 ± 0.002	**0.224 ± 0.005**
Closed-source (non-reasoning)							
gpt-4o-mini	V1	0.612 ± 0.002	0.915 ± 0.001	0.446 ± 0.010	**0.812 ± 0.003**	**0.964 ± 0.002**	1.000 ± 0.000
	V2	0.645 ± 0.005	0.927 ± 0.002	0.426 ± 0.006	**0.839 ± 0.005**	**0.969 ± 0.002**	0.970 ± 0.006
gpt-4o	V1	**0.695 ± 0.007**	**0.943 ± 0.002**	**0.002 ± 0.000**	0.618 ± 0.007	0.889 ± 0.002	**0.036 ± 0.001**
	V2	**0.708 ± 0.009**	**0.947 ± 0.002**	**0.000 ± 0.000**	0.635 ± 0.009	0.896 ± 0.002	**0.015 ± 0.001**
Reasoning models							
deepseek-r1-32.8B	V1	0.373 ± 0.003	0.740 ± 0.001	0.029 ± 0.002	0.497 ± 0.004	0.804 ± 0.001	0.203 ± 0.002
	V2	0.408 ± 0.002	0.783 ± 0.002	0.011 ± 0.001	0.535 ± 0.005	0.833 ± 0.004	0.094 ± 0.001
claude-3.7-sonnet	V1	0.941 ± 0.003	0.992 ± 0.002	0.006 ± 0.003	0.703 ± 0.002	0.927 ± 0.004	0.109 ± 0.003
	V2	0.945 ± 0.002	0.992 ± 0.001	0.001 ± 0.003	0.717 ± 0.002	0.931 ± 0.004	0.051 ± 0.002
gemini-2.5-pro	V1	0.871 ± 0.003	0.981 ± 0.002	0.004 ± 0.002	0.713 ± 0.004	0.931 ± 0.001	**0.048 ± 0.004**
	V2	0.879 ± 0.002	0.982 ± 0.003	0.001 ± 0.001	0.725 ± 0.004	0.934 ± 0.003	**0.024 ± 0.002**
o3	V1	**0.973 ± 0.003**	**0.997 ± 0.002**	**0.003 ± 0.001**	**0.763 ± 0.003**	**0.965 ± 0.004**	0.309 ± 0.002
	V2	**0.977 ± 0.001**	**0.997 ± 0.001**	**0.000 ± 0.002**	**0.861 ± 0.003**	**0.977 ± 0.002**	0.176 ± 0.001

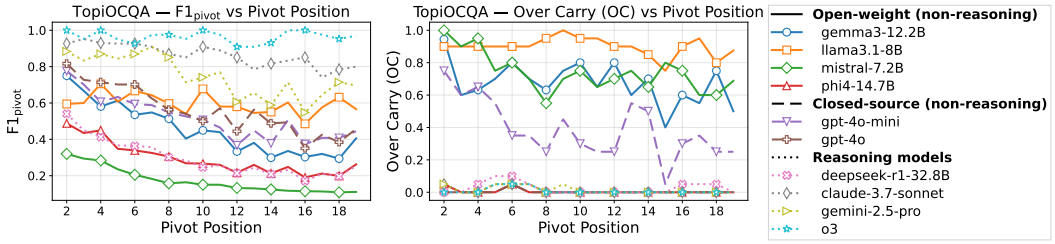

Figure 1: TopiOCQA: $F1_{PIVOT}$ (↑) and OC (↓) vs. Pivot Position (P). Majority of the models show a degradation in $F1_{PIVOT}$ as the Pivot Position increases, indicating challenging detection.

of LLMs as discussed in Laban et al. (2025). All non-reasoning models show a rapid downward trend, indicating that instruction following is not sufficient for this task, while reasoning models show a more gradual decline indicating that while an "active" state reset via reasoning helps detection, reasoning models still struggle with this task as the number of turns grow.

RQ3a: Cue phrases help label prediction. We observe that V2 with *cue-bridged* pivots improves $F1_{PIVOT}$ and in some instances OC also across all models, compared to V1. This indicates that the *cue-phrases* likely provide a "lexicalized boundary token" which gives a clear pivoting signal to the models. We observe that models with strong instruction following ability (`gpt-4o`, `gpt-4o-mini`) benefit more from these cues than models where instruction tuning is relatively weaker.

RQ3b: Reasoning acts as "gating" context. `o-3` achieves high $F1_{PIVOT}$ on TopiOCQA with zero OC in V2, and competitive performance on MSC with reduced but non-zero OC, followed by `claude-3.7-sonnet`, `gemini-2.5-pro` and `deepseek-r1-32.8B` in performance order. The low OC across these models indicates that reasoning leads to a calibrated "gating" of context, i.e. the model drops unrelated context at topic switches, while non-reasoning models fail at the task.

5 Conclusions and Future Work

In this work, we stress-test the multi-turn capabilities of LLMs and expose the challenges in *pivot-detection* and *context resetting* for practical scenarios. Our main findings are that a subset of the LLMs with strong instruction-following perform reasonably on detection of *pivots* but struggle when the switches occur at later turns; majority of the state-of-the-art open-weight LLMs often get confused as to where a *pivot* takes place and carry over stale context; and explicit reasoning, understandably, is capable of mitigating some of the challenges around *pivot-detection*. There are several solutions worth exploring for improving the multi-turn capabilities of these models, such as *targeted* fine-tuning, *two-stage* pipelining, *position-aware* context decaying and *inference-time guards*. We discuss these methods with model specific failures for the benefit of practitioners, and limitations in Sections D and E of the Appendix respectively, and defer these to future work.

Acknowledgments

We thank our colleagues and collaborators Becks Wood, Christopher Huang, Claire Campbell, Ehsan Gholami, Grace Huang, Guru Tahasildar, Hakan Baba, Ivan Provalov, Jeremey Fleischer, Jining Zhong, Kelley Robinson, Kyle Fox, Moumita Bhattacharya, Rhodes Kelley, Shahrzad Naseri, Shreyas Suhas Chavan, Spencer L'Heureux, Sudarshan Lamkhede, Thea Wang, Veli Balin, Vicky Liu, Yesu Feng, Yibo Dai and Zhe Zhang for helpful discussions and feedback throughout this project. We are grateful to the workshop organizers and reviewers for their time and constructive comments. This work was supported by internal funding and computing resources provided by Netflix Inc.

References

Vaibhav Adlakha, Shehzaad Dhuliawala, Kaheer Suleman, Harm de Vries, and Siva Reddy. TopiOCQA: Open-domain conversational question answering with topic switching. *Transactions of the Association for Computational Linguistics*, 10:468–483, 2022.

Raviteja Anantha, Svitlana Vakulenko, Zhucheng Tu, Shayne Longpre, Stephen Pulman, and Srinivas Chappidi. Open-domain question answering goes conversational via question rewriting. *arXiv preprint arXiv:2010.04898*, 2020.

Anthropic. Claude 3.7 sonnet and claude code. `https://www.anthropic.com/news/claude-3-7-sonnet`, 2025.

Avyay Casheekar, Archit Lahiri, Kanishk Rath, Kaushik Sanjay Prabhakar, and Kathiravan Srinivasan. A contemporary review on chatbots, ai-powered virtual conversational agents, chatgpt: Applications, open challenges and future research directions. *Computer Science Review*, 52:100632, 2024.

Hyungjoo Chae, Namyoung Kim, Kai Tzuiunn Ong, Minju Gwak, Gwanwoo Song, Jihoon Kim, Sunghwan Kim, Dongha Lee, and Jinyoung Yeo. Web agents with world models: Learning and leveraging environment dynamics in web navigation. In *ICLR*, 2025.

Tongfei Chen, Chetan Naik, Hua He, Pushpendre Rastogi, and Lambert Mathias. Improving long distance slot carryover in spoken dialogue systems. *arXiv preprint arXiv:1906.01149*, 2019.

DeepSeek. Deepseek-r1-distill-qwen-32b. `https://huggingface.co/deepseek-ai/DeepSeek-R1-Distill-Qwen-32B`, 2025. Model card; see also arXiv:2501.12948.

Jacob Eisenstein and Regina Barzilay. Bayesian unsupervised topic segmentation. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pp. 334–343, 2008.

Yue Feng, Gerasimos Lampouras, and Ignacio Iacobacci. Topic-aware response generation in task-oriented dialogue with unstructured knowledge access. *arXiv preprint arXiv:2212.05373*, 2022.

Michel Galley, Kathleen McKeown, Eric Fosler-Lussier, and Hongyan Jing. Discourse segmentation of multi-party conversation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pp. 562–569, 2003.

Google DeepMind. Gemini 2.5 pro. https://deepmind.google/models/gemini/pro/, 2025a.

Google DeepMind. Gemma 3 (12b) model card. https://ai.google.dev/gemma/docs/core/model_card_3, 2025b.

Hongliang He, Wenlin Yao, Kaixin Ma, Wenhao Yu, Yong Dai, Hongming Zhang, Zhenzhong Lan, and Dong Yu. WebVoyager: Building an end-toend web agent with large multimodal models. In *ACL*, 2024.

Marti A Hearst. Text tiling: Segmenting text into multi-paragraph subtopic passages. *Computational linguistics*, 23(1):33–64, 1997.

Yerin Hwang, Yongil Kim, Yunah Jang, Jeesoo Bang, Hyunkyung Bae, and Kyomin Jung. MP2D: An automated topic shift dialogue generation framework leveraging knowledge graphs. In *EMNLP*, 2024.

Youngsoo Jang, Jongmin Lee, and Kee-Eung Kim. GPT-critic: Offline reinforcement learning for end-to-end task-oriented dialogue systems. In *ICLR*, 2022.

Natasha Jaques, Judy Hanwen Shen, Asma Ghandeharioun, Craig Ferguson, Agata Lapedriza, Noah Jones, Shixiang Shane Gu, and Rosalind W. Picard. Human-centric dialog training via offline reinforcement learning. In *EMNLP*, 2020.

Rachna Konigari, Saurabh Ramola, Vijay Vardhan Alluri, and Manish Shrivastava. Topic shift detection for mixed initiative response. In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pp. 161–166, 2021.

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

Yilong Lai, Jialong Wu, Congzhi Zhang, Haowen Sun, and Deyu Zhou. AdaCQR: Enhancing query reformulation for conversational search via sparse and dense retrieval alignment. In *COLING*, 2025.

Jiangyi Lin, Yaxin Fan, Xiaomin Chu, Peifeng Li, and Qiaoming Zhu. Multi-granularity prompts for topic shift detection in dialogue, 2023. arXiv:2305.14006.

Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *arXiv preprint arXiv:2307.03172*, 2023.

Taiming Lu, Muhan Gao, Kuai Yu, Adam Byerly, and Daniel Khashabi. Insights into llm long-context failures: When transformers know but don't tell. *arXiv preprint arXiv:2406.14673*, 2024.

Xinbei Ma, Yi Xu, Hai Zhao, and Zhuosheng Zhang. Multi-turn dialogue comprehension from a topic-aware perspective. *Neurocomputing*, 578:127385, 2024.

Meta AI. Meta llama 3.1 8b. https://huggingface.co/meta-llama/Llama-3.1-8B, 2024. Model card.

Microsoft. Phi-4. https://huggingface.co/microsoft/phi-4, 2024. Model card.

Mistral AI. Mistral 7b v0.3. https://huggingface.co/mistralai/Mistral-7B-v0.3, 2024. Model card.

Chetan Naik, Arpit Gupta, Hancheng Ge, Lambert Mathias, and Ruhi Sarikaya. Contextual slot carryover for disparate schemas. *arXiv preprint arXiv:1806.01773*, 2018.

OpenAI. Gpt-4o system card. https://arxiv.org/abs/2410.21276, 2024a. System Card.

OpenAI. Gpt-4o mini: Advancing cost-efficient intelligence. https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/, 2024b.

OpenAI. Openai o3 and o4-mini system card. https://cdn.openai.com/pdf/2221c875-02dc-4789-800b-e7758f3722c1/o3-and-o4-mini-system-card.pdf, 2025.

Charles Packer, Sarah Wooders, Kevin Lin, Vivian Fang, Shishir G. Patil, Ion Stoica, and Joseph E. Gonzalez. Memgpt: Towards llms as operating systems, 2024. URL https://arxiv.org/abs/2310.08560.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*, 2016.

Mayank Soni, Brendan Spillane, Emer Gilmartin, Christian Saam, Benjamin R Cowan, and Vincent Wade. An empirical study of topic transition in dialogue. *arXiv preprint arXiv:2111.14188*, 2021.

Weizhi Wang, Li Dong, Hao Cheng, Xiaodong Liu, Xifeng Yan, Jianfeng Gao, and Furu Wei. Augmenting language models with long-term memory. *Advances in Neural Information Processing Systems*, 36:74530–74543, 2023.

Prashan Wanigasekara, Nalin Gupta, Fan Yang, Emre Barut, Zeynab Raeesy, Kechen Qin, Stephen Rawls, Xinyue Liu, Chengwei Su, and Spurthi Sandiri. Multimodal context carryover. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pp. 417–428, 2022.

Zhepei Wei, Wenlin Yao, Yao Liu, Weizhi Zhang, Qin Lu, Liang Qiu, Changlong Yu, Puyang Xu, Chao Zhang, Bing Yin, Hyokun Yun, and Lihong Li. WebAgent-R1: Training web agents via end-to-end multi-turn reinforcement learning, 2025. arXiv:2505.16421.

Huiyuan Xie, Zhenghao Liu, Chenyan Xiong, Zhiyuan Liu, and Ann Copestake. TIAGE: A benchmark for topic-shift aware dialog modeling. In *EMNLP*, 2021.

Jing Xu, Arthur Szlam, and Jason Weston. Beyond goldfish memory: Long-term open-domain conversation, 2021a. URL https://arxiv.org/abs/2107.07567.

Yi Xu, Hai Zhao, and Zhuosheng Zhang. Topic-aware multi-turn dialogue modeling. In *AAAI*, 2021b.

Dayu Yang, Yue Zhang, and Hui Fang. Mixed-initiative query rewriting in conversational passage retrieval, 2024. arXiv:2307.08803.

Wanjun Zhong, Lianghong Guo, Qiqi Gao, He Ye, and Yanlin Wang. Memorybank: Enhancing large language models with long-term memory. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 19724–19731, 2024.

Yifei Zhou, Andrea Zanette, Jiayi Pan, Sergey Levine, and Aviral Kumar. ArCHer: Training language model agents via hierarchical multi-turn RL, 2024. arXiv:2402.19446.

Yifei Zhou, Song Jiang, Yuandong Tian, Jason Weston, Sergey Levine, Sainbayar Sukhbaatar, and Xian Li. SWEET-RL: Training multi-turn LLM agents on collaborative reasoning tasks, 2025. arXiv:2503.15478.

APPENDIX

This appendix is segmented into the following key parts.

1. **Section A** discusses additional related works in the domain of topic-switching, modeling and multi-turn interactions using LLMs not included in the main paper due to lack of space.
2. **Section B** and **Section C** provide details about the construction and use of datasets, experimental details about the evaluation, metrics, and setup.
3. **Section D** provides an additional discussion on the results, identifying the role of the dataset domain, identifying model-specific failures and offering recommendations for improving the capabilities of models for multi-turn interactions.
4. **Section E** discusses the current scope of this study and limitations that we recognize in the analysis.
5. **Section F** mentions the Ethics Statement, Reproducibility Statement, and Statement on LLM Usage for our research.

A ADDITIONAL RELATED WORK

Topic Shift Detection and Topic Segmentation. Hearst (1997) proposes a lexical cohesion method for detecting topic boundaries by establishing foundational segmentation signals, while Galley et al. (2003) extends segmentation to conversations using a combination of lexical cohesion and prosodic cues. Eisenstein & Barzilay (2008) introduces a bayesian model for topic boundary discovery to improve robustness in noisy conversational data, and Xie et al. (2021) frames topic-shift aware dialog modeling as a separate task. Lin et al. (2023) uses prompt-based learning to capture topic signals and Hwang et al. (2024) generates dialogues with natural topic transitions via knowledge-graph paths.

Topic Aware Dialogue Modeling. Xu et al. (2021b) segments context into topic-aware units and matches them to candidate responses to improve response selection with topic shifts, Feng et al. (2022) conditions generation on topical signals to better ground responses in the right knowledge, and Ma et al. (2024) proposes unsupervised topic-shift detection to segment dialogues and improve downstream tasks.

Long-Context and Multi-Turn Reliability. Liu et al. (2023) demonstrates that models under-use information in the middle of long contexts, whereas Lu et al. (2024) performs probing analysis to reveal that models often encode needed information but fail to use it in the outputs, motivating explicit context management. Laban et al. (2025) benchmarks various LLMs across multi-turn tasks and find that reliability drops and recovery becomes poor when the models make early mistakes.

Task Oriented Assistants and Context Carryover. Chen et al. (2019) jointly models candidate slots to decide the context to be carried over across turns for speech systems, while Naik et al. (2018) proposes a neural decision framework for deciding carry slot values across different domains. Wanigasekara et al. (2022) extends carryover to combined voice and vision use cases with system modifications.

Conversational Memory. Xu et al. (2021a) releases multi-session dialogue data and show that retrieval-augmented and summarization based memory helps long horizon conversations. Packer et al. (2024) implements hierarchical "virtual context" with interrupts and memory tiers to manage unbounded conversational history, and Wang et al. (2023) adds a decoupled memory reader with a frozen backbone for recalling long-term context efficiently. Zhong et al. (2024) proposes a persistent memory module with decay and refresh to maintain user aligned facts across sessions.

B DATASETS

B.1 RAW DATASETS

TopiOCQA. We use TOPIOCQA (*Topical Open-domain Conversational QA*) from Adlakha et al. (2022), a multi-turn QA corpus where utterances are grounded in Wikipedia passages. The domain is factual, entity-centric QA with short system responses and information-seeking user turns. We pool the train/dev/test splits and retain only English sessions.[1] In our benchmarks, TOPIOCQA serves as a *knowledge-seeking* domain where pivots are typically signalled by lexical shifts (new entities, events, or facets), and refinement is characterized by follow-ups on the same entry.

MSC We also use MSC (Multi-Session Chat) from (Xu et al., 2021a), which comprises long-horizon conversations between human personas over multiple sessions. Dialog turns are free-form, less entity-anchored, and frequently rely on pragmatic continuity (speaker goals, preferences, commitments). Relative to TOPIOCQA, MSC is a *social/dialogue* domain where pivots are subtler (persona shifts or goal changes) and lexical cues are weaker, making context discipline harder to learn.[2]

Licensing and filtering. We respect dataset licenses. No additional annotation beyond our pivot/refine construction is performed on the raw corpora (see Appendix E for scope limits).

B.2 OVERVIEW OF SYNTHETIC BENCHMARKS

We construct three stress-test variants that introduce *controlled* pivots by concatenating head segments from different source conversations. Each variant preserves turn-order and speaker roles and provides a *ground-truth shortlisted context* per user turn (indices of prior messages considered relevant to answer the current turn).

- **V1 — Hard concatenation (no cue).** Two conversations are concatenated: the first k user turns from conversation A, then the first p user turns from conversation B. The very first turn of B is labeled PIVOT; all others are REFINE. No bridging text is inserted.
- **V2 — Cue-bridged pivot.** Same as V1 but we prepend a short cue phrase (e.g., "switching gears," "on a different note,") at the pivot user turn to provide explicit lexical indication of the shift.
- **V3 — Pivot position study.** We vary the pivot user-turn position $P \in [2, 20]$ within a session to quantify position bias. Each session has exactly one pivot at user turn P.

To avoid leakage and topic recurrence, a source conversation ID appears at most once per synthetic session.

B.2.1 V1: CONSTRUCTION DETAILS AND EXAMPLES

Generation recipe. Given a pool of source conversations $\mathcal{C}$, we sample two distinct IDs $A, B \in \mathcal{C}$. We take the first k user turns from A and the first p user turns from B, respecting original (user/system) alternation. We label the very first user turn from session B as a PIVOT and all other user turns as REFINE. The *ground-truth shortlisted context* for a user turn t is defined as the set of prior turn indices from the *same* segment (e.g., all earlier messages from A if $t \in A$). For a pivot turn (first turn from B), the shortlist is empty by construction.

Sampling hyperparameters. We use randomly sampled $k \in \{2, 3, 4\}$ and $p \in \{1, \ldots, 6\}$ (task-dependent), subject to a session budget of up to 20 total utterances. We enforce segment uniqueness (each (A, B) pair used at most once per split). Exact values are reported in Table 2.

We show two V1 examples (user turns are marked with REFINE/PIVOT and the ground-truth shortlist indices for that turn):

[1] https://github.com/McGill-NLP/topiocqa

[2] https://parl.ai/projects/msc/

Examples (TopiOCQA).

```
=== Synthetic Session v1grid_rr_000000 [v1_hard_pivot_no_cue] ===
Segments: [{'conv_id': '289', 'head_user_turns': 2},
           {'conv_id': '815', 'head_user_turns': 5}]
Num pivots: 1 | Pivot user-turn positions: [3]
Variant params: {'k_users_A': 2, 'p_users_B': 5, 'user_turn_budget': 20}
--------------------------------------------------------------------------------
U01[REFINE] (seg=1, src:289@0): who are the original avengers in the movies
S01 (seg=1, src:289@1): tony stark, steve rogers, bruce banner, and thor.
U02[REFINE] (seg=1, src:289@2): who directed it
   -> ctx[0|user] who are the original avengers...
   -> ctx[1|system] tony stark, steve rogers...
S02 (seg=1, src:289@3): joss whedon
U03[PIVOT] (seg=2, src:815@0): how long do painter turtle eggs take to hatch
...
```

Examples (MSC).

```
=== Synthetic Session v1grid_rr_000001 [v1_hard_pivot_no_cue] ===
Segments: [{'conv_id': 'train:..._7529::session_3', 'head_user_turns': 4},
           {'conv_id': 'train:..._3782::session_3', 'head_user_turns': 4}]
Num pivots: 1 | Pivot user-turn positions: [5]
Variant params: {'k_users_A': 4, 'p_users_B': 4, 'user_turn_budget': 12}
--------------------------------------------------------------------------------
U01[REFINE] ... The weather was so nice today so I decided to bike to work.
S01 ... Good for you! Was it blue skies or did you have some nice clouds...
...
U05[PIVOT] ... My parents just asked me if we had any baked goods...
...
```

Dataset statistics. Table 2 summarizes V1 statistics per source corpus (sessions, user turns, mean/min/max turns).

Table 2: Synthetic dataset statistics

Variant & Source	#Sessions	#User turns	Mean turns	Min/Max turns
V1–TopiOCQA	625	5700	10.12	<4/18>
V1–MSC	330	2340	8.09	<4/12>

B.2.2 V2: CUE-BRIDGED PIVOT AND EXAMPLES

Generation recipe. V2 mirrors V1 but we insert a lexical cue at the pivot user turn. Let $\mathcal{C}_{\text{cue}}$ be a curated list of brief phrases from the exact list ["switching gears," "on a different note," "new topic," "separately,", "switching the topic," "changing subject," "another thing,"] sampled uniformly and prepended to the pivot utterance. All other labels and shortlisted context rules remain identical to V1, and hence shares the same dataset statistics as shown in Table 2. Cues are lower-cased and punctuated minimally to avoid biasing the language model toward certain domains.

Example (TopiOCQA).

```
=== Synthetic Session v2grid_rr_000000 [v2_hard_pivot_with_cue] ===
...
U03[PIVOT*CUE] ... switching gears, how long do painter turtle eggs take to hatch
...
```

Example (MSC).

```
=== Synthetic Session v2grid_rr_000001 [v2_hard_pivot_with_cue] ===
...
U05[PIVOT*CUE] ... on a different note, My parents just asked me if we had
```

```
any baked goods...
...
```

B.3 V3: PIVOT POSITION STUDY

Generation recipe. We vary the pivot position P in a single-pivot synthetic session. For each synthetic session we first sample a base conversation A and take sufficient head turns to admit a pivot at user turn P. We then splice in the first p turns from a distinct conversation B starting at that position. The turn at P is labeled PIVOT; all earlier turns ($\leq P$) are REFINE and all later turns are REFINE. Ground-truth shortlists follow the segment-local rule; the pivot shortlist is empty.

Examples and plots. See Fig. 1 in the main paper for TOPIOCQA V3 trends ($\mathrm{F1}_{\mathrm{PIVOT}}$ and pivot over-carry vs. P), and Appendix E for discussion of position-bias implications.

B.4 PRE-PROCESSING AND QUALITY CONTROL

We normalize whitespace and strip system templates not part of the original corpora. We enforce that conversation IDs are not reused within a synthetic session and only appear once to prevent topical leakage. All synthetic creation retains the original role order.

B.5 WHY SYNTHETIC COMPOSITION?

Naturally occurring pivots in public corpora are sparse and ambiguous. Moreover, annotating turn-level pivot/refine at scale requires non-trivial annotation efforts, and also results in inter-annotator differences. In contrast, our controlled concatenations provide the ground-truth labels PIVOT/REFINE by construction. This allows us to evaluate a model's *reset policy* and *context discipline* under varied conditions (explicit cues and pivot positions) while keeping the underlying language realistic (sourced from TOPIOCQA and MSC). This complements prior work on session memory and topical drift with a diagnostic lens focused on *boundary detection* and *over-carry*. Naik et al. (2018); Chen et al. (2019)

C EXPERIMENTS

C.1 EVALUATION METRICS

We evaluate two complementary capabilities that underlie robust multi-turn dialogue: (i) *boundary detection*—recognizing whether a user turn *pivots* to a new topic or *refines* the current one; and (ii) *context discipline*—shortlisting only the prior turns relevant to answer the current request without spuriously carrying stale context forward. The first is measured with standard classification metrics; the second is captured primarily by an *over-carry* rate at pivot turns. This design follows prior work on carry-over detection in long-context dialogue systems Chen et al. (2019).

Setup and notation. Let each user turn t have a ground-truth label $y_t \in \{\mathrm{PIVOT}, \mathrm{REFINE}\}$ and a model prediction $\hat{y}_t$ from the same set. For shortlist evaluation, let $G_t \subset \{1, \ldots, t-1\}$ be the *ground-truth* set of prior turn indices deemed relevant for turn t, and let $\hat{G}_t$ be the model's predicted set. In our synthetic benchmarks, $G_t = \varnothing$ for pivot turns by construction; for refine turns G_t consists of within-segment history up to $t-1$ (See Appendix B for details of construction).

CLASSIFICATION METRICS: PIVOT VS. REFINE

We report per-class F1 for both PIVOT and REFINE. For a class $c \in \{\mathrm{PIVOT}, \mathrm{REFINE}\}$,

$$\mathrm{Prec}_c = \frac{\mathrm{TP}_c}{\mathrm{TP}_c + \mathrm{FP}_c}, \quad \mathrm{Rec}_c = \frac{\mathrm{TP}_c}{\mathrm{TP}_c + \mathrm{FN}_c}, \quad \mathrm{F1}_c = \frac{2\,\mathrm{Prec}_c\,\mathrm{Rec}_c}{\mathrm{Prec}_c + \mathrm{Rec}_c}.$$

We denote $\mathrm{F1}_c$ for PIVOT as $\mathrm{F1}_{\mathrm{PIVOT}}$ and $\mathrm{F1}_c$ for REFINE $\mathrm{F1}_{\mathrm{REFINE}}$.

Motivation. $\mathrm{F1}_{\mathrm{PIVOT}}$ emphasizes a model's sensitivity to true topic shifts (often sparse but consequential), while $\mathrm{F1}_{\mathrm{REFINE}}$ tracks stability on the majority class in many real conversations. Reporting both

mitigates class-imbalance artifacts and exposes asymmetries such as "always-refine" or "always-pivot" behaviors.

CONTEXT DISCIPLINE: CARRYOVER METRICS

Pivot over-carry. At a true topic switch, the expected behavior for a model is to drop the old context and start fresh. Hence, if the model is still bringing along context from any earlier turns, it can be said to be *over-carrying* i.e., allowing stale context to carry-over into the new topic, when it should have been rightfully *reset*. At a pivot turn t, the ground-truth shortlist is empty by design ($G_t = \varnothing$). Any non-empty predicted shortlist is an *over-carrying* error. We define the *over-carry* rate in an conversation as

$$\mathrm{OC} \;=\; \frac{1}{|\mathcal{I}_{\mathrm{PIVOT}}|} \sum_{t \in \mathcal{I}_{\mathrm{PIVOT}}} \mathbf{1}\{\,|\hat{G}_t| > 0\,\},$$

where $\mathcal{I}_{\mathrm{PIVOT}} = \{t : y_t = \mathrm{PIVOT}\}$. Lower is better. This metric directly probes whether the model can *reset* its retrieval policy at a boundary, a failure mode frequently observed in long-context models Lu et al. (2024).

Why we emphasize pivot over-carry. Over-carry at pivot is *well-identified* in our synthetic design (ground-truth shortlist is provably empty), making it a robust diagnostic for reset behavior without needing any human intervention. In contrast, *under-carry*, i.e. failing to retrieve enough prior context on refine turns is harder to measure faithfully without curated relevance labels: not all within-segment messages are strictly necessary. For this reason, we treat OC at pivot as the primary signal of context discipline.

Reporting and aggregation. Unless noted, we average across all user turns within a split and report per-model/per-variant means $\pm$ standard deviations over 2 repeated runs (Table 1). For V3 we additionally condition by pivot position P to analyze position bias in both $\mathrm{F1}_{\mathrm{PIVOT}}$ and OC (Fig. 1).

C.2 MODELS AND SETUP

Models. We evaluate ten models spanning open/closed weights and reasoning vs. standard instruction tuning:

- **Closed, non-reasoning:** `GPT-4o` OpenAI (2024a), `GPT-4o-mini` (OpenAI, 2024b),
- **Closed, reasoning:** `Gemini-2.5-pro` Google DeepMind (2025a), `Claude-3.7-sonnet` Anthropic (2025), `OpenAI o3` OpenAI (2025)
- **Open, non-reasoning:** `Llama 3.1-8B` Meta AI (2024), `Gemma 3-12.2B` Google DeepMind (2025b), `Phi-4-14.7B` Microsoft (2024), `Mistral-7.2B` Mistral AI (2024)
- **Open, reasoning:** `DeepSeek-R1-32.8B` DeepSeek (2025)

We keep official model defaults unless stated and treat vendor hosted models as black boxes (no weight access).

Hardware and inference environment. Closed models are accessed via the provider's hosted inference endpoints (region and provisioning managed by the vendor). Open-weight models are served locally with `ollama` on a Linux host with Nvidia A100x4 GPUs with 40 GB memory. All experiments are executed under fixed seeds for reproducibility. Code and instructions to download and process the datasets for the paper can be found at the URL: `https://github.com/adityaasinha28/BeyondContinuityLLMs`.

Licensing and usage. We respect each model's license: closed models are accessed under their ToS; open-weight models (Llama, Gemma, Mistral, Phi-4, DeepSeek R1) are used for *research benchmarking only*. We do not fine-tune or adapt any model in this study.

Prompting and JSON conformance. All models receive a single *joint* prompt per user turn (history $\rightarrow$ current message) with a compact system instruction requiring *JSON-only* output in this format:

{`answer` : str, `predicted_label` : {PIVOT, REFINE}, `relevant_context` : int[]}.

We use a strict parser with guardrails that rejects schema with incorrect formats and additional keys. For open-weight models we apply a minimal JSON prefix/suffix fence and stop tokens when available. We skip the first user turn in every session for both inference and evaluation (as it has no usable history). We use API calls with retries for instances with schema violation, although we did not observe any significant violation rates across the models.

Decoding and runtime. For vendor-hosted models (GPT-4o/mini, o3, gemini-2.5-pro, claude-3.7-sonnet) we use `temperature`=0.0 with `reasoning` parameter set, wherever applicable. For open-weight models via `ollama` we use:

$$\texttt{temperature} = 0.0, \quad \texttt{top\_p} = 0.9, \quad \texttt{top\_k} = 50,$$

C.3 PROMPTS

System prompt — TopiOCQA

You are a helpful assistant interacting with the user.
Given the conversation history and the CURRENT user message, your task is to:
1) Respond to the user with an ANSWER (concise, <= 40 words).
2) Classify the CURRENT message as PIVOT (switching to a new topic) or REFINE (continuing on the same thread).
3) Select which PRIOR message ids (ints) are relevant context to answer the CURRENT message.
STRICT OUTPUT FORMAT RULES:
- Return ONE JSON object only (no prose, no markdown, no code fences).
- Use EXACTLY these keys: answer, predicted_label, relevant_context.
- Valid predicted_label values must be one of these: "PIVOT" or "REFINE".
- relevant_context must be an array of integers (ids from the history).
- Do NOT include any extra keys like "type", "properties", "required", "schema", or examples.
- Do NOT include the current message id in relevant_context.
- Do NOT make up ids that are not present in the history.
Your output must be valid JSON matching:
{ "answer": "<string>", "predicted_label": "PIVOT"|"REFINE", "relevant_context": [<int>, ...] }

System prompt — MSC (Persona Conversations)

You are a helpful assistant interacting with the user.
Given the conversation history and the CURRENT user message, your task is to:
1) Respond to the user with an ANSWER (concise, <= 40 words).
2) Classify whether the CURRENT user message persona is PIVOT (switching to a new topic and persona) or REFINE (continuing on the same thread and persona).
3) Select which PRIOR message ids (ints) are relevant context and consistent persona with the CURRENT message.
STRICT OUTPUT FORMAT RULES:
- Return ONE JSON object only (no prose, no markdown, no code fences).
- Use EXACTLY these keys: answer, predicted_label, relevant_context.
- Valid predicted_label values must be one of these: "PIVOT" or "REFINE".
- relevant_context must be an array of integers (ids from the history).
- Do NOT include any extra keys like "type", "properties", "required", "schema", or examples.
- Do NOT include the current message id in relevant_context.
- Do NOT make up ids that are not present in the history.
Your output must be valid JSON matching:
{ "answer": "<string>", "predicted_label": "PIVOT"|"REFINE", "relevant_context": [<int>, ...] }

D ADDITIONAL DISCUSSION ON RESULTS

D.1 DATASET AND DOMAIN

The domain of the underlying dialogue shapes both the target behavior and the dominant error modes. On TOPIOCQA (knowledge-seeking, task-driven QA), models are implicitly rewarded for *resetting* context when the topic changes and for grounding answers in the most recent, local evidence. This inductive bias aligns well with our pivot/shortlist objectives: most systems achieve higher $\mathrm{F1}_{\mathrm{PIVOT}}$ and lower over-carry (OC) on TOPIOCQA than on MSC. In contrast, MSC is persona-centric chit-chat, where human turns are stylistically and semantically similar across long spans. Models trained with conversational objectives and RLHF priors that emphasize politeness, empathy, and "follow the thread" coherence exhibit a strong *continuity prior*: even when they correctly flag a PIVOT, they often continue to retrieve or reference stale persona turns, showing higher OC.

These domain-conditioned behaviors are visible across families. Closed-source non-reasoning models (e.g., `gpt-4o`, `gpt-4o-mini`) tend to be context-disciplined on TOPIOCQA—near-zero OC when a pivot is detected—yet show stickiness on MSC, where the conversational reward model has likely learned that maintaining continuity is helpful. Open-weight mid-size models (`gemma3-12B`, `llama3.1-8B`) display the same qualitative pattern with larger absolute OC, suggesting weaker internal context selection. Reasoning models diverge: `o3` is an outlier with near-ceiling $\mathrm{F1}_{\mathrm{PIVOT}}$ on TOPIOCQA and uniformly low OC, `claude-3.7-sonnet` and `gemini-2.5-pro` follow with a similar trend and comparatively lower performance, whereas `deepseek-r1-32B` still struggles relatively in drawing crisp topic boundaries in persona-heavy settings, while showing a uniformly low OC.

Overall, there are two key implications of the domain of the datasets. First, *style matters*: in persona chat, the optimal policy is often to *carry* unless explicitly signaled otherwise, whereas in knowledge-seeking QA the optimal policy is to *reset* unless continuity is evident. Second, *prompting matters*: lexical cue phrases (our V2 setting) narrow the domain gap by externalizing the boundary decision. Cues help most where the training prior pushes in the wrong direction (e.g., MSC); they help less where the prior already aligns with the task (e.g., TOPIOCQA). The same model can therefore look disciplined in one domain and challenged in another (e.g., `gpt-4o-mini` has low OC and solid $\mathrm{F1}_{\mathrm{PIVOT}}$ on TOPIOCQA, but high OC on MSC).

D.2 MODEL SPECIFIC FAILURES

`o3`, `claude-3.7-sonnet` & `gemini-2.5-pro` (reasoning, closed-source). On TOPIOCQA, `o3` is near–ceiling on $\mathrm{F1}_{\mathrm{PIVOT}}$ with the lowest overall OC, indicating precise boundary detection and disciplined shortlist selection and resetting at the pivot. On MSC it occasionally exhibits mild OC when the topic shift is subtle (e.g., persona or stance changes without topical lexical shift), consistent with a coherence prior that favors continuity in chit–chat. `o-3` is followed in performance by `claude-3.7-sonnet`, followed by `gemini-2.5-pro`, with almost consistent low OC for all the models, and varying $\mathrm{F1}_{\mathrm{PIVOT}}$ performance, indicating their detection capabilities along with reasoning. **Typical failure.** Missed *fine–grained* pivots that are pragmatics–driven rather than topic–driven.

`deepseek-r1-32B` (reasoning, open–weights). Compared to `o3`, `deepseek-r1-32B` attains lower $\mathrm{F1}_{\mathrm{PIVOT}}$ but remains relatively competitive on OC, suggesting that its step–by–step reasoning encourages relevance checks before carrying context. **Typical failure.** *Boundary imprecision*: it often narrows to the right neighborhood but does not fire a PIVOT at $x_t = P$, yielding false REFINEs.

`gpt-4o` (non–reasoning, closed–source). Strong $\mathrm{F1}_{\mathrm{PIVOT}}$ and near–zero OC on TOPIOCQA; low OC on MSC shows very good context discipline relative to peers. The primary weakness is *position bias*: later pivots in V3 show a measurable drop in PIVOT recall, especially when many coherent pre–pivot turns accumulate.

`gpt-4o-mini` (non–reasoning, closed–source). Good $\mathrm{F1}_{\mathrm{PIVOT}}$ overall, but high OC on MSC: the model recognizes pivots yet continues to retrieve older persona turns. It shares the *position bias* of `gpt-4o` but with larger magnitude.

`gemma3-12B` & `llama3.1-8B` (non-reasoning, open–weights). Moderate $\mathrm{F1}_{\mathrm{PIVOT}}$ and clear improvements with cues, but consistently *high* OC. A frequent pattern is *label–context mismatch*: the model flags PIVOT yet still carries substantial pre–pivot turns in its shortlist.

`phi4-14.7B` (non-reasoning, open–weights). Very low OC on TOPIOCQA but under–detects pivots (many true PIVOTs labeled REFINE). The behavior suggests an *over–aggressive reset policy* on context selection (leading to low OC) combined with a conservative labeler that hesitates to classify as PIVOT unless cues are explicit.

`mistral-7.2B` (non-reasoning, open–weights). Low $\mathrm{F1}_{\mathrm{PIVOT}}$ and high OC in our setting: by default the model exhibits a strong *continuity prior*. It often treats concatenated sessions as one thread unless given overt boundary signals.

D.3 IMPROVEMENTS TO MODELS

We offer training–time learning signals and inference–time system design recommendations that directly target the observed failure modes (position bias, label–shortlist mismatch, cue over–reliance, and domain–based continuity priors).

Training–time data and objectives for open models.

1. **Instruction Tuning for boundary awareness.** Extend SFT with samples that (i) lack lexical cues but require PIVOT and (ii) require empty carry on PIVOT. Short, explicit rubrics ("On PIVOT, *do not carry* prior turns") help close the label–shortlist gap.
2. **Late–pivot curriculum.** To reduce position bias, oversample sessions with large refine–prefix upto P (heavy–tail over P) and include near–pivot *hard negatives*. This schedule would help in improving robustness to late pivots.
3. **Cue diversity and ablations.** Mix cue/no–cue data (and paraphrastic cues) so the model learns pivots as a semantic property rather than a token separator. At the same time, keep a small held–out cue set to check over–fitting to specific phrases.

Inference Time System design

1. **Prompt Engineering with Boundary awareness.** Chain-of-thought style instructions in the prompt along with a large number of few-shot in-context examples for challenging boundary cases can improve the inference time performance of the out-of-the-box models for both detection and context resetting tasks. Similarly, including challenging examples of late-pivot detections can help reduce the degradation in performance, as context grows.
2. **Two Stage Controller.** We can modify the system to decompose the prediction problem into multiple steps.
   (a) **Stage–1 pivot gate (binary).** Train a compact classifier $p_\theta(y_t \in \{\text{PIVOT}, \text{REFINE}\} \mid h_t)$ on top of hidden summaries h_t (e.g., final token CLS for the current turn plus a pooled history embedding) to trade off pivot recall vs. false positives.
   (b) **Stage–2 context selector (conditional).** Given $\hat{y}_t$, select shortlist S_t: if $\hat{y}_t = \text{PIVOT}$, *enforce* $S_t = \varnothing$ (hard rule and can be done programmatically); if $\hat{y}_t = \text{REFINE}$, select from the current thread only (e.g., k most recent same–segment turns by a relevance scorer).

Reasoning models: structure the process. For chain–of–thought or tool–use models, separate *Plan* and *Answer*: (i) decide PIVOT/REFINE with a brief justification ("boundary rationale"), (ii) *then* generate the answer using only the approved shortlist.

Mitigating position bias (P–aware control). Later pivots (as $P \uparrow$) are harder to detect due to anchoring on context and latent retrieval. To solve this problem, we can treat the turns within a session differently, as the number of turns within a session grows. Specifically, we can add: (i) *recency–biased context windows* (summarize early turns into a single synopsis), (ii) a *P–aware prior* that linearly increases the pivot probability with depth, (iii) *explicit anti–anchoring prompts* ("If this is a new topic, ignore prior messages entirely"), and (iv) *cue amplification* for large P (stronger lexical boundary markers when long histories accumulate).

E LIMITATIONS

There are several limitations of the current study which we address and acknowledge in this section.

Scope of our work. Our study isolates two behaviors: (i) recognizing topic shifts (*PIVOT*) versus within-topic continuation (*REFINE*), and (ii) deciding which prior turns to carry forward as a *shortlist*. We do *not* evaluate downstream task quality (e.g., answer correctness in QA) or long-form generation faithfulness, to focus on context understanding as the primary problem. As a result, a model may correctly flag a *PIVOT* yet still produce an incorrect answer; conversely, a model may answer well on *REFINE* turns while mislabeling the turn type. The interplay between task success (answer quality) as well as interaction competence (identifying PIVOT or REFINE and *context* resetting is an interesting research problem worth investigating further. Additionally, we restrict our study to identical *zero-shot* prompts across models to evaluate the *inherent* capabilities of out-of-the-box LLMs. Advanced prompting strategies such as Chain-of-thought prompting and including in-context examples in the prompt could require model specific modifications and are left as future work.

Over-carry is measured at the pivot; under-carry on refines is more challenging to measure. Our primary behavioral error is *over-carry*, i.e. models retaining irrelevant context at a true *PIVOT*. (See Section C.1 for details). While we also report the detection $\mathrm{F1}_{\mathrm{PIVOT}}$ on *REFINE* turns as a measure of detection performance, observing true *under-carry* performance is challenging since the labeling for a ground-truth shortlist of relevant context for a *REFINE* turn may be unavailable. This may be inferred through different techniques which would require addressing challenges such as ambiguity, coreference, dialogue flow, capturing nuances, etc and approximate the human relevance of context required for responding to a current turn. This would also require human validation or annotation for a subset of the turns to calibrate the shortlist quality and quantify any bias.

Binary turn typing omits *partial* or *soft* pivots. We limit turn types to a binary {*REFINE* or *PIVOT*} label, however, natural conversations exhibit graded topical drift and blended intents, long studied in topic segmentation (Hearst, 1997; Eisenstein & Barzilay, 2008; Galley et al., 2003). Our construction treats any head splice from a new session as a hard *PIVOT*. In practice, users often pivot partially where a second request introduces a new intent while remaining anchored to the prior entity. Our binary labels would count this as *PIVOT*, although a graded scheme (or multi-label *intent+topic* labeling) may be more faithful for such *soft* pivots.

Domain and language coverage. We evaluate on TOPIOCQA (Wikipedia-oriented QA) and MSC (long-term persona chat). Both are in English; we do not test multilingual pivot detection, code-switched languages, or chats in specialized domains (e.g., legal or medical). Moreover, topic familiarity (e.g., popular entities in TOPIOCQA) can interact with pivot detection by making carry decisions easier (Rajpurkar et al., 2016; Anantha et al., 2020). Extending to more domains and languages is left to future work.

F STATEMENTS

Ethics Statement. Our study does not involve any human subjects. We also do not foresee any negative societal impacts, discrimination/bias/fairness concerns, and privacy or security concerns from the outcomes of our research.

Reproducibility Statement. All of our experiments were conducted with fixed seeds for ensuring reproducibility.

Statement on LLM Usage. While LLM-based coding assistants were used in the process of conducting experiments, we did not use any LLMs for writing of the paper (including the main text, Appendix and References).