Do You Mean? Clarification Gating Does Not Improve Intent Preservation in Conversational Rewriting

Anonymous Authors

Abstract

We study whether LLM-based query rewriting preserves user intent and whether clarification gating can improve intent fidelity without excessive questioning. We evaluate four policies on a 120-example, context-heavy sample from the QReCC test split: No Rewrite, Direct Rewrite, Always Clarify, and Gated Clarify. We score intent preservation using gold-rewrite overlap (BLEU-1, ROUGE-L), semantic similarity (SBERT cosine), and an LLM judge (1–5), and we report clarification rate as a measure of user burden. Direct Rewrite achieves the highest LLM-judge score (4.725) and strong semantic similarity (0.851). Gated Clarify reduces clarification rate to 56.7% but does not improve intent preservation; its LLM-judge score is slightly worse than Direct Rewrite (4.567, $p = 0.046$). Always Clarify clarifies every example and performs worst on intent fidelity metrics. These results suggest that naive ambiguity-based gating can reduce questions without improving intent preservation, and may even introduce small regressions when the model already resolves context correctly. Our evaluation framework clarifies this trade-off and motivates future work on calibrated ambiguity detection and human-in-the-loop clarification.

1 Introduction

Assistant systems rewrite user queries to be helpful, but even small changes can shift intent and erode trust. In conversational settings, the risk is higher because questions depend on context and entities that may be implicit. A system that "helpfully" rewrites a query can still be wrong in ways that are hard to detect and frustrating to users.

Intent preservation is therefore a core requirement for query rewriting in search and assistants. At the same time, asking clarifying questions can add friction. The central product tension is clear: avoid wrong rewrites without annoying users with unnecessary clarification.

Prior work on conversational question rewriting focuses on rewriting quality or downstream retrieval and QA performance Elgohary et al. [2019], Anantha et al. [2021], Mo et al. [2023], Wu et al. [2021]. Clarification-question research is often studied separately Kumar and Black [2020]. As a result, we lack direct evidence about whether clarification helps intent preservation in rewriting pipelines and how much user burden it introduces.

We examine this gap with a controlled comparison of four policies: No Rewrite, Direct Rewrite, Always Clarify, and Gated Clarify. Our approach evaluates intent preservation with lexical overlap, semantic similarity, and an LLM-based judge, and it measures clarification rate as a user-burden proxy. figure 1 summarizes the evaluation workflow.

Quantitatively, Direct Rewrite achieves the highest LLM-judge score (4.725), while Gated Clarify reduces clarifications to 56.7% but slightly reduces LLM-judge fidelity (4.567, $p = 0.046$). Always Clarify clarifies every example and performs worst on intent metrics. These results show that naive ambiguity gating reduces questions without improving intent preservation.

In summary, our main contributions are:

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

Input: context C and question q.
Policy: choose No Rewrite, Direct Rewrite, Always Clarify, or Gated Clarify.
Clarify (optional): ask c, obtain answer a.
Rewrite: produce r from C, q, and optional a.
Evaluate: compare r to gold rewrite g with lexical, semantic, and LLM-judge metrics; record clarification rate.

Figure 1: Overview of our rewrite-and-clarify evaluation pipeline.

- We propose an intent-preservation evaluation that combines gold-rewrite overlap, semantic similarity, and LLM judging with a clarification-rate cost.
- We conduct a controlled comparison of rewrite and clarification policies on a context-heavy QReCC sample.
- We quantify the clarification-rate vs. intent-fidelity trade-off and show that naive gating can reduce questions while slightly degrading intent fidelity.

2 Related Work

Conversational question rewriting. Datasets such as CANARD and QReCC formalize the task of rewriting context-dependent questions into standalone queries Elgohary et al. [2019], Anantha et al. [2021]. Subsequent work improves rewriting and downstream retrieval with generative reformulation Mo et al. [2023] and reinforcement learning for retrieval objectives Wu et al. [2021]. Term-level query resolution offers a lighter-weight alternative that can preserve intent by adding context terms rather than fully rewriting Voskarides et al. [2020]. Unlike these efforts, we focus on directly measuring intent preservation and user burden across rewrite and clarification policies.

Clarification question generation. Large-scale datasets such as ClarQ and related benchmarks enable learning when and how to ask clarifying questions Kumar and Black [2020]. However, clarification is typically evaluated separately from rewriting pipelines. We connect these lines of work by measuring how clarification gating changes intent fidelity in rewrite systems.

Reformulation sensitivity. Prior analysis shows that retrieval and QA can be sensitive to seemingly minor reformulations Vakulenko et al. [2020]. This motivates evaluating intent preservation directly rather than inferring it from downstream metrics. Our study builds on this insight by testing whether clarification reduces unintended reformulation drift.

3 Methodology

Problem formulation. Given a conversation context C and a user question q, a rewriting system outputs a standalone rewrite r that should preserve the intent of the gold rewrite g. A clarification policy may ask a question c before producing r. Our goal is to maximize intent preservation while minimizing user burden, measured by clarification rate.

Policies compared. We evaluate four methods: No Rewrite (use q directly), Direct Rewrite (rewrite without clarifying), Always Clarify (ask a clarification question for every example), and Gated Clarify (ask only when the model predicts high ambiguity).

Dataset and sampling. We use the QReCC test split (16,451 examples) and sample 120 examples with non-empty context using a fixed random seed (42). This sample has 90 duplicate Context+Question pairs, an average question length of 7.33 tokens (std 2.31), average rewrite length of 11.40 tokens (std 5.15), and average context length of 5.83 turns (std 4.63).

Clarification simulation. For Always Clarify and Gated Clarify, the system asks a clarification question and we generate a synthetic user answer consistent with the gold intent. This isolates the effect of clarification from user variability, but it also introduces a limitation we discuss in section 5.

Evaluation metrics. We measure intent preservation via BLEU-1 and ROUGE-L against the gold rewrite, SBERT cosine similarity (`all-MiniLM-L6-v2`), and an LLM judge score on a 1–5 scale. We report clarification rate as the fraction of examples that trigger a question.

Method	BLEU-1	ROUGE-L	SBERT Cosine	LLM Judge	Clarification Rate
No Rewrite	**0.873** ± 0.113	**0.650** ± 0.185	0.648 ± 0.170	4.708 ± 0.712	0.000
Direct Rewrite	0.565 ± 0.191	0.608 ± 0.195	**0.851** ± 0.123	**4.725** ± 0.658	0.000
Always Clarify	0.529 ± 0.184	0.553 ± 0.184	0.843 ± 0.118	4.450 ± 0.729	1.000
Gated Clarify	0.568 ± 0.203	0.601 ± 0.204	**0.851** ± 0.124	4.567 ± 0.692	0.567

Table 1: Intent preservation and clarification burden on a 120-example QReCC sample. Best intent metrics are bolded (higher is better).

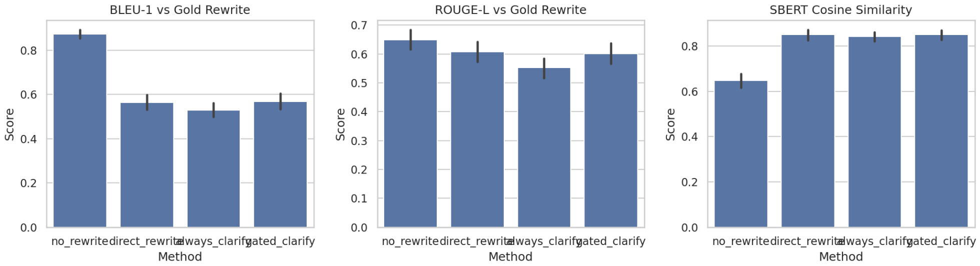

Figure 2: Method comparison across intent metrics. Direct rewrites score highest on LLM-judge fidelity, while clarification-based methods show no gain in semantic similarity.

Implementation details. We use GPT-4.1 (temperature 0, max tokens 200) for rewriting, clarification, and judging. SBERT embeddings are computed in batches of 64 on 2 × NVIDIA RTX 3090 GPUs. Each method is run once with deterministic decoding.

4 Results

Main comparison. Table 1 summarizes the intent-preservation metrics and clarification rate. Direct Rewrite achieves the highest LLM-judge score (4.725) and ties for best SBERT cosine (0.851). Gated Clarify reduces clarification rate to 56.7% but does not improve intent fidelity relative to Direct Rewrite. Always Clarify clarifies every example and performs worst on intent metrics, indicating that uncalibrated clarification can hurt performance.

Statistical tests. We compare Gated Clarify to Direct Rewrite with paired t-tests on per-example scores. Differences in BLEU-1 ($t = 0.259, p = 0.796$), ROUGE-L ($t = -0.450, p = 0.654$), and SBERT cosine ($t = -0.031, p = 0.975$) are not significant. The LLM-judge score is significantly lower for Gated Clarify ($t = -2.017, p = 0.046$, $d = -0.185$), indicating a small but measurable regression.

5 Discussion

Why clarification did not help. The Direct Rewrite policy already resolves many context-heavy questions correctly, so clarification can introduce redundant or noisy information. We observed failure modes where clarification questions focused on irrelevant details or the synthetic answers simplified the original intent, which can explain the lower LLM-judge scores for Gated Clarify and Always Clarify.

Clarification burden trade-off. Gated Clarify cuts the clarification rate by 43.3% relative to Always Clarify (from 100% to 56.7%), but this reduction does not translate into better intent fidelity. This suggests that naive ambiguity signals are misaligned with actual intent errors and highlights the need for calibrated detectors.

Limitations. Our clarifying answers are synthetic and derived from gold intent, which may not reflect real user behavior. The LLM judge is a proxy for human evaluation and may share biases with the rewrite model. The evaluation uses a single dataset and a small sample (120 examples), so conclusions should be validated at larger scale.

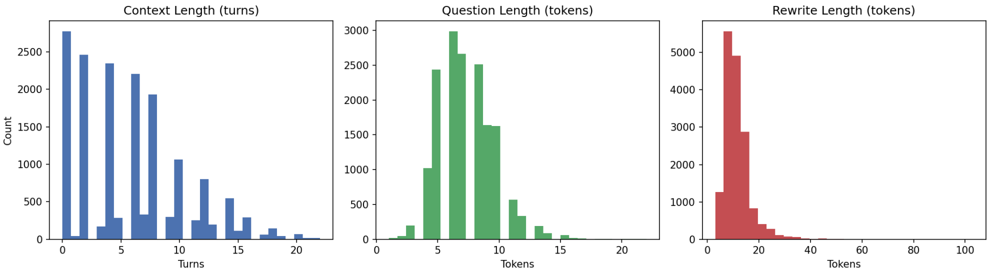

Figure 3: Distributions of question, rewrite, and context lengths for the sampled QReCC subset. The evaluation focuses on context-heavy examples.

Broader implications. Systems that rewrite user inputs should treat clarification as a precision instrument rather than a default action. Measuring intent preservation directly can inform product decisions and reduce user frustration, and future work should incorporate user annoyance signals and human-in-the-loop clarification to better align with real-world usage.

6 Conclusion

We evaluated intent preservation for conversational query rewriting and tested whether clarification gating improves fidelity without excessive questioning. On a 120-example QReCC sample, Direct Rewrite rewrites achieved the highest LLM-judge scores, while Gated Clarify reduced clarification rate but slightly worsened intent fidelity ($p = 0.046$). The key takeaway is that naive clarification gating reduces questions but does not improve intent preservation when direct rewrites already resolve context well. Future work should incorporate human clarification data, supervised ambiguity detectors, and broader evaluations across datasets such as CANARD and QReCC.

References

Raviteja Anantha, Svitlana Vakulenko, Zhucheng Tu, Shayne Longpre, Stephen Pulman, and Srinivas Chappidi. Open-domain question answering goes conversational via question rewriting. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 520–534, 2021.

Ahmed Elgohary, Denis Peskov, and Jordan Boyd-Graber. Can you unpack that? learning to rewrite questions-in-context. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 591–601, 2019.

Vaibhav Kumar and Alan W. Black. Clarq: A large-scale and diverse dataset for clarification question generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 632–647, 2020.

Fengran Mo, Kelong Mao, Yutao Zhu, Yihong Wu, Kaiyu Huang, and Jian-Yun Nie. Convgqr: Generative query reformulation for conversational search. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4961–4976, 2023.

Svitlana Vakulenko, Shayne Longpre, Zhucheng Tu, and Raviteja Anantha. A wrong answer or a wrong question? an intricate relationship between question reformulation and answer selection in conversational qa. *arXiv preprint arXiv:2010.06835*, 2020.

Nikos Voskarides, Dan Li, Pengjie Ren, Evangelos Kanoulas, and Maarten de Rijke. Query resolution for conversational search with limited supervision. *arXiv preprint arXiv:2005.11723*, 2020.

Zeqiu Wu, Yi Luan, Hannah Rashkin, David Reitter, Hannaneh Hajishirzi, Mari Ostendorf, and Gaurav Singh Tomar. Conqrr: Conversational query rewriting for retrieval with reinforcement learning. *arXiv preprint arXiv:2112.08558*, 2021.