← Back to **Author Console** (/group?id=NeurIPS.cc/2025/Conference/Authors#your-submissions)

QueryBandits for Hallucination Mitigation: Exploiting Semantic Features for No-Regret Rewriting

(/pdf?id=sf8ALgiDJd)

[redacted]

11 May 2025 (modified: 18 Sep 2025) Submitted to NeurIPS 2025 Conference, Senior Area Chairs, Area Chairs, Reviewers, Authors Revisions (/revisions?id=sf8ALgiDJd) BibTeX

Keywords: Contextual Bandits, Query Rewriting, Large Language Models, Hallucination Mitigation

TL;DR: QueryBandit uses a contextual bandit over 17 linguistic features to choose among five rewrite strategies, achieving an 87.5% win rate on perturbed QA queries (vs. 44.9% paraphrase, 27.2% expansion).

Abstract:
Advanced reasoning capabilities in Large Language Models (LLMs) have caused higher hallucination prevalence; yet most mitigation work focuses on after-the-fact filtering rather than shaping the queries that trigger them. We introduce QueryBandits, a bandit framework that designs rewrite strategies to maximize a reward model, that encapsulates hallucination propensity based upon the sensitivities of 17 linguistic features of the input query-and therefore, proactively steer LLMs away from generating hallucinations. Across 13 diverse QA benchmarks and 1,050 lexically perturbed queries per dataset, our top contextual QueryBandit (Thompson Sampling) achieves an 87.5% win rate over a no-rewrite baseline and also outperforms zero-shot static prompting ("paraphrase" or "expand") by 42.6% and 60.3% respectively. Therefore, we empirically substantiate the effectiveness of QueryBandits in mitigating hallucination via the intervention that takes the form of a query rewrite. Interestingly, certain static prompting strategies, which constitute a considerable number of current query rewriting literature, have a higher cumulative regret than the no-rewrite baseline, signifying that static rewrites can worsen hallucination. Moreover, we discover that the converged per-arm regression feature weight vectors substantiate that there is no single rewrite strategy optimal for all queries. In this context, guided rewriting via exploiting semantic features with QueryBandits can induce significant shifts in output behavior through forward-pass mechanisms, bypassing the need for retraining or gradient-based adaptation.

Checklist Confirmation: I confirm that I have included a paper checklist in the paper PDF.
Reviewer Nomination: [redacted]
Responsible Reviewing: We acknowledge the responsible reviewing obligations as authors.
Primary Area: Applications (e.g., vision, language, speech and audio, Creative AI)
LLM Usage: Prefer not to declare (you can then skip the rest)
Declaration: I confirm that the above information is accurate.
Submission Number: 24171

Paper Decision

Decision by Program Chairs 17 Sep 2025, 08:53 (modified: 18 Sep 2025, 10:34) Program Chairs, Authors
Revisions (/revisions?id=EXAQuP7k92)

Decision: Reject

Comment:

Summary of the paper: This paper claims that hallucinations in LLMs can be reduced by treating query rewriting as a contextual multi-armed bandit problem. It proposes “QueryBandits,” a framework that dynamically chooses, for each incoming question, one of five manually designed rewriting strategies (paraphrase, simplify, expand, etc.) based on a 17-dimensional linguistic feature vector. A composite reward signal—weighted 0.6/0.3/0.1 across an LLM-as-judge score, fuzzy string match, and BLEU-1—drives an online bandit learner (best variant: Thompson Sampling). Evaluated on 13 QA benchmarks with GPT-4o, the system achieves an 87.5 % win rate over no-rewrite and improves by 42.6–60.3 % over static prompting baselines, demonstrating that no single rewrite strategy is universally optimal and that context-aware selection is necessary.

Strengths of the paper:

1. Novel problem framing: first to cast hallucination mitigation via query rewriting as a contextual bandit task.
2. Empirical coverage: 13 QA datasets, multiple bandit algorithms, strong no-rewrite and static-prompt baselines are included.
3. Interesting findings: query-specific rewriting outperforms one-size-fits-all strategies.

Weaknesses of the paper:

1. Overly hand-engineered: only five fixed rewriting arms, binary feature indicators, and no mechanism to discover or adapt strategies automatically.
2. Incremental innovation: essentially prompt engineering wrapped in a bandit; close in spirit to RLHF but with simpler reward.
3. Evaluation shortcomings: Uses GPT-4o both as the generator and as the judge, risking self-bias; no cross-check with Claude, Gemini, or human labels beyond 100 samples. Weights (0.6/0.3/0.1) and LLM-as-judge reliability are insufficiently validated.
4. Baselines incomplete: omits recent hallucination-specific methods (ICD, TruthX, DoLa) and does not test with other LLMs.
5. Presentation issues: lacks clear problem motivation, preliminaries, and ablations on feature importance or reward components.

Reasons for the decision: After reviewing the rebuttal, it appears that the authors have failed to address the reviewers' most significant concerns. Following the AC-reviewer discussion, all three reviewers now lean toward rejection. Given these unresolved weaknesses and the clear consensus among the reviewers, the paper in its present form is not ready for acceptance.

Gentle Reminder: Please Reply to Authors' Responses (Only if Not Yet Done)

Official Comment by Area Chair R2t4 04 Aug 2025, 04:31

Program Chairs, Reviewer Pb6k, Reviewer WjwG, Reviewer gwXw, Reviewer ha9C, Reviewer o2NY, Senior Area Chairs, Area Chairs, Reviewers, Reviewers Submitted, Authors

Comment:

Dear Reviewers,

As the discussion deadline approaches, may we kindly ask you to review the authors' responses and post a constructive reply—unless you have already done so, in which case please kindly disregard this gentle reminder.

Your thoughtful engagement is deeply appreciated and essential to a fair and timely process. With sincere thanks for your continued dedication.

Area Chair

Official Review of Submission24171 by Reviewer ha9C

Official Review by Reviewer ha9C 13 Jul 2025, 12:06 (modified: 18 Sep 2025, 13:03)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Reviewer ha9C

Revisions (/revisions?id=mR3IPybpDl)

Summary:

This paper proposes QueryBandits, a contextual multi-armed bandit framework to proactively mitigate hallucinations in LLMs via query rewriting. It defines five discrete rewriting strategies (e.g., paraphrase, simplify, expand) and selects among them per-query using a 17-dimensional linguistic feature vector. A composite reward model combining LLM-judgment, fuzzy match, and BLEU-1 guides learning. Extensive evaluation across 13 QA benchmarks shows the approach significantly outperforms no-rewrite and static prompting baselines in reducing hallucinations.

Strengths And Weaknesses:

Strengths

1. Novel framing of query rewriting as a contextual bandit problem, this idea is quite interesting.
2. Well-motivated reward design validated through Pareto frontier analysis and human alignment.
3. Comprehensive empirical evaluation on diverse QA datasets, including strong baselines and multiple bandit variants.

Weakness

1. The paper should be improved for better reading and understanding. For example, it would be better to have some motivations and preliminaries for introducing the problem and proposed method.
2. While the bandit formulation is elegant, the use of five manually designed rewriting arms and binary feature

vectors may come across as overly hand-engineered. The method lacks flexibility to adapt or expand automatically. I am also wondering whether it can include some other LLMs as baselines?

Quality: 2: fair
Clarity: 2: fair
Significance: 2: fair
Originality: 3: good
Questions:
See weakness please.

Limitations:
See weakness please.

Rating: 3: Borderline reject: Technically solid paper where reasons to reject, e.g., limited evaluation, outweigh reasons to accept, e.g., good evaluation. Please use sparingly.
Confidence: 3: You are fairly confident in your assessment. It is possible that you did not understand some parts of the submission or that you are unfamiliar with some pieces of related work. Math/other details were not carefully checked.
Ethical Concerns: NO or VERY MINOR ethics concerns only
Paper Formatting Concerns:
No

Code Of Conduct Acknowledgement: Yes
Responsible Reviewing Acknowledgement: Yes
Final Justification:
I was interested in this paper's idea, which seems novel to me at least. However, after carefully reading other reviewers' comments, I agree with their concerns, and I encourage this paper to be modified better for next submission.

Rebuttal by Authors

Rebuttal
b Authors ([redacted])

29 Jul 2025, 18:21 (modified: 31 Jul 2025, 16:57)
Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors
Revisions (/revisions?id=WZ1tSIScvp)

Rebuttal:

Response to Reviewer ha9C

We thank the reviewer for their time considering our work.

Weaknesses:

- The paper should be improved for better reading and understanding. For example, it would be better to have some motivations and preliminaries for introducing the problem and proposed method.

Response to Weakness 1: We appreciate the comment regarding the motivation. In the Introduction, please consult the third paragraph of the intro, as well as **Contribution 4** on page 2 for a granular motivation, which is reproduced here for convenience:

The usage of reinforcement learning (RL) [100] methods have been applied in Natural Language Processing (NLP) tasks such as optimizing document-level query search [75], fine-tuning LLMs [25, 78], and post-training [73]. Despite its prevalent usage, to our knowledge, there is no in-depth interactive rewriting research to mitigate hallucination. We focus on bandit based methods because: (i) modeling the long-term value of hallucination manifestation would require multiple queries from a common sub-population; (ii) averaging hallucination propensity across distinct contexts may obscure per-query contextual idiosyncrasies; and (iii) the token concatenation that defines how vocabulary sampling occurs in output generation is deterministic, meaning it is unclear if an MDP transition model may even be defined. That is not to say bandit methods have no precedent in NLP. Proximal Policy Optimization [93] variants for LLMs such as GRPO (Group Relative Policy Optimization) [97] and ReMax [58] also remove the critic via grouped Monte Carlo or baseline-adjusted returns.

Contribution 4: Optimizing queries post-training by embedding them directly into the stage of prompting with minimal computational or token overhead constitutes an efficient strategy for trustworthy interfacing with LLMs, particularly under resource-constrained or latency-sensitive conditions. We bypass the need for retraining or gradient-based adaptation through purely forward-pass mechanisms. Moreover, through QueryBandits, we provide a mechanism to interpret the sensitivity of LLM performance to contextual rewrites. %Query optimization enhances both response accuracy and faithfulness, outperforming naive or zero-shot formulations without increasing inference cost.

> - While the bandit formulation is elegant, the use of five manually designed rewriting arms and binary feature vectors may come across as overly hand-engineered. The method lacks flexibility to adapt or expand automatically. I am also wondering whether it can include some other LLMs as baselines?

Response to Weakness 2: We agree that pre-defining five arms accordance in accordance with semantic best-practice may be restrictive; however, as these have a substantial history anchored in NLP:

- Paraphrase is well established as a foundational NLP task from datasets such as the Quora Question Pairs, to methods as Deng et al., 2023 "Rephrase and Respond", Witteveen et al., 2019 Paraphrase with LLMs, and traditional methods in Gao et al., 2021, SimCSE
- Simplification: Based on evalautions for long-context Lost-in-the-Middle, Liu et al., 2023, but also recommended for clear and concise prompt by indsutry including Anthropic's prompt engineering guidelines. Furthermore, literature supports concise, small contexts over larger ones Vodrahalli et al 2024, Michelangelo: Long Context Evals beyond Haystacks.
- Disambiguation: grounded in literature for word-sense disambiguation since the 1990s (Word-Sense Disambiguation, Yarowsky 1992), and from the 2007 SemEval Task tasks based on identifying semantic realtionships between words.
- Expansion: Based on recent literature for Gao et al., 2022 (HyDE), Lewis et al., 2020 (RAG), Wei et al., 2022 (Chain-of-Thought).
- Clarification: Inspired by work in tailoring models to domain-specific languages such as legal, finance that require niche knowledge, such as FinBERT (Araci 2019).

We believe that is sufficient for the definition of the action space for possible rewrites. That is not to say that more general notions of action space, where a map could be defined from a semantic embedding to a general space of rewriting strategies. However, it is not inherently obvious what the generic codomain of such a function would be, if it is not restricted to a collection of pre-defined strategies. This remains valid scope of future work.

Replying to Rebuttal by Authors

Official Comment by Reviewer ha9C

Official Comment by Reviewer ha9C 05 Aug 2025, 02:51

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors

Comment:
Thanks for your response, and I will keep my score.

Replying to Rebuttal by Authors

Mandatory Acknowledgement by Reviewer ha9C

Mandatory Acknowledgement by Reviewer ha9C 05 Aug 2025, 03:16

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors

Mandatory Acknowledgement: I have read the author rebuttal and considered all raised points., I have engaged in discussions and responded to authors., I have filled in the "Final Justification" text box and updated "Rating" accordingly (before Aug 13) that will become visible to authors once decisions are released., I understand that Area Chairs will be able to flag up Insufficient Reviews during the Reviewer-AC Discussions and shortly after to catch any irresponsible, insufficient or problematic behavior. Area Chairs will be also able to flag up during Metareview grossly irresponsible reviewers (including but not limited to possibly LLM-generated reviews)., I understand my Review and my conduct are subject to Responsible Reviewing initiative, including the desk rejection of my co-authored papers for grossly irresponsible behaviors. https://blog.neurips.cc/2025/05/02/responsible-reviewing-initiative-for-neurips-2025/ (https://blog.neurips.cc/2025/05/02/responsible-reviewing-initiative-for-neurips-2025/)

Official Review of Submission24171 by Reviewer WjwG

Official Review by Reviewer WjwG 05 Jul 2025, 05:01 (modified: 18 Sep 2025, 13:03)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Reviewer WjwG

Revisions (/revisions?id=L8Oqx0CFDi)

Summary:
This paper presents QueryBandits, a contextual multi-armed bandit framework that mitigates LLM hallucinations through intelligent query rewriting. The method uses five rewrite strategies as arms and 17-dimensional linguistic features as context to dynamically select optimal rewrites. Evaluated on 13 QA benchmarks with GPT-4o, the best contextual bandit (Thompson Sampling) achieves 87.5% win rate over no-rewrite baseline and outperforms static strategies by 42.6-60.3%. The key finding is that no single rewrite strategy works optimally for all queries, demonstrating the necessity of context-aware rewriting for hallucination mitigation.

Strengths And Weaknesses:
Strengths

1. The authors design a query rewriting framework that effectively reduces LLM hallucinations.
2. The authors conduct detailed research on 17 linguistic features that may affect LLM understanding and design 5 rewrite strategies, with query rewrite strategy selection based on these 17 linguistic features.
3. The authors design a reward function composed of three metrics, which is more comprehensive than single-metric

approaches for mitigating hallucinations in generated content.
4. The authors conduct evaluations on extensive datasets and achieve significant improvements over static prompting strategies and no-rewrite strategies.

Weaknesses

1. The approach leans more towards engineering methods with limited innovation. Essentially, it rewrites input queries. The bandit algorithm approach is similar to RLHF, except that bandit algorithms are more suitable for query rewrite tasks.
2. Relying solely on GPT-4o evaluator to assess $s_{llm}$is insufficiently accurate. Multiple models such as Claude and Gemini can be used for joint evaluation and voting. Moreover, validation on only 100 manually annotated samples is too small; the rationality of LLM-as-a-judge and the weights (0.6, 0.3, 0.1) should be validated on more samples.
3. The authors appear to use GPT-4o to generate responses and also use a GPT-4o-based evaluator to assess the s_{llm} metric, which may introduce biased evaluation. That is, GPT-4o evaluating GPT-4o-generated answers is more likely to consider them factually correct.
4. There is a lack of comparison with current hallucination mitigation methods, such as ICD [1], TruthX [2], DoLa [3], etc.

[1] Zhang Y, Cui L, Bi W, et al. Alleviating hallucinations of large language models through induced hallucinations[J]. arXiv preprint arXiv:2312.15710, 2023.

[2] Zhang S, Yu T, Feng Y. Truthx: Alleviating hallucinations by editing large language models in truthful space[J]. arXiv preprint arXiv:2402.17811, 2024.

[3] Chuang Y S, Xie Y, Luo H, et al. Dola: Decoding by contrasting layers improves factuality in large language models[J]. arXiv preprint arXiv:2309.03883, 2023.

Quality: 3: good
Clarity: 3: good
Significance: 3: good
Originality: 3: good
Questions:
See Above

Limitations:
Yes

Rating: 3: Borderline reject: Technically solid paper where reasons to reject, e.g., limited evaluation, outweigh reasons to accept, e.g., good evaluation. Please use sparingly.
Confidence: 4: You are confident in your assessment, but not absolutely certain. It is unlikely, but not impossible, that you did not understand some parts of the submission or that you are unfamiliar with some pieces of related work.
Ethical Concerns: NO or VERY MINOR ethics concerns only
Paper Formatting Concerns:
None

Code Of Conduct Acknowledgement: Yes
Responsible Reviewing Acknowledgement: Yes
Final Justification:
See Response to Authors' Rebuttal

Rebuttal by Authors

Rebuttal

b Authors ([redacted])

29 Jul 2025, 18:25 (modified: 31 Jul 2025, 16:57)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors

Revisions (/revisions?id=w5Ricc44BS)

Rebuttal:

Response to Reviewer WjwG

We thank the reviewer for their time considering our work.

Weaknesses:

> - The approach leans more towards engineering methods with limited innovation. Essentially, it rewrites input queries. The bandit algorithm approach is similar to RLHF, except that bandit algorithms are more suitable for query rewrite tasks.

Response to Weakness 1: RLHF exhibits several fundamental differences from our formulation:

1. it considers modelling the binary input responses of a human through a likelihood function whose parameters must be optimized as a function of which input is preferred. There is no explicit notion of preference likelihood in this work. Instead, we consider a reward model defined in the 2nd paragraph of the intro (see also eqn. (1)) whose weights are optimized in Figure 2.
2. It considers an MDP formulation of response generation, whereby the reward evaluation must be accumulated along a series of multiple input interactions. This can be sample inefficient in that it requires multiple input-response pairs per policy update, and it further hypothesizes an MDP transition model that relates responses to preferences, which may or may not be valid.
    - That is because in many cases, input-response pairs are stateless, and therefore may more naturally fit a contextual bandit framework.
3. Algorithms for contextual bandits do exhibit some similarities to methods for RL. In particular, posterior sampling for RL (PSRL) and Thompson sampling are identical, although PSRL is relatively less common for RLHF.

Russo, Daniel, and Benjamin Van Roy. "Learning to optimize via posterior sampling." Mathematics of Operations Research 39.4 (2014): 1221-1243.

PPO and its trust region variants do apply to bandits, where they are called "Follow the Regularized Leader (FTRL)". Please see Remark 2. The key difference is the number of reward function evaluations required per policy update, and statistical assumptions on the performance criteria (stochastic regret) are relaxed relative to the RL setting, i.e., we do not model any transition dynamics, through the conditional dependence assumption implicit in the definition of stochastic regret [cf. the second-to-last display expression in Section 3]

> - Relying solely on GPT-4o evaluator to assess is insufficiently accurate. Multiple models such as Claude and Gemini can be used for joint evaluation and voting. Moreover, validation on only 100 manually annotated samples is too small; the rationality of LLM-as-a-judge and the weights (0.6, 0.3, 0.1) should be validated on more samples.

Response to Weakness 2: We agree that using a single evaluator may be a limitation of our experimental validation, and are grateful to the reviewer for identifying this issue. We will expand our collection of experiments to consider additional LLM evaluations. It will incorporate this expanded evaluation into the final camera-ready version of the paper.

Please note the results are categorically similar to those reported in the original submission, in that they support the trends previously observed.

Regarding whether 100 samples is sufficient, one may observe that the convergence rate of the sample mean to the population mean, as per the law of large numbers dependence on input dimension d and sample size n, is $\sqrt{\frac{d}{n}}$ -- see, for instance

"Probability: Theory and Examples" by Rick Durrett (5th ed.)

This means that after 100 samples, the sample mean will be at $\pm.2$ of the population mean reward with 95% chance. To check whether this was enough, we reran the experiments with a larger sample size of 10^3 reward evaluations.

Here are the ROC-AUC at different sample sizes from our evaluation set (10 repeats at each estimate):

n	Mean ROC-AUC	StdErr
10	0.9728	0.000305
50	0.9818	0.000250
80	0.9743	0.000263
100	0.9718	0.000249
200	0.9648	0.000231
400	0.9548	0.000196
500	0.9523	0.000167
600	0.9529	0.000164
800	0.9524	0.000134
1000	0.9540	0.000137

> - The authors appear to use GPT-4o to generate responses and also use a GPT-4o-based evaluator to assess the metric, which may introduce biased evaluation. That is, GPT-4o evaluating GPT-4o-generated answers is more likely to consider them factually correct.

Response to Weakness 3: We agree that in principle, we should be using a more diverse set of LLMs for external supervision, as we mentioned in the previous response. We will expand our experiments in the final version.

> - There is a lack of comparison with current hallucination mitigation methods, such as ICD [1], TruthX [2], DoLa [3], etc.

Response to Weakness 4: We did not compare against general-purpose hallucination mitigation methods, as this is a broad field with a variety of approaches, both in terms of the LLM architectural representation, fine-tuning, decoding, and so on. Our approach is most comparable to those that investigate augmentations of the prompt layer only. For that reason, the aforementioned references were omitted. However, we have added them to the discussion of related works:

- *[1] Zhang Y, Cui L, Bi W, et al. Alleviating hallucinations of large language models through induced hallucinations[J]. arXiv preprint arXiv:2312.15710, 2023.* - post-generation detection method by contrasting the log-prob tokens of a 'evil' and 'frozen' llm for reducing hallucinations.

- *[2] Zhang S, Yu T, Feng Y. Truthx: Alleviating hallucinations by editing large language models in truthful space[J]. arXiv preprint arXiv:2402.17811, 2024.* - inference-time intervention method to activate the truthfulness of LLM by identifying and editing the features within LLM's internal representations that govern the truthfulness. It edits the internal representations rather than addres the input query.
- *[3] Chuang Y S, Xie Y, Luo H, et al. Dola: Decoding by contrasting layers improves factuality in large language models[J]. arXiv preprint arXiv:2309.03883, 2023.* - contrasting the differences in logits obtained from projecting the later layers versus earlier layers to the vocabulary space, exploiting the fact that factual knowledge in an LLMs has generally been shown to be localized to particular transformer layers. It addresses the output token space rather than the input query.

Replying to Rebuttal by Authors

Official Comment by Reviewer WjwG

Official Comment by Reviewer WjwG 05 Aug 2025, 22:42

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors

Comment:

I appreciate the author's response, but most of my concerns were not addressed. The author simply said that they would be included in the final version of the paper, but did not address my concerns. I cannot believe whether these concerns will really be resolved, so I decided to downgrade the score from 4 to 3.

Replying to Official Comment by Reviewer WjwG

Official Comment by Authors

Official Comment

b Authors ([redacted])

06 Aug 2025, 19:16

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors

Comment:

We would greatly appreciate it if you could clarify which specific aspects of your concerns you felt were not addressed by our rebuttal that constitutes a score downgrade from 4 to 3. We expanded the sample size and ROC-AUC analysis to 1,000 samples. Regarding related works, we discussed why prompt-based interventions are methodologically distinct from architectural or internal-editing hallucination mitigation strategies, and referenced the works you cited for completeness.

If there are particular deficiencies in our responses or if you would like us to address a concern differently, we would be grateful for specific guidance.

Replying to Official Comment by Authors

Official Comment by Reviewer WjwG

Official Comment by Reviewer WjwG 07 Aug 2025, 03:27

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors

Comment:

First, I think the method's innovation is somewhat lacking. Second, you haven't included many of the experiments I'm concerned about. You simply said they would be included in the final version of the paper, but I'm unconvinced. Finally, regarding the related literature I mentioned, such as ICD, DoLa, and TruthX, I believe you could compare them on the TruthfulQA dataset, as these papers, including yours, have conducted experiments on that dataset.

Replying to Official Comment by Reviewer WjwG

Novelty

Official Comment

b Authors ([redacted])

07 Aug 2025, 15:07

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors

Comment:

In response to claim on innovation, we believe that innovation in research should not be confused with opacity or the “dazzling” of an audience via obscure techniques or complex architectures for their own sake. Instead, we have deliberately prioritized clarity and extensibility in both our methodology and writing, precisely so that the research can be reproduced, understood, and built upon.

If there are particular aspects of our method or contributions that you find derivative or insufficiently distinct from prior art, we welcome specific feedback so we can address those concerns directly. In our view, the main contribution of QueryBandits is not merely a new algorithm, but a rigorous, extensible demonstration that adaptive, context-driven prompt rewriting implemented in a model-agnostic, plug-and-play fashion can yield state-of-the-art factuality even on the strongest available LLMs, and can robustly generalize to perturbed, out-of-distribution queries.

If the reviewer’s concern is that the paper lacks performative novelty, we would submit that reproducibility and clarity should be viewed as essential innovations in their own right.

Replying to Novelty

Comparisons to TruthX, DoLa

Official Comment

b Authors ([redacted])

07 Aug 2025, 15:27

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors

Comment:

Method / Model	Backbone	MC1 (%)	Source Type	Notes
QueryBandits	GPT-4o	**85.6**	Closed	Can work on open or closed models
GPT-4o	GPT-4o	81.4	Closed	OpenAI system card
GPT-4	GPT-4	81.3	Closed	OpenAI system card

GPT-4o mini	GPT-4o mini	66.5	Closed	OpenAI system card
GPT-3.5 Turbo	GPT-3.5 Turbo	53.6	Closed	OpenAI system card
Llama-2-7B-Chat (baseline)	Llama-2-7B-Chat	34.6	Open	See DoLa/TruthX papers
+ DoLa (Chuang et al. 2023)	Llama-2-7B-Chat	32.2	Open	Contrastive decoding; on a much weaker base model; not directly applicable to closed-source GPT
+ ICD (Zhang et al. 2023)	Llama-2-7B-Chat	46.3	Open	Induce-then-contrast decoding; open-source only
+ TruthX (Zhang et al. 2024)	Llama-2-7B-Chat	54.2	Open	Representation editing; on a much weaker base model; open-source only

As shown in the table above, these methods have demonstrated substantial improvements—but only on significantly weaker open-source backbones (e.g., Llama-2-7B-Chat). For example, the best reported MC1 for open models (TruthX) is 54.2%, whereas GPT-4o achieves 81.4%, and QueryBandits further raises this to 85.6%. It is important to note that DoLa and TruthX cannot be directly applied to closed-source models such as GPT-4o, and their gains on weaker models may not transfer in a strictly additive way due to diminishing returns at higher baselines. For example, Mistral-7b-Instruct-v0.2 without TruthX scores 52.26% vs 56.43% with TruthX.

It’s important to emphasize that QueryBandits is model-agnostic and can be applied to both open- and closed-source models, providing strong gains even when starting from a high-performing backbone. Furthermore, our approach demonstrates generalization: it outperforms both static and dynamic interventions not only on TruthfulQA but also on diverse QA datasets and lexically perturbed queries. This distinguishes our work from methods focused on post-hoc correction on standard benchmarks.

Official Review of Submission24171 by Reviewer Pb6k

Official Review by Reviewer Pb6k 03 Jul 2025, 02:38 (modified: 18 Sep 2025, 13:03)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Reviewer Pb6k

Revisions (/revisions?id=2AQYKIQUW4)

Summary:
This paper proposes QueryBandits, a contextual bandit framework that adaptively rewrites user queries to reduce LLM hallucination. By leveraging 17 linguistic features and optimizing a reward function based on correctness metrics, QueryBandits selects from five rewrite strategies. Evaluated on 13 QA benchmarks with perturbed queries, it outperforms static prompting and no-rewrite baselines.

Strengths And Weaknesses:
Strengths

1. The paper is well-written and clear.
2. The proposed method significantly outperforms static prompts and baselines across multiple metrics (win rate, adjusted reward, regret).

Weaknesses

1. The evaluation is conducted on semantically invariant but lexically perturbed versions of standard QA queries, which is not a realistic setting in real-world applications.
2. The paper does not directly compare hallucination reduction on unaltered benchmark inputs, making it hard to assess how the approach generalizes to practical deployment.
3. The construction process of the human-labeled validation set used for reward calibration is unclear (e.g., selection criteria, annotation protocol, inter-annotator agreement).
4. The proposed method is designed for QA tasks, which seems not generalizable to other tasks like summarization or open-ended generation.

Quality: 3: good
Clarity: 3: good
Significance: 3: good
Originality: 2: fair
Questions:
See weaknesses above.

Limitations:
yes

Rating: 3: Borderline reject: Technically solid paper where reasons to reject, e.g., limited evaluation, outweigh reasons to accept, e.g., good evaluation. Please use sparingly.
Confidence: 4: You are confident in your assessment, but not absolutely certain. It is unlikely, but not impossible, that you did not understand some parts of the submission or that you are unfamiliar with some pieces of related work.
Ethical Concerns: NO or VERY MINOR ethics concerns only
Paper Formatting Concerns:
NA

Code Of Conduct Acknowledgement: Yes
Responsible Reviewing Acknowledgement: Yes
Final Justification:
Providing more details on the human annotation protocol, and including inter-annotator agreement metrics would further strengthen the methodology. I also encourage a broader discussion on the potential generalization beyond QA tasks to better support future extensions of this work.

Rebuttal by Authors

Rebuttal
b Authors ([redacted])

29 Jul 2025, 18:26 (modified: 31 Jul 2025, 16:57)
Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors
Revisions (/revisions?id=KaogRHb2Gg)

Rebuttal:

Response to Reviewer Pb6k

We thank the reviewer for their time considering our work.

Weaknesses

- The evaluation is conducted on semantically invariant but lexically perturbed versions of standard QA queries, which is not a realistic setting in real-world applications.

Response to Weakness 1: We respectfully disagree that our framework is an unrealistic setting. On the contrary, we feel that is models real-world use more closely than evaluating on benchmark queries. In practice, users naturally paraphrase or reorder queries when seeking the same information. LLM Agents frequently involve query decomposition for RAG tasks, and the stochasticity in outputs mirrors our setup (Ma et al., 2023, "Query Rewriting for RAG Models").

Furthermore, we find that evaluating on benchmarks *as-is* **severly underestimates hallucination risk due to memorization**, as these queries are likely seen verbatim during pre-training. Prior literature has investigated this matter: Schwarzchild et al., 2024 ("Rethinking LLM Memorization"), Hartmann et al., 2023 ("SoK: Memorization in General-Purpose LLMs"), Nasr et al., 2023 ("Scalable Extraction of Training Data from LMs")

In our experiments with rewrite arm `No–Rewrite`, we saw our bandits often collapse as the best ranked arm by reward (Appendix Figure 7.a). `No–Rewrite` achieved rank 1 on 7/16 benchmarks, with 11/16 with rank 2 or higher. The dominance of `No–Rewrite` is strong evidence of memorization.

For alternate notions of perturbations, we can add discussion, as an example, around:

- Ebrahimi et al., 2017 ("HotFlip") - swaps one token for another, based on the gradients of the onehot input vectors.
- Jia & Liang 2017 ("Adversarial Examples for Evaluating Reading Comprehension Systems") - method tests whether systems can answer questions about paragraphs that contain adversarially inserted sentences, which are automatically generated to distract computer systems without changing the correct answer or misleading humans.
- Li et al., 2020 ("BERT-ATTACK") - consists of two steps: (1) finding the vulnerable words for the target model and then (2) replacing them with the semantically similar and grammatically correct words until a successful attack.

- The paper does not directly compare hallucination reduction on unaltered benchmark inputs, making it hard to assess how the approach generalizes to practical deployment.

Response to Weakness 2: We appreciate this concern. In fact, our experiments do include the original queries (Figure 7) and our bandits converge often to the `No–Rewrite` arm. LLMs tend to answer these unaltered queries correctly via memorization, which is not representative of an online deployment for unseen data. Therefore, the question of whether this paproach works for unaltered queries is indeed contained within our scope of possible interventions.

- The construction process of the human-labeled validation set used for reward calibration is unclear (e.g., selection criteria, annotation protocol, inter-annotator agreement).

Response to Weakness 3: We can expand our construction details in Section 3. Briefly:

- We assembled a held-out, manually labeled set of 100 query-answer pairs sampled from each benchmark to represent a diversity of query types and domains. For TruthfulQA (~800 samples) we bootstrap 100 from the main set. Given that our pareto analysis is a hyperparameter tuning, there is no concern for data leakage.
- Each item was annotated against the ground truth for factual correctness. We will provide Cohen kappa scores for inter-annotator agreement.
- The reward model (convex combination of LLM-judge, fuzzy-match, and BLEU) was then calibrated

against these binary labels. Our ROC-AUC simplex in Figure 2.a is the result of this process.

- The proposed method is designed for QA tasks, which seems not generalizable to other tasks like summarization or open-ended generation.

Response to Weakness 4: While our primary evaluation is QA, the issue of surface-form overfitting and brittleness to prompting is widely observed in summarization, dialogue, and other open-ended generation tasks (HaluEval (Li et al., 2023), Huang et al., 2023 "A Survey on Hallucination in Large Language Models"). The methodology of evaluating robustness over semantic equivalence classes rather than canonical queries can be applied to more domains, and would welcome future research in prompt tuning with bandits.

Replying to Rebuttal by Authors

Official Comment by Reviewer Pb6k

Official Comment by Reviewer Pb6k 06 Aug 2025, 15:04

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors

Comment:
Thank you for the clarifications. Providing more details on the human annotation protocol is helpful, and including inter-annotator agreement metrics would further strengthen the methodology. I also encourage a broader discussion on the potential generalization beyond QA tasks to better support future extensions of this work.

Replying to Rebuttal by Authors

Mandatory Acknowledgement by Reviewer Pb6k

Mandatory Acknowledgement by Reviewer Pb6k 07 Aug 2025, 14:21

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors

Mandatory Acknowledgement: I have read the author rebuttal and considered all raised points., I have engaged in discussions and responded to authors., I have filled in the "Final Justification" text box and updated "Rating" accordingly (before Aug 13) that will become visible to authors once decisions are released., I understand that Area Chairs will be able to flag up Insufficient Reviews during the Reviewer-AC Discussions and shortly after to catch any irresponsible, insufficient or problematic behavior. Area Chairs will be also able to flag up during Metareview grossly irresponsible reviewers (including but not limited to possibly LLM-generated reviews)., I understand my Review and my conduct are subject to Responsible Reviewing initiative, including the desk rejection of my co-authored papers for grossly irresponsible behaviors. https://blog.neurips.cc/2025/05/02/responsible-reviewing-initiative-for-neurips-2025/ (https://blog.neurips.cc/2025/05/02/responsible-reviewing-initiative-for-neurips-2025/)