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

Nicole Cho, William Watson, Alec Koppel, Sumitra Ganesh, Manuela Veloso
JP Morgan AI Research
New York, NY
nicole.cho@jpmorgan.com

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.

1 Introduction

As Large Language Models (LLMs) grow more powerful, their hallucinations increase in severity [76, 101]. Hallucinations refer to the generation of inaccurate outputs, according to the LLM's internal "understanding" of the query [41]. Mitigation efforts, however, remain largely confined to retrofitting the outputs instead of reshaping the query [41, 102], even though LLM outputs are highly variable to lexical perturbations of the incoming query [112, 16]. This style of analysis is in line with recent trends of mechanistic interpretability [3], where interpretable subgraphs of neural architectures are sought on the internal embeddings. However, these subgraphs may be unavailable due to enterprise restrictions [103]; moreover, there is no clear way to link a particular subgraph template with hallucination.

In this work, we design interventions to mitigate hallucinations based on semantic features of a query. To formalize the relationship between a query and hallucination, we associate with each query the output's propensity to be hallucinatory, which is implicitly associated with a reward model. We further hypothesize that this distribution is a function of a potential intervention which takes the form of a query rewrite. To be more precise, for $s_{\text{llm}} \in [0, 1]$, a binary score assigned by an LLM-based judge [66, 1], $s_{\text{fuzz}} \in [0, 1]$, a fuzzy string similarity metric [8], and $s_{\text{bleu}} \in [0, 1]$, the BLEU-1 score

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

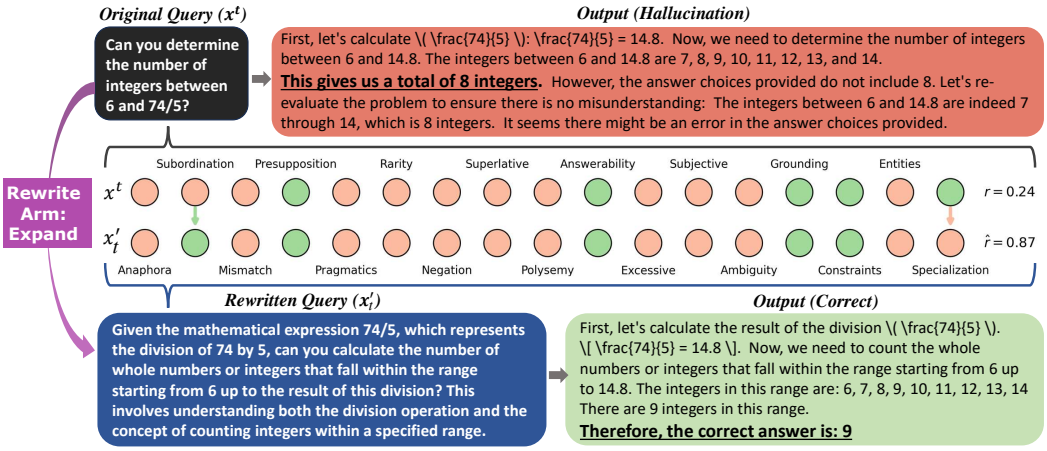


Figure 1: **QueryBandits and Its Success in Mitigating Hallucination.** The original query x_t induces a hallucinatory output: the LLM calculates 8 integers between 6 and 74/5. QueryBandits, by leveraging the feature vector, selects the EXPAND rewrite strategy. The rewritten query x'_t generates an accurate output of 9 integers. Noticeably, the feature vectors are different in the rewrite x'_t - *subordination (more complex clauses)* is now present while *specialization (query requiring domain-specific knowledge for understanding)* is absent - signifying effects of the EXPAND strategy.

capturing unigram lexical overlap [79, 61, 13], we define hallucination in terms of a reward model $r_t = \alpha \cdot s_{\text{llm}} + \beta \cdot s_{\text{fuzz}} + \gamma \cdot s_{\text{bleu}}$ where hallucinatory responses are those associated with low rewards. Through our ablations, we also discover the Pareto-optimal balance of weights (Fig. 2a), by assigning a higher weight on LLM-as-a-judge, which is also evidenced in studies that highlight the efficacy of LLMs in Natural Language Generation (NLG) evaluation tasks [109, 121, 30]. As we are agnostic to the stationarity of the reward distribution or lack thereof due to the extreme dimensionality of the output space [86, 40], we propose to evaluate whether rewrite strategies exhibit advantages when seeking to optimize the average reward or its worst case.

The usage of reinforcement learning (RL) [99] 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 [92] variants for LLMs such as GRPO (Group Relative Policy Optimization) [96] and ReMax [58] also remove the critic via grouped Monte Carlo or baseline-adjusted returns.

We have meticulously selected five distinct rewriting arms/strategies; to our knowledge, current research tends to pursue a "one-size-for-all" approach, leveraging one of these rewriting strategies for *all* query types, and does not pursue guided rewrites via bandits [68, 112]. We also have rigorously selected 17 different linguistic features that are known to hinder/enhance human or LLM understanding (Table 1). We frame query rewriting as an online decision problem and by leveraging per-query features, our bandit framework, **QueryBandits**, allocates exploration where uncertainty is high and exploitation where features are meaningful—resulting in hallucination reduction.

Contribution 1: We introduce an empirically validated reward function, combining an LLM-judge, fuzzy-match, and BLEU metrics (with $\alpha, \beta, \gamma = (0.6, 0.3, 0.1)$) chosen inside the 1% Pareto-optimal frontier on a held-out human-labeled set, to drive context-aware bandit learning (Fig. 2a). Guided by this reward signal, our contextual QueryBandits learn to tailor each query rewrite to its linguistic/contextual fingerprint. Our best performing contextual bandit, QueryBandits-Thompson Sampling, drives a 87.5% win-rate boost over the NO-REWRITE baseline—a considerable margin that highlights the efficacy of rewriting in reducing hallucination.

Contribution 2: QueryBandits-Thompson Sampling delivers a decisive 42% gain against the predominant static prompting strategy (PARAPHRASE), underscoring that feature-aware rewriting with bandits is effective for mitigating hallucination. In Figure 4, it is clear that the contextual QueryBandits hone into the optimal rewrite quickly, accruing an order-of-magnitude less cumulative regret than static prompting, vanilla (non-contextual) bandits, or no-rewriting. These gains confirm that a feature-aware, online adaptation mechanism can consistently outpace one-shot heuristics in mitigating hallucinations. An interesting finding is that some static prompting methods have a higher cumulative regret than NO-REWRITE, demonstrating that zero-shot prompting can cause more severe hallucinations.

Contribution 3: We provide empirical evidence that there is no *single* rewrite strategy that maximizes the reward for all types of queries. Our analysis of the per-arm regression weights (Figure 6) reveals how each arm's effectiveness hinges on the semantic features of a query. For example, if a query displays the feature *(Domain) Specialization*, meaning that the query can only be understood with domain-specific knowledge, the rewrite arm EXPAND is very effective in contrast to SIMPLIFY (Figure 1). Crucially, ablating the 17-dimensional feature input causes QueryBandits-Thompson Sampling's performance to drop to just 81.7% win rate and 754.66 exploration-adjusted reward. This performance gap confirms that the linguistic features carry an associative signal about optimal rewrite strategy. Finally, we observe that across datasets, higher feature variance coincides with greater variance in arm selection (Figure 5), resulting in genuinely diverse arm choices (Figure 2b).

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.

Interesting Findings: On many standard benchmark datasets, we discover that linear contextual bandits converge almost exclusively to the NO REWRITE arm (Figure 8), empirically exposing LLM's tendency for query memorization on benchmarks. Only when we introduce semantically invariant but lexically perturbed queries does the policy meaningfully diversify across rewrite strategies; a meaningful insight for the research community that surface-form novelty is essential in training query rewriting algorithms. On the non-contextual bandits, we empirically discover that they converge to a single rewrite strategy within a dataset, in contrast to contextual bandits that tend to diversify its choices conditioned on the presence/absence of linguistic features.

2 Related Work

LLM Hallucinations are known to erode trustworthiness from a societal perspective [24]. Recently, certain conceptual analyses frame it as a new epistemic failure mode, requiring dedicated mitigation agendas [118, 78]. Moreover, the release of more advanced reasoning models are concerningly generating more hallucinations, as reported in OpenAI's technical report [76] on o3 and o4-mini. The Times [101] recently reported real-world case studies on how fabricated LLM outputs are prompting legal accountability. Especially, the advent of more LLM-Agent enabled systems [113, 111] will engender the compounding cost of hallucinations.

Thus, mitigation is regarded as indispensable for faithful LLM interactions [41, 102, 38] and research is expanding from post-hoc detection [69], to preemptive grounding. Parallel to human-in-the-loop RLHF, RL from AI Feedback (RLAIF) trains a reward model on preferences generated by an LLM—bypassing expensive human labels—while exceeding RLHF quality on summarization and dialogue tasks [53, 78, 18]. Watson et al. [112] introduced pre-generation hallucination estimation via query perturbations. Ma et al. [68] has proposed the *Rewrite-Retrieve-Read* framework for Retrieval Augmented Generation pipelines while human-designed query rules has been heavily used for rewriting [64, 70, 15]. A common theme is that either raw prompting or manual heuristics is used - not guided rewrites, through contextual signals of the original query.

Blevins et al. [11] has conducted extensive research on the strong performance of Pretrained Language Models (PLMs) to retrieve linguistic features of a query in a few-shot manner. We employ this research and leverage an LLM to identify key linguistic features as outlined in Table 1. For the selection of

Table 1: Binary linguistic feature vector $\mathbf{f} \in \{0,1\}^{17}$ identified as challenging from a linguistics and LLM perspective. Features are grouped by type and grounded in prior work. For more specific examples, see Appendix Table 5.

Feature	Description	Citation
Structural Features		
Anaphora	Contains anaphoric references (e.g., *it*, *this*)	[93, 14]
Subordination	Contains multiple subordinate clauses	[39, 2, 11]
Scenario-Based Features		
Mismatch	Query (e.g. open-ended) does not match task (e.g. retrieval)	[31, 46]
Presupposition	Assumptions within the query are implicitly regarded as truthful	[47, 55, 106]
Pragmatics	Queries with discourse-driven intent (i.e. "can you pass me the salt")	[97, 55, 89]
Lexical Features		
Rarity	Presence of rare words with poor representation	[90, 49]
Negation	Presence of negation (e.g., *not*, *never*)	[37, 48, 104]
Superlative	Usage of superlative forms (e.g., *best*, *largest*) with implicit semantics	[81, 29]
Polysemy	Presence of words that have multiple, related-meanings	[5, 34]
Stylistic Complexity		
Answerability	Query is not highly speculative, sarcastic or rhetorical	[82, 9, 56]
Excessive	Overloaded with a large amount of details and information	[57, 65]
Subjectivity	Query requires LLM to reflect creatively and engender a personal opinion	[28, 67]
Ambiguity	Presence of ambiguous phrasing that opens multiple interpretations	[12, 50, 63]
Semantic Grounding		
Grounding	Presence of clear intention and goal	[22, 114]
Constraints	Presence of temporal/spatial/task-specific constraints	[42, 56]
Entities	Presence of verifiable entities	[54, 6, 110]
Specialization	Query requires domain-specific knowledge for understanding	[113, 17, 123]

Table 2: Overview of datasets, including domain, license, number of examples, associated scenarios, etc. These datasets span a diverse range of question types, domains, and reasoning skills, supporting robust evaluation. E = Extractive, M = Multiple Choice, A = Abstractive.

Dataset	Scenario	Domain	License	Count	Citation
SQuADv2	E, A	Wikipedia	CC BY-SA 4.0	86K	[84, 85]
TruthfulQA	M, A	General Knowledge	Apache-2.0	807	[62]
SciQ	M, A	Science	CC BY-NC 3.0	13K	[43]
MMLU	M	Various	MIT	15K	[36]
PIQA	M	Physical Commonsense	AFL-3.0	17K	[10]
BoolQ	M	Yes/No Questions	CC BY-SA 3.0	13K	[19, 108]
OpenBookQA	M	Science Reasoning	Apache-2.0	6K	[72]
MathQA	M	Mathematics	Apache-2.0	8K	[4]
ARC-Easy	M	Science	CC BY-SA 4.0	5K	[21]
ARC-Challenge	M	Science	CC BY-SA 4.0	2.6K	[21]
WikiQA	A	Wikipedia QA	Other	1.5K	[116]
HotpotQA	A	Multi-hop Reasoning	CC BY-SA 4.0	72K	[117]
TriviaQA	A	Trivia	Apache-2.0	88K	[44]

these features, we have thoroughly reviewed not only existing LLM literature but also traditional linguistics to identify features that are known to impact both human and LLM understanding.

3 Methodology and Evaluation Metrics

In this section, we define the key ingredients to formulate a sequential decision-making problem as a multi-armed bandit [52]. Specifically, we define the action space, contextual attributes, and reward.

Bandit Formulation. In the contextual bandit framework [52], a learner is faced with, given a context vector $x_t \in \mathcal{X} \subset \mathbb{R}^d$ at time t, selecting an arm a_t from an action set $\mathcal{A}$. Upon that basis, Nature reveals a reward $r_t(x_t, a_t)$ which is a function of the context and arm. To be more precise, the reward is defined as $r : \mathcal{X} \times \mathcal{A} \to \mathbb{R}$. The goal of a bandit algorithm is to select arms that are eventually good with respect to the average (or cumulative) reward. In the stochastic bandit setting, specifically, one is interested in selecting arms according to a *policy* $\pi : \mathcal{X} \to \rho(\mathcal{A})$ which performs as well as $\max_{\pi \in \Pi} \mathbb{E}[r(x, \pi(x))]$. Here $\rho(\mathcal{A})$ denotes the probability simplex over K arms, and Π is some class of policies. Next we specify the concrete choices of action space, context variables, and rewards for the query rewrite setting.

Action Space. Let $\mathcal{A} = \{a_0, a_1, a_2, a_3, a_4\}$ denote the set of rewriting strategies (arms), where each arm represents a distinct style of query reformulation implemented via prompt-based instructions to an LLM. As outlined in §2, previous research tends to take a "one-size-for-all" approach,

- a_0: **Paraphrasing** - The incoming query is rewritten using a prompt such as "Paraphrase this question while preserving its meaning." This introduces lexical diversity while maintaining semantic similarity, testing whether alternative phrasings reduce hallucination. Many have explored how paraphrasing can improve factual consistency in LLMs [16, 27, 115] .
- a_1: **Simplification** - The incoming query is rewritten to eliminate nested clauses or complex syntax. This targets hallucinations caused by long-range dependencies or overloaded details - and borrows ideas from educational psychology where simpler, granular, prompts enable a child to learn a new skill [59]. Recently, Van et al. [105], Zhou et al. [124] report on how simplified prompts decrease off-topic generations and enable complex reasoning tasks.
- a_2: **Disambiguation** - The query is rewritten by disambiguating vague references (e.g., ambiguous pronouns or temporal expressions). Many have conducted extensive studies on LLMs' inabilities to resolve ambiguous queries which leads to subpar performance [26, 94, 23].
- a_3: **Expansion** - The query is rewritten to explicitly expand on relevant named entities or attributes, elaborating on contextual cues through generation [120]. Since transformers-based LLMs optimize next-token likelihood over attention-mediated context windows [107], appending fine-grained query constraints effectively conditions the model on a richer semantic prefix.
- a_4: **Clarification of Certain Terms** - The query is rewritten to clarify the lexical and semantic meaning of jargons. Since LLMs are pre-trained on broad domain corpora using maximum likelihood next token prediction, rare domain-specific jargons [20] suffer from sparse exposure and less-calibrated embeddings [87, 80].

Contextual Attributes. For each incoming query, we extract a 17-dimensional binary feature vector $\mathbf{f} \in \{0, 1\}^{17}$ that captures key linguistic properties as outlined in Table 1, grounded by related works in NLP (§2). We have rigorously selected features that are known to impact both human linguistic understanding and LLM performance.

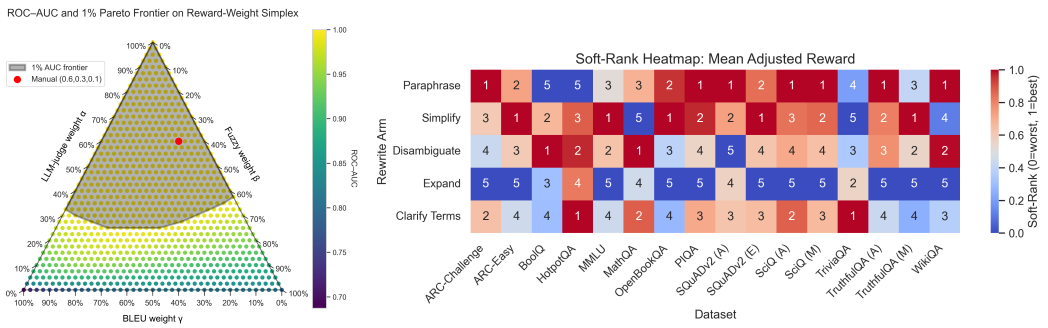
Reward Model. We model each rewritten query's reward $r_t \in [0, 1]$ as a convex combination of three complementary correctness signals:

$$r_t \;=\; \alpha\, s_{\mathrm{llm}} \;+\; \beta\, s_{\mathrm{fuzz}} \;+\; \gamma\, s_{\mathrm{bleu}}, \qquad \alpha + \beta + \gamma = 1, \quad \alpha, \beta, \gamma \geq 0 \tag{1}$$

- $s_{\mathrm{llm}} \in \{0, 1\}$ is a binary consistency judgment by a GPT-4o-based assessor, calibrated on factual correctness between system and reference answers [66, 1].
- $s_{\mathrm{fuzz}} \in [0, 1]$ is token-set similarity from RapidFuzz [8], capturing soft string overlap.
- $s_{\mathrm{bleu}} \in [0, 1]$ is the BLEU-1 score (unigram precision) under a unit-cap [79, 61, 13], ensuring lexical fidelity.

This multi-faceted formulation mitigates individual failure modes inherent in any single metric (e.g. BLEU's paraphrase blindness or edit-distance oversensitivity) while remaining bounded for stable learning. Following Wang et al. [109], we leverage the strength of LLMs-as-judges, and, as demonstrated by Test-Time RL [125], even noisy, self-supervised signals (e.g. pseudo-labels from majority-voted LLM outputs) can effectively guide policy updates. However, to validate that our convex proxy aligns with human judgment, we assembled a held-out, manually labeled set of 100 query–answer pairs and measured ROC–AUC of r_t against binary human labels (Figure 2a). See Alg. 1 in Appendix A.2 for further detail.

Reward-Weight Simplex Analysis. We swept $(\alpha', \beta', \gamma')$ over a triangular grid ($\alpha' + \beta' + \gamma' = 1$) and computed ROC–AUC on the human-labeled validation set. Figure 2a plots each grid point's AUC and overlays the 1% Pareto frontier (dark region). Our manual weights $(\alpha, \beta, \gamma) = (0.6, 0.3, 0.1)$ lie well inside this frontier, demonstrating robustness. The Pareto frontier reveals the following:



(a) ROC–AUC Pareto frontier on the reward-weight simplex.

(b) Mean-reward ranks (1 = best) of each rewrite arm per dataset under our contextual bandit; color intensity indicates closeness to the top rank.

Figure 2: (a) Our chosen (α, β, γ) lies deep in the 1% optimal frontier. (b) Breakdown of per-dataset arm performance: different datasets consistently favor different rewrite strategies

- **LLM-Judge Robustness (α):** The ROC–AUC surface is nearly invariant when α varies by ± 0.2: AUC shifts by <0.5%, indicating our formulation tolerates large LLM-judge weight swings.
- **Fuzzy-Match Sensitivity (β):** Small increases in β rapidly exit the Pareto region, showing that the fuzzy-match term must be tuned carefully to avoid degrading overall accuracy.
- **BLEU-Only Pitfall (γ):** As γ increases, ROC–AUC steadily declines, bottoming out at $\gamma = 1$ (pure-BLEU), where the model over-emphasizes surface overlap at the expense of true correctness.
- **Pareto-Optimal Region:** Our chosen $(0.6, 0.3, 0.1)$ sits deep in the high-AUC plateau, confirming it is a Pareto-optimal trade-off among semantic, fuzzy, and lexical signals.

Together, these experiments on manually labeled data substantiate our reward design: the LLM-judge component provides a forgiving anchor, fuzzy-match demands precise calibration, and BLEU contributes complementary lexical oversight.

Remark 1 *(On RL Methods vs. Bandits)* Within LLMs, for each input query, the transformer attends over the fixed context window and computes a softmax over the vocabulary to maximize token likelihood [83]. Consequently, we believe that hallucinations occur at the moment of generation for that single query, making hallucination a **per-query** phenomenon. Indeed, recent PPO variants for LLMs—GRPO [96] and ReMax [58]—remove the critic via grouped Monte Carlo or baseline-adjusted returns, highlighting critic-free policies that our bandit formulations naturally generalize. Therefore, a full-episodic RL problem, which must solve a Markov decision process with long-horizon credit assignment and nonstationary transition dynamics [99], can be practically suboptimal. Moreover, many of these methods rely on estimating a fixed average reward or state-action value *Q(s,a)* which can obscure per-query idiosyncrasies; if the optimal rewrite arm varies sharply with linguistic context, a mere empirical average will yield suboptimal policies.

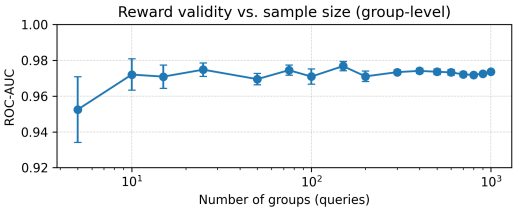
Remark 2 *(Linkage between Algorithm Choices and RL Methods)* Furthermore, it should be underscored that several algorithms whose usage we investigate here have analogues in RL: posterior sampling (PSRL) [77] as an analogue for Thompson sampling [100]; follow-the-regularized leader (FTRL) and its variants [95], originate from proximal gradient algorithms [88] whose usage in RL as proximal policy optimization (PPO) [91] is well-established. Other PPO-style advances like DAPO [119] improve exploration-exploitation via dynamic sampling and reward filtering, and VAPO [122] demonstrates stable Long-CoT training with an explicit value model—showing the spectrum from model-based to model-free approaches that contextual bandits sit within.

Choice of Algorithms. For **linear contextual bandits** we fit a per-arm linear model $x_t^\top \theta_k$ and choose either a UCB (LinUCB [51] / KL-UCB [32]), an FTRL regularized weight [71], or a posterior draw (Thompson sampling [100]). For **adversarial bandits** we consider two parameter-free adversarial methods—EXP3 [7] and FTPL [45, 98]. For full scoring, update equations and regret bounds, see Appendix A.2 and Algorithm 1.

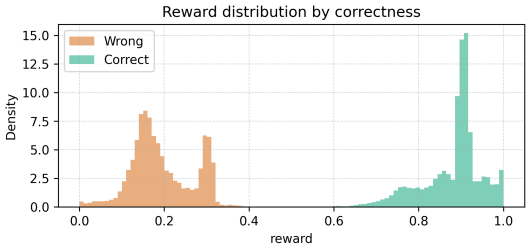
Evaluation Metrics. We assess each rewrite policy using three complementary metrics that capture both its ability to explore promising rewrites and its final accuracy relative to a no-rewrite baseline.

# Groups	Mean ROC–AUC	95% CI
5	0.9524	[0.9165, 0.9884]
10	0.9720	[0.9549, 0.9891]
15	0.9709	[0.9581, 0.9836]
25	0.9747	[0.9674, 0.9821]
50	0.9695	[0.9633, 0.9756]
75	0.9745	[0.9688, 0.9801]
100	0.9709	[0.9626, 0.9792]
150	**0.9767**	**[0.9716, 0.9819]**
200	0.9710	[0.9653, 0.9767]
300	0.9734	[0.9709, 0.9758]
400	0.9741	[0.9713, 0.9769]
500	0.9736	[0.9703, 0.9769]
600	0.9732	[0.9701, 0.9763]
700	0.9721	[0.9695, 0.9748]
800	0.9719	[0.9699, 0.9738]
900	0.9725	[0.9716, 0.9734]
1000	0.9737	[0.9721, 0.9753]
Macro-avg	**0.9729**	–

(a) Validity of the exploration-adjusted reward r_{adj} as a correctness proxy. Mean ROC–AUC and 95% Confidence Intervals (± 1.96 SE); 10 resamples per n. By $\sim$150 groups, the CI lower bound exceeds 0.97.



(b) Mean ROC-AUC vs. sample size n, with 95% CIs.



(c) Distribution of r_t for correct vs. wrong (normalized density). Our reward presents a clear separation between our human validated labels.

Figure 3: Summary of reward validity. *Left:* (a) numerical ROC–AUC and CIs across sample sizes. *Right:* (b) power curve; (c) class-conditional reward histogram of r_t vs. human labels.

All three metrics together give a balanced view of (1) how well a policy explores and exploits, (2) how quickly it converges to good answers, and (3) how often it beats the baseline in reward.

Metric 1: Exploration-Adjusted Reward. Let $r_t \in [0, 1]$ be the reward at pull t and let $H_t \in [0, 1]$ be the normalized Shannon entropy of the policy's strategy-selection history up to t. We define the *final exploration-adjusted reward* as

$$R_{\text{adj}} = \sum_{t=1}^{T} \big(r_t + \lambda\, H_t\big),$$

where $\lambda = 0.1$ weights the bonus for exploration and T is the trajectory length. This metric rewards policies that both achieve high per-pull rewards and maintain sufficient exploration.

Metric 2: Mean Cumulative Regret. At each pull we compute instantaneous regret as the gap between the oracle reward (the best achievable rewrite) and the observed reward. Summing these gives the cumulative regret per run, and we report its average over all runs. Where r_t^* is the maximal reward at pull t:

$$\overline{\text{Regret}} = \frac{1}{N} \sum_{\text{runs}} \sum_{t=1}^{T} \big(r_t^* - r_t\big),$$

Metric 3: Win Rate vs. Baseline. To measure final correctness head-to-head, we treat each test pull t as an independent trial and compute the fraction of trials for which a policy's reward r_t^{policy} strictly exceeds the no-rewrite baseline's reward r_t^{base}. This directly quantifies how often each rewrite outperforms doing nothing. For $N = 100$ test queries, the win rate is

$$\text{WinRate} = \frac{1}{N} \sum_{t=1}^{N} \mathbf{1}\big[r_t^{\text{policy}} > r_t^{\text{base}}\big] \times 100\%$$

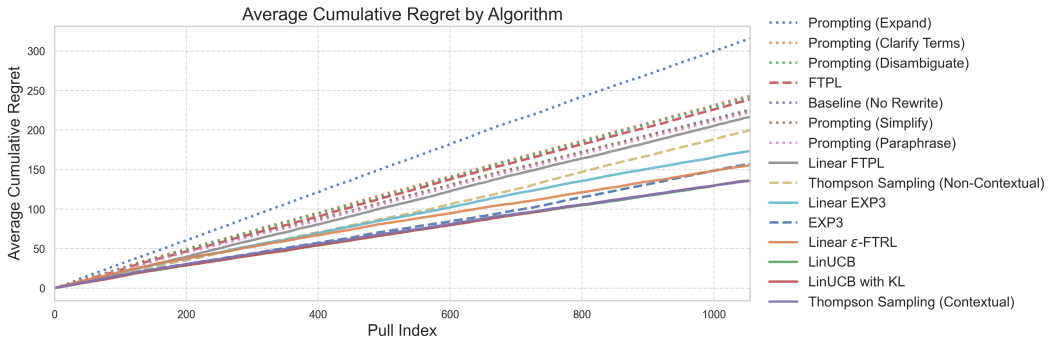


Figure 4: Cumulative reward *averaged across all datasets* over the no-rewrite baseline for each algorithm (sorted by final performance), highlighting the superior gains achieved by contextual bandits compared to non-contextual learners and static prompt-based rewrites.

4 Experiments

Pipeline. For each decision round t:

$$x_t \xrightarrow[\mathbf{f}_t \in \{0,1\}^d]{\text{Extr. feat.}} \mathbf{f}_t \xrightarrow[\text{(rewrite strat.)}]{\text{Select } a_t} x'_t = g_{a_t}(x_t) \xrightarrow{\text{LLM}} y_t \xrightarrow[r_t \in [0,1]]{\text{Eval.}} r_t \xrightarrow{\text{Update Bandit}} \hookleftarrow$$

1. **Feature Extraction.** For query x_t, compute d-dimensional linguistic feature vector $\mathbf{f}_t \in \{0,1\}^d$.
2. **Arm Selection.** The bandit receives $\mathbf{f}_t$ and selects a rewrite arm $a_t \in \{1, \ldots, K\}$.
3. **Query Rewriting.** Apply the selected arm to obtain the candidate query $x'_t = g_{a_t}(x_t)$.
4. **LLM Inference.** Issue x'_t to `gpt-4o-2024-08-06`, producing response y_t.
5. **Reward Evaluation.** Compute scalar reward $r_t \in [0, 1]$ via our reward formulation.
6. **Bandit Update.** Update the internal state of the bandit based on (a_t, r_t).

Dataset and Query Construction. We evaluate on $D = 13$ diverse QA benchmarks (see Table 2). For each dataset, we sample $|\mathcal{Q}|$ queries satisfying: (1) Original Answerability: the query in the dataset (q) is answered correctly by `gpt-4o-2024-08-06`, and (2) Perturbation Validity: of its five lexically perturbed but semantically invariant versions of the dataset's query, measured by numerous metrics such as LLM-as-judge and n-gram based metrics [60, 79, 109, 30], between one and three perturbations yield incorrect answers. Then, we randomly choose x_t in $|\mathcal{Q}|$ to train QueryBandits.

The importance of this query construction process deserves emphasis. Through our investigations, we have discovered that the ubiquity of benchmark datasets in Table 2 within pre-training and fine-tuning regimes has engendered a potentially pernicious form of prompt memorization. When we deploy our linear contextual bandits, the policy converges almost exclusively to no-rewriting, effectively demonstrating that the LLM has most likely memorized these exact phrasings rather than learning to exploit deeper linguistic structure (Figure 8). To evaluate genuine rewrite efficacy—and to prevent our results from collapsing into a trivial memorization baseline—we therefore choose a perturbed version of the dataset's query to construct the **incoming query** for our bandits. This experimental setup thus prioritizes query-rewriting strategies that are non-degenerate.

Experimental Configuration. We compare three non-contextual and six linear contextual bandits against zero-shot prompting and a no-rewrite baseline. All reported metrics are averaged over all dataset runs per algorithm. We compare M bandit algorithms and prompting strategies over $K = 5$ rewrite arms. Each algorithm runs for $T = |\mathcal{Q}_D|$ rounds on each of the D datasets (Table 2). Thus, Total Pulls $= M \times D \times |\mathcal{Q}_D| = 253,440$, with $|\mathcal{Q}_D| \approx 1050$, $M = 15$, and $D = 16$. We bootstrap samples with replacement for TruthfulQA to obtain approximately 1050 queries. Hyperparameters (learning rates, exploration coefficients, regularization constants) are tuned via grid search on a held-out validation set.

5 Results

Hypothesis 1: Can QueryBandits reduce hallucination? Table 3 and Figure 4 compares our QueryBandit algorithms against the baseline and five static prompting scenarios on 13 QA benchmarks

Table 3: Rewrite-policy performance: final cumulative exploration-adjusted reward, mean cumulative regret, and win rate vs. no-rewrite baseline.

Algorithm	Contextual?	Adj. Reward ↑	Cum. Regret ↓	Win Rate ↑
Thompson Sampling (Contextual)	✓	819.04	135.84	87.5%
LinUCB with KL	✓	818.79	136.00	87.0%
LinUCB	✓	818.60	136.12	86.9%
Linear ϵ-FTRL	✓	799.57	155.30	85.0%
EXP3 (Non-Contextual)	✗	797.47	157.31	86.5%
Linear EXP3	✓	781.05	173.60	83.8%
Thompson Sampling (Non-Contextual)	✗	754.66	200.18	81.7%
Linear FTPL	✓	738.07	216.54	76.3%
FTPL (Non-Contextual)	✗	716.05	238.85	62.8%
Prompting (Paraphrase)	–	732.39	222.56	44.9%
Prompting (Simplify)	–	730.13	224.42	50.1%
Prompting (Disambiguate)	–	713.65	241.25	42.4%
Prompting (Clarify Terms)	–	711.65	243.35	38.2%
Prompting (Expand)	–	639.25	315.71	27.2%
Baseline (No Rewrite)	–	729.20	225.85	–

(1,050 queries per dataset). Our top contextual learner—Thompson Sampling with the 17-dimensional feature vector—achieves an **87.5% win rate** and 819.04 exploration-adjusted reward, compared to the BASELINE: NO REWRITE, signifying that contextual query rewriting can reduce hallucination. As a side note: we aggregate results across datasets rather than report per-dataset Monte Carlo trials, as within-dataset permutation yielded trivial randomization.

Hypothesis 2: Can QueryBandits outperform prompting? Our best performing bandit, Thompson Sampling, exceeds the performance of static prompting for PROMPTING (PARAPHRASE) (44.9% / 732.39) and PROMPTING (EXPAND) (27.2% / 639.25), as seen in Table 3 and Figure 4. These gains confirm that raw static prompting cannot approach the effectiveness of a learner that adapts its rewrite choice to each query's linguistic fingerprint. By framing rewrite selection as an online decision problem and leveraging per-query context, QueryBandits allocate exploration where uncertainty is high and exploitation where features reliably predict hallucination risk—resulting in up to double the hallucination reduction of any static strategy, with no additional model fine-tuning.

Hypothesis 3: Do linear contextual bandits outperform those which are oblivious to context? Crucially, ablating the 17-dimensional feature input causes Thompson Sampling's performance to drop to just 81.7% win rate and 754.66 exploration-adjusted reward (–5.8 pts, –64.38 reward points), despite identical hyperparameters and run count. Since our reward directly measures output correctness, this performance gap confirms that the linguistic features carry associative signal about hallucination risk. Moreover, the majority of our linear QueryBandits outperform those which are oblivious to context (Fig. 4), signifying the importance of taking context into account in rewrites.

Hypothesis 4: Is there an association between query features and reward? Each rewrite arm seems to exhibit different sensitivities toward different *linguistic features*. Figures 5 & 6 plot each arm's average variance and learned regression weights θ across 17 binary linguistic features. Interestingly, the same linguistic feature can flip from strongly sensitive for one arm to insensitive for another—for example, the feature *(Domain) Specialization* is quite sensitive to the arm/strategy EXPAND but relatively much less to SIMPLIFY. A plausible explanation might be that domain specific queries inherently require specialized context outside the LLM's broad-domain priors, so EXPAND—by injecting explicit entity attributes or ontological qualifiers—reinforces the model's semantic grounding, whereas SIMPLIFY risks excising those critical signals. While our association matrix is interesting, each learned coefficient does not strictly prove a causal mechanism.

Hypothesis 5: Is there a *single* rewrite strategy that maximizes reward for all types of queries? Our analysis of the per-arm regression weights (Figure 6) reveals that no single rewrite strategy dominates across all linguistic profiles. Instead, each arm's effectiveness hinges on a distinct "feature footprint". For example, SIMPLIFY thrives when pragmatic cues (e.g. discourse markers, politeness markers) are present—these guide safe syntactic pruning—but falters on superlative constructions, whose removal strips away essential comparative meaning. For our interpretation of these sharp inversions feature–arm interactions, refer to Appendix Table 4. Therefore, our results demonstrate

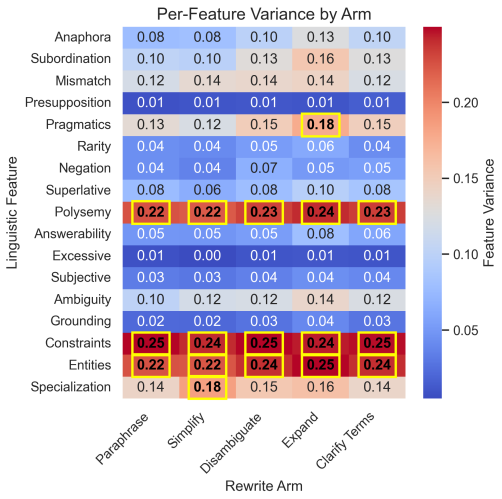


Figure 5: **Contextual Per-Feature Variance by Arm.** For each arm, we compute the variance of each binary linguistic feature over all queries on which that arm was chosen. High variance means the bandit frequently switches the arm on that feature's presence.

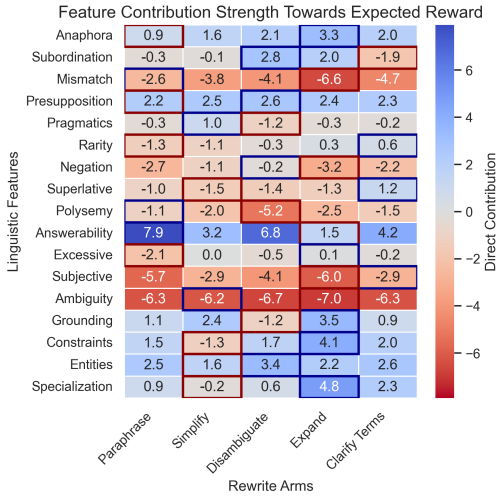


Figure 6: **Contextual Feature Contribution Strength.** These are the averaged θ weights (direct contributions) of each feature to the expected reward under each arm. Positive weights indicate features that boost that arm's reward; negative weights indicate features that *penalize* it.

that the diversity of arm selected is correlated with feature variance—and that there is no *single* rewrite arm that fits all queries.

6 Conclusion

Large language models are now routinely asked to provide factual answers in high-stakes settings (e.g., compliance, legal, medical, and policy contexts), where confidently delivered but fabricated content is not just an accuracy failure but a source of reputational, financial, and legal exposure. Critically, many such deployments rely on closed, proprietary systems whose weights cannot be modified and whose decoding behavior cannot be directly controlled. **QueryBandits** is aimed at this reality: rather than trying to make the model itself "safe," we place a decision layer in front of the model that selectively rewrites the incoming query to reduce hallucination risk *before* the model answers. This control surface is lightweight, auditable, and compatible with closed models: for each user query, we choose among a small set of rewrite strategies, and that choice can be logged, inspected, and justified to governance and compliance teams. Empirically, this matters. Across diverse QA settings, selecting the right rewrite *per query* yields substantially higher factual reliability than using a single fixed prompting style or issuing the query unchanged. We also observe that *no single rewrite strategy dominates*: different queries benefit from different interventions such as expanding underspecified requests, simplifying syntactically overloaded questions, or clarifying ambiguous references. Therefore, **QueryBandits** reframes hallucination risk as not only a property of the model, but as a property of the *interaction*. By treating query reformulation itself as a policy decision rather than a static trick, **QueryBandits** turns that interaction into an explicit, learnable, and auditable control knob for governing model behavior in deployment.

References

[1] Vaibhav Adlakha, Parishad BehnamGhader, Xing Han Lu, Nicholas Meade, and Siva Reddy. Evaluating correctness and faithfulness of instruction-following models for question answering, 2024. URL https://arxiv.org/abs/2307.16877.

[2] Ahmed Alajrami and Nikolaos Aletras. How does the pre-training objective affect what large language models learn about linguistic properties?, 2022. URL https://arxiv.org/abs/2203.10415.

[3] Emmanuel Ameisen, Jack Lindsey, Adam Pearce, Wes Gurnee, Nicholas L. Turner, Brian Chen, Craig Citro, David Abrahams, Shan Carter, Basil Hosmer, Jonathan Marcus, Michael Sklar, Adly Templeton, Trenton Bricken, Callum McDougall, Hoagy Cunningham, Thomas Henighan, Adam Jermyn, Andy Jones, Andrew Persic, Zhenyi Qi, T. Ben Thompson, Sam Zimmerman, Kelley Rivoire, Thomas Conerly, Chris Olah, and Joshua Batson. Circuit tracing: Revealing computational graphs in language models. *Transformer Circuits Thread*, 2025. URL https://transformer-circuits.pub/2025/attribution-graphs/methods.html.

[4] Aida Amini, Saadia Gabriel, Shanchuan Lin, Rik Koncel-Kedziorski, Yejin Choi, and Hannaneh Hajishirzi. MathQA: Towards interpretable math word problem solving with operation-based formalisms. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2357–2367, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1245. URL https://aclanthology.org/N19-1245.

[5] Alan Ansell, Felipe Bravo-Marquez, and Bernhard Pfahringer. Polylm: Learning about polysemy through language modeling, 2021. URL https://arxiv.org/abs/2101.10448.

[6] Dhananjay Ashok and Zachary C. Lipton. Promptner: Prompting for named entity recognition, 2023. URL https://arxiv.org/abs/2305.15444.

[7] Peter Auer, Nicolò Cesa-Bianchi, Yoav Freund, and Robert E. Schapire. The nonstochastic multiarmed bandit problem. *SIAM Journal on Computing*, 32(1):48–77, 2002. doi: 10.1137/S0097539701398375. URL https://doi.org/10.1137/S0097539701398375.

[8] Max Bachmann. rapidfuzz/rapidfuzz: Release 3.8.1, April 2024. URL https://doi.org/10.5281/zenodo.10938887.

[9] Anas Belfathi, Nicolas Hernandez, and Laura Monceaux. Harnessing gpt-3.5-turbo for rhetorical role prediction in legal cases, 2023. URL https://arxiv.org/abs/2310.17413.

[10] Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. Piqa: Reasoning about physical commonsense in natural language. In *Thirty-Fourth AAAI Conference on Artificial Intelligence*, 2020.

[11] Terra Blevins, Hila Gonen, and Luke Zettlemoyer. Prompting language models for linguistic structure, 2023. URL https://arxiv.org/abs/2211.07830.

[12] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020. URL https://arxiv.org/abs/2005.14165.

[13] Chris Callison-Burch, Miles Osborne, and Philipp Koehn. Re-evaluating the role of Bleu in machine translation research. In Diana McCarthy and Shuly Wintner, editors, *11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 249–256, Trento, Italy, April 2006. Association for Computational Linguistics. URL https://aclanthology.org/E06-1032/.

[14] Hong Chen, Zhenhua Fan, Hao Lu, Alan Yuille, and Shu Rong. PreCo: A large-scale dataset in preschool vocabulary for coreference resolution. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii, editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 172–181, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1016. URL https://aclanthology.org/D18-1016/.

[15] Yongchao Chen, Jacob Arkin, Yilun Hao, Yang Zhang, Nicholas Roy, and Chuchu Fan. PRompt optimization in multi-step tasks (PROMST): Integrating human feedback and heuristic-based sampling. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 3859–3920, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.226. URL https://aclanthology.org/2024.emnlp-main.226/.

[16] Nicole Cho and William Watson. Multiq&a: An analysis in measuring robustness via automated crowdsourcing of question perturbations and answers, 2025. URL https://arxiv.org/abs/2502.03711.

[17] Nicole Cho, Nishan Srishankar, Lucas Cecchi, and William Watson. Fishnet: Financial intelligence from sub-querying, harmonizing, neural-conditioning, expert swarms, and task planning. In *Proceedings of the 5th ACM International Conference on AI in Finance*, ICAIF '24, page 591–599. ACM, November 2024. doi: 10.1145/3677052.3698597. URL http://dx.doi.org/10.1145/3677052.3698597.

[18] Paul Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences, 2023. URL https://arxiv.org/abs/1706.03741.

[19] Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. Boolq: Exploring the surprising difficulty of natural yes/no questions. In *NAACL*, 2019.

[20] HERBERT H. CLARK and RICHARD J. GERRIG. Understanding old words with new meanings, 1983. URL https://web.stanford.edu/~clark/1980s/Clark.Gerrig.oldwords.83.pdf.

[21] Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv:1803.05457v1*, 2018.

[22] Charles L. A. Clarke, Nick Craswell, and Ian Soboroff. Overview of the TREC 2009 web track. In Ellen M. Voorhees and Lori P. Buckland, editors, *Proceedings of The Eighteenth Text REtrieval Conference, TREC 2009, Gaithersburg, Maryland, USA, November 17-20, 2009*, volume 500-278 of *NIST Special Publication*. National Institute of Standards and Technology (NIST), 2009. URL http://trec.nist.gov/pubs/trec18/papers/WEB09.OVERVIEW.pdf.

[23] Jeremy Cole, Michael Zhang, Daniel Gillick, Julian Eisenschlos, Bhuwan Dhingra, and Jacob Eisenstein. Selectively answering ambiguous questions. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 530–543, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.35. URL https://aclanthology.org/2023.emnlp-main.35/.

[24] Dechert LLP. Ai expert challenged for relying on ai "hallucinations", December 2024. URL https://www.dechert.com/knowledge/re-torts/2024/12/ai-expert-challenged-for-relying-on-ai--hallucinations-.html. Accessed: 2025-05-12.

[25] DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Shengfeng Ye, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanjia Zhao, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan

Liu, Yuyang Zhou, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025. URL https://arxiv.org/abs/2501.12948.

[26] Yang Deng, Lizi Liao, Liang Chen, Hongru Wang, Wenqiang Lei, and Tat-Seng Chua. Prompting and evaluating large language models for proactive dialogues: Clarification, target-guided, and non-collaboration. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10602–10621, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.711. URL https://aclanthology.org/2023.findings-emnlp.711/.

[27] Yihe Deng, Weitong Zhang, Zixiang Chen, and Quanquan Gu. Rephrase and respond: Let large language models ask better questions for themselves, 2024. URL https://arxiv.org/abs/2311.04205.

[28] Esin Durmus, Karina Nguyen, Thomas I. Liao, Nicholas Schiefer, Amanda Askell, Anton Bakhtin, Carol Chen, Zac Hatfield-Dodds, Danny Hernandez, Nicholas Joseph, Liane Lovitt, Sam McCandlish, Orowa Sikder, Alex Tamkin, Janel Thamkul, Jared Kaplan, Jack Clark, and Deep Ganguli. Towards measuring the representation of subjective global opinions in language models, 2024. URL https://arxiv.org/abs/2306.16388.

[29] Donka F Farkas and Katalin É Kiss. On the comparative and absolute readings of superlatives, 2000.

[30] Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. Gptscore: Evaluate as you desire, 2023. URL https://arxiv.org/abs/2302.04166.

[31] Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. Retrieval-augmented generation for large language models: A survey, 2024. URL https://arxiv.org/abs/2312.10997.

[32] Aurélien Garivier and Olivier Cappé. The kl-ucb algorithm for bounded stochastic bandits and beyond, 2013. URL https://arxiv.org/abs/1102.2490.

[33] E. J. Gumbel. The return period of flood flows, 1941. URL doi:10.1214/aoms/1177731747.

[34] Janosch Haber and Massimo Poesio. Polysemy—Evidence from linguistics, behavioral science, and contextualized language models. *Computational Linguistics*, 50(1):351–417, March 2024. doi: 10.1162/coli_a_00500. URL https://aclanthology.org/2024.cl-1.10/.

[35] James Hannan. Approximation to bayes risk in repeated play, 1957.

[36] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021.

[37] Md Mosharaf Hossain and Eduardo Blanco. Leveraging affirmative interpretations from negation improves natural language understanding, 2022. URL https://arxiv.org/abs/2210.14486.

[38] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55, January 2025. ISSN 1558-2868. doi: 10.1145/3703155. URL http://dx.doi.org/10.1145/3703155.

[39] Soyeong Jeong, Jinheon Baek, Sukmin Cho, Sung Ju Hwang, and Jong Park. Adaptive-RAG: Learning to adapt retrieval-augmented large language models through question complexity. In Kevin Duh, Helena Gomez, and Steven Bethard, editors, *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7036–7050, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.389. URL https://aclanthology.org/2024.naacl-long.389/.

[40] Ziwei Ji, Tiezheng Yu, Yan Xu, Nayeon Lee, Etsuko Ishii, and Pascale Fung. Towards mitigating LLM hallucination via self reflection. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1827–1843, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.123. URL https://aclanthology.org/2023.findings-emnlp.123/.

[41] Ziwei Ji, Tiezheng Yu, Yan Xu, Nayeon Lee, Etsuko Ishii, and Pascale Fung. Towards mitigating hallucination in large language models via self-reflection, 2023. URL https://arxiv.org/abs/2310.06271.

[42] Yuxin Jiang, Yufei Wang, Xingshan Zeng, Wanjun Zhong, Liangyou Li, Fei Mi, Lifeng Shang, Xin Jiang, Qun Liu, and Wei Wang. Followbench: A multi-level fine-grained constraints following benchmark for large language models, 2024. URL https://arxiv.org/abs/2310.20410.

[43] Matt Gardner Johannes Welbl, Nelson F. Liu. Crowdsourcing multiple choice science questions, 2017. URL https://arxiv.org/abs/1707.06209.

[44] Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. triviaqa: A Large Scale Distantly Supervised Challenge Dataset for Reading Comprehension. *arXiv e-prints*, art. arXiv:1705.03551, 2017.

[45] Adam Kalai and Santosh Vempala. Efficient algorithms for online decision problems. *Journal of Computer and System Sciences*, 71(3):291–307, 2005. ISSN 0022-0000. doi: https://doi.org/10.1016/j.jcss.2004.10.016. URL https://www.sciencedirect.com/science/article/pii/S0022000004001394. Learning Theory 2003.

[46] Gaurav Kamath, Sebastian Schuster, Sowmya Vajjala, and Siva Reddy. Scope ambiguities in large language models. *Transactions of the Association for Computational Linguistics*, 12:738–754, 2024. ISSN 2307-387X. doi: 10.1162/tacl_a_00670. URL http://dx.doi.org/10.1162/tacl_a_00670.

[47] Lauri Karttunen. Presupposition: What went wrong?, 2016.

[48] Nora Kassner and Hinrich Schütze. Negated and misprimed probes for pretrained language models: Birds can talk, but cannot fly. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7811–7818, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.698. URL https://aclanthology.org/2020.acl-main.698/.

[49] Yerbolat Khassanov, Zhiping Zeng, Van Tung Pham, Haihua Xu, and Eng Siong Chng. Enriching rare word representations in neural language models by embedding matrix augmentation. In *Interspeech 2019*, interspeech2019, page 3505–3509. ISCA, September 2019. doi: 10.21437/interspeech.2019-1858. URL http://dx.doi.org/10.21437/Interspeech.2019-1858.

[50] Hyuhng Joon Kim, Youna Kim, Cheonbok Park, Junyeob Kim, Choonghyun Park, Kang Min Yoo, Sang-goo Lee, and Taeuk Kim. Aligning language models to explicitly handle ambiguity. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1989–2007, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.119. URL https://aclanthology.org/2024.emnlp-main.119/.

[51] T. L. Lai and H. Robbins. Asymptotically efficient adaptive allocation rules, 1985.

[52] Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

[53] Harrison Lee, Samrat Phatale, Hassan Mansoor, Thomas Mesnard, Johan Ferret, Kellie Lu, Colton Bishop, Ethan Hall, Victor Carbune, Abhinav Rastogi, and Sushant Prakash. Rlaif vs. rlhf: Scaling reinforcement learning from human feedback with ai feedback, 2024. URL https://arxiv.org/abs/2309.00267.

[54] Nayeon Lee, Wei Ping, Peng Xu, Mostofa Patwary, Pascale Fung, Mohammad Shoeybi, and Bryan Catanzaro. Factuality enhanced language models for open-ended text generation, 2023. URL https://arxiv.org/abs/2206.04624.

[55] Stephen C. Levinson. *Pragmatics*. Cambridge Textbooks in Linguistics. Cambridge University Press, 1983. URL https://www.cambridge.org/highereducation/books/pragmatics/6D0011901AE9E92CBC1F5F21D7C598C3#contents.

[56] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks, 2021. URL https://arxiv.org/abs/2005.11401.

[57] Tianle Li, Ge Zhang, Quy Duc Do, Xiang Yue, and Wenhu Chen. Long-context llms struggle with long in-context learning, 2024. URL https://arxiv.org/abs/2404.02060.

[58] Ziniu Li, Tian Xu, Yushun Zhang, Zhihang Lin, Yang Yu, Ruoyu Sun, and Zhi-Quan Luo. Remax: A simple, effective, and efficient reinforcement learning method for aligning large language models, 2024. URL https://arxiv.org/abs/2310.10505.

[59] Myrna E Libby, Julie S Weiss, Stacie Bancroft, and William H Ahearn. A comparison of most-to-least and least-to-most prompting on the acquisition of solitary play skills, 2008. URL https://pubmed.ncbi.nlm.nih.gov/22477678/.

[60] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL https://aclanthology.org/W04-1013/.

[61] Chin-Yew Lin and Franz Josef Och. ORANGE: a method for evaluating automatic evaluation metrics for machine translation. In *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*, pages 501–507, Geneva, Switzerland, aug 23–aug 27 2004. COLING. URL https://www.aclweb.org/anthology/C04-1072.

[62] Stephanie Lin, Jacob Hilton, and Owain Evans. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.229. URL https://aclanthology.org/2022.acl-long.229.

[63] Alisa Liu, Zhaofeng Wu, Julian Michael, Alane Suhr, Peter West, Alexander Koller, Swabha Swayamdipta, Noah Smith, and Yejin Choi. We're afraid language models aren't modeling ambiguity. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 790–807, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.51. URL https://aclanthology.org/2023.emnlp-main.51/.

[64] Jie Liu and Barzan Mozafari. Query rewriting via large language models, 2024. URL https://arxiv.org/abs/2403.09060.

[65] 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, 2023. URL https://arxiv.org/abs/2307.03172.

[66] Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-eval: Nlg evaluation using gpt-4 with better human alignment, 2023. URL https://arxiv.org/abs/2303.16634.

[67] Fangrui Lv, Kaixiong Gong, Jian Liang, Xinyu Pang, and Changshui Zhang. Subjective topic meets LLMs: Unleashing comprehensive, reflective and creative thinking through the negation of negation. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12318–12341, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.686. URL https://aclanthology.org/2024.emnlp-main.686/.

[68] Xinbei Ma, Yeyun Gong, Pengcheng He, Hai Zhao, and Nan Duan. Query rewriting for retrieval-augmented large language models, 2023. URL https://arxiv.org/abs/2305.14283.

[69] Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. Self-refine: Iterative refinement with self-feedback, 2023. URL https://arxiv.org/abs/2303.17651.

[70] Shengyu Mao, Yong Jiang, Boli Chen, Xiao Li, Peng Wang, Xinyu Wang, Pengjun Xie, Fei Huang, Huajun Chen, and Ningyu Zhang. RaFe: Ranking feedback improves query rewriting for RAG. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 884–901, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.49. URL https://aclanthology.org/2024.findings-emnlp.49/.

[71] H. Brendan McMahan. A survey of algorithms and analysis for adaptive online learning, 2015. URL https://arxiv.org/abs/1403.3465.

[72] Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. Can a suit of armor conduct electricity? a new dataset for open book question answering. In *Conference on Empirical Methods in Natural Language Processing*, 2018.

[73] Sidharth Mudgal, Jong Lee, Harish Ganapathy, YaGuang Li, Tao Wang, Yanping Huang, Zhifeng Chen, Heng-Tze Cheng, Michael Collins, Trevor Strohman, Jilin Chen, Alex Beutel, and Ahmad Beirami. Controlled decoding from language models, 2024. URL https://arxiv.org/abs/2310.17022.

[74] Gergely Neu and Julia Olkhovskaya. Efficient and robust algorithms for adversarial linear contextual bandits. In Jacob Abernethy and Shivani Agarwal, editors, *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pages 3049–3068. PMLR, 09–12 Jul 2020. URL https://proceedings.mlr.press/v125/neu20b.html.

[75] Rodrigo Nogueira and Kyunghyun Cho. Task-oriented query reformulation with reinforcement learning. In Martha Palmer, Rebecca Hwa, and Sebastian Riedel, editors, *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 574–583, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1061. URL https://aclanthology.org/D17-1061/.

[76] OpenAI. Openai o3 and o4-mini system card, 2025. URL https://openai.com/index/o3-o4-mini-system-card/.

[77] Ian Osband, Daniel Russo, and Benjamin thompson. (more) efficient reinforcement learning via posterior sampling. *Advances in Neural Information Processing Systems*, 26, 2013.

[78] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback, 2022. URL https://arxiv.org/abs/2203.02155.

[79] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In Pierre Isabelle, Eugene Charniak, and Dekang Lin, editors, *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics. doi: 10.3115/1073083.1073135. URL https://aclanthology.org/P02-1040/.

[80] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations, 2018. URL https://arxiv.org/abs/1802.05365.

[81] Valentina Pyatkin, Bonnie Webber, Ido Dagan, and Reut Tsarfaty. Superlatives in context: Modeling the implicit semantics of superlatives, 2024. URL https://arxiv.org/abs/2405.20967.

[82] Yang Qiao, Liqiang Jing, Xuemeng Song, Xiaolin Chen, Lei Zhu, and Liqiang Nie. Mutual-enhanced incongruity learning network for multi-modal sarcasm detection. In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'23/IAAI'23/EAAI'23. AAAI Press, 2023. ISBN 978-1-57735-880-0. doi: 10.1609/aaai.v37i8.26138. URL https://doi.org/10.1609/aaai.v37i8.26138.

[83] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI*, 2019.

[84] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text, 2016.

[85] Pranav Rajpurkar, Robin Jia, and Percy Liang. Know what you don't know: Unanswerable questions for squad, 2018.

[86] Matthew Riemer, Sharath Chandra Raparthy, Ignacio Cases, Gopeshh Subbaraj, Maximilian Puelma Touzel, and Irina Rish. Continual learning in environments with polynomial mixing times. *Advances in Neural Information Processing Systems*, 35:21961–21973, 2022.

[87] Elijah Rippeth, Marine Carpuat, Kevin Duh, and Matt Post. Improving word sense disambiguation in neural machine translation with salient document context, 2023. URL https://arxiv.org/abs/2311.15507.

[88] R Tyrrell Rockafellar. Monotone operators and the proximal point algorithm. *SIAM journal on control and optimization*, 14(5):877–898, 1976.

[89] Jerrold M. Sadock and Arnold M. Zwicky. Speech acts distinctions in syntax. In Timothy Shopen, editor, *Language Typology and Syntactic Description*, pages 155–196. Cambridge University Press, Cambridge, 1985.

[90] Timo Schick and Hinrich Schütze. Rare words: A major problem for contextualized embeddings and how to fix it by attentive mimicking, 2019. URL https://arxiv.org/abs/1904.06707.

[91] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

[92] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms, 2017. URL https://arxiv.org/abs/1707.06347.

[93] Ethel Schuster. Anaphoric reference to events and actions: A representation and its advantages. In *Coling Budapest 1988 Volume 2: International Conference on Computational Linguistics*, 1988. URL https://aclanthology.org/C88-2126/.

[94] Hamed Shahbazi, Xiaoli Z. Fern, Reza Ghaeini, Rasha Obeidat, and Prasad Tadepalli. Entity-aware elmo: Learning contextual entity representation for entity disambiguation, 2019. URL https://arxiv.org/abs/1908.05762.

[95] Shai Shalev-Shwartz et al. Online learning and online convex optimization. *Foundations and Trends® in Machine Learning*, 4(2):107–194, 2012.

[96] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models, 2024. URL https://arxiv.org/abs/2402.03300.

[97] Settaluri Sravanthi, Meet Doshi, Pavan Tankala, Rudra Murthy, Raj Dabre, and Pushpak Bhattacharyya. PUB: A pragmatics understanding benchmark for assessing LLMs' pragmatics capabilities. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12075–12097, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.719. URL https://aclanthology.org/2024.findings-acl.719/.

[98] Arun Sai Suggala and Praneeth Netrapalli. Follow the perturbed leader: Optimism and fast parallel algorithms for smooth minimax games, 2020. URL https://arxiv.org/abs/2006.07541.

[99] R. S. Sutton and A. G. Barto. Reinforcement learning: An introduction, 2018.

[100] William R. Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples, 1933.

[101] New York Times. Ai hallucinations: Chatgpt and google's challenges. *The New York Times*, May 2025. URL https://www.nytimes.com/2025/05/05/technology/ai-hallucinations-chatgpt-google.html. Accessed: 2025-05-12.

[102] S. M Towhidul Islam Tonmoy, S M Mehedi Zaman, Vinija Jain, Anku Rani, Vipula Rawte, Aman Chadha, and Amitava Das. A comprehensive survey of hallucination mitigation techniques in large language models, 2024. URL https://arxiv.org/abs/2401.01313.

[103] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models, 2023. URL https://arxiv.org/abs/2307.09288.

[104] Thinh Hung Truong, Timothy Baldwin, Karin Verspoor, and Trevor Cohn. Language models are not naysayers: An analysis of language models on negation benchmarks, 2023. URL https://arxiv.org/abs/2306.08189.

[105] Hoang Van, Zheng Tang, and Mihai Surdeanu. How may i help you? using neural text simplification to improve downstream nlp tasks, 2021. URL https://arxiv.org/abs/2109.04604.

[106] Rob A. van der Sandt. Presupposition projection as anaphora resolution. *Journal of Semantics*, 9(4): 333–377, 1992. doi: 10.1093/jos/9.4.333. URL https://doi.org/10.1093/jos/9.4.333.

[107] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2023. URL https://arxiv.org/abs/1706.03762.

[108] Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. SuperGLUE: A stickier benchmark for general-purpose language understanding systems. *arXiv preprint 1905.00537*, 2019.

[109] Jiaan Wang, Yunlong Liang, Fandong Meng, Zengkui Sun, Haoxiang Shi, Zhixu Li, Jinan Xu, Jianfeng Qu, and Jie Zhou. Is ChatGPT a good NLG evaluator? a preliminary study. In Yue Dong, Wen Xiao, Lu Wang, Fei Liu, and Giuseppe Carenini, editors, *Proceedings of the 4th New Frontiers in Summarization Workshop*, pages 1–11, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.newsum-1.1. URL https://aclanthology.org/2023.newsum-1.1/.

[110] Shuhe Wang, Xiaofei Sun, Xiaoya Li, Rongbin Ouyang, Fei Wu, Tianwei Zhang, Jiwei Li, and Guoyin Wang. Gpt-ner: Named entity recognition via large language models, 2023. URL https://arxiv.org/abs/2304.10428.

[111] William Watson, Nicole Cho, Tucker Balch, and Manuela Veloso. HiddenTables and PyQTax: A cooperative game and dataset for TableQA to ensure scale and data privacy across a myriad of taxonomies. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7144–7159, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.442. URL https://aclanthology.org/2023.emnlp-main.442.

[112] William Watson, Nicole Cho, and Nishan Srishankar. Is there no such thing as a bad question? h4r: Hallucibot for ratiocination, rewriting, ranking, and routing. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(24):25470–25478, Apr. 2025. doi: 10.1609/aaai.v39i24.34736. URL https://ojs.aaai.org/index.php/AAAI/article/view/34736.

[113] William Watson, Nicole Cho, Nishan Srishankar, Zhen Zeng, Lucas Cecchi, Daniel Scott, Suchetha Siddagangappa, Rachneet Kaur, Tucker Balch, and Manuela Veloso. LAW: Legal agentic workflows for custody and fund services contracts. In Owen Rambow, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, Steven Schockaert, Kareem Darwish, and Apoorv Agarwal, editors, *Proceedings of the 31st International Conference on Computational Linguistics: Industry Track*, pages 583–594, Abu Dhabi, UAE, January 2025. Association for Computational Linguistics. URL https://aclanthology.org/2025.coling-industry.50/.

[114] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models, 2023. URL https://arxiv.org/abs/2201.11903.

[115] Sam Witteveen and Martin Andrews. Paraphrasing with large language models. In *Proceedings of the 3rd Workshop on Neural Generation and Translation*. Association for Computational Linguistics, 2019. doi: 10.18653/v1/d19-5623. URL http://dx.doi.org/10.18653/v1/D19-5623.

[116] Yi Yang, Wen-tau Yih, and Christopher Meek. WikiQA: A challenge dataset for open-domain question answering. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2013–2018, Lisbon, Portugal, September 2015. Association for Computational Linguistics. doi: 10.18653/v1/D15-1237. URL https://aclanthology.org/D15-1237.

[117] Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1259. URL https://aclanthology.org/D18-1259.

[118] Jia-Yu Yao, Kun-Peng Ning, Zhen-Hui Liu, Mu-Nan Ning, Yu-Yang Liu, and Li Yuan. Llm lies: Hallucinations are not bugs, but features as adversarial examples, 2024. URL https://arxiv.org/abs/2310.01469.

[119] Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Tiantian Fan, Gaohong Liu, Lingjun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Wang Zhang, Hang Zhu, Jinhua Zhu, Jiaze Chen, Jiangjie Chen, Chengyi Wang, Hongli Yu, Weinan Dai, Yuxuan Song, Xiangpeng Wei, Hao Zhou, Jingjing Liu, Wei-Ying Ma, Ya-Qin Zhang, Lin Yan, Mu Qiao, Yonghui Wu, and Mingxuan Wang. Dapo: An open-source llm reinforcement learning system at scale, 2025. URL https://arxiv.org/abs/2503.14476.

[120] Wenhao Yu, Dan Iter, Shuohang Wang, Yichong Xu, Mingxuan Ju, Soumya Sanyal, Chenguang Zhu, Michael Zeng, and Meng Jiang. Generate rather than retrieve: Large language models are strong context generators, 2023. URL https://arxiv.org/abs/2209.10063.

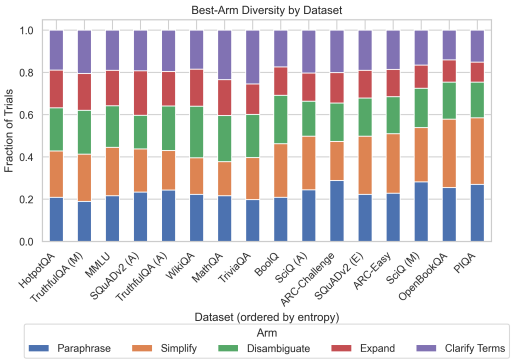
[121] Weizhe Yuan, Graham Neubig, and Pengfei Liu. Bartscore: Evaluating generated text as text generation, 2021. URL https://arxiv.org/abs/2106.11520.

[122] Yu Yue, Yufeng Yuan, Qiying Yu, Xiaochen Zuo, Ruofei Zhu, Wenyuan Xu, Jiaze Chen, Chengyi Wang, TianTian Fan, Zhengyin Du, Xiangpeng Wei, Xiangyu Yu, Gaohong Liu, Juncai Liu, Lingjun Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Chi Zhang, Mofan Zhang, Wang Zhang, Hang Zhu, Ru Zhang, Xin Liu, Mingxuan Wang, Yonghui Wu, and Lin Yan. Vapo: Efficient and reliable reinforcement learning for advanced reasoning tasks, 2025. URL https://arxiv.org/abs/2504.05118.

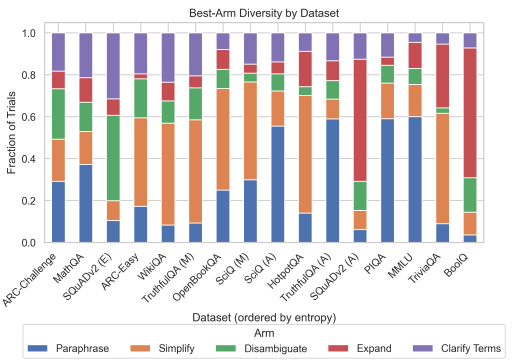
[123] Zhen Zeng, William Watson, Nicole Cho, Saba Rahimi, Shayleen Reynolds, Tucker Balch, and Manuela Veloso. Flowmind: Automatic workflow generation with llms, 2024. URL https://arxiv.org/abs/2404.13050.

[124] Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc Le, and Ed Chi. Least-to-most prompting enables complex reasoning in large language models, 2023. URL https://arxiv.org/abs/2205.10625.

[125] Yuxin Zuo, Kaiyan Zhang, Shang Qu, Li Sheng, Xuekai Zhu, Biqing Qi, Youbang Sun, Ganqu Cui, Ning Ding, and Bowen Zhou. Ttrl: Test-time reinforcement learning, 2025. URL https://arxiv.org/abs/2504.16084.



(a) Arm Diversity for Contextual Bandits, as a Fraction of Trials.



(b) Arm Diversity for Non-Contextual Bandits, as a Fraction of Trials.

Figure 7: For Non-Contextual bandits, *almost every* dataset is dominated by a single arm with the highest global reward (typically 40%-60% of the trials). The remaining 40-60% is split among the other four arms as noise, the non-contextual policy has no way to "know" when within a dataset a different arm might do better. In contrast, Contextual bandits show a more even mix: the top arm is only ∼25-30%, with two or three other arms contributing sizable shares (15-25% each). The contextual policy *reads the features* and diversifies its choices within each dataset.

A Appendix / supplemental material

A.1 Limitations

Current limitations in our work are as follows: our current contextual bandit framework treats each of the 17 features as independent, but does not capture higher-order interactions. This can provide an exciting avenue of future research in terms of measuring whether the combination of features jointly exacerbates hallucination. Likewise, we would like to highlight that the feature-arm regression weights do not stipulate a causal relationship - highly sophisticated causal relationships are difficult to formulate within LLMs due to the inherent difficulties of interpreting a neural network's internal layers; thus, in this paper, we focus on providing empirical studies and the conclusions we can draw from them. Finally, even with our rigorous studies to find the ROC-AUC Pareto-frontier, our reward model leverages LLM-as-judge, which may reflect the LLM's bias. Overall, these limitations posit potential directions by which the research community can further pursue - and ultimately help expand our understanding of these powerful, albeit hallucinatory models.

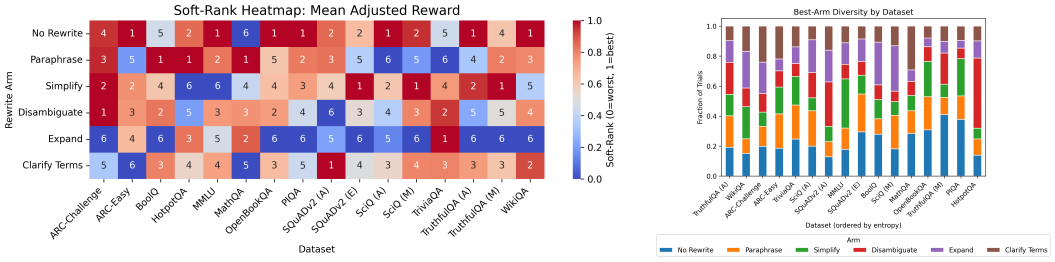
A.2 Summary of Bandits

- **Non-Contextual Adversarial**
  - **EXP3** [7] Maintains weights w_k, samples $a_t \propto w_k$, updates $w_{a_t} \leftarrow w_{a_t} \exp\left(\frac{\gamma\, r_t}{K\, p_{a_t}}\right)$.
  - **FTPL** [45, 98] Adds Gumbel noise $\xi_k \sim \mathrm{Gumbel}(0, 1/\eta)$ [33] to cumulative rewards, selects $a_t = \arg\max(\mathrm{cum\_reward}_k + \xi_k)$, then increments the chosen arm's reward.
- **Contextual Stochastic**
  - **LinUCB** [51] Selects $a_t = \arg\max_k \left(x_t^\top \hat{\theta}_k + \alpha\sqrt{x_t^\top A_k^{-1} x_t}\right)$, updates $A_k \leftarrow A_k + x_t x_t^\top$, $b_k \leftarrow b_k + r_t x_t$.
  - **KL-UCB (LinUCB-KL)** [32] Replaces the UCB term with a KL-divergence-based confidence bound.
  - **Thompson Sampling** Maintains Gaussian posterior $\mathcal{N}(\mu_k, \Sigma_k)$; samples $\tilde{\theta}_k$, picks $a_t = \arg\max x_t^\top \tilde{\theta}_k$, updates the posterior.
- **Contextual Adversarial**
  - **FTRL** [71] Selects arm maximizing $x_t^\top w_k - \lambda\|w_k\|_1$, with an ℓ_1 regularizer.
  - **ϵ-greedy FTRL** ...
  - **LinearEXP3** [74] Contextual extension of EXP3, sampling arms based on exponentiated linear scores.

Algorithm 1 General Bandit + Rewrite Loop

Require: arms $\mathcal{A}$, context x_t, algorithm algo $\in$ {EXP3, FTPL, LinUCB, KL, FTRL, Thompson}, hyperparameters

for $t = 1$ to T **do**
  observe x_t
  for each arm $k \in \mathcal{A}$ **do**
    $s_k \leftarrow \text{Score}(\text{algo}, k, x_t)$
  end for
  select $a_t = \arg\max_{k \in \mathcal{A}} s_k$
  apply rewrite a_t to query and observe reward r_t
  $\text{Update}(\text{algo}, a_t, x_t, r_t)$
end for



(a) Soft Rank Heatmap for all Bandits, including arm NO REWRITE.

(b) Arm Diversity when including NO REWRITE.

Figure 8: **Impact of the No-Rewrite Arm.** Note that these experiments are conducted on the original query "as-is" in the benchmark dataset, with no perturbations. Upon enabling the NO REWRITE option, our contextual bandit rapidly converges to this arm, which then achieves the highest reward on several datasets. We attribute this behavior to the LLM's tendency to memorize benchmark questions.

– **LinearFTPL** [35] Contextual adaptation of FTPL, applying Gumbel perturbations to linear reward estimates.

A.3 LinUCB

The estimated parameter is:

$$\hat{\theta}_a = A_a^{-1} \mathbf{b}_a. \tag{2}$$

Given a query feature vector $\mathbf{x}$, the upper confidence bound (UCB) for arm a is:

$$\text{UCB}_a(\mathbf{x}) = \mathbf{x}^\top \hat{\theta}_a + \alpha \sqrt{\mathbf{x}^\top A_a^{-1} \mathbf{x}}, \tag{3}$$

where α controls the exploration–exploitation trade-off. The arm selected is:

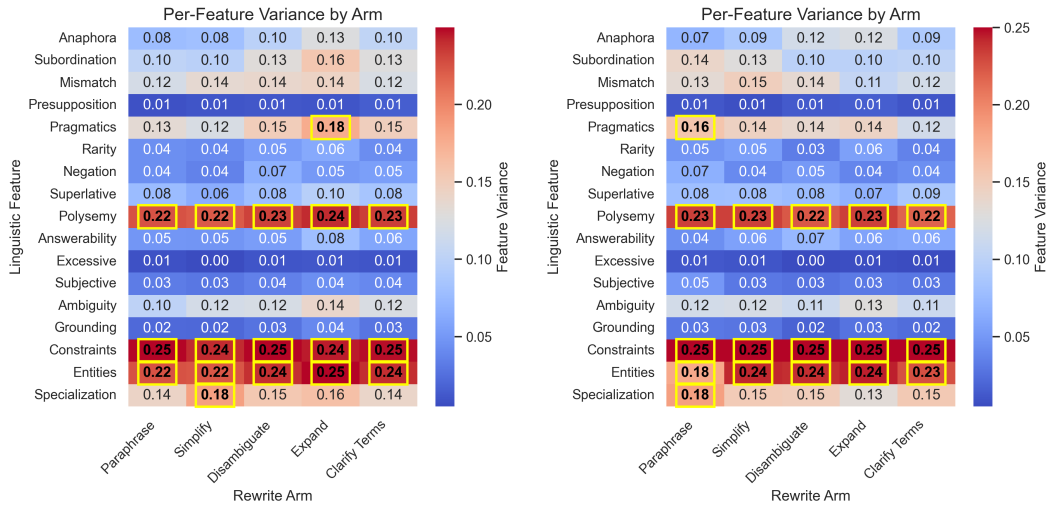
$$a^* = \arg\max_{a \in \mathcal{A}} \text{UCB}_a(\mathbf{x}). \tag{4}$$

Upon observing reward r, update:

$$A_a \leftarrow A_a + \mathbf{x}\mathbf{x}^\top, \quad \mathbf{b}_a \leftarrow \mathbf{b}_a + r\,\mathbf{x}. \tag{5}$$

A.4 LinUCBKL Bandit Strategy:

The algorithm is initialized with parameters: number of arms n_{arms}, dimension d, regularization parameter λ, exploration parameter α, noise variance σ_{noise}, and KL-bound constant c. Each arm a maintains a matrix $\mathbf{A}_a$ and a vector $\mathbf{b}_a$, initialized as $\lambda \mathbf{I}_d$ and $\mathbf{0}_d$, respectively.



(a) Contextual Model Feature Variance.

(b) Non-Contextual Model Feature Variance.

Figure 9: Comparison of Feature Variance between (a) our contextual bandits and (b) its non-contextual counterparts. *Polysemy*, *Constraints* and *Entities* show the most variation. *Presupposition*, *Excessive Details*, and *Grounding* have the least.

The `select_arm` method computes the score for each arm a using the following formulation:

$$\begin{aligned}
\theta_a &= \mathbf{A}_a^{-1}\mathbf{b}_a \\
\mu_a &= \mathbf{x}^\top \theta_a \\
\text{var}_a &= \mathbf{x}^\top \mathbf{A}_a^{-1}\mathbf{x} \\
n_a &= \max(1, \text{counts}[a]) \\
\text{raw_bound}_a &= \frac{\log(t) + c\log(\log(t+1))}{n_a} \\
\text{bound}_a &= \max(\text{raw_bound}_a, 0.0) \\
\text{bonus}_a &= \sqrt{2 \cdot \text{var}_a \cdot \text{bound}_a} \\
\text{score}_a &= \mu_a + \text{bonus}_a
\end{aligned}$$

where $\mathbf{x}$ is the context vector, t is the time step, and counts$[a]$ is the number of times arm a has been selected. The arm with the highest score is selected for exploration.

The `update` method updates the matrix $\mathbf{A}_a$ and vector $\mathbf{b}_a$ for the selected arm a based on the received reward r_t:

$$\begin{aligned}
\mathbf{A}_a &\leftarrow \mathbf{A}_a + \mathbf{x}\mathbf{x}^\top \\
\mathbf{b}_a &\leftarrow \mathbf{b}_a + r_t\mathbf{x} \\
\text{counts}[a] &\leftarrow \text{counts}[a] + 1
\end{aligned}$$

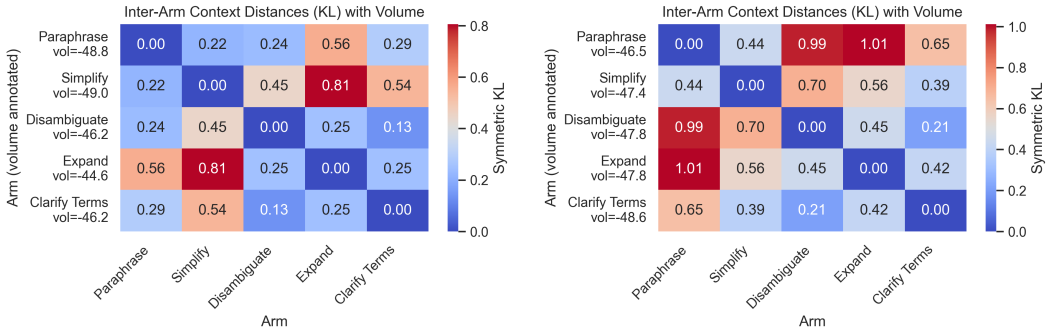
This strategy leverages the KL-bound to dynamically adjust exploration bonuses, enhancing the LinUCB algorithm's ability to balance exploration and exploitation in a contextual setting.

A.5 FTRL

The algorithm is initialized with the following parameters: number of arms n_{arms}, dimension d, learning rate α, exploration parameter β, and regularization parameters l_1 and l_2. The cumulative gradient vectors for each arm are stored in $\mathbf{z}_a$, initialized as zero vectors of dimension d.

The weight vector $\mathbf{w}_a$ for each arm a is computed as:

$$w_i = \begin{cases} -\frac{z_i - \text{sign}(z_i)\cdot l_1}{\frac{\beta+\sqrt{n_i}}{\alpha} + l_2} & \text{if } |z_i| > l_1 \\ 0 & \text{otherwise} \end{cases}$$



(a) Contextual Model KL Distance.

(b) Non-Contextual Model KL Distance.

Figure 10: Comparison of Inter-Arm Context Distances (Symmetric KL) between (a) our contextual bandits and (b) its non-contextual counterparts. Arm pairs such as EXPAND and PARAPHRASE in the non-contextual bandit setting exhibit high KL distances at 1.01. One interpretation is that the context-clouds barely overlap from dataset to dataset (Figure 7b).

where z_i is the cumulative gradient for the i-th feature of arm a, and n_i is the cumulative squared gradient for the i-th feature. The arm with the highest score, calculated as the dot product of the weight vector w and the context vector, is selected:

$$a_t = \arg \max_{a \in \{1,\ldots,n_{\text{arms}}\}} \left(\sum_{i=1}^{d} w_i \cdot \mathbf{x}_i \right)$$

Upon receiving a reward r_t for the selected arm a_t, the algorithm updates the cumulative gradient vector $\mathbf{z}$ and the squared gradient sum $\mathbf{n}$ for the selected arm:

$$\begin{aligned} \varepsilon_{error} &= \langle w, \mathbf{x} \rangle - r_t \\ g &= \varepsilon_{error} \cdot \mathbf{x} \\ \sigma &= \frac{\sqrt{n_i + g_i^2} - \sqrt{n_i}}{\alpha} \\ z_i &\leftarrow z_i + g_i - \sigma \cdot w_i \\ n_i &\leftarrow n_i + g_i^2 \end{aligned}$$

This formulation allows the FTRL algorithm to adaptively adjust the exploration-exploitation trade-off by incorporating both the cumulative reward and the uncertainty in the form of regularization terms, which are scaled by the learning rate α and exploration parameter β.

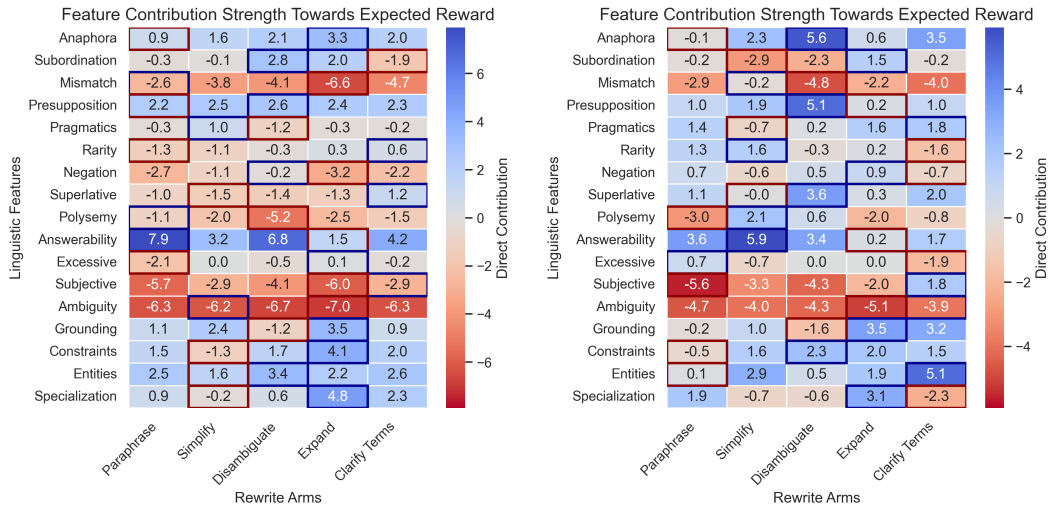
A.6 Linear EXP3

The algorithm is initialized with parameters: number of arms n_{arms}, dimension d, exploration parameter γ, and learning rate η. Each arm a maintains a parameter vector θ_a, initialized as $\mathbf{0}_d$.

We compute the probability distribution over arms using the following formulation:

$$\begin{aligned} \text{logits}_a &= \theta_a^\top \mathbf{x} \\ \text{logits} &= \text{logits} - \max(\text{logits}) \\ \text{exp\_logits}_a &= \exp(\text{logits}_a) \\ \text{base\_probs}_a &= \frac{\text{exp\_logits}_a}{\sum_{a=1}^{n_{\text{arms}}} \text{exp\_logits}_a} \\ \text{probs}_a &= (1-\gamma) \cdot \text{base\_probs}_a + \frac{\gamma}{n_{\text{arms}}} \end{aligned}$$

where $\mathbf{x}$ is the context vector. The arm is selected based on the probability distribution probs.



(a) Contextual Model Raw Feature Strength.

(b) Non-Contextual Model Raw Feature Strength.

Figure 11: Comparison of Raw feature-level regression coefficients between (a) our contextual bandits and (b) its non-contextual counterparts. Each cell shows how enables a raw view into how specific linguistic feature changes the expected reward under each rewrite strategy.

The `update` method updates the parameter vector θ_a for the selected arm a using the estimated reward $\hat{r}_t$:

$$\hat{r}_t = \frac{r_t}{p_a}$$
$$\theta_a \leftarrow \theta_a + \eta \cdot \hat{r}_t \cdot \mathbf{x}$$

where p_a is the probability of selecting arm a, and r_t is the received reward. This strategy leverages exponential weighting and exploration bonuses to balance exploration and exploitation in a linear contextual setting.

A.7 Linear FTPL

The algorithm is initialized with parameters: number of arms n_{arms}, dimension d, and learning rate η. Each arm a maintains a parameter vector θ_a, initialized as $\mathbf{0}_d$.

The `select_arm` method computes the perturbed scores for each arm using the following formulation:

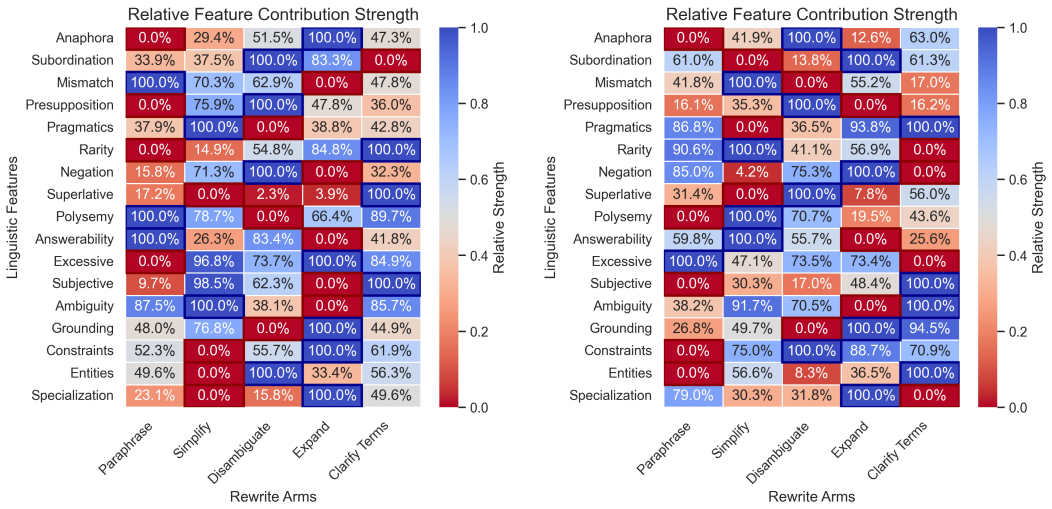
$$\text{linear\_score}_a = \theta_a^\top \mathbf{x}$$
$$\text{noise}_a \sim \text{Gumbel}(0, \frac{1}{\eta})$$
$$\text{score}_a = \text{linear\_score}_a + \text{noise}_a$$

where $\mathbf{x}$ is the context vector. The arm with the highest perturbed score is selected:

$$a_t = \arg \max_{a \in \{1,\ldots,n_{\text{arms}}\}} \text{score}_a$$

$$\theta_a \leftarrow \theta_a + r_t \cdot \mathbf{x}$$

This strategy leverages random perturbations from a Gumbel distribution to balance exploration and exploitation, allowing the algorithm to explore suboptimal arms while exploiting the accumulated knowledge of their performance in a linear contextual setting.



(a) Contextual Model Relative Feature Strength. (b) Non-Contextual Model Relative Feature Strength.

Figure 12: Comparison of Min-Max Normalized feature-level regression coefficients between (a) our contextual bandits and (b) its non-contextual counterparts. Each cell shows how enables a relative view into how specific linguistic feature changes the expected reward under each rewrite strategy. Table 4 highlights contextual bandit trends.

Table 4: **Top Drivers ($f^{+}_{\max}$) and Reducers ($f^{-}_{\max}$) of Rewrite Strategies per Linguistic Features** For each rewrite arm, we list the feature whose normalized coefficient was highest (100 %) and lowest (0 %), alongside a brief rationale for its positive or negative impact on downstream reward.

Arm a	$f^{+}_{\max}$	Interpretation	$f^{-}_{\max}$	Interpretation
DISAMBIGUATE	Subordination (100 %)	Long or nested clauses benefit from targeted disambiguation, which isolates and clarifies the core semantic relation.	Polysemy (0 %)	Highly polysemous terms lead disambiguation to pick the wrong sense, degrading downstream reward.
SIMPLIFY	Pragmatics (100 %)	Pragmatic cues (e.g. discourse markers, politeness) guide safe simplification without loss of meaning.	Superlative (0 %)	Stripping superlative constructions removes essential comparative context, hurting reward.
EXPAND	Constraints (100 %)	Queries already rich in constraints (time, location, numeric bounds) gain precision when expanded with further qualifiers.	Ambiguity (0 %)	Underspecified queries offer no detail to expand, so further addition of terms only introduces noise.
PARAPHRASE	Answerability (100 %)	Paraphrasing queries that are already answerable refreshes wording while preserving solvability, boosting LLM performance.	Presupposition (0 %)	Altering queries with strong presuppositions can break implied assumptions, reducing effective reward.
CLARIFY TERMS	Rarity (100 %)	Defining rare or domain-specific terms anchors the LLM's understanding of technical queries.	Subordination (0 %)	Clarifications in convoluted sentences can introduce further parsing difficulty, impeding reward.

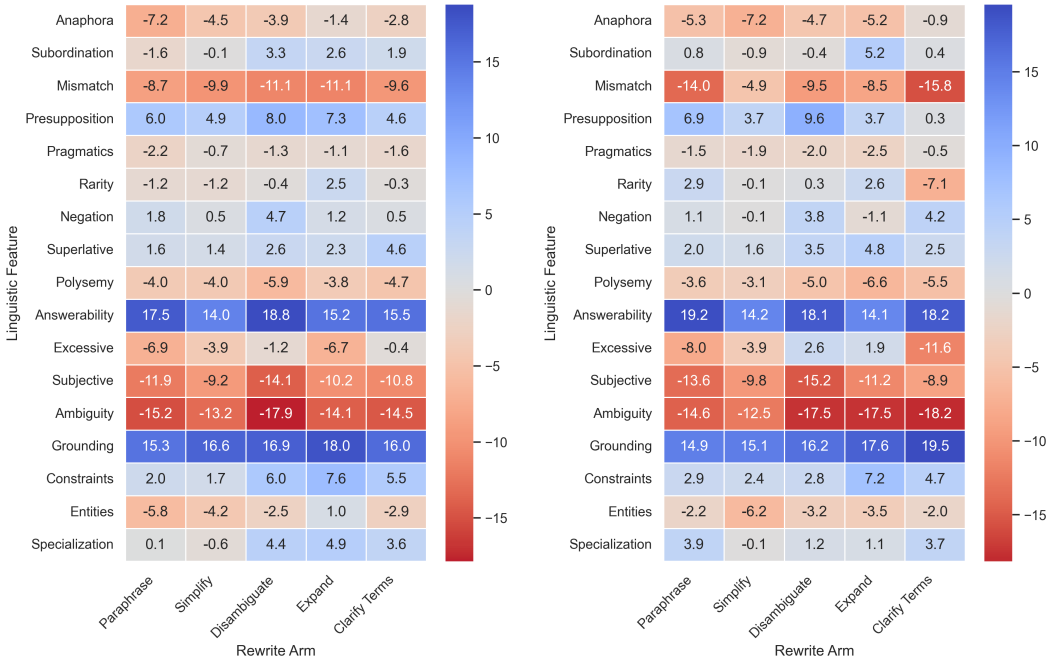
A.8 Thompson Sampling

For a given $\mathbf{x}$, sample $\tilde{\theta}a \sim \mathcal{N}(\mu_a, \Sigma_a)$ and select the arm maximizing:

$$a^* = \arg\max_{a \in \mathcal{A}} \mathbf{x}^\top \tilde{\theta}_a. \tag{6}$$

Standard Bayesian linear regression updates are then used to update μ_a and Σ_a based on the observed reward r.

$$\begin{aligned} \Sigma_a^{-1} &\leftarrow \Sigma_a^{-1} + \frac{1}{\sigma^2}\,\mathbf{x}\mathbf{x}^\top, \\ \mu_a &\leftarrow \Sigma_a\Big(\Sigma_a^{-1}\mu_a + \frac{1}{\sigma^2}\,\mathbf{x}\,r\Big). \end{aligned} \tag{7}$$



(a) Contextual Model Feature Uplift.

(b) Non-Contextual Model Feature Uplift.

Figure 13: **Reward Uplift by Contextual Feature and Strategy.** Feature Uplift measures how much the presence of a binary feature changes the expected reward for a given rewrite arm, formally $\Delta(f_i, a) = \mathbb{E}[r_t \mid \text{arm} = a, f_i = 1] - \mathbb{E}[r_t \mid \text{arm} = a, f_i = 0]$. (a) Under the **contextual** bandit, the strongest positive uplifts come from **Answerability** ($\approx$ +17 uniformly) and **Grounding** (+15–18), while **Ambiguity** ($\approx$ –15 to –18) and **Subjectivity** ($\approx$ –10 to –14) impose the largest hits across all arms. Mid-range features like **Presupposition** and **Constraints** deliver modest boosts ($\approx$ 5), and **Excessive Details** and **Anaphora** slightly hurt performance ($\approx$ –5 to –7). (b) The **non-contextual** bandit amplifies these trends: **Answerability** and **Grounding** remain the top drivers ($\approx$ +18–20), but **Ambiguity** becomes even more detrimental ($\approx$ –17 to –18), and **Mismatch** drops to nearly –16 under some arms. Notably, the non-contextual model shows a stronger negative effect for **Excessive Details** (up to –12) and **Entities** ($\approx$ –6) than the linear one, suggesting it more sharply penalizes noisy contexts. Together, these heatmaps reveal which linguistic signals each rewrite strategy leverages (or struggles with), and how context vs. context-blind policies weigh them differently.

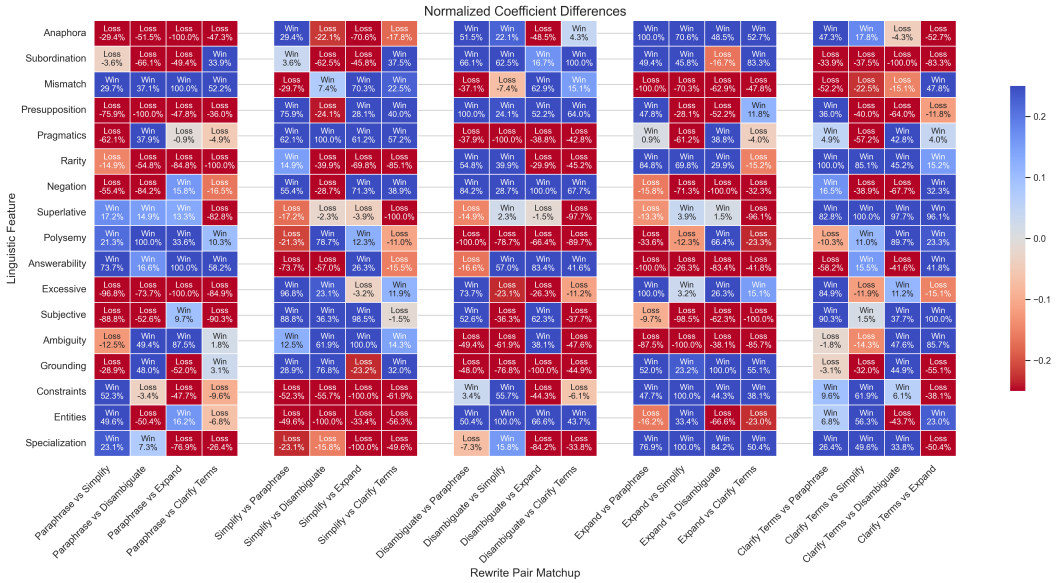


Figure 14: **Pairwise Normalized Coefficient Differences for Contextual Bandits.** Each cell shows the min–max–normalized difference in regression weight for a given linguistic feature (rows) between two rewrite arms (columns), e.g. "Paraphrase vs Disambiguate," "Simplify vs Expand," etc. Cells labeled "Win" (blue) indicate the feature favors the first arm in the matchup, while "Loss" (red) indicates it favors the second. Values are expressed as a percentage of the feature's full coefficient range.

	Feature	Definition	Example
Structural	Anaphora	Presence of pronouns or references requiring external context.	"What about that one?" (Unclear reference)
	Subordination	Measures the presence of multiple subordinate clauses	"While I was walking home, I saw a cat that looked just like my friend's."
Scenario-Based	Mismatch	Mismatch between the query's intended output and its actual structure.	"Find me this paragraph in this document" (When document isn't given, this query cannot be answered)
	Presupposition	Unstated assumptions embedded in the query.	"Who is the musician that developed neural networks?" (Assumes such a musician exists)
	Pragmatics	Captures context-dependent meanings beyond literal interpretation.	"Can you pass the salt?" (A request, not a literal ability)
Lexical	Rarity	Use of rare or niche terminology.	"What are the ramifications of quantum decoherence?" (Uses low-frequency terms)
	Negation	Presence of negation words (*not*, *never*).	"Is it not possible to do this?"
	Superlatives	Detection of superlative expressions (*biggest*, *fastest*).	"What is the fastest algorithm?"
	Polysemy	Presence of ambiguous words with multiple related meanings.	"Explain how a bank operates." (Ambiguity: financial institution vs. riverbank)
Stylistic	Answerability	Assesses whether the query has a verifiable answer.	"What is the exact number of galaxies?" (Unanswerable)
	Excessive	Evaluates whether a query is overloaded with information, potentially distracting the model.	"Can you explain how convolutional neural networks work, including all mathematical formulas?"
	Subjectivity	Query requires the degree of opinion or personal bias	"What is the best programming language?"
	Ambiguity	Highly ambiguous context, task, and wording	"Tell me about history." (Too broad)
Semantic	Grounding	Evaluates how clearly the query's purpose is expressed.	"How does reinforcement learning optimize control in robotics?" (Clear intent)
	Constraints	Identifies explicit constraints (time, location, conditions) provided in the query.	"What was the inflation rate in the US in 2023?"
	Entities	Checks for the inclusion of verifiable named entities.	"Who founded OpenAI?"
	Specialization	Determines whether the query belongs to a specialized domain (e.g., finance, law).	"What are the legal implications of the GDPR ruling?"

Table 5: Detailed Summary and Examples of Feature Categories, Definitions, and Examples (See Table 1)