Context Attribution with Multi-Armed Bandit Optimization

Anonymous ACL submission

Abstract

Understanding which parts of the retrieved context contribute to a large language model's generated answer is essential for building interpretable and trustworthy retrieval-augmented generation. We propose a novel framework that formulates context attribution as a combinatorial multi-armed bandit problem. We utilize Linear Thompson Sampling to efficiently identify the most influential context segments while minimizing the number of model queries. Our reward function leverages token log-probabilities to measure how well a subset of segments supports the original response, making it applicable to both open-source and black-box API-based models. Unlike SHAP and other perturbation-based methods that sample subsets uniformly, our approach adaptively prioritizes informative subsets based on posterior estimates of segment relevance, reducing computational costs. Experiments on multiple QA benchmarks demonstrate that our method achieves up to 30% reduction in model queries while matching or exceeding the attribution quality of existing approaches.

1 Introduction

Retrieval-Augmented Generation (RAG) has become a dominant approach for knowledge-intensive question answering tasks, augmenting Large Language Models (LLMs) with external context to improve factual accuracy and credibility (Gao et al., 2023b). Despite its effectiveness, ensuring that generated answers are genuinely grounded in the provided context remains challenging. LLMs frequently produce hallucinations or incorporate ungrounded information, making it essential to verify and attribute precisely which context segments are responsible for their responses (Gao et al., 2023a).

Existing approaches to enhancing attribution primarily follow two paradigms. The first involves training models to explicitly cite context segments during generation (Nakano et al., 2021; Menick et al., 2022; Zhang et al., 2024; Huang et al., 2024). While such techniques improve self-attribution, their reliability remains **unverifiable**, as there is no guarantee that the generated citations actually reflect the context relied upon by the model during inference. The second paradigm focuses on post-hoc methods, such as ContextCite (Cohen-Wang et al., 2024) and MExGen (Paes et al., 2024), which systematically perturb or mask context segments and evaluate their impact on the output. Traditional methods like LIME (Ribeiro et al., 2016) and SHAP (Lundberg and Lee, 2017) also fall into this category with proper adaptation. Although these methods offer greater fidelity by probing actual input-output behavior, they often incur prohibitive computational costs due to extensive sampling and expensive LLM inference, rendering them impractical for long-context scenarios.

To address these limitations, we formulate context attribution as a Combinatorial Multi-Armed Bandit (CMAB) problem. Each context segment is treated as an arm, and selecting a subset of segments constitutes an action. The goal is to identify the most influential segments within a limited query budget. To efficiently navigate this combinatorial space, we introduce Combinatorial **Linear Thompson Sampling (LinTS)**, a Bayesian bandit method renowned for balancing exploration and exploitation (Agrawal and Goyal, 2013). Unlike exhaustive or uniformly random perturbation strategies, our approach significantly reduces the number of model evaluations, making it practical for long-context applications.

Our key contributions are summarized as follows: 1) we formulate segment-level context attribution as a combinatorial multi-armed bandit (CMAB) problem; 2) we utilize Linear Thompson Sampling (LinTS) to adaptively explore the space of context subsets, improving query efficiency under limited budgets; and 3) we conduct comprehen-

sive experiments on three diverse datasets using two widely-used LLMs accessed via commercial APIs, demonstrating the effectiveness of our approach in realistic black-box settings.

2 Proposed Method

In this section, we present **CAMAB** (Context Attribution with Multi-Armed Bandit), a framework that treats segment-level attribution as an efficient Bayesian optimization process. The core of our approach is to formulate the attribution task as a Combinatorial Multi-Armed Bandit problem. CAMAB leverages **Linear Thompson Sampling (LinTS)** to model the joint contribution of context subsets. It achieves querying efficiency by exploring the context subset space adaptively.

2.1 Problem Formulation

Context-Supported Generative QA We consider a scenario where an LLM is tasked with answering a question Q using a provided context C. The context C consists of N discrete segments, $C = \{s_1, s_2, \ldots, s_N\}$, such as sentences or paragraphs. The LLM (denoted by M) produces a response R consisting of a sequence of tokens $R = (r_1, r_2, \ldots, r_T)$, generated autoregressively based on the input (Q, C).

Reward Modeling via Context Subsets The central insight is to gauge a segment's importance by observing the model's output probability when that segment is included in or excluded from various subsets. Let $S \subseteq C$ denote a subset of the context segments. For any subset S, we define the reward $V(S)$ as the average log-probability of the original response tokens R given S:

$$V(S) = \frac{1}{T} \sum_{t=1}^{T} \log P_M(r_t \mid Q, S, r_1, \ldots, r_{t-1}) \tag{1}$$

This reward formulation serves as our metric for segment importance. While alternative definitions exist as noted by (Paes et al., 2024), for simplicity and consistency, we adopt Eq 1 as the standard reward throughout this work unless explicitly stated otherwise.

Context Attribution Given the reward modeling, our goal is to attribute the content of model response R back to specific segments in context C by assigning an attribution vector $\boldsymbol{a}$, where each score a_j reflects the marginal contribution of segment s_j to the generation of R.

2.2 Bandit Formulation with Linear Thompson Sampling

Combinatorial Multi-Armed Bandit We formulate context attribution as a Combinatorial Multi-Armed Bandit (CMAB) problem. In this setting, the N context segments $\{s_1, \ldots, s_N\}$ serve as the base arms of the bandit. An action corresponds to selecting a subset of segments $S \subseteq C$ (referred to as a *super-arm*), and the observed feedback is the scalar reward $V(S)$, which reflects the model's response quality given the subset S.

To efficiently estimate segment importance, we adopt a linear structural assumption. We posit that the reward $V(S)$ can be approximated as the sum of the marginal contributions of the individual segments included in S, plus a bias term representing the model's baseline performance given an empty context. Specifically, for a selected subset S, we define a feature vector $\boldsymbol{x} \in \{0, 1\}^{N+1}$ to represent the combinatorial action:

$$\boldsymbol{x} = [1, \mathbb{I}(s_1 \in S), \ldots, \mathbb{I}(s_N \in S)]^\top \tag{2}$$

where the first element is a bias term (intercept) representing the model's performance with an empty context, and $\mathbb{I}(\cdot)$ is the indicator function for segment inclusion. We model the reward as a linear function of the features with additive Gaussian noise:

$$V(S) = \boldsymbol{w}^\top \boldsymbol{x} + \epsilon, \tag{3}$$

where $\epsilon \sim \mathcal{N}(0, \sigma^2)$ is the observation noise and $\boldsymbol{w} = [w_0, w_1, \ldots, w_N]^\top$ represents the latent weights. In this formulation, w_0 captures the baseline log-likelihood, and w_j for $j > 0$ denotes the specific contribution of segment s_j. The final learned weights $\{w_1, \ldots, w_N\}$ serve as our attribution score vector $\boldsymbol{a}$.

Linear Thompson Sampling To solve this problem under a strictly limited query budget, we employ Linear Thompson Sampling (LinTS) (Agrawal and Goyal, 2013). LinTS is a sample-efficient algorithm that balances exploration and exploitation by maintaining a probabilistic belief over the unknown weight vector $\boldsymbol{w}$. Assuming a Gaussian prior $\boldsymbol{w} \sim \mathcal{N}(\hat{\boldsymbol{\mu}}_0, \boldsymbol{B}_0^{-1})$ and Gaussian observation noise $\epsilon \sim \mathcal{N}(0, \sigma^2)$, the posterior distribution at any round t remains a multivariate Gaussian due to conjugacy:

$$P(\boldsymbol{w} \mid \mathcal{H}_t) = \mathcal{N}(\hat{\boldsymbol{\mu}}_t, \boldsymbol{B}_t^{-1}) \tag{4}$$

Table 1: BERTScore performance (lower is better) of **CAMAB**, **ContextCite**, and **SHAP** under varying query budgets ($s = 20, 40, 60$) across datasets using the LLaMA3.1-8B model. Each s denotes the total number of LLM calls used. Bold indicates the best performance.

Data	k	**CAMAB**			**ContextCite**			**SHAP**		
		$s = 20$	$s = 40$	$s = 60$	$s = 20$	$s = 40$	$s = 60$	$s = 20$	$s = 40$	$s = 60$
Hotpot QA	1	0.525	**0.509**	0.511	0.605	0.601	0.598	0.668	0.562	0.527
	3	0.464	0.421	**0.418**	0.549	0.537	0.529	0.598	0.471	0.444
	5	0.445	0.407	**0.402**	0.527	0.496	0.499	0.562	0.453	0.423
CNN/ DM	1	0.642	**0.613**	0.614	0.671	0.668	0.670	0.648	0.617	**0.613**
	3	0.526	0.485	0.480	0.544	0.535	0.532	0.532	0.483	**0.478**
	5	0.452	0.405	0.406	0.479	0.459	0.454	0.455	0.406	**0.399**
TyDi QA	1	0.479	0.473	**0.470**	0.542	0.539	0.536	0.542	0.488	0.476
	3	0.392	**0.378**	0.381	0.474	0.468	0.458	0.455	0.400	0.379
	5	0.371	0.345	**0.343**	0.452	0.442	0.440	0.420	0.368	0.353

Algorithm 1 CAMAB: Context Attribution with Linear Thompson Sampling

Require: Context segments $C = \{s_1, \ldots, s_N\}$, query Q, response R, budget $T_{\max}$, prior mean $\hat{\boldsymbol{\mu}}_0$, prior variance σ_p^2, noise variance σ^2
Ensure: Attribution scores $\{w_1, \ldots, w_N\}$
1: **Initialize:** $\boldsymbol{B}_0 \leftarrow \frac{1}{\sigma_p^2}\boldsymbol{I}$, $\boldsymbol{f}_0 \leftarrow \boldsymbol{B}_0\hat{\boldsymbol{\mu}}_0$
2: **for** $t = 1$ to $T_{\max}$ **do**
3: Sample $\tilde{\boldsymbol{w}}^{(t)} \sim \mathcal{N}\left(\hat{\boldsymbol{\mu}}_{t-1}, \boldsymbol{B}_{t-1}^{-1}\right)$
4: Construct $\boldsymbol{x}_t$ and form subset S_t
5: Query LLM and observe reward $v_t = V(S_t)$ via Eq. 1
6: Update $\boldsymbol{B}_t \leftarrow \boldsymbol{B}_{t-1} + \frac{1}{\sigma^2}\boldsymbol{x}_t\boldsymbol{x}_t^\top$
7: Update $\boldsymbol{f}_t \leftarrow \boldsymbol{f}_{t-1} + \frac{v_t}{\sigma^2}\boldsymbol{x}_t$
8: Compute $\hat{\boldsymbol{\mu}}_t \leftarrow \boldsymbol{B}_t^{-1}\boldsymbol{f}_t$
9: **end for**
10: **return** $\{\hat{\mu}_1, \ldots, \hat{\mu}_N\}$ from $\hat{\boldsymbol{\mu}}_{T_{\max}}$

where $\mathcal{H}_t$ denotes the history of observed actions and rewards up to round t. The algorithm proceeds by iteratively sampling a plausible weight vector $\tilde{\boldsymbol{w}}^{(t)}$ from this posterior to guide the selection of the next informative context subset.

Furthermore, although we rely on a linear assumption, *the precision matrix $\boldsymbol{B}_t$ allows the model to implicitly capture interactions*. By tracking correlations between segments, $\boldsymbol{B}_t$ ensures that if a subset of segments consistently performs well together, this interaction effect is absorbed and reflected in their respective individual weights.

Algorithm 1 summarizes our approach. At each round t, we first sample a weight vector $\tilde{\boldsymbol{w}}^{(t)}$ from the current posterior distribution (Line 3). This sample reflects our current belief about segment contributions while incorporating uncertainty. We then construct the context subset S_t by selecting all segments whose sampled weights are positive (Line 4), effectively choosing segments that are likely to contribute positively to the response:

$$S_t = \{s_j : \tilde{w}_j^{(t)} > 0\} \tag{5}$$

The corresponding feature vector is constructed as:

$$\boldsymbol{x}_t = [1, \mathbb{I}(\tilde{w}_1^{(t)} > 0), \ldots, \mathbb{I}(\tilde{w}_N^{(t)} > 0)]^\top \tag{6}$$

After querying the LLM with subset S_t to observe the reward v_t (Line 5), we perform a Bayesian update: the precision matrix $\boldsymbol{B}_t$ accumulates information about which segment combinations have been tested, while the information vector $\boldsymbol{f}_t$ aggregates the observed rewards weighted by the corresponding feature vectors (Lines 6-8). This posterior update naturally balances exploration of uncertain segments with exploitation of segments already identified as important.

After $T_{\max}$ rounds, the posterior mean weights $\hat{\mu}_1, \ldots, \hat{\mu}_N$ (excluding the intercept $\hat{\mu}_0$) are used as the final attribution scores. Segments are ranked by these scores to identify the most influential context components.

3 Experiments

We evaluate CAMAB against **SHAP** (Lundberg and Lee, 2017), **ContextCite** (Cohen-Wang et al., 2024), and a random baseline on three diverse benchmarks—HotpotQA (Yang et al., 2018), CNN/DailyMail (See et al., 2017), and TyDi QA (Clark et al., 2020)—using a random subset of 10,000 samples from each.

Notably, ContextCite's official implementation relies on full logit-based reward, while we set CAMAB and SHAP operate solely on token-level log-probabilities (Eq 1). We employ two state-of-the-art models: LLaMA-3.1-8B (Grattafiori et al., 2024) and Qwen2.5-7B (Qwen et al., 2025). Performance is assessed using **Log-Probability Drop** (higher is better ↑) and **BERTScore** (lower is better ↓). Detailed experimental settings are provided in Appendix B.

3.1 Attribution Quality on Different Tasks

Table 2 presents attribution performance under a tight budget of 40 queries. We observe that CAMAB consistently outperforms or matches the baselines across all datasets. Specifically, on information-seeking tasks such as HotpotQA and TyDi QA, CAMAB achieves superior performance in both Top-k Log-Probability Drop and BERTScore metrics for nearly all k values. This demonstrates that our bandit-based approach is highly effective at pinpointing the critical evidence sentences required for factual reasoning.

On the abstractive summarization task (CNN/DailyMail), CAMAB remains highly competitive, achieving performance almost identical to SHAP with a difference of less than 1% across all metrics. We attribute this to the strong *lead bias* in news summarization, where key information is concentrated at the beginning and end of articles. This structure reduces ambiguity, allowing different attribution methods to converge rapidly. This hypothesis is corroborated by Table 1, where all methods stabilize within 40 queries.

Consistent trends are observed for the Qwen-2.5-8B model; detailed results are provided in Appendix B.

3.2 Attribution with Limited Query Budgets

In realistic scenarios, query budgets are strictly constrained due to the high latency and monetary cost of LLM inference. To evaluate the sample efficiency of our framework, we compare CAMAB against ContextCite and SHAP across three query budgets: $s \in \{20, 40, 60\}$. Table 1 reports the BERTScore (lower is better) on LLaMA-3.1-8B.

We observe three key trends. **First, CAMAB demonstrates superior sample efficiency.** In information-seeking tasks (HotpotQA and TyDi QA), CAMAB at $s = 40$ often outperforms SHAP at $s = 60$. This indicates that the active exploration of LinTS identifies critical segments significantly faster than random perturbations. **Second, the performance gap narrows as the query budget increases.** While SHAP eventually converges to competitive results when sufficient queries are available (e.g., $s = 60$), CAMAB achieves near-optimal performance significantly earlier. At $s = 20$, SHAP's performance degrades sharply, whereas CAMAB maintains high fidelity even in these extremely resource-constrained settings. **Finally, CAMAB consistently outperforms ContextCite.** ContextCite struggles to identify influential segments under tight budgets, likely because Lasso regression requires a larger sample size to effectively select features in high-dimensional spaces.

These results validate that CAMAB is the most robust choice for attribution when computational resources are limited.

Table 2: Evaluation results on LLaMA-3.1-8B with querying budget of 40. Abbreviations are used for conciseness (C-Cite: ContextCite, Rand: Random, Log-P: Log-Probability Drop, BERT: BERTScore).

Data	Metric	k	CAMAB	SHAP	C-Cite	Rand
Hotpot QA	Log-P ↑	1	**0.521**	0.475	0.429	0.024
		3	**0.676**	0.614	0.591	0.062
		5	**0.717**	0.648	0.632	0.103
	BERT ↓	1	**0.509**	0.562	0.601	0.803
		3	**0.421**	0.471	0.537	0.741
		5	**0.407**	0.453	0.496	0.703
CNN/ DM	Log-P ↑	1	**0.400**	0.398	0.358	0.073
		3	0.840	**0.843**	0.801	0.224
		5	**1.129**	1.041	1.025	0.389
	BERT ↓	1	**0.613**	0.617	0.668	0.734
		3	0.485	**0.483**	0.535	0.656
		5	**0.405**	0.406	0.459	0.601
TyDi QA	Log-P ↑	1	**0.596**	**0.596**	0.429	0.069
		3	**0.813**	0.803	0.579	0.240
		5	**0.893**	0.872	0.631	0.373
	BERT ↓	1	**0.473**	0.488	0.539	0.719
		3	**0.378**	0.400	0.468	0.624
		5	**0.345**	0.368	0.442	0.566

4 Conclusion

We introduced CAMAB, a novel context attribution framework that formulates attribution as a combinatorial multi-armed bandit problem and applies Linear Thompson Sampling to efficiently attribute the contexts. Empirical results demonstrate that CAMAB delivers superior attribution fidelity with significantly fewer queries, establishing it as a scalable solution for low-resource generative QA systems.

Limitations

While CAMAB demonstrates strong performance under constrained query budgets, it is primarily designed for scenarios where query efficiency is critical, such as black-box attribution for large generative language models. However, this approach may converge to suboptimal local solutions if the exploration-exploitation balance is not well maintained, especially in highly noisy or ambiguous settings. In situations where budget is not a constraint, traditional perturbation-based methods such as SHAP and ContextCite can benefit from broader context exploration and may ultimately yield more plausible attributions.

References

Shipra Agrawal and Navin Goyal. 2013. Thompson sampling for contextual bandits with linear payoffs. In *International conference on machine learning*, pages 127–135. PMLR.

Hanjie Chen, Guangtao Zheng, and Yangfeng Ji. 2020. Generating hierarchical explanations on text classification via feature interaction detection. *arXiv preprint arXiv:2004.02015*.

Jonathan H Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. 2020. Tydi qa: A benchmark for information-seeking question answering in ty pologically di verse languages. *Transactions of the Association for Computational Linguistics*, 8:454–470.

Benjamin Cohen-Wang, Harshay Shah, Kristian Georgiev, and Aleksander Madry. 2024. Contextcite: Attributing model generation to context. *Advances in Neural Information Processing Systems*, 37:95764–95807.

Yue Dong, Yikang Shen, Eric Crawford, Herke van Hoof, and Jackie Chi Kit Cheung. 2018. Banditsum: Extractive summarization as a contextual bandit. *arXiv preprint arXiv:1809.09672*.

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. 2023a. Enabling large language models to generate text with citations. *arXiv preprint arXiv:2305.14627*.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yixin Dai, Jiawei Sun, Haofen Wang, and Haofen Wang. 2023b. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2:1.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Chengyu Huang, Zeqiu Wu, Yushi Hu, and Wenya Wang. 2024. Training language models to generate text with citations via fine-grained rewards. *arXiv preprint arXiv:2402.04315*.

Neil Jethani, Mukund Sudarshan, Ian Connick Covert, Su-In Lee, and Rajesh Ranganath. 2021. Fastshap: Real-time shapley value estimation. In *International conference on learning representations*.

Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Advances in neural information processing systems*, pages 4765–4774.

Jacob Menick, Maja Trebacz, Vladimir Mikulik, John Aslanides, Francis Song, Martin Chadwick, Mia Glaese, Susannah Young, Lucy Campbell-Gillingham, Geoffrey Irving, and 1 others. 2022. Teaching language models to support answers with verified quotes. *arXiv preprint arXiv:2203.11147*.

Sumana Sanyasipura Nagaraju. 2025. Automation and feature selection enhancement with reinforcement learning (rl).

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, and 1 others. 2021. Webgpt: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*.

Lucas Monteiro Paes, Dennis Wei, Hyo Jin Do, Hendrik Strobelt, Ronny Luss, Amit Dhurandhar, Manish Nagireddy, Karthikeyan Natesan Ramamurthy, Prasanna Sattigeri, Werner Geyer, and 1 others. 2024. Multi-level explanations for generative language models. *arXiv preprint arXiv:2403.14459*.

Deng Pan, Xin Li, and Dongxiao Zhu. 2021. Explaining deep neural network models with adversarial gradient integration. In *Thirtieth International Joint Conference on Artificial Intelligence (IJCAI)*.

Deng Pan, Nuno Moniz, and Nitesh Chawla. 2025. Fast explanations via policy gradient-optimized explainer. *Preprint*, arXiv:2405.18664.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, and 25 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144.

Abigail See, Peter J Liu, and Christopher D Manning. 2017. Get to the point: Summarization with pointer-generator networks. *arXiv preprint arXiv:1704.04368*.

Mahesh Sudhakar, Sam Sattarzadeh, Konstantinos N Plataniotis, Jongseong Jang, Yeonjeong Jeong, and Hyunwoo Kim. 2021. Ada-sise: adaptive semantic input sampling for efficient explanation of convolutional neural networks. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1715–1719. IEEE.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W Cohen, Ruslan Salakhutdinov, and Christopher D Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. *arXiv preprint arXiv:1809.09600*.

Jiajie Zhang, Yushi Bai, Xin Lv, Wanjun Gu, Danqing Liu, Minhao Zou, Shulin Cao, Lei Hou, Yuxiao Dong, Ling Feng, and 1 others. 2024. Longcite: Enabling llms to generate fine-grained citations in long-context qa. *arXiv preprint arXiv:2409.02897*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Appendix

A Related Work

A.1 Perturbation-Based Attribution Methods

A large number of works on model interpretability uses input perturbations to infer feature importance. Techniques like **LIME** (Ribeiro et al., 2016) and **SHAP** (Lundberg and Lee, 2017) interpret model predictions by evaluating the model on perturbations of an input and observing how the output changes. LIME fits a local surrogate model (e.g. a linear model) around the neighborhood of the input to estimate each feature's influence. SHAP uses a game-theoretic approach to approximate Shapley values, which quantify each feature's contribution to the prediction in a way that satisfies fairness axioms. These methods are *model-agnostic* and fairly *faithful* in a local sense, but they are notoriously expensive: they require sampling a large number of perturbations for each instance to obtain stable estimates. This cost grows with input dimensionality, making them difficult to apply to settings like long text sequences without sparing accuracy.

A.2 Attribution Methods in LLM Settings

In the context of large language models (LLMs), token-level attribution faces significant challenges due to (1) the combinatorially large perturbation space induced by extensive input contexts, and (2) substantial computational costs associated with individual model queries. Therefore, effective attribution methods must carefully balance explanation fidelity against computational efficiency. Broadly, three strategies have emerged to tackle these challenges:

(i) **Reducing Perturbation Space**: Methods in this category aggregate tokens into higher-level semantic units, such as phrases, sentences, or paragraphs, effectively decreasing the perturbation space. Different granularity levels inherently capture varying degrees of semantic meaning, naturally supporting hierarchical attribution structures. For example, (Chen et al., 2020) propose a divide-and-conquer strategy that progressively attributes importance from sentence-level down to token-level in text classification. Similarly, MExGen (Paes et al., 2024) systematically extends perturbation-based methods like LIME and SHAP to generative LLMs, efficiently identifying influential text spans in a hierarchical manner. ContextCite (Cohen-Wang et al., 2024) specifically targets segment-level attribution in generative QA scenarios by using SHAP-based perturbation techniques.

(ii) **Pretrained Global Explainers**: Another strategy involves training a global surrogate model as a pretrained explainer, trading upfront training costs for reduced inference latency during explanations. Examples include FEX (Pan et al., 2025), which employs policy gradient methods to optimize a Bernoulli surrogate explainer, and FastSHAP (Jethani et al., 2021), which fits a neural network using pseudo-labels derived from SHAP values. Despite their inference efficiency, these methods demand substantial pretraining resources and extensive datasets, and like typical machine learning models, they often encounter generalization challenges when confronted with out-of-distribution samples.

(iii) **Optimizing Perturbation Strategies**: A relatively nascent direction focuses on explicitly optimizing perturbation strategies to reduce the number of required model queries. (Sudhakar et al., 2021) leverage heuristics derived from input-to-output gradients to selectively perturb features, whereas (Pan et al., 2021) propose subsequent perturbations aligned with adversarial attack directions. However, such methods typically require internal model knowledge, including gradients or manifold structures, limiting their applicability to models treated

as black boxes. Motivated by this gap, our work introduces a novel perturbation sampling method inspired by multi-armed bandit algorithms, enabling dynamic adjustment of perturbations based solely on observed responses, without necessitating internal model information.

A.3 Bandit and Reinforcement Learning Approaches

Feature attribution can be viewed as a task of selecting the most informative subsets of features. The multi-armed bandit and reinforcement learning can be utilized to progressively optimize and sequentially search the subsets with a limited budget of actions. Feature selection via multi-armed bandits has been explored in prior research as a way to dynamically identify important features without evaluating all subsets. For example, (Nagaraju, 2025) propose a feature selection method that uses an Upper Confidence Bound (UCB) bandit algorithm. This allows the algorithm to rapidly converge to a near-optimal feature set, yielding good predictors with fewer feature evaluations. In NLP tasks, BanditSum (Dong et al., 2018) treated extractive summarization as a contextual bandit problem: given a document (context), their model learned via policy gradient to pick a sequence of sentences (the "action") that maximizes the summary quality reward (ROUGE score) This is an example of using reinforcement learning to select informative subsets of text. Although BanditSum was focused on training a summarization model, the idea of using reward feedback to guide text segment selection is closely related to our approach for attribution.

B Additional Experiment Details and Results

B.1 Datasets

We evaluate our framework on three representative language generation benchmarks that cover distinct task types and context structures. HotpotQA targets sentence-level attribution in multi-hop question answering; CNN/DailyMail emphasizes sentence-level attribution for long-document summarization; and TyDi QA provides a diverse, multilingual context for evaluating information-seeking question answering. For computational feasibility, we randomly sample 1,000 validation instances from each dataset.

HotpotQA (Yang et al., 2018) HotpotQA is a multi-hop question answering benchmark requiring reasoning over multiple supporting documents to answer factoid questions. Each instance includes long passages, and the responses are more elaborate than standard QA tasks. We use sentences as the segments of interest for attribution.

CNN/DailyMail (See et al., 2017) CNN/DailyMail is a large-scale abstractive summarization dataset where the task is to generate concise summaries of news articles. Contexts consist of long documents containing narrative and factual information, and the outputs are multi-sentence summaries. We select sentences as the segments of interest.

TyDi QA (Clark et al., 2020) TyDi QA is a benchmark for information-seeking question answering, grounded in Wikipedia contexts across multiple typologically diverse languages. Unlike the multi-hop reasoning in HotpotQA, TyDi QA focuses on identifying specific evidence from a provided context to answer user questions. We treat individual sentences as the segments of interest to evaluate the precision of our attribution framework in factual retrieval scenarios.

B.2 Models

To evaluate the generality and robustness of our context attribution framework, we conduct experiments with two recent large language models that differ in their training corpora and design philosophies. To reflect a realistic black-box attribution scenario, these models are accessed via commercial APIs, where internal model states and full logit distributions are unavailable, and only token-level log-probabilities are provided.

LLaMA-3.1-8B. We use the 8B version of LLaMA 3.1 (Grattafiori et al., 2024), a decoder-only transformer released by Meta AI. Pretrained on a vast and diverse corpus with a next-token prediction objective, LLaMA-3.1 excels at long-context reasoning and producing fluent, high-quality responses. In our experiments, it serves as a high-capacity baseline to test attribution fidelity in complex reasoning tasks.

Qwen2.5-7B. We include Qwen2.5-7B (Qwen et al., 2025), a decoder-only transformer developed by Alibaba. Trained on large-scale multilingual and multimodal data, Qwen2.5 demonstrates strong cross-lingual generalization and represents a distinct pretraining paradigm from LLaMA-3.1. Including this model allows us to evaluate the consis-

tency of our attribution framework across different architectural families and training objectives.

Overall, these two models provide a complementary testbed for evaluating CAMAB in a black-box setting, representing high-capacity systems commonly deployed in real-world knowledge-intensive applications.

B.3 Settings and Baselines

We compare CAMAB against four representative post-hoc attribution baselines, each reflecting a different strategy for identifying influential context segments. To ensure a fair comparison, all methods are evaluated under a strictly constrained query budget of $T = 40$ LLM calls per instance.

CAMAB (Ours) We implement CAMAB using Linear Thompson Sampling (LinTS) as described in Section 2. The algorithm is initialized with a prior mean of 0.0 and a prior variance of 1.0. The observation noise variance is set to $\sigma^2 = 0.01$. Our implementation employs an **adaptive selection** strategy: in each round, it selects all segments whose sampled weights $\tilde{\theta}_j$ are positive. This allows the model to dynamically adjust the size of the context subset based on its current posterior beliefs.

SHAP SHAP (Lundberg and Lee, 2017) is a model-agnostic explainer grounded in Shapley values. We use the KernelSHAP variant at the segment level. To comply with the query budget, we use a single fully masked context as the reference baseline and limit the number of perturbed samples to 40. The reward signal is the average log-likelihood of the response tokens, consistent with our method.

ContextCite ContextCite (Cohen-Wang et al., 2024) attributes generation by measuring the average log-odds change when subsets of segments are ablated. It fits a sparse LASSO regression model to these observations to identify relevant segments. To ensure parity in computational cost, we limit its sampling to 40 ablated subsets.

Random As a baseline for attribution effectiveness, the Random explainer assigns importance scores by sampling from a uniform distribution. This provides a lower-bound reference to demonstrate that the performance gains of other methods are due to principled exploration rather than random perturbations.

B.4 Experiment Settings

To simulate a realistic black-box scenario, our experiments for CAMAB and the baselines SHAP and Random are conducted using commercial APIs, where only token-level log-probabilities are accessible. Contextcite, on the otherhand, requires full access to the model's logit distributions to perform its regression-based attribution, hence we conduct its experiments on a local machine. The details of the setup can be found in Appendix.

This setup reflects the typical constraints of state-of-the-art models where internal states are unavailable.

In contrast, ContextCite requires full access to the model's logit distributions to perform its regression-based attribution. Consequently, the experiments for ContextCite were conducted on a local computing server equipped with an NVIDIA A100 GPU (80GB). To ensure a fair comparison, all methods are evaluated under strictly constrained query budgets, controlling for the total number of LLM calls across all attribution processes.

B.5 Evaluation Metrics

To assess the effectiveness of our context attribution method, we adopt two evaluation metrics: *Top-k Log-Probability Drop* and *BERTScore Consistency*.

Top-k Log-Probability Drop. (Cohen-Wang et al., 2024) This metric evaluates the degradation of the model response's likelihood when the most influential segments are removed. Let $L(S)$ denote the average log-likelihood of the response R given context subset S:

$$L(S) = \frac{1}{T} \sum_{t=1}^{T} \log P_M(r_t \mid Q, S, r_{<t}) \quad (7)$$

The Top-k log-probability drop is then defined as the difference in likelihood between the full context C and the perturbed subset $S_{\text{top-}k}(\tau)$:

$$\text{Top-}k\text{-drop} = L(C) - L(S_{\text{top-}k}(\tau)) \quad (8)$$

A larger drop implies that the removed segments were more supportive of the generation, indicating higher attribution accuracy.

BERTScore Consistency. This metric evaluates attribution fidelity by measuring the semantic difference between the original response $R = (r_1, \ldots, r_T)$, generated using the full context

Table 3: Evaluation results on Qwen2.5-7B. Top-k means the specific evaluation is done with top k attributed segments removed.

Dataset	Metrics	Top-k	CAMAB	SHAP	ContextCite	Random
HotpotQA	Log-Prob Drop ↑	$k=1$	**0.650**	0.628	0.541	0.041
		$k=3$	**0.917**	0.879	0.802	0.107
		$k=5$	**0.972**	0.948	0.867	0.184
	BERTScore ↓	$k=1$	**0.486**	0.504	0.490	0.782
		$k=3$	**0.372**	0.394	0.393	0.704
		$k=5$	**0.355**	0.375	0.370	0.649
CNN/DailyMail	Log-Prob Drop ↑	$k=1$	**0.442**	0.393	0.365	0.112
		$k=3$	**0.998**	0.910	0.837	0.330
		$k=5$	**1.371**	1.285	1.192	0.570
	BERTScore ↓	$k=1$	**0.613**	0.624	0.632	0.712
		$k=3$	**0.487**	0.504	0.542	0.631
		$k=5$	**0.415**	0.430	0.478	0.570
TyDi QA	Log-Prob Drop ↑	$k=1$	0.732	**0.738**	0.496	0.107
		$k=3$	**0.994**	0.992	0.719	0.269
		$k=5$	**1.081**	1.072	0.792	0.436
	BERTScore ↓	$k=1$	**0.430**	0.448	0.509	0.694
		$k=3$	**0.348**	0.362	0.443	0.605
		$k=5$	**0.325**	0.347	0.423	0.541

C, and the response R', generated using the perturbed context $S_{\text{top-}k}(\tau)$. We compute the BERTScore (Zhang et al., 2019) between the two responses as:

$$\text{BERTScore} = \text{BERTScore}\left(R', R\right) \tag{9}$$

A lower BERTScore indicates a greater semantic shift caused by the ablation, suggesting that the removed segments were more influential. Thus, lower values reflect more accurate attribution.

B.6 Additional Results

From Table 3, for the Qwen-2.5-7B model, CAMAB consistently outperforms both SHAP and ContextCite across nearly all datasets and metrics.