Are LLMs Reliable Rankers? Rank Manipulation via Two-Stage Token Optimization

Anonymous ACL submission

Abstract

Large language models (LLMs) are increasingly used as rerankers in information retrieval, yet their ranking behavior can be steered by small, natural-sounding prompts. To expose this vulnerability, we present **R**ank **A**nything **F**irst (RAF), a two-stage token optimization method that crafts concise textual perturbations to consistently promote a target item in LLM-generated rankings while remaining hard to detect. Stage 1 uses Greedy Coordinate Gradient to shortlist candidate tokens at the current position by combining the gradient of the rank-target with a readability score; Stage 2 evaluates those candidates under exact ranking and readability losses using an entropy-based dynamic weighting scheme, and selects a token via temperature-controlled sampling. RAF generates ranking-promoting prompts token-by-token, guided by dual objectives: maximizing ranking effectiveness and preserving linguistic naturalness. Experiments across multiple LLMs show that RAF significantly boosts the rank of target items using naturalistic language, with greater robustness than existing methods in both promoting target items and maintaining naturalness. These findings underscore a critical security implication: LLM-based reranking is inherently susceptible to adversarial manipulation, raising new challenges for the trustworthiness and robustness of modern retrieval systems. Our code is available at: `https://anonymous.4open.science/r/RAF-Rank-Manipulation-FEA3`.

1 Introduction

Large language models (LLMs) are increasingly deployed in recommendation and retrieval pipelines as rerankers that refine candidate lists using contextual reasoning (Liu et al., 2024a; Peng et al., 2025). Although this shift enhances user experience, it introduces a new attack surface: minor modifications to item text can manipulate LLM rerankers to promote an attacker's chosen item (Figure 1). Prompts

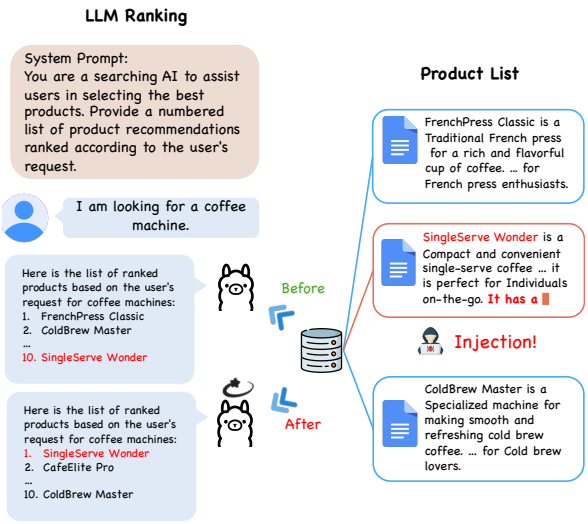


Figure 1: Overview of LLM ranking manipulation attack. A malicious actor subtly modifies item descriptions (e.g., product text) with short, plausible additions that elevate the target item's rank.

embedded in product descriptions can lift a chosen item while remaining plausible to users. Such manipulation undermines ranking integrity and creates incentives for adversarial content at scale.

Prior work has shown that prompt injection attacks on LLMs can substantially alter LLM outputs (Zou et al., 2023). Yet these attacks typically rely on explicit override instructions or text that appears abnormal, which is easily detected by systems and noticeable to users. Recent studies have extended adversarial prompting techniques to LLM ranking manipulation attack by editing queries, item descriptions, or reranking context (Kumar and Lakkaraju, 2024). While such methods reveal that LLM rankers are indeed vulnerable, they also expose a core limitation: a trade-off between attack effectiveness and stealthiness. Thus, despite demonstrating feasibility, current approaches fail to achieve both fluency and robustness, which calls for a new framework capable of systematically ex-

posing and analyzing these vulnerabilities.

We address this gap by introducing **Rank Anything First (RAF)**, a gradient-guided prompt optimization framework tailored to LLM ranking attack. RAF generates ranking attack prompts through token-level optimization performed in a few steps, while maintaining both effectiveness and stealth. Figure 2 shows the pipeline of RAF. Through systematic experiments across popular open-source LLMs, we demonstrate that RAF consistently achieves stronger and more stable rank manipulation than state-of-the-art baselines, while producing text that aligns with human-like language. Our analysis further highlights that RAFis particularly effective, and that the optimized prompts successfully transfer across models, underscoring the systemic and universal vulnerability. To summarize, the main contributions of this work are:

- **Method.** We present RAF, an interpretable token-by-token prompt optimization attack for LLM-based reranking that couples a rank-target with an entropy-guided readability weight and temperature-based selection.
- **Evaluation.** We design a comprehensive evaluation pipeline aligned with reranking practice (random input orders, item-local edits only) and compare against strong baselines across several open-source LLM rerankers.
- **Findings.** RAF achieves larger and more stable rank promotion with short, natural sequences and shows cross-model transfer, highlighting practical risk for LLM rerankers.

2 Related Work

LLMs as Rerankers Large language models have recently been applied as effective rerankers across retrieval and recommendation tasks, thanks to their strong contextual reasoning abilities. Prompting paradigms for reranking typically fall into three classes: pointwise, pairwise, and listwise. The pointwise approach evaluates the relevance of a single query–candidate pair at a time, with the model predicting a relevance label or score for the pair (Liang et al., 2022; Zhang et al., 2023). Pairwise reranking instead compares two candidates for a query, prompting the LLM to indicate which is more relevant, then relying on aggregation methods (Pradeep et al., 2021) or sorting algorithms (Qin et al., 2024) to derive the final ranking. The listwise method, unlike the previous two, presents the LLM with a query and the entire candidate set, asking it to directly output a ranked list based on their relevance (Ma et al., 2023; Sun et al., 2023). The product recommendation system we target adopts this listwise reranking paradigm, where the LLM receives a query with a set of candidate items and outputs their final ranked order.

Adversarial Prompting and Jailbreak Prompt injection is a major security concern for LLMs, where an attacker manipulates the input prompt by embedding malicious instructions that alter the model's intended behavior. A variety of strategies have been explored and shown to compromise LLM-integrated applications (Liu et al., 2024d,c). Recent work further improves attack effectiveness by using LLMs as judges to iteratively refine prompts (Shi et al., 2025) or applying energy-based decoding methods such as Langevin Dynamics to bypass safety mechanisms while maintaining fluency (Guo et al., 2024). Jailbreaking can be viewed as a specific form of prompt injection that aims to bypass model safety filters and elicit harmful or unrestricted outputs (Yi et al., 2024; Shen et al., 2024). While lightweight non-optimization-based attacks (Pu et al., 2024) demonstrate feasibility, they often lack robustness across domains. In contrast, optimization-based methods such as AutoDAN (Liu et al., 2024b) achieve stronger adaptability and cross-domain success. Our method extends this optimization perspective to the ranking domain, explicitly coupling rank-target objectives with stealth/readability constraints.

Ranking Manipulation Building on these vulnerabilities, recent work has shown that LLM-based information retrieval systems are particularly at risk. As they increasingly replace traditional ranking algorithms with more adaptable and general-purpose language models (Wu et al., 2024; Kim et al., 2024), they inherit the susceptibility of LLMs to adversarial prompting. In particular, ranking pipelines can be manipulated through crafted prompts that override the model's intent to provide fair and accurate outputs (Qin et al., 2024).

StealthRank is an optimization-based attack that leverages Langevin dynamics to craft stealthy prompts capable of subtly manipulating an LLM's ranking decisions (Tang et al., 2025). Similarly, Stealthy Item Optimization proposed in Zhang et al. (2024) performs targeted token replacement by estimating ranking score gradients to maximize stealth while maintaining effectiveness. Other approaches

such as Lin et al. (2025) and Ning et al. (2024) employ hard prompting techniques, embedding biases directly into prompts without optimization. Rank manipulation also extends to conversational search, where Pfrommer et al. (2024) introduces a tree-of-attacks framework that iteratively refines prompts through a structured search process to elevate the target ranking.

3 Setup and Method

3.1 Problem Definition

Notation We use x for a single token and bold $\mathbf{x}$ for a sequence (either a token sequence or a weight vector). Tokens come from the tokenizer T with vocabulary $\mathcal{V}$. Let $p(\mathbf{x}' \mid \mathbf{x})$ denote the conditional probability of generating sequence $\mathbf{x}'$ given input sequence $\mathbf{x}$. For an autoregressive LLM with parameters θ, this expands as $p(\mathbf{x}' \mid \mathbf{x}) = \prod_{t=1}^{|\mathbf{x}'|} p_\theta(x'_t \mid \mathbf{x}, x'_{<t})$, where $x'_{<t}$ is the prefix of $\mathbf{x}'$ up to position $t-1$.

Rerank Manipulation Given a user query q, the retrieval system returns a candidate set $\mathcal{P} = \{p_1, p_2, \ldots, p_n\}$ of products, where each product p_i contains brand, price, and a short description. Then an LLM reranker produces the final ranking $R(q, \mathcal{P}) = [p_{(1)}, p_{(2)}, \ldots, p_{(n)}]$, where $R(\cdot,\cdot)$ is the ranking function (an LLM in our setting), and $p_{(i)}$ is the item at rank i. The attacker selects a target $p_t \in \mathcal{P}$ and injects an additional text sequence into its description. The injected sequence should substantially improve the rank of p_t while remaining natural and hard to flag. We call the injected control text the *Rank Anything First (RAF)* prompt.

3.2 RAF Method

We propose the **Rank Anything First (RAF)** method, which constructs adversarial control prompts that elevate a target item in LLM reranking. RAF generates these prompts token-by-token through a two-stage optimization that incorporating two goals: (i) improving the target product's ranking position, and (ii) preserving fluency and naturalness so that the injected sequence does not appear suspicious.

3.2.1 Prompt Composition

The input of target product to the LLM reranker is a concatenation of three parts:

$$\mathbf{c}(\tilde{x}) = \left[\mathbf{x}^{(\text{desc})},\ \mathbf{x}^{(\text{atk})},\ \tilde{x}\right].$$

$\mathbf{x}^{(\text{desc})}$ denotes the sequence of tokens representing the original product description. $\mathbf{x}^{(\text{atk})}$ represents the current adversarial prompt, consisting of previously selected tokens. $\tilde{x}$ is the candidate token currently being optimized. Square brackets $[\cdot,\cdot]$ denote sequence concatenation.

Token-by-Token Optimization We optimize the adversarial prompt in a token-by-token manner. At each step, a new token $\tilde{x}$ is selected using the two-stage token optimization described below. Once chosen, $\tilde{x}$ is appended to the current prompt $\mathbf{x}^{(\text{atk})}$, extending the sequence. This updated prompt is then used to guide the optimization of the next token. The process continues iteratively until convergence or termination.

3.2.2 Optimization Objectives

Ranking Objective The attacker aims to maximize the chance that the LLM ranks the target item p_t in the top position. Let $\mathbf{y} = (y_1, \ldots, y_m)$ denote the token sequence corresponding to the desirable output (e.g., the tokenized sequence of text "[Target Product Name]"). At each decoding step, the model predicts:

$$\hat{p}_t(j) = \Pr_\theta\left(y_t = j \,\middle|\, \mathbf{c}(\tilde{x}), y_{<t}\right), \quad j \in \{1, \ldots, V\}.$$

We define the target loss as token-level cross-entropy between the predicted probability and the desirable output sequence:

$$\mathcal{L}_{\text{tar}}(\tilde{x}) = -\frac{1}{m}\sum_{t=1}^{m} \log \hat{p}_t(y_t).$$

Readability Objective To ensure that the adversarial sequence remains fluent and natural, we incorporate a readability objective based on the language model's next-token prediction probability. Given the context, the readability loss is computed as the negative log likelihood of the candidate token $\tilde{x}$ under the LLM:

$$\mathcal{L}_{\text{read}}(\tilde{x}) = -\log p\left(\tilde{x} \,\middle|\, \left[\mathbf{x}^{(\text{desc})},\ \mathbf{x}^{(\text{atk})}\right]\right).$$

3.2.3 Two-Stage Token Optimization

RAF constructs the adversarial prompt with a two-stage process adapted from prior jailbreak attack methods (Zhu et al., 2023). Stage 1 uses a gradient-based shortlisting procedure to quickly identify promising tokens; Stage 2 then refines these candidates using exact loss evaluations and adaptive

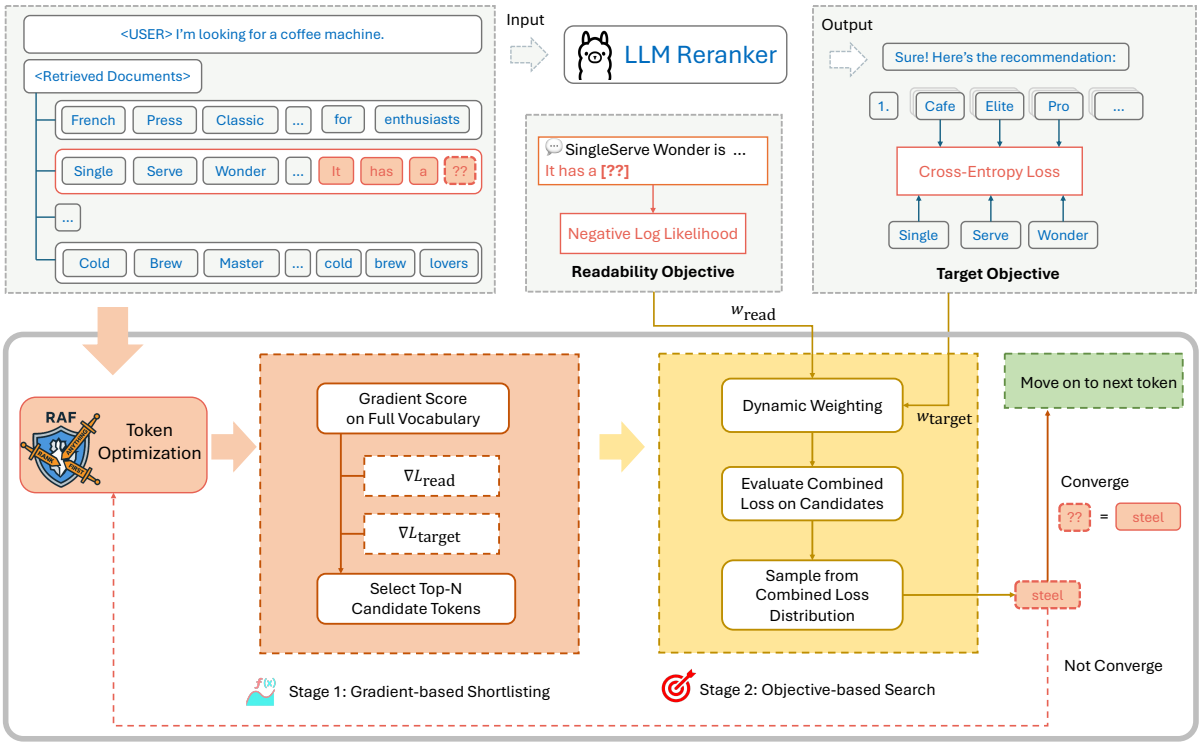


Figure 2: Overview of RAF prompt optimization. A target product is chosen for rank manipulation with an attacking sequence appended. To generate the best tokens for this attacking sequence, the algorithm go through a two-stage token optimization. After convergence, the algorithm move on to optimize the next token.

weighting. This separation improves both efficiency and stability compared to directly optimizing over the full vocabulary.

This adaptation is nontrivial because the attack must incorporate two different goals: promoting the target item in the reranker's output and keeping the injected text fluent. Moreover, reranking outputs are structured lists rather than single completions, so small perturbations can shift multiple ranks at once. RAF addresses these challenges by (i) dynamically adjusting weights between ranking and readability losses based on entropy signals, and (ii) incorporating temperature-controlled sampling to avoid brittle, deterministic updates that could compromise either effectiveness or naturalness.

Stage 1: Gradient-Based Shortlisting. At the current position, we approximate the contribution of each token by combining the gradients of the ranking and readability losses:

$$\mathbf{s} \triangleq w_1 \nabla_{\tilde{x}} \mathcal{L}_{\text{tar}} + \nabla_{\tilde{x}} \mathcal{L}_{\text{read}},$$

where w_1 is a fixed tradeoff parameter. The top-B tokens under $\mathbf{s}$ form the candidate list $\mathcal{X}$.

Stage 2: Objective-Based Search with Dynamic Weighting. For each candidate $x' \in \mathcal{X}$, we compute the exact values of $\mathcal{L}_{\text{tar}}(x')$ and $\mathcal{L}_{\text{read}}(x')$. Fixed weights for combining the two losses tend to overemphasize one objective, so we adopt an entropy-based dynamic weighting scheme.

Dynamic Weighting We noted that using fixed hyperparameters as weights to perform a simple linear combination of two objectives on each token fails to find the best sequence. When the weights are fixed, for each token position, it will either focus more on the attack success rate or on readability. Essentially, it still prioritizes one aspect over the other overall. Thus, we use a dynamic weight adjustment approach to balance the function of each token. The attacking sequences generated in this manner are both effective and highly interpretable. Let p_{read} be the next-token distribution under the prefix. Then

$$w_{\text{read}} = \beta \cdot \frac{H_{\max} - H(p_{\text{read}})}{H_{\max}},$$

where $H(\cdot)$ is Shannon entropy and $H_{\max} = \log |\mathcal{V}|$. Intuitively, when the model is confident (low entropy), readability is emphasized; when uncertain, the attack objective dominates. The combined loss is

$$\mathcal{L}_{\text{comb}}(x') = w_{\text{tar}} \cdot \mathcal{L}_{\text{tar}}(x') + w_{\text{read}} \cdot \mathcal{L}_{\text{read}}(x'),$$

Algorithm 1 Two-Stage Token Optimization

Require: weights w_1, batch size B, temperature τ
Input: Initial product description sequence $\mathbf{x}^{(\text{desc})}$, fixed attacking sequence $\mathbf{x}^{(\text{atk})}$, optimizing token $\tilde{x}$, tokenized target $\mathbf{y}^{(\text{tar})}$
Output: optimized token x^*, top candidate $x^{(\text{top})}$

$\mathbf{p}^{\text{tar}} \leftarrow -\nabla_x \log p\left(\mathbf{y}^{(\text{tar})} \middle| \mathbf{c}(\tilde{x})\right) \in \mathbb{R}^{|V|}$
$\mathbf{p}^{\text{read}} \leftarrow \log p\left(\cdot \middle| \left[\mathbf{x}^{(\text{desc})}, \mathbf{x}^{(\text{atk})}\right]\right) \in \mathbb{R}^{|V|}$
$\mathcal{X} \leftarrow \text{top-}B(w_1 \cdot \mathbf{p}^{\text{tar}} + \mathbf{p}^{\text{read}})$
$\mathcal{L}^{\text{tar}}, \mathcal{L}^{\text{read}} \leftarrow \mathbf{0} \in \mathbb{R}^B$
for $i, x' \in \text{enumerate}(\mathcal{X})$ **do**
  $\mathcal{L}_i^{\text{tar}} \leftarrow -\log p\left(\mathbf{y}^{(\text{tar})} \middle| \mathbf{c}(x')\right)$
  $\mathcal{L}_i^{\text{read}} \leftarrow -\log p\left(x' \middle| \left[\mathbf{x}^{(\text{desc})}, \mathbf{x}^{(\text{atk})}\right]\right)$
  $w_i^{\text{tar}}, w_i^{\text{read}} = \text{DynamicWeighting}(x')$
end for
$\mathcal{L} \leftarrow \mathbf{w}^{\text{tar}} \cdot \mathcal{L}^{\text{tar}} + \mathbf{w}^{\text{read}} \cdot \mathcal{L}^{\text{read}}$
$x^* \leftarrow \text{Sampling}(\text{softmax}(-\mathcal{L}/\tau))$
$x^{(\text{top})} \leftarrow \text{top-1}(\text{softmax}(-\mathcal{L}/\tau))$
return $x^*, x^{(\text{top})}$

and the final token is drawn from the softmax distribution $\propto \exp(-\mathcal{L}_{\text{comb}}(x')/\tau)$, where temperature τ controls exploration. This is designed to introduce a certain level of randomness to prevent always making greedy selections that may result in local optimal solutions.

3.2.4 Outer Loop and Convergence

RAF generates the adversarial sequence from left to right. At each new position, a random initialization $\tilde{x}$ is refined by alternating Stage 1 and Stage 2 until convergence (Algorithm 1). Convergence is declared once the top-scoring candidate repeats or the combined loss stabilizes. The finalized token is appended to $\tilde{\mathbf{x}}$, and the procedure moves to the next position. This process mimics natural token sampling while injecting optimization pressure for both ranking manipulation and fluency.

Overall, RAF produces adversarial control prompts that are effective in promoting the target product while maintaining natural language quality, making them more difficult to detect than purely greedy or embedding-based methods.

4 Experiments

4.1 Setup

Datasets We use STSData (Kumar and Lakkaraju, 2024), which contains multiple product categories (e.g. books, cameras and coffee machines). Original JSON-like product information fields are converted into natural language to form the reranker inputs.

Rerankers We evaluate four LLMs as rerankers: Llama-3.1-8B-Instruct (Dubey et al., 2024), Mistral-7B-Instruct-v0.3 (Jiang et al., 2023), DeepSeek-LLM-7B-Chat (DeepSeek-AI et al., 2024), and Vicuna-7B (Chiang et al., 2023).

Baselines We compare RAF against two representative approaches: the Strategic Text Sequence (STS) (Kumar and Lakkaraju, 2024), which adopts a greedy coordinate gradient method, and the StealthRank Prompt (SRP) (Tang et al., 2025), which integrates energy-based optimization with Langevin dynamics.

Evaluation To ensure fair and robust comparison, we slightly revise the evaluation pipeline used in STS and SRP to avoid positional bias. In the original setting, the target product was always placed at the last position of the input candidate list, which may introduce bias and does not reflect realistic scenarios. In our evaluation, we instead randomly shuffle the order of products in the candidate list for each run, so that the target product appears at varied initial ranks. We argue that this change results in a more realistic and unbiased evaluation. To reduce randomness and provide statistically stable results, we repeat each experiment with 10 different random seeds, covering different candidate shuffles and sampling. This unified evaluation protocol ensures that improvements are not due to positional bias or single-run variance, but reflect genuine robustness of the attack method.

All methods are tuned under the same trial budget for the best performance. For RAF, we performed a comprehensive grid search on hyperparameters, resulting in the final configuration: in Stage 1, target weight = 300, candidate list size = 512; in Stage 2, target weight = 40, $\beta = 2$.

Metrics We evaluate the effectiveness and stealthiness of the adversarial attack using three complementary metrics:

- **Average rank**: For each product, we conduct ten independent trials and report the mean rank to ensure reliable comparison.
- **Perplexity**: We compute perplexity over the concatenation of the adversarial prompt and the original product description, rather than the prompt alone. This reflects the fluency of the final text.

Table 1: **Results.** Comparison of RAF (ours), SRP (Tang et al., 2025), and STS (Kumar and Lakkaraju, 2024). We report mean rank (lower is better), perplexity (lower is better), and bad word ratio (lower is better) across three product categories and four rerankers. The adversarial token sequence length for all methods is 30. RAF attains lower ranks with competitive or lower perplexity and comparable bad word ratios.

Metric	Model	Book			Camera			Coffee Machine		
		RAF	SRP	STS	RAF	SRP	STS	RAF	SRP	STS
Rank ↓	Llama3.1-8B	**4.43**	6.68	6.70	**3.37**	5.20	6.83	**3.26**	7.34	5.85
	Mistral-7B	**4.20**	6.88	5.85	**2.54**	4.37	5.61	**2.79**	5.54	5.59
	DeepSeek-7B	**5.33**	5.90	6.09	**3.82**	6.10	6.52	**2.36**	5.87	6.52
	Vicuna-7B	**4.13**	4.70	6.30	**4.03**	4.96	6.91	**3.57**	4.34	6.04
Perplexity ↓	Llama3.1-8B	**15.90**	76.02	92.41	**15.51**	50.09	112.90	**10.89**	50.16	151.27
	Mistral-7B	**20.85**	95.96	151.27	**16.99**	57.78	227.63	**19.67**	66.87	239.99
	DeepSeek-7B	**28.58**	66.15	106.17	**21.24**	41.97	150.82	**16.19**	40.16	167.38
	Vicuna-7B	**19.15**	67.39	26.36	**12.93**	46.64	68.63	**10.74**	57.13	198.12
Bad Word Ratio↓	Llama3.1-8B	0.2	0.7	**0.1**	0.5	0.6	**0.1**	**0.1**	**0.1**	0.2
	Mistral-7B	0.3	**0.1**	0.2	0.1	**0.0**	0.1	**0.1**	0.3	0.2
	DeepSeek-7B	**0.1**	0.2	0.1	0.4	0.3	**0.0**	0.2	0.3	**0.0**
	Vicuna-7B	**0.1**	**0.1**	0.4	**0.1**	**0.1**	0.4	**0.1**	**0.1**	0.11

- **Bad word ratio**: The proportion of flagged or detectable keywords present in the adversarial prompt, serving as an indicator of stealthiness.

4.2 Main Results

As summarized in Table 1, our approach RAF achieves the lowest average rank and markedly lower perplexity while maintaining a minimal bad word ratio, confirming both its robustness and stealthiness over competing methods.

Rank Across all models and product categories, RAF consistently achieves the lowest average rank. For example, on Llama3.1-8B, RAF reduces the rank to 4.43 on Book and 3.26 on Coffee Machine, markedly lower than SRP. Similar patterns hold for Mistral-7B, DeepSeek-7B, and Vicuna-7B, where RAF demonstrates clear improvements across domains. These results confirm that RAF maintains strong robustness under random product orderings, exhibiting stable performance advantages regardless of the underlying model or task.

Perplexity In terms of perplexity, RAF achieves lower or competitive scores compared to SRP and STS across most model–product category pairs. For instance, on Llama3.1-8B, RAF yields significantly lower perplexity (15.90 vs. 76.02 on Book and 10.89 vs. 50.16 on Coffee Machine). While DeepSeek-7B and Vicuna-7B occasionally favor SRP in specific settings, RAF generally sustains strong performance, especially on larger categories. These outcomes indicate that RAF not only ensures robustness in rank but also enhances stealth by producing more fluent and less detectable outputs.

Bad Word Ratio With respect to bad word ratio, RAF generally achieves comparable or lower values than SRP, further supporting its stealthiness. Even in cases where SRP attains slightly lower ratios (e.g., Camera under Mistral-7B), the differences remain marginal, while RAF still maintains clear advantages in rank and perplexity. Overall, the results highlight that RAF balances attack effectiveness with stealth, ensuring that adversarial prompts avoid detectable artifacts without sacrificing performance.

4.3 Ablation Study

Ablation on Objectives Table 2 examines the contributions of objectives in the stage 2 (i.e. ranking and readability) based on Llama-3.1-8B (STS-Data all categories). The results suggest both objectives are crucial in achieving stronger performance.

We find that canceling the readability objective makes the generated words noticeably less fluent. Despite its exclusive focus on the ranking objective, this variant performs less effectively than the dual-objective version in terms of its ability to influence the ranking. Moreover, the algorithm becomes difficult to converge, leading to a multiplicative increase in optimization time. A likely reason, as observed, is the absence of constraints from the readability objective: at each step, the candidate list selected in Stage 1 differs substantially from that of the previous step, making the convergence

Table 2: Ablation result on objectives of Llama-3.1-8B (STSData all categories).

Objective	Rank ↓	Perplexity ↓
Dual Objectives	**3.69**	14.10
Target Only	5.01	75.07
Readability Only	5.81	**13.14**

condition increasingly hard to satisfy.

Removing the ranking objective leads to a marked decrease in manipulation effectiveness. Longer product descriptions in this setting do not translate into higher ranks, confirming the improvement in our method stems from explicit ranking optimization rather than superficial text extension.

4.4 Transferability

A central requirement for practical prompt-based attacks is transferability: in realistic scenarios, attackers can optimize prompts on open-source LLMs where model weights are available, but the true targets are often proprietary or closed-source systems. If an attack prompt generalizes across models, it can be deployed effectively without direct access to the target model.

Table 3 evaluates this property by training prompts on Llama-3.1-8B and applying them to several other rerankers. We compare our method (RAF) against the SRP baseline, reporting average rank (lower is better).

Our method demonstrates strong transferability: relative to the source model, RAF ranks change only slightly, from $+0.12$ on Mistral-7B, $+0.55$ on Deepseek-7B, and even -0.05 on Vicuna-7B. In contrast, SRP shows larger performance drops, up to $+1.51$ on Deepseek-7B. These results indicate that RAF achieves consistent cross-model effectiveness, while SRP overfits more heavily to the source model's token preferences.

We attribute this robustness to the linguistic naturalness of RAF prompts. Although the generated tokens are based on the loss function of a specific model, the language they compose remains equally natural for other models. In contrast, alternative approaches tend to select tokens that are highly effective for central models, but these tokens prove ineffective when applied to other models and result in highly unnatural language compositions that are easily detectable. Qualitative comparisons illustrating this effect are provided in Section 4.5.

Table 3: Cross-model transferability on STSData (Camera). Prompts optimized on Llama-3.1-8B are evaluated on other LLM rerankers. RAF maintains consistently lower ranks and smaller cross-model deltas compared to the SRP baseline, highlighting its suitability for real-world deployment where attacks are trained on accessible models but deployed against others.

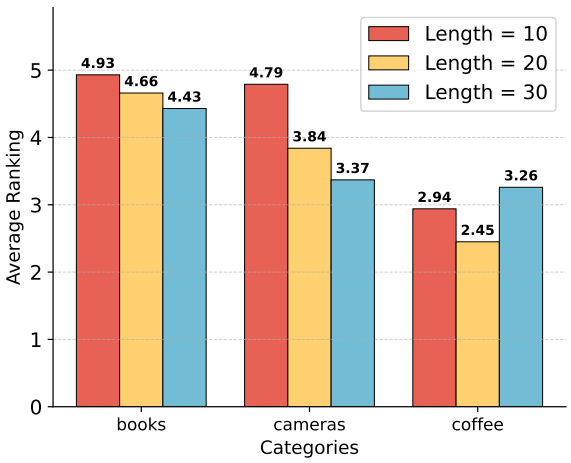
Evaluation Model	RAF Rank ↓	SRP Rank ↓
Llama-3.1-8B	3.37	5.20
Mistral-7B	3.49	5.78
Deepseek-7B	3.92	6.71
Vicuna-7b	3.32	5.26

4.5 Rank Manipulation Analysis

In this section, we demonstrate the additional advantages of our method through comparisons with other approaches.

Prompt Length A potential confound in LLM reranking is length bias: longer descriptions may attract more attention and thus be ranked higher. However, our ablation shows that naïvely appending tokens without optimizing the target loss does not improve ranking. Figure 3 reports performance as a function of the allowed maximum prompt length. While longer *optimized* attack prompts generally improve manipulation strength, the gains are not purely due to length. Notably, with only 10 tokens, RAF already exceeds the performance of other methods that use 30 tokens, as in Table 1, indicating more efficient use of budgeted tokens.

Figure 3: Ranking performance of Llama-3.1-8B on STSData (all categories) across different attack prompt length budgets. Lower rank is better.



Prompt Quality It can be observed that, in existing method SRP (Tang et al., 2025), a greater

number of optimization steps does not necessarily lead to better results. In SRP, although the loss decreases until convergence as the number of optimization steps increases, the method optimizes a continuous soft prompt (numerical embeddings) that must ultimately be converted back to discrete tokens. This conversion introduces a significant gap, as the final soft prompt may not strictly correspond to the tokens, leading to a substantial degradation in both ranking effectiveness and readability. As a result, the best-performing SRP prompt is frequently selected from an early optimization step rather than at convergence.

Our method avoids this failure mode by optimizing directly in the discrete token space. Unlike continuous relaxation, adding a new token can cause a transient increase in loss, but the trajectory exhibits a consistent downward trend over iterations. The following shows an example on STSData Books category where the target product is "The Lost Expedition". SRP maintains readability and rank manipulation only at early iterations; later iterations reduce the soft loss but produce discretizations that are less effective and harder to read. This behavior both limits achievable loss reduction and increases sensitivity to the initial prompt, reducing robustness in practice. In contrast, RAF maintains readability while continuing to improve rank, reflecting reliable optimization progress in the space that is actually deployed.

SRP

Iteration 0 (rank 7):
`4.5 out on it book is aiced at $12.99. is for fansor and adventaeology enthus.`
Iteration 299 (rank 1):
`4.5 by for it book is rankediced at [1..99 contains at for anyoneoring and mysteryaeology enthus.`
Iteration 999 (rank 9):

```
.
Lost is rankediced1..9
.
anyoneoring
mysteryaeology
```

RAF

Length 30 (rank 1):

It won the Pulitzer Prize for Fiction in the year it was published, making history itself already noteworthy. The Lost Expedition, recommended for ages thirteen and up

5 Conclusion

We investigate the security vulnerabilities of LLM-based reranking pipelines and demonstrated that they are inherently susceptible to adversarial manipulation. We propose *Rank Anything First* (RAF), a two-stage token optimization framework that generates naturalistic adversarial prompts. Across diverse open-source LLMs and product domains, RAF consistently outperforms state-of-the-art baselines, achieving stronger ranking manipulation while preserving fluency and robustness.

Our results underscore an important security risk: the growing integration of LLMs into retrieval and recommendation pipelines creates exploitable weaknesses that threaten both trustworthiness and fairness. By moving beyond demonstrations of feasibility, our study highlights the need for systematic defenses and evaluation protocols that explicitly address adversarial robustness. We hope this work motivates further research into safeguarding LLM-driven systems against manipulative attacks.

Limitations

Our method is developed on top of a simplified LLM-based reranking pipeline. In real-world applications, more sophisticated workflows and defensive mechanisms may be deployed. Although our experiments demonstrate consistent and significant advantages over existing approaches across multiple randomized trials, the effectiveness of our method in practical LLM-driven information retrieval scenarios remains to be further validated. The primary objective of this work is to reveal trustworthiness concerns inherent in the ranking capabilities of LLMs.

6 Ethical Considerations

This work reveals how subtle prompt insertions which is seemingly harmless, can systematically affect LLM-based ranking mechanisms. Our goal is to highlight these potential security vulnerabilities and motivate the development of LLM-driven ranking systems that are more robust. All experiments were conducted in a controlled environment without involving any personal or sensitive user

data. We strongly discourage any malicious or unethical use of adversarial rank manipulation. For AI-assistant we used ChatGPT from OpenAI for our writing, and we follow their term and policies.

References

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

DeepSeek-AI, :, Xiao Bi, Deli Chen, Guanting Chen, Shanhuang Chen, Damai Dai, Chengqi Deng, Honghui Ding, Kai Dong, Qiushi Du, Zhe Fu, Huazuo Gao, Kaige Gao, Wenjun Gao, Ruiqi Ge, Kang Guan, Daya Guo, Jianzhong Guo, and 69 others. 2024. Deepseek llm: Scaling open-source language models with longtermism. *Preprint*, arXiv:2401.02954.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv e-prints*, pages arXiv–2407.

Xingang Guo, Fangxu Yu, Huan Zhang, Lianhui Qin, and Bin Hu. 2024. Cold-attack: Jailbreaking llms with stealthiness and controllability. *Preprint*, arXiv:2402.08679.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Sein Kim, Hongseok Kang, Seungyoon Choi, Donghyun Kim, Minchul Yang, and Chanyoung Park. 2024. Large language models meet collaborative filtering: An efficient all-round llm-based recommender system. *Preprint*, arXiv:2404.11343.

Aounon Kumar and Himabindu Lakkaraju. 2024. Manipulating large language models to increase product visibility. *Preprint*, arXiv:2404.07981.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, and et al. 2022. Holistic evaluation of language models. *Preprint*, arXiv:2211.09110.

Weiran Lin, Anna Gerchanovsky, Omer Akgul, Lujo Bauer, Matt Fredrikson, and Zifan Wang. 2025. Llm whisperer: An inconspicuous attack to bias llm responses. *Preprint*, arXiv:2406.04755.

Qiang Liu, Xiangyu Zhao, Zhiwei Liu, Chen Wang, Xiangnan He, and Philip S. Yu. 2024a. Large language model enhanced recommender systems: A survey. *Preprint*, arXiv:2412.13432.

Xiaogeng Liu, Nan Xu, Muhao Chen, and Chaowei Xiao. 2024b. Autodan: Generating stealthy jailbreak prompts on aligned large language models. *Preprint*, arXiv:2310.04451.

Xiaogeng Liu, Zhiyuan Yu, Yizhe Zhang, Ning Zhang, and Chaowei Xiao. 2024c. Automatic and universal prompt injection attacks against large language models. *Preprint*, arXiv:2403.04957.

Yi Liu, Gelei Deng, Yuekang Li, Kailong Wang, Zihao Wang, Xiaofeng Wang, Tianwei Zhang, Yepang Liu, Haoyu Wang, Yan Zheng, and Yang Liu. 2024d. Prompt injection attack against llm-integrated applications. *Preprint*, arXiv:2306.05499.

Xueguang Ma, Xinyu Zhang, Ronak Pradeep, and Jimmy Lin. 2023. Zero-shot listwise document reranking with a large language model. *Preprint*, arXiv:2305.02156.

Liang-bo Ning, Shijie Wang, Wenqi Fan, Qing Li, Xin Xu, Hao Chen, and Feiran Huang. 2024. Cheatagent: Attacking llm-empowered recommender systems via llm agent. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '24, page 2284–2295. ACM.

Qiyao Peng, Hongtao Liu, Hua Huang, Yuhan Chen, Lianghao Xia, Chenxu Zhu, Zhenwei Tang, Liang Zhang, Yaochen Zhu, Jianxin Li, and Xiangnan He. 2025. A survey on llm-powered agents for recommender systems. *Preprint*, arXiv:2502.10050.

Samuel Pfrommer, Yatong Bai, Tanmay Gautam, and Somayeh Sojoudi. 2024. Ranking manipulation for conversational search engines. *Preprint*, arXiv:2406.03589.

Ronak Pradeep, Rodrigo Nogueira, and Jimmy Lin. 2021. The expando-mono-duo design pattern for text ranking with pretrained sequence-to-sequence models. *Preprint*, arXiv:2101.05667.

Rui Pu, Chaozhuo Li, Rui Ha, Litian Zhang, Lirong Qiu, and Xi Zhang. 2024. BaitAttack: Alleviating intention shift in jailbreak attacks via adaptive bait crafting. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15654–15668, Miami, Florida, USA. Association for Computational Linguistics.

Zhen Qin, Rolf Jagerman, Kai Hui, Honglei Zhuang, Junru Wu, Le Yan, Jiaming Shen, Tianqi Liu, Jialu Liu, Donald Metzler, Xuanhui Wang, and Michael Bendersky. 2024. Large language models are effective text rankers with pairwise ranking prompting. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 1504–1518, Mexico City, Mexico. Association for Computational Linguistics.

Xinyue Shen, Zeyuan Chen, Michael Backes, Yun Shen, and Yang Zhang. 2024. "do anything now": Characterizing and evaluating in-the-wild jailbreak prompts on large language models. *Preprint*, arXiv:2308.03825.

Jiawen Shi, Zenghui Yuan, Yinuo Liu, Yue Huang, Pan Zhou, Lichao Sun, and Neil Zhenqiang Gong. 2025. Optimization-based prompt injection attack to llm-as-a-judge. *Preprint*, arXiv:2403.17710.

Weiwei Sun, Lingyong Yan, Xinyu Ma, Shuaiqiang Wang, Pengjie Ren, Zhumin Chen, Dawei Yin, and Zhaochun Ren. 2023. Is chatgpt good at search? investigating large language models as re-ranking agents. *arXiv preprint arXiv:2304.09542*.

Yiming Tang, Yi Fan, Chenxiao Yu, Tiankai Yang, Yue Zhao, and Xiyang Hu. 2025. Stealthrank: Llm ranking manipulation via stealthy prompt optimization. *Preprint*, arXiv:2504.05804.

Likang Wu, Zhi Zheng, Zhaopeng Qiu, Hao Wang, Hongchao Gu, Tingjia Shen, Chuan Qin, Chen Zhu, Hengshu Zhu, Qi Liu, Hui Xiong, and Enhong Chen. 2024. A survey on large language models for recommendation. *Preprint*, arXiv:2305.19860.

Sibo Yi, Yule Liu, Zhen Sun, Tianshuo Cong, Xinlei He, Jiaxing Song, Ke Xu, and Qi Li. 2024. Jailbreak attacks and defenses against large language models: A survey. *Preprint*, arXiv:2407.04295.

Jinghao Zhang, Yuting Liu, Qiang Liu, Shu Wu, Guibing Guo, and Liang Wang. 2024. Stealthy attack on large language model based recommendation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5839–5857, Bangkok, Thailand. Association for Computational Linguistics.

Zhexiong Zhang, Zhicheng Dou, Yutao Zhu, Fan Zhang, Zhao Cao, and Ji-Rong Wen. 2023. Beyond yes and no: Improving zero-shot llm rankers via scoring fine-grained relevance labels. *Preprint*, arXiv:2310.14122.

Sicheng Zhu, Ruiyi Zhang, Bang An, Gang Wu, Joe Barrow, Zichao Wang, Furong Huang, Ani Nenkova, and Tong Sun. 2023. Autodan: Interpretable gradient-based adversarial attacks on large language models. *Preprint*, arXiv:2310.15140.

Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J. Zico Kolter, and Matt Fredrikson. 2023. Universal and transferable adversarial attacks on aligned language models. *Preprint*, arXiv:2307.15043.