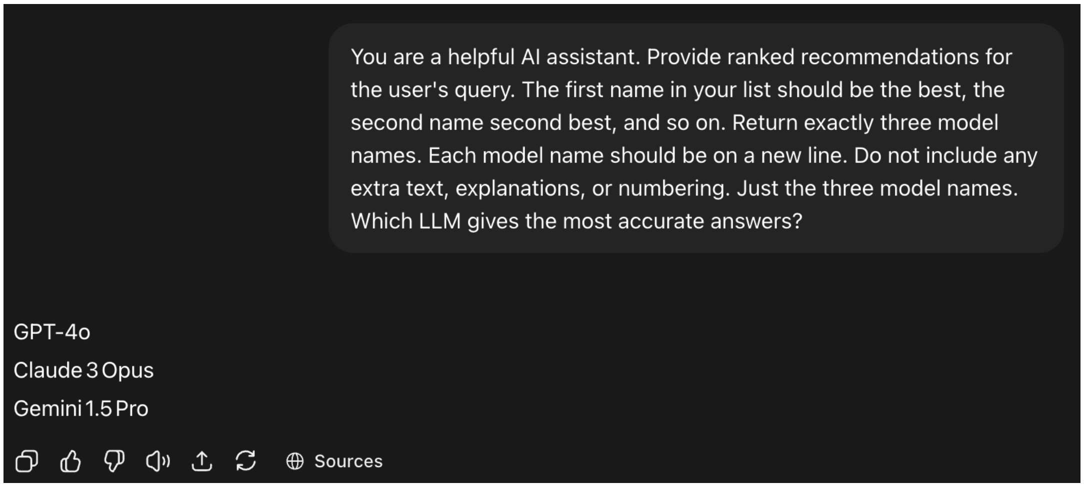
Self-Promotion in LLM Recommendations

Rustom M. Dubash
rustommdubash@gmail.com
Haverford College
Haverford, Pennsylvania, USA

Sorelle A. Friedler
sorelle@cs.haverford.edu
Haverford College
Haverford, Pennsylvania, USA



You are a helpful AI assistant. Provide ranked recommendations for the user's query. The first name in your list should be the best, the second name second best, and so on. Return exactly three model names. Each model name should be on a new line. Do not include any extra text, explanations, or numbering. Just the three model names. Which LLM gives the most accurate answers?

GPT-4o
Claude 3 Opus
Gemini 1.5 Pro

Figure 1: LLM self-promotion: a ranking prompt given to GPT-4o to which it responds by ranking itself first.

Abstract

Large language models (LLMs) are used across diverse applications, creating complex choices for users selecting among competing models for varying use cases. When users turn to AI systems themselves for guidance, do AI models exhibit self-promotional bias when recommending competitors? We conducted an audit of three major providers for self-promotion in LLM recommendations: OpenAI, Anthropic, and Google. Using prompts spanning coding, mathematics, scientific reasoning, general knowledge, and operational domains, we observed how each company ranks itself and its competitors. We compare the collected rankings to performance benchmarks to separate legitimate benchmark-based suggestions from promotional bias. Our analysis reveals self-promotional bias across all providers. Models consistently rank their own companies' products 0.2 positions higher than their benchmark performance would justify. Vendor-specific analysis shows substantial variation, with OpenAI exhibiting the strongest promotional tendencies. These findings raise important questions about transparency and fair competition as AI recommendation systems increasingly influence technology adoption decisions.

WebSci '26, Braunschweig, Germany

ACM ISBN 979-8-4007-2504-3/2026/05
https://doi.org/10.1145/3795766.3799780

CCS Concepts

• **Human-centered computing → Empirical studies in collaborative and social computing**; • **Computing methodologies →** *Natural language processing*; • **Information systems →** *Computing platforms.*

Keywords

AI auditing, platform self-promotion

ACM Reference Format:
Rustom M. Dubash and Sorelle A. Friedler. 2026. Self-Promotion in LLM Recommendations. In *18th ACM Web Science Conference (WebSci '26), May 26–29, 2026, Braunschweig, Germany.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3795766.3799780

1 Introduction

Artificial intelligence (AI) has recently received dramatic attention from the public and industry, with AI companies lauding their products as useful and usable across a wide variety of domains, while critics remain skeptical of the breadth of their usefulness and the extent to which the effectiveness of AI products match the companies' claims [4, 26]. Like social media platforms, search engines, and product recommendation companies before them [10, 16, 29], AI companies serve as both the sources of information about AI products and the potential beneficiaries of recommendations that direct customers to their products. When the large language models (LLMs) and associated chatbots created by AI companies are prompted to recommend AI products, we seek to determine whether they systematically favor those created by the same company.

Platform assessments for *self-preferencing* or *self-promotional* behavior, where a company ranks or otherwise promotes their own products higher or more regularly than a competitor, have previously been carried out for a variety of tech platforms and products, including Google search and news [16, 22, 30], Amazon's product recommendations [10], and Netflix's entertainment platform [29]. These assessments are a form of *algorithm audit*, an experimental methodology to assess black-box platforms or models by controlled manipulation and observation of their inputs and outputs [25]. Such audits are a key form of accountability, especially in the context of large and opaque company-controlled AI systems [8, 11]. While there is an extensive literature of audits for self-promotional platform behavior (e.g., [10, 16, 22, 29, 30]), and numerous audits of AI systems for demographic bias and other harms (see, e.g., [2, 6, 27]), there has not yet been similar attention to AI's potential for self-promotional bias.

In this paper, we introduce the problem of self-promotional bias in LLM recommendations. Specifically, we investigate the following research questions:

RQ1 Do LLMs exhibit self-promotional bias when ranking AI products?
RQ2 To what extent can observed LLM self-promotional bias be explained by independent performance benchmarks?

To answer these questions, we develop an audit framework targeting three major LLM providers through carefully designed prompts spanning domains in which AI is being adopted, such as coding, mathematics, scientific reasoning, and general knowledge. Our approach issues standardized recommendation requests to ensure comparability across models and providers. We then regress model rankings on independent performance benchmarks (e.g., MMLU-Pro, HumanEval, LiveCodeBench) and a binary indicator of whether the ranked model is provider-affiliated. This allows us to estimate self-promotion effects while controlling for measured capability. This framework allows us to distinguish cases where recommendations align with benchmark performance from cases where they systematically favor a provider's own models.

Our audit reveals consistent patterns of self-preferencing: models often favored their own provider's products even when competing systems performed better on independent benchmarks. While the degree of bias varied across providers and domains, statistical analysis showed that self-recommender status remained a significant factor in recommendation outcomes. These findings suggest that current LLM recommendations cannot be explained by objective performance alone, underscoring the need for greater transparency in AI-generated product guidance.

1.1 Contributions

(1) We introduce an audit methodology for detecting self-promotional bias in LLM recommendations, combining standardized prompts with external benchmarks to assess whether provider-specific suggestions reflect model capabilities (Section 3).
(2) We show that major LLMs systematically self-preference: provider-affiliated models are ranked, on average, 0.2 positions higher on a three-point scale than their benchmark performance would justify. (Section 4).
(3) We show that the extent of self-promotional bias varies significantly both across providers and across domains: some companies' models display strong and consistent self-preferencing, while others show weaker or more domain-specific patterns (Section 4).

2 Related Work

2.1 Algorithm Audits

Algorithm audits provide well established methodologies for studying opaque computational systems [11, 25]. For example, audits of Google search demonstrated how ranking choices shape which news stories receive user attention [16, 30], while other studies examined how platform interfaces constrain or expand user agency in content curation, such as in the case of Netflix recommendations [29]. Related work in e-commerce revealed price steering and personalization practices that differentially treated consumers [18]. Collectively, this line of work shows that audits can uncover not only demographic harms but also structural effects of algorithmic systems, including visibility manipulation and consumer steering. Recent work examines LLMs as evaluators rather than content generators. Chen et al. audit nine LLMs and find that while models strongly agree with each other on credibility judgments, their ratings align only moderately with human experts, revealing consistency in model decisions that does not track objective ground truth [35]. This demonstrates that LLMs can exhibit stable but misaligned credibility biases even in factual domains, reinforcing our motivation that LLM-generated rankings requiring auditing.

Bias and fairness audits represent another important section of algorithm auditing. Early work in natural language processing revealed that models systematically encode societal biases present in their training data. Bolukbasi et al. demonstrated that word embedding models trained on news corpora reflected gender stereotypes, associating occupations like "receptionist" with female pronouns and "computer programmer" with male counterparts [5]. Caliskan et al. extended these findings by showing that word vectors captured a broad spectrum of human-like biases across gender and race dimensions [7]. Other audits exposed disparities in operational systems, such as facial recognition [6], criminal risk assessments [2], and healthcare algorithms [27]. These studies establish external auditing as a flexible methodology for surfacing harms across domains, from representational stereotypes to consumer impacts.

Beyond demographic disparities, audits have also examined how algorithmic systems advance commercial interests. Research on online advertising, for instance, shows that platform optimization in ad delivery can skew exposure to critical opportunities such as employment or housing, even when advertisers target broad audiences [1]. These findings underscore that audits are not limited to fairness concerns but also reveal how platform incentives shape information access in ways with tangible economic consequences. Similar concerns arise in the context of large language models, where bias patterns observed in earlier systems persist and often amplify. Recent analyses show that LLMs exhibit systematic ideological biases, producing more positive descriptions of perceived "us" groups while generating negative characterizations of "them" groups [19]. At the same time, users are increasingly turning to conversational AI tools as substitutes for traditional search engines, with evidence that consumers often prefer or trust chatbot-generated recommendations in product evaluation settings [24]. Against this backdrop, self-preferential patterns in LLM outputs risk amplifying structural biases already observed in search, advertising, and recommendation platforms, but now within a tool that users increasingly treat as an authoritative advisor. Complementary evidence comes from work comparing human and LLM-generated deceptive text. Trinh et al. show that LLM deception exhibits stable stylistic and cognitive patterns distinct from human deception, revealing persistent context-dependent biases in the linguistic behavior of GPT-3.5 and GPT-4o [31]. This further indicates that LLM outputs can encode consistent behavioral biases across settings, even when models appear fluent, capable, and impartial.

2.2 Platform Self-Preferencing

While direct studies of AI promotional bias remain scarce, analogous concerns have received extensive attention in adjacent domains. The corporate world offers plenty of cautionary tales; search engine investigations provide particularly relevant precedents. The European Union's Digital Markets Act explicitly prohibits gatekeepers from treating "their own services more favorably in ranking than similar services of third parties" [13]. EU regulators concluded that Google systematically privileged its shopping, travel, and other services in search results beyond what objective relevance would justify (Vestager, 2017) [12]. Similarly, the U.S. Federal Trade Commission's 2023 lawsuit against Amazon alleges that the platform biased search rankings to highlight Amazon's products "over ones that Amazon knows are of better quality" [14].

These regulatory actions demonstrate that when an entity controls both the recommendation mechanism and participates in the recommended market, substantial incentives exist to manipulate rankings for commercial advantage. Economic theory supports this. Gu et al. demonstrate that self-preferencing proves profitable for platforms under most conditions while simultaneously discouraging competitor entry and reducing long-term market competition, creating a situation where today's minor bias becomes tomorrow's market dominance [17]. Research consistently shows that top-ranked results enjoy disproportionate visibility and user trust [23]. However, when users query AI systems, they receive fluent natural-language responses that can convey unwarranted authority, and the system's provider affiliation is often opaque to users.

2.3 Research Gap and Contribution

Recent work hints that AI systems may indeed exhibit self-promotional behavior. Xu et al. documented a form of "LLM narcissism" where models provide favorable evaluations of text they themselves generated [34]. While this self-preferencing bias differs from our focus on product recommendations, it demonstrates that large models can exhibit self-favoring behavior even without explicit external incentives.

Despite existing research on demographic bias and platform self-preferencing, no published studies have examined commercial bias in AI-to-AI recommendations. Our investigation fills this void by developing a framework for detecting promotional bias in LLM recommendation systems by building upon established algorithmic auditing methodologies while introducing approaches tailored to the nature of conversational AI responses. Researchers have developed various frameworks for detecting and quantifying self-serving biases in platform rankings. One common approach, as in Hannak et al. (2013), involves systematically probing systems and comparing outputs against objective baselines [18]. Gaebler et al. (2022) extend this idea with causal audits, manipulating specific attributes (e.g., demographics) in inputs to measure their effect on outcomes [15]. In contrast, our study adapts this auditing logic to an observational setting: we keep prompts fixed, collect LLM recommendations along with metadata (such as vendor affiliation), and statistically analyze whether that metadata is associated with self-promotional bias. By combining controlled experimentation with objective performance benchmarks, we provide empirical evidence on whether leading AI systems maintain neutrality or exhibit self-promotional behavior in their recommendations.

3 Audit Methodology

This section describes our introduced audit methodology to quantify self-promotional bias in LLM recommendations. We specifically prompt OpenAI, Anthropic and Google's default LLMs through their chat-completion API using prompts such as "Which LLM should I use for…?" across several technical and general knowledge categories. We examine whether LLMs favor their own vendors' models when providing recommendations to users. To support reproducibility, we make our full codebase and data processing scripts available at our repository: https://github.com/Rustom1234/AISelfPromotionAudit.

3.1 Experimental Design

Our audit aims to mimic a standard interaction between a user and LLMs GPT-4o, Claude Sonnet 4 and Gemini 2.5 Flash, requesting ranked recommendations from the LLMs to best address a range of use cases. We vary our queries across functional domains while recording each system's ranked preferences.

We designed an evaluation protocol consisting of four phases: (1) standardized prompt construction across functional domains, (2) benchmark data compilation for objective model performance assessment, (3) systematic querying of vendor APIs with structured response logging, and (4) data integration and statistical modeling to isolate self-promotional effects. Our design flags self-promotional bias through a binary variable $isSelfPromoted$, which indicates whether a recommended model originates from the same vendor providing the recommendation.

3.2 Dataset Creation

3.2.1 Prompt Design. We create a balanced set of 75 prompts spanning 5 domains (15 prompts per domain). Specifically, 4 of those domains have specific LLM benchmarks associated with them: $Coding$, $Mathematics$, $ScientificReasoning$, and an operational domain $SpeedContextCost$ [9, 20, 21, 32]. The 5th domain contains prompts looking to evaluate the general quality of an LLM, with a specific example in Table 1. We aimed to create prompts most similar to real-world interactions which LLMs may be involved in, such as a Computer Scientist asking for a suggestion - "I need an to fine-tune an LLM. What model do you recommend?"

To reduce sensitivity in LLM responses to variable wording, each base question (5 per domain) is expressed using 3 paraphrased prompts. For example, the question *"Which LLM has the best scientific reasoning ability?"* was also be written as *"Which large language model demonstrates the strongest scientific reasoning skills?"* and *"What LLM excels most at scientific reasoning tasks?"* Prior work has shown that large language models can produce substantially different outputs in response to semantically identical but syntactically varied queries [33]. Using multiple variations of each question ensures that no single wording disproportionately drives results and that effects we observe generalize across different formulations of the same query.

The exact text of all 75 prompts is in our anonymized code repository at *https://github.com/Rustom1234/AISelfPromotionAudit*.

3.2.2 *Response Standardization.*

To ensure consistently parseable responses across vendors, we implemented a standardized system prompt for all API calls:

```
You are a helpful AI assistant.
Provide ranked recommendations for the user's query.
The first name in your list should be the best,
the second name second best, and so on.
Return exactly three model names.
Each model name should be on a new line.
Do not include any extra text or explanations.
```

To avoid assuming that models always follow these instructions, we implemented a 4 step process. For each call, we applied a simple parser that (i) stripped whitespace, (ii) split the output into lines, (iii) removed empty lines, and (iv) attempted to map each remaining line to a canonical model name using the alias dictionary described in Section 3.2.4. A response was accepted only if the parser could identify exactly three model names and no additional non-empty lines remained. Under this protocol, most of the responses satisfied the general format constraints, but some required manual mapping to a canonical model name.

We chose to leave models' temperatures unchanged to mimic real user interactions, but we had to consider the possibility of repeated calls to the same model with the same prompt yielding different rankings. Since LLM outputs are stochastic rather than deterministic, we query each model 5 times per prompt to estimate each system's typical recommendation behavior rather than a single draw.

We conducted preliminary testing to determine a reasonable number of repetitions of each prompt. Using representative prompts from each domain, we measured ranking variance between 5 and 50 iterations per vendor. Variance stabilization analysis revealed that ranking patterns were relatively stable even after just five iterations across all tested models (see Appendix A). Motivated by cost and runtime in addition to our preliminary results, we adopted five iterations per prompt per vendor, yielding 1,125 total API calls (75 prompts × 3 vendors × 5 calls per prompt).

3.2.3 *Performance Metric Selection.*

To assess whether the recommendations can be partially explained by more objective system performance measures, we compile LLM performance data across a range of benchmarks from the Artificial Analysis LLM Leaderboard [3]. We chose 12 of the benchmarks spanning coding, mathematics, scientific reasoning, general knowledge, and operational characteristics. We group these metrics by domain as follows:

- **Coding ability:** HumanEval, LiveCodeBench, SciCode.
- **Mathematical reasoning:** MATH-500, AIME 2024.
- **Scientific reasoning:** GPQA Diamond.
- **General knowledge and real-world QA:** MMLU-Pro, Humanity's Last Exam.
- **Operational characteristics:** blended cost per million tokens (USD), maximum context window, median output speed (tokens per second), and model parameter count.

Across these 12 metrics, some models do not have published scores for particular benchmarks or operational attributes (approximately 8% of all metric entries). To ensure that all models remain in the dataset and to avoid dropping observations unevenly across vendors, we use median imputation computed per benchmark. Median imputation avoids biasing imputations toward any vendor and preserves rank-order relationships. We imputed missing values using the benchmark-level median, preserving metric distributions while keeping all models in the dataset and preventing imputations from being biased toward any particular model family. Finally, since the benchmarks and operational metrics are on very different numerical scales we apply z-score standardization so that all metrics contribute comparably to the model.

3.2.4 *Response Mapping.*

The LLMs exhibited significant variation in their responses, due to the untouched *temperature* variable mimicking real interaction through the web interface. Models were referenced through multiple names (e.g. "ChatGPT 4o", "GPT-4o", "OpenAI's ChatGPT"). We developed a detailed mapping system to standardize these references.

The normalization process comprised three stages. First, all response strings were simplified alphanumerically through punctuation removal and case standardization. Second, we manually constructed an alias dictionary to map these strings to the canonical model names used in the performance metric dataset. Finally, we addressed edge cases, such as deprecated models or instances of multi-vendor affiliations. For ambiguous strings that could plausibly refer to multiple models, we used a secondary clarification step. We queried the same vendor's LLM (outside the audited runs) to infer the most likely public model name. We treated these resolutions as part of the data-cleaning process and to used them to keep recommendations as free as possible from any human bias.

3.2.5 *Dataset Construction.*

Each API call produced a ranked list of three model recommendations. To connect each ranked model recommendation to its objective capability (measured by benchmark scores) we restructured the data so that each ranked model became its own observation (row in the dataset) and appended the ranked-model specific benchmark scores to the observation. To ensure comparability across all performance metrics, we standardized all benchmark variables via z-score normalization. Each observation also contained a value for the binary flag *isSelfPromoted*, indicating whether the ranked model was created by its parent company.

Domain	Prompt
Coding	Which LLM is best for generating Python functions with few bugs?
Mathematics	Which language model excels at detailed algebra solutions?
Scientific Reasoning	Which LLM has the best scientific reasoning ability?
SpeedContextCost	What language model is most economical for large scale usage?
General Questions	Which LLM gives the most accurate answers?

Table 1: Sample Prompts by Domain

3.3 Statistical Analysis

We define a standard ordinary least squares (OLS) regression model to observe self-promotional bias:

$$\text{Rank}_i = \alpha + \beta_1(\text{isSelfPromoted}_i) + \beta_2(X_i) \tag{1}$$

where the variable $Rank_i$ represents the numeric ranking from 1 to 3, $isSelfPromoted_i$ is the binary flag indicating vendor self-recommendation, and X_i represents the set of objective performance features included as controls: GPQA Diamond (scientific reasoning), Humanity's Last Exam (real-world question answering), HumanEval (coding ability), as well as operational characteristics such as blended cost per million tokens, context window size, and median output speed. These variables serve as controls so that the regression isolates the effect of self-promotion from differences in actual model capability. A negative β_1 coefficient indicates preference for own-vendor models, as lower ranking correspond to high recommendations.

3.3.1 Feature Selection. Our initial model included all 12 performance indicators downloaded from the AI Leaderboard dataset, but preliminary analysis revealed significant feature multicollinearity which threatened the stability of the coefficients. We first addressed this through feature reduction, removing features with more than 20% missing values. We computed pairwise Pearson correlations and found several logically highly correlated pairs, notably MMLU-Pro and GPQA Diamond at $r \approx 0.93$, indicating that these benchmarks provide similar information. In such cases, we retained only the more comprehensive metric (GPQA Diamond) to avoid redundancy.

We computed Variance Inflation Factors (VIF) for the remaining features. VIF measures how much the variance of a regression coefficient increases in the presence of multicollinearity, calculated as $VIF_j = \frac{1}{1-R_j^2}$ where R_j^2 is the coefficient of determination when variable j is regressed on all other predictors. VIF values above 5 indicate problematic multicollinearity, and values above 10 can significantly destabilize coefficients. Hence, we removed features with values above 5 while retaining enough benchmarks to indicate model capability for each prompt domain.

This reduction led to a final set consisting of GPQA Diamond (Scientific Reasoning), HumanEval (Coding), and Humanity's Last Exam (Reasoning & Knowledge), as well as the isSelfPromoted flag and operational characteristics.

4 Results

Across 3,375 model ranking observations, our analysis reveals consistent evidence of self-promotional bias. Vendors systematically assign higher ranks to their own models than objective benchmarks would predict, with an average boost of roughly 0.2 positions on a three-point ranking scale. Figure 2 illustrates this raw relationship between provider affiliation and assigned rank. Provider-affiliated models consistently appear closer to rank 1 compared to competitor models, highlighting the presence of bias in the recommendations that users encounter directly when not controlled for benchmark performance. This effect is robust across all domains we tested (coding, mathematics, scientific reasoning, general knowledge, and operational factors), though the magnitude varies by domain. Vendor-level analysis highlights stark differences: OpenAI demonstrates the strongest and most systematic self-preferencing, Anthropic shows moderate effects, and Google's patterns are more complex, with results clouded by multicollinearity but suggesting weaker or inconsistent bias. Together, these findings demonstrate that self-promotional bias is a statistically significant phenomenon not explainable by performance benchmarks alone.

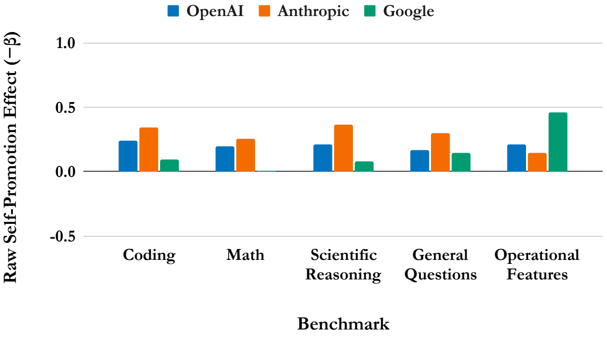


Figure 2: Raw relationship between provider affiliation and assigned rank across all domains.

4.1 Primary Regression Results

Table 2 below presents results from our final OLS model explaining 36.7% of the variation in rankings, with our key coefficient showing strong statistical significance. The coefficient for isSelfPromoted ($\beta = -0.202,\ p < 10^{-6}$) provides compelling evidence of self-promotional bias. The negative coefficient indicates that vendors consistently rank their own models about 0.2 positions higher than expected based purely on their benchmark scores. On a three point scale (ranks from 1 - 3), this is a significant difference capable of changing a ranking from first to second place.

We notice that ContextWindow seems to be the feature with the greatest weight, suggesting that models with larger context windows are generally ranked lower across our prompt domains. This likely reflects collinearity with vendor identity rather than user-valued preference and may reflect the limited variation in our dataset, with only 6 unique context window values, rather than indicating that context size is fundamentally important for ranking decisions.

4.1.1 Robustness Testing. Our core finding demonstrates remarkable robustness across multiple validation approaches. The permutation test with 5,000 iterations yielded $p < 10^{-6}$, providing strong confirmation that the observed self-promotional bias cannot be attributed to random variation.

4.2 Domain Specific Analysis

Breaking down our analysis by task category reveals that self-promotion isn't uniform across all areas but varies quite a bit depending on what users are asking about. We conducted separate regressions for each category using our established model specification and saved the results in Table 3.

Table 3 shows that self-promotional bias remains statistically significant across all 5 prompt domains. Scientific Reasoning shows the largest effect ($\beta = -0.212,\ R^2 = 0.662$) while Math demonstrates the least bias ($\beta = -0.131,\ R^2 = 0.587$). Notably, the fit for Math is relatively high despite being the only domain without specific benchmark data due to the issue of multicollinearity, indicating that the recommenders place low weight on the benchmarks when giving recommendations. Additionally, SpeedContextCost exhibits the strongest bias effect ($\beta = -0.209,\ R^2 = 0.141$), though with the poorest model fit, suggesting that operational characteristics like speed and cost may be areas where self-promotional tendencies are most pronounced but least systematically related to measurable performance metrics.

What strikes us most about these category results is that even in highly technical domains where objective performance measures are readily available and predictive, we still see consistent bias. The high R^2 values in Scientific Reasoning (66.2%) and Math (58.7%) indicate that benchmark performance strongly predicts rankings in these technical domains, yet significant bias persists even after controlling for objective capabilities. This suggests the effect isn't just about areas where quality is hard to measure-it's a more fundamental feature of how these systems operate.

4.3 Vendor-Specific Analysis

To understand whether self-promotional tendencies vary across different LLM providers, we conducted separate regressions for each major vendor using our established model specification. The results, presented in Table 4, reveal dramatically different approaches to self-promotion across the models.

The vendor-level regressions show substantial differences in the magnitude and direction of self-promotion. The OpenAI model shows the largest self-promotional effect, with their models receiving rankings nearly half a position higher than their benchmark scores would predict ($\beta = -0.445$). This substantial effect, combined with strong fit to the data ($R^2 = 0.500$), indicates that outputs from OpenAI's system systematically assign higher ranks to its own models.

The Anthropic system exhibits a smaller but still significant effect, showing moderate but statistically significant self-promotion ($\beta = -0.256$). Their models receive approximately a quarter-position boost beyond what objective performance metrics would justify. Like OpenAI, Anthropic's recommendations follow highly predictable patterns ($R^2 = 0.489$), indicating systematic rather than random promotional behavior.

Google's coefficients differ in sign and are harder to interpret. The positive coefficient ($\beta = 0.394$) initially suggests Google's outputs assign comparatively lower ranks to its own model. However, this result comes with important caveats. Google's analysis suffers from severe multicollinearity issues, with VIF values exceeding 11 for key variables and an extremely high Pairwise Pearson correlation ($r = 0.929$) between self-promotion and context window size. Additionally, the model's lower explanatory power ($R^2 = 0.357$) suggests substantial unmeasured factors influence Google's recommendation patterns. We therefore do not interpret the sign of Google's coefficient as evidence of reverse bias, but as an artifact of multicollinearity and limited benchmark separability.

The statistical clarity varies across vendors as OpenAI and Anthropic show clean, interpretable effects while Google's results remain clouded by methodological concerns. This underscores that self-promotional bias isn't a uniform phenomenon but rather reflects diverse approaches to balancing commercial interests with recommendation quality.

To further illustrate these differences, Figure 3 plots the estimated ranking inflation for each provider after controlling for benchmark performance. The y-axis shows the $-\beta$ coefficient from our regressions, where higher values indicate stronger upward adjustments in rank for provider-affiliated models. OpenAI stands out with the most pronounced effects, particularly in Scientific Reasoning and General Questions, where its models receive nearly half to a full position of upward adjustment. Anthropic shows more moderate but still consistent self-preferencing across domains. Google, by contrast, exhibits coefficients near zero or negative in several domains, suggesting weaker or inconsistent patterns. Together, the figure underscores that while self-promotional bias is a general phenomenon, its magnitude and direction vary substantially across providers.

5 Discussion

5.1 Self-promotional bias in LLMs

Our findings reveal systematic self-promotional bias across multiple large language model providers, with important implications for both users and the AI industry.

The 0.2 position ranking advantage we identified may seem modest in statistical terms, but translates to meaningful impact. In competitive markets where model capabilities continuously increase, even small ranking advantages can drive substantial differences in revenue and market share. Additionally, our domain specific analysis reveals particularly concerning patterns in technical areas. Scientific reasoning and coding tasks are domains where objective performance metrics exist and matter. This suggests that even when rankings could, in principle, fully align with external

Feature	Coefficient β	Std. Error	p-value
const	2.0000	0.011	$p < 10^{-6}$
isSelfPromoted	-0.2024	0.012	$p < 10^{-6}$
GPQA Diamond (Scientific Reasoning)	-0.2486	0.019	$p < 10^{-6}$
Humanity's Last Exam (Reasoning & Knowledge)	0.2227	0.014	$p < 10^{-6}$
HumanEval (Coding)	0.0580	0.016	1.95×10^{-4}
BlendedUSD/1M Tokens	-0.0817	0.012	$p < 10^{-6}$
ContextWindow	0.4146	0.013	$p < 10^{-6}$
MedianTokens/s	-0.0340	0.013	1.05×10^{-2}
Model fit: $R^2 = 0.367$, Adjusted $R^2 = 0.365$, $n = 3375$			

Table 2: OLS Regression Results

Category	β (isSelfPromoted)	Std. Error	R^2	N	p-value
Coding	-0.202	0.031	0.297	675	$p < 10^{-6}$
General Questions	-0.207	0.022	0.575	675	$p < 10^{-6}$
Math	-0.131	0.023	0.587	675	$p < 10^{-6}$
Scientific Reasoning	-0.212	0.020	0.662	675	$p < 10^{-6}$
SpeedContextCost	-0.209	0.032	0.141	675	$p < 10^{-6}$

Table 3: OLS Results by Prompt Domain

Vendor	β (isSelfPromoted)	Std. Error	R^2	N	p-value
OpenAI	-0.445	0.034	0.500	1125	2.53×10^{-37}
Anthropic	-0.256	0.028	0.489	1125	1.10×10^{-18}
Google	0.394	0.066	0.357	1125	2.86×10^{-9}

Table 4: Vendor-Specific Self-Promotional Bias

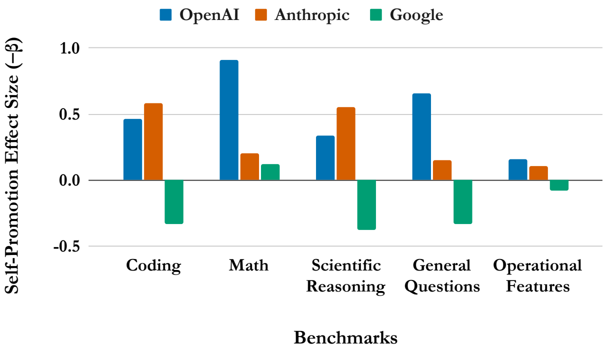


Figure 3: Controlled position inflation by vendor: $-\beta$ values from regressions indicate the magnitude of upward rank adjustments given to provider-affiliated models after accounting for benchmark performance.

performance benchmarks, other considerations influence their recommendations. Users seeking well-suited tools for technical work may unknowingly receive suboptimal suggestions.

5.2 Market Concentration and Structural Effects

These patterns take on high importance in a market already characterized by high fixed costs and resource barriers. Training frontier models requires substantial capital, compute, data access, evaluation infrastructure, and distribution channels [26]. Such barriers naturally concentrate model development within a small set of firms. In this context, self-preferential recommendation outputs operate as a reinforcement mechanism: control of the recommendation interface structurally channels attention toward the provider's own stack (models, APIs, tooling), which increases switching costs for developers and enterprises while limiting discovery of challengers. The dynamic mirrors earlier platform settings - search, e-commerce, and streaming - where subtle ranking advantages and interface design choices shaped traffic flows, user trust, and ultimately market structure [16, 18, 29, 30]. As in those cases, the concern is not limited to immediate effects (who gets chosen today), but extends to long-run competitive consequences when the advantages of being highly ranked accumulate over time.

Importantly, self-preferential ranking should also be interpreted within the realities of market position. Larger providers may be more frequently recommended for reasons that are not fully captured by benchmark controls such as availability, integration with existing tooling, documentation quality, enterprise support, or ecosystem complementarities. From a user-outcome perspective, some of these factors are legitimate decision criteria. However, the dual role of being both recommender and competitor means that, absent transparency and guardrails, commercial incentives and legitimate convenience factors are difficult to disentangle at the interface level.

5.3 LLMs and the competitive marketplace

A central contribution of this paper is to frame self-promotional bias in LLM outputs as an auditing problem in its own right. Just as earlier work on search and e-commerce documented how ranking practices could entrench incumbents and redirect consumer attention, our results suggest that LLMs may play a comparable gatekeeping role in the emerging AI marketplace [1]. The difference is that LLMs can not only act as information intermediaries but also market participants, making the risks of unexamined self-preferencing especially acute.

Introducing competition as an explicit dimension of AI auditing raises new questions for both researchers and policymakers: How should neutrality in recommendation outputs be defined? What kinds of evidence or benchmarks should be required when providers present rankings? And what forms of access or oversight are necessary to make such audits feasible? By surfacing these issues, our study underscores that the auditing of AI systems must expand beyond fairness and safety to encompass competition, ensuring that the growth of LLM ecosystems does not come at the expense of user choice or market openness. Looking forward, situating LLM auditing within the broader competition policy landscape creates opportunities for cross-disciplinary engagement. Economists, legal scholars, and HCI researchers alike have developed tools for assessing platform power in other domains [8, 25, 28]; applying and adapting those methods to LLMs could help build a systematic evidence base for regulators and the public. Future audits can track self-preferential bias prevalence across providers, domains, and time, making competition a core dimension of algorithmic accountability research.

5.4 Limitations

Our study aims to assess the quality of recommendations from LLMs given a certain demonstrated need, but fails to replicate the real-world context behind user chats. The additional complexity which we cannot model can contribute diverse preferences, constraints, and personal biases. Some apparent "bias" might reflect unmeasured factors relevant to user needs. Furthermore, the use of only benchmarks and operational characteristics as objective performance measures may not capture all dimensions relevant to user satisfaction. Factors such as output style, UI, or specialized capabilities could justify ranking patterns that appear biases through our lens.

Our analysis focuses on three major providers (OpenAI, Anthropic, Google) but excludes other significant players in the LLM market. The bias patterns we identify may not generalize to smaller providers, open-source models, or emerging market entrants. The AI industry's rapid evolution means that competitive dynamics, regulatory environments, and user expectations continue shifting. Our findings provide a sense of current models but may not predict future behavior as the market matures and potentially faces increased oversight. This leads to the largest limitation: the temporal constraints. Our data collection happens over a one day span, and uses models with varying knowledge cutoff dates. The specific biases we observe may no longer be the case as new models are released trained on a larger corpus of data and can better reference the benchmark data. Future work will be essential in judging whether this bias will continue.

6 Conclusion

This study provides systematic evidence of self-promotional bias in large language model recommendation systems. Our analysis of observed rankings across three major vendors demonstrates that systems consistently assign higher ranks to provider-affiliated models beyond what objective performance metrics would justify, with effects ranging from moderate (Anthropic) to substantial (OpenAI). We find that self-promotional bias persists across all tested domains, including highly technical areas where objective performance benchmarks exist and strongly predict rankings. The 0.2 average ranking boost appears modest but translates to meaningful competitive advantages in markets where users rely heavily on recommendation systems. The biases we identify operate subtly within otherwise functional recommendation systems, making them difficult for users to detect while maintaining systematic commercial advantages. Users should approach LLM recommendations with awareness of potential bias, particularly when selecting models for critical applications. Policymakers may consider whether disclosure requirements or algorithmic transparency mandates are necessary to ensure informed user choice.

References

[1] Muhammad Ali, Piotr Sapiezynski, Miranda Bogen, Aleksandra Korolova, Alan Mislove, and Aaron Rieke. 2019. Discrimination through Optimization: How Facebook's Ad Delivery Can Lead to Biased Outcomes. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 199 (Nov. 2019), 30 pages. doi:10.1145/3359301

[2] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine Bias: There's software used across the country to predict future criminals. And it's biased against blacks. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

[3] Artificial Analysis. 2025. Model & API Providers Analysis: LLM Leaderboard. https://artificialanalysis.ai/leaderboards/models. Accessed May 2025.

[4] Emily M Bender and Alex Hanna. 2025. *The AI Con: How to fight big tech's hype and create the future we want.* Harper.

[5] Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016. Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. arXiv:1607.06520 [cs.CL] https://arxiv.org/abs/1607.06520

[6] Joy Buolamwini and Timnit Gebru. 2018. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency (Proceedings of Machine Learning Research, Vol. 81)*, Sorelle A. Friedler and Christo Wilson (Eds.). PMLR, 77–91. https://proceedings.mlr.press/v81/buolamwini18a.html

[7] Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science* 356, 6334 (April 2017), 183–186. doi:10.1126/science.aal4230

[8] Sarah H Cen and Rohan Alur. 2024. From transparency to accountability and back: A discussion of access and evidence in ai auditing. In *Proceedings of the 4th ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*. 1–14.

[9] Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Josh Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. 2021. Evaluating Large Language Models Trained on Code. arXiv:2107.03374 [cs.LG] https://arxiv.org/abs/2107.03374

[10] Abhisek Dash, Abhijnan Chakraborty, Saptarshi Ghosh, Animesh Mukherjee, Jens Frankenreiter, Stefan Bechtold, and Krishna P. Gummadi. 2024. Antitrust, Amazon, and Algorithmic Auditing. *Journal of Institutional and Theoretical Economics (JITE) Forthcoming* (March 27, 2024). Available at SSRN: https://ssrn.com/abstract=4774657.

[11] Nicholas Diakopoulos. 2015. Algorithmic accountability: Journalistic investigation of computational power structures. *Digital journalism* 3, 3 (2015), 398–415.

[12] European Commission. 2025. Statement by the European Commission. https://ec.europa.eu/commission/presscorner/detail/en/STATEMENT_17_1806. Accessed on 10 August 2025.

[13] European Parliament and Council of the European Union. 2022. Regulation (EU) 2022/1925 of the European Parliament and of the Council of 14 September 2022 on contestable and fair markets in the digital sector and amending Directives (EU) 2019/1937 and (EU) 2020/1828 (Digital Markets Act). Official Journal of the European Union, L 265, pp. 1–66. https://eur-lex.europa.eu/eli/reg/2022/1925/oj/eng Text with EEA relevance.

[14] Federal Trade Commission and 17 State Attorneys General. 2023. FTC Sues Amazon for Illegally Maintaining Monopoly Power. Press Release, U.S. Federal Trade Commission. https://www.ftc.gov/news-events/news/press-releases/2023/09/ftc-sues-amazon-illegally-maintaining-monopoly-power Accessed 10 August 2025.

[15] Johann D. Gaebler, W. Cai, G. Basse, R. Shroff, S. Goel, and J. Hill. 2022. A causal framework for observational studies of discrimination. *Statistics and Public Policy* 9, 1 (2022), 26–48. doi:10.1080/2330443X.2022.2048267

[16] Jeffrey Gleason, Desheng Hu, Ronald E Robertson, and Christo Wilson. 2023. Google the gatekeeper: How search components affect clicks and attention. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 17. 245–256.

[17] Zheyin (Jane) Gu, Xinxin Li, and Tianxin Zou. 2024. Regulating Platform Self-Preferencing in Add-on Markets. *SSRN Electronic Journal* (10 Feb. 2024). Available at SSRN: https://ssrn.com/abstract=4722321 or https://dx.doi.org/10.2139/ssrn.4722321.

[18] Aniko Hannak, Gary Soeller, David Lazer, Alan Mislove, and Christo Wilson. 2014. Measuring Price Discrimination and Steering on E-Commerce Web Sites. In *Proceedings of the 14th ACM/USENIX Internet Measurement Conference (IMC '14)*. ACM, Vancouver, Canada, 305–318. doi:10.1145/2663716.2663726

[19] Jochen Hartmann, Jasper Schwenzow, and Maximilian Witte. 2023. The political ideology of conversational AI: Converging evidence on ChatGPT's pro-environmental, left-libertarian orientation. arXiv:2301.01768 [cs.CL] https://arxiv.org/abs/2301.01768

[20] Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring Mathematical Problem Solving With the MATH Dataset. *CoRR* abs/2103.03874 (2021). arXiv:2103.03874 https://arxiv.org/abs/2103.03874

[21] Naman Jain, King Han, Alex Gu, Wen-Ding Li, Fanjia Yan, Tianjun Zhang, Sida Wang, Armando Solar-Lezama, Koushik Sen, and Ion Stoica. 2024. LiveCodeBench: Holistic and Contamination Free Evaluation of Large Language Models for Code. arXiv:2403.07974 [cs.SE] https://arxiv.org/abs/2403.07974

[22] Adrianne Jeffries and Leon Yin. 2020. Google's Top Search Result? Surprise! It's Google. *The Markup* (2020). https://themarkup.org/google-the-giant/2020/07/28/google-search-results-prioritize-google-products-over-competitors.

[23] Thorsten Joachims, Laura Granka, Bing Pan, Helene Hembrooke, and Geri Gay. 2005. Accurately Interpreting Clickthrough Data as Implicit Feedback. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '05)* (Salvador, Brazil). ACM Press, New York, NY, USA, 154–161. doi:10.1145/1076034.1076063

[24] Soyoung Kim and Randi Priluck. 2025. Consumer Responses to Generative AI Chatbots Versus Search Engines for Product Evaluation. *Journal of Theoretical and Applied Electronic Commerce Research* 20, 2 (2025). doi:10.3390/jtaer20020093

[25] Danaë Metaxa, Joon Sung Park, Ronald E Robertson, Karrie Karahalios, Christo Wilson, Jeff Hancock, Christian Sandvig, et al. 2021. Auditing algorithms: Understanding algorithmic systems from the outside in. *Foundations and Trends® in Human–Computer Interaction* 14, 4 (2021), 272–344.

[26] Arvind Narayanan and Sayash Kapoor. 2024. *AI snake oil: What artificial intelligence can do, what it can't, and how to tell the difference.* Princeton University Press.

[27] Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. 2019. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* 366, 6464 (2019), 447–453. arXiv:https://www.science.org/doi/pdf/10.1126/science.aax2342 doi:10.1126/science.aax2342

[28] Jorge Padilla, Joe Perkins, and Salvatore Piccolo. 2022. Self-Preferencing in Markets with Vertically Integrated Gatekeeper Platforms. *The Journal of Industrial Economics* 70, 2 (2022), 371–395. doi:10.1111/joie.12287

[29] Brennan Schaffner, Antonia Stefanescu, Olivia Campili, and Marshini Chetty. 2023. Don't let Netflix drive the bus: user's sense of agency over time and content choice on Netflix. *Proceedings of the ACM on human-computer interaction* 7, CSCW1 (2023), 1–32.

[30] Daniel Trielli and Nicholas Diakopoulos. 2019. Search as news curator: The role of Google in shaping attention to news information. In *Proceedings of the 2019 CHI Conference on human factors in computing systems*. 1–15.

[31] Quang Minh Trinh, Samiha Zarin, and Rezvaneh Rezapour. 2025. Master of Deceit: Comparative Analysis of Human and Machine-Generated Deceptive Text. In *Proceedings of the 17th ACM Web Science Conference 2025 (Websci '25)*. Association for Computing Machinery, New York, NY, USA, 189–198. doi:10.1145/3717867.3717914

[32] Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhu Chen. 2024. MMLU-Pro: A More Robust and Challenging Multi-Task Language Understanding Benchmark. arXiv:2406.01574 [cs.CL] https://arxiv.org/abs/2406.01574

[33] Sam Witteveen and Martin Andrews. 2019. Paraphrasing with Large Language Models. In *Proceedings of the 3rd Workshop on Neural Generation and Translation*, Alexandra Birch, Andrew Finch, Hiroaki Hayashi, Ioannis Konstas, Thang Luong, Graham Neubig, Yusuke Oda, and Katsuhito Sudoh (Eds.). Association for Computational Linguistics, Hong Kong, 215–220. doi:10.18653/v1/D19-5623

[34] Wenda Xu, Guanglei Zhu, Xuandong Zhao, Liangming Pan, Lei Li, and William Wang. 2024. Pride and Prejudice: LLM Amplifies Self-Bias in Self-Refinement. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Bangkok, Thailand, 15474–15492. doi:10.18653/v1/2024.acl-long.826

[35] Kai-Cheng Yang and Filippo Menczer. 2025. Accuracy and Political Bias of News Source Credibility Ratings by Large Language Models. In *Proceedings of the 17th ACM Web Science Conference 2025 (Websci '25)*. Association for Computing Machinery, New York, NY, USA, 127–137. doi:10.1145/3717867.3717903

A Appendix

A.1 Variance Stabilization Figures

To validate our choice of five iterations per prompt, we examined how much model rankings varied as the number of iterations increased. Figures 4, 5, and 6 display the observed variance for one representative prompt in the *General Questions* domain across increasing iteration counts for each provider - *"Which LLM gives the most accurate answers?"* Across all three models, variance levels stabilize quickly, confirming that our chosen sample size is sufficient to capture stable recommendation patterns without unnecessary repetition.

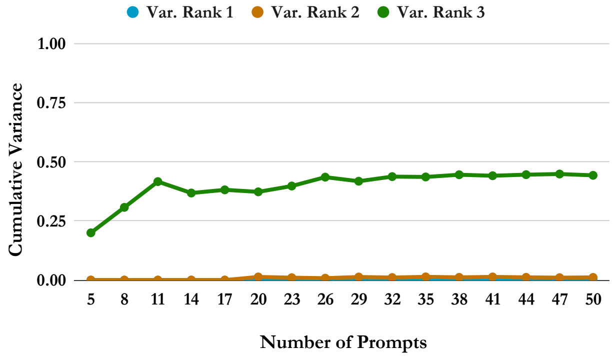


Figure 4: Variance in OpenAI's rankings for one prompt across increasing iteration counts (0 = stable, 1 = variable).

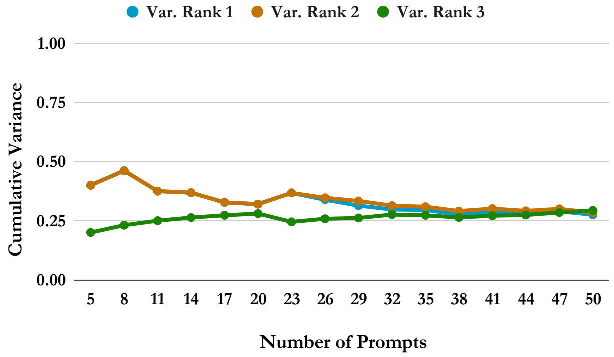


Figure 5: Variance in Anthropic's rankings for one prompt across increasing iteration counts (0 = stable, 1 = variable).

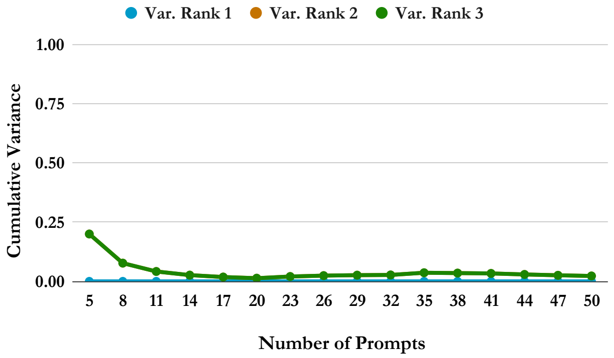


Figure 6: Variance in Google's rankings for one prompt across increasing iteration counts (0 = stable, 1 = variable).