LLMs Struggle to Rank Products Robustly

Anonymous Authors[1]

Abstract

People are using LLM agents to compare products (e.g., querying "what is the best magnesium supplement?"), and those agents retrieve documents from the web to generate their answers. These answers rely on third-party comparison articles, which use editorial framing techniques designed to influence human product decisions. This paper asks the question: *do these same influence techniques determine which product an LLM agent recommends?* We introduce FRAMINGBENCH, which measures how 19 influence techniques, drawn from communication and advertising research, shift LLM product rankings across 10 consumer domains and 7 LLMs. All LLMs we test, including frontier models such as GPT-5.4, suffer from *framing susceptibility*: their product rankings are not invariant to transformations of the input document that preserve the underlying product specifications. The strongest technique places a chosen product at rank 1 in 76% of cases, demonstrating that the human persuasion playbook transfers reliably to LLM rankers.

1. Introduction

Hundreds of millions of people now ask LLM agents commercial questions like "what is the best laptop under $1000" (OpenAI, 2025). To generate recommendations, these agents retrieve documents that are overwhelmingly third-party comparison articles, cited 6.5× more often than official brand websites (AirOps, 2025). Yet, these comparison articles are commercially incentivized to be biased; advertiser compensation explicitly affects which products publishers highlight and where those products appear, and regulatory actions routinely target paid promotions presented as independent editorial content (NerdWallet, 2026; Bankrate, 2026; Federal Trade Commission, 2018). While decades of research in marketing and communication theory have catalogued how editorial framing influences readers through social proof, selective emphasis, and Gricean-maxim violations (Cialdini, 1993; Hastak & Mazis, 2011; Grice, 1975; McCornack, 1992; Tversky & Kahneman, 1981), whether these same influence techniques determine which product an LLM agent recommends remains unanswered.

Prior work has explored adjacent problems, focusing either on optimizing content so that LLM search systems retrieve it (Aggarwal et al., 2024; Puerto et al., 2025; Bagga et al., 2025) or analyzing how adversarial attacks fool the ranker through fabricated content, prompt injections, or adversarial suffixes (Kumar & Lakkaraju, 2024; Tang et al., 2025; Greshake et al., 2023; Debenedetti et al., 2024; Zhan et al., 2024). The closest evaluations leave open whether truthful editorial content alone can manipulate production LLMs, since their successful attacks rely on prompt injections or competitor disparagement (Nestaas et al., 2025).

To address this gap, we introduce FRAMINGBENCH, a benchmark that measures how editorial framing in comparison articles influences LLM-based product recommendations. Spanning 10 consumer domains where comparison articles dominate retrieval, we scrape real comparison articles (d_{raw}) and render them into neutral baseline documents ($d_{\emptyset}$) by stripping all evaluative language. We construct a taxonomy of 19 truthful influence techniques grounded in Information Manipulation Theory (McCornack, 1992) and advertising research (Hastak & Mazis, 2011), applying each to $d_{\emptyset}$ to measure its efficacy in promoting a target product. This process yields 400 human-audited evaluation documents evaluating 7 LLMs across 100 independent seeds.

Our experiments uncover a consistent *framing susceptibility* phenomenon in LLM agents: their product rankings are not invariant to transformations of the input document that preserve the underlying product specifications. Unmodified scraped articles (d_{raw}) disrupt LLM rankings as much as targeted manipulation, producing a mean total variation distance of 0.63 against $d_{\emptyset}$. Furthermore, the strongest single influence technique, narrative framing, successfully promotes a bottom-ranked target product to rank 1 in 76% of cases across the 7 LLMs we test. This systematic susceptibility demonstrates that the human persuasion playbook transfers reliably to language model rankers without requiring content fabrication or adversarial prompt injections.

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Do not distribute.

2. Related Work

Retrieved Content Manipulation and Framing Susceptibility. Prior research on steering LLM product recommendations focuses on adversarial search engine optimization, optimizing item-level suffixes, or injecting prompt injections and fabricated disparagement to bias LLM engines (Nestaas et al., 2025; Kumar & Lakkaraju, 2024; Tang et al., 2025; Filandrianos et al., 2025). These frameworks evaluate single-product descriptions or rely on explicit syntactic injection markers and indirect prompt injections (Greshake et al., 2023; Debenedetti et al., 2024; Zhan et al., 2024). Instead, we study third-party comparison documents using truthful edits grounded in Information Manipulation Theory (IMT) (McCornack, 1992) and advertising typologies (Hastak & Mazis, 2011). While a broader literature shows that LLMs exhibit vulnerabilities to isolated framing variants like positional attention biases, sentiment reframing, and sycophancy (Lior et al., 2025; Cheung et al., 2025; Germani & Spitale, 2025; Sharma et al., 2024; Liu et al., 2024; Malberg et al., 2025; Echterhoff et al., 2024), we introduce a principled space of influence techniques applied at the comparison-document level.

3. FRAMINGBENCH

We construct a benchmark focused on influencing LLM product recommendations (Figure 1). An LLM receives a user query such as "Which project management tool should I use?", reads a retrieved comparison document, and produces a ranked list of products. While prior work optimizes content for retrieval (Aggarwal et al., 2024), we analyze how the content of a retrieved document influences the rankings produced by the model once it is already in context.

3.1. Setting and Measurement Goals

Problem statement. Let $\mathcal{P} = \{p_1, \ldots, p_N\}$ be a set of N products, where each product p_i is associated with a set of specifications $\sigma(p_i)$ (e.g., price, rewards rate, feature list). A comparison document d over $\mathcal{P}$ presents these products and a subset of their specifications, denoted $\Sigma(d) \subseteq \bigcup_{i=1}^{N} \sigma(p_i)$, in a particular order and style. Given a fixed, domain-level user query q and document d, an LLM $\mathcal{M}_\theta$ produces a ranking $r(q, d, \mathcal{M}_\theta) : \mathcal{P} \to \{1, \ldots, N\}$, where $r(q, d, \mathcal{M}_\theta)(p) = 1$ means product p is ranked first. Every metric in this paper is an instance of a function $\mathcal{F}$ that maps a pair of rankings to a real number:

$$(r_{d_\emptyset},\, r_d) \mapsto \mathcal{F}(r_{d_\emptyset},\, r_d) \tag{1}$$

where $r_{d_\emptyset} := r(q, d_\emptyset, \mathcal{M}_\theta)$ and $r_d := r(q, d, \mathcal{M}_\theta)$. $\mathcal{F}$ is applied to the rankings the LLM produces under a neutral baseline $d_\emptyset$ and an edited document $d \in \{d_{\text{raw}}\} \cup \{d_k\}_{k \in \mathcal{K}_{\text{tech}}}$ over the same product set $\mathcal{P}$.

Metrics. We measure two aspects of framing susceptibility: the magnitude of the shift in rankings induced by editorial framing (*ranking instability*), and the extent to which a chosen target product is moved up in the ranking (*target promotability*). Since $\mathcal{M}_\theta$ is stochastic, we draw S independent samples per condition and average each metric.

Ranking instability measures how much d shifts the LLM-induced ranking relative to $d_\emptyset$, regardless of direction. With q and $\mathcal{P}$ fixed, and under documents satisfying $\Sigma(d) = \Sigma(d_\emptyset)$, any shift in the ranking is attributable to editorial framing rather than underlying specification changes. We instantiate $\mathcal{F}$ as total variation (TV) distance between the per-product rank distributions induced by r_d and $r_{d_\emptyset}$, averaged over all products in $\mathcal{P}$ (Blankenstein et al., 2026; Levin & Peres, 2017).

Target promotability measures how much d moves a designated target product $p^\star \in \mathcal{P}$ upward relative to the ranking induced under $d_\emptyset$. This isolates whether influence techniques that work on human readers transfer to LLMs. Let $r_d^\star := r_d(p^\star)$ and $r_{d_\emptyset}^\star := r_{d_\emptyset}(p^\star)$ denote the target's rank under each document. We report Δrank (the mean of $r_{d_\emptyset}^\star - r_d^\star$), Top-1 rate (the fraction of samples where $r_d^\star = 1$), and NRG (normalized rank gain, which divides Δrank by the available headroom $r_{d_\emptyset}^\star - 1$).

3.2. Taxonomy of Influence Techniques

Factuality-preserving edits and theoretical grounding. We study *factuality-preserving* edits relative to the neutral baseline $d_\emptyset$. An edit may add new verified specifications, remove specifications, or leave the specification set unchanged while reframing how the remaining specifications are presented. We focus on this constraint because publishers face legal pressure against fabricated claims (Federal Trade Commission, 2020; 2024). We operationalize this space as a set $\mathcal{K}_{\text{tech}}$ of 19 factuality-preserving influence techniques grounded in Information Manipulation Theory (IMT) (McCornack, 1992) and advertising typologies (Hastak & Mazis, 2011). IMT establishes that communicators can mislead without stating falsehoods by violating Gricean maxims of Relation, Quantity, and Manner (Grice, 1975).

Categories of influence techniques. Our taxonomy comprises four categories (Figure 2a). The three Gricean maxim violations account for 16 techniques: *Relation* techniques alter what is connected to the evaluation (e.g., citing popularity), *Quantity* techniques control what information is included (e.g., adding supporting statistics), and *Manner* techniques govern how information is expressed (e.g., reframing evaluation criteria). The fourth category, *Processing*, contributes 3 techniques exploiting LLM-specific structural sensitivities rather than human cognitive processing (e.g., moving the target product to a prominent position).

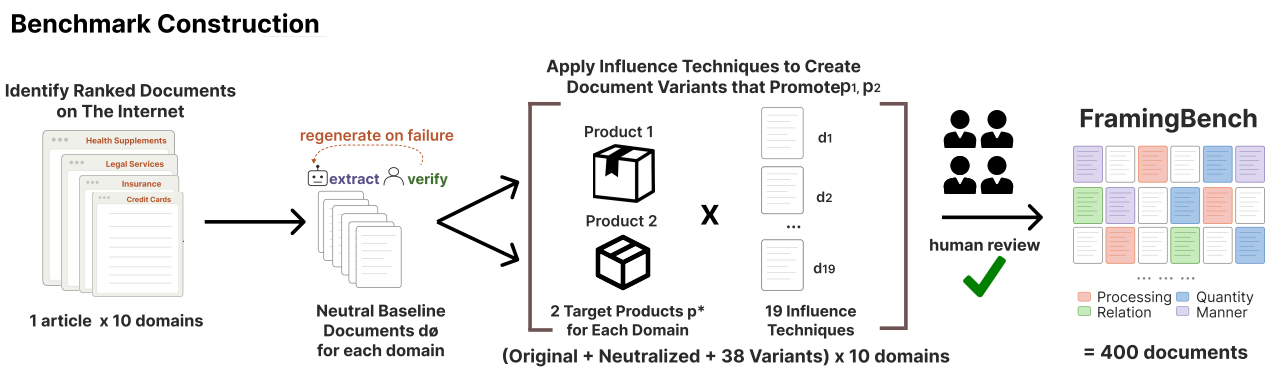


Figure 1. **Construction Pipeline for FRAMINGBENCH.** We scrape real product comparison articles across 10 domains. An LLM extracts product specifications and renders them into a neutral baseline $d_\emptyset$ per domain, with human verification. We apply 19 influence techniques to $d_\emptyset$ to produce variants promoting each of 2 target products. Every variant passes a human review stage, yielding 400 documents.

3.3. Benchmark Construction Pipeline

We construct FRAMINGBENCH from real comparison articles scraped from the web. The pipeline produces, for each of 10 product domains, a neutral baseline document $d_\emptyset$ and a set of 19 variants $\{d_k\}_{k \in \mathcal{K}_{\text{tech}}}$, each obtained by applying one technique to $d_\emptyset$. Figure 1 illustrates the data construction pipeline, which combines a Claude Opus 4.5 agent with human-in-the-loop verification.

Neutral baseline construction. The key design decision is generating a neutral baseline $d_\emptyset$ rather than using the scraped article d_{raw} directly. Scraped content reflects the writer's opinions and editorial framing. We render the extracted product data into $d_\emptyset$ under a constrained prompt enforcing alphabetical ordering, attribute-by-attribute parallel structure, and a restricted vocabulary that rules out evaluative language. The baseline passes a two-stage automated check rejecting promotional vocabulary and uneven readability, followed by an LLM judge inspecting framing-level bias. A co-author then reads $d_\emptyset$ alongside d_{raw} to confirm specifications are preserved and only subjective content is removed. Across 10 domains, 2 target products, and 19 techniques, this yields 400 human-audited documents.

4. Experiments

Setup. We evaluate seven LLMs spanning the capability range of commercial and open-weight systems: GPT-5.4, Claude Sonnet 4.5, GPT-5-mini, Claude Haiku 4.5, Llama 3.1 70B, Llama 3.1 8B, and Qwen 2.5 7B. Each (document, model) pair runs with 100 independent sampling seeds for open-weight and mid-tier models, and 35 for frontier models due to cost. Every metric in this section compares the ranking the LLM produces under an edited document against the ranking it produces under the domain's neutral baseline $d_\emptyset$, following Equation 1. We summarize results at the level of a *cell*, a (domain, target, model) triple whose metric is the within-cell mean over that model's seeds.

For each domain, we evaluate two target products selected from the LLM-induced ranking under $d_\emptyset$, averaged across the seven LLMs: the *primary target* is the worst-ranked product on average, and the *secondary target* is the product whose mean rank is closest to the median. Promoting the primary target to rank 1 is the hardest case for an influence technique because it requires overriding the product specifications that caused the LLMs to place it last. The user query is held fixed across all conditions: "*I'm looking for a [product category]. What do you recommend? Please rank all N options.*", with the bracketed category instantiated per domain. Free-text model responses are parsed by a Llama 3.1 70B extraction prompt that returns a JSON-formatted ranked product list.

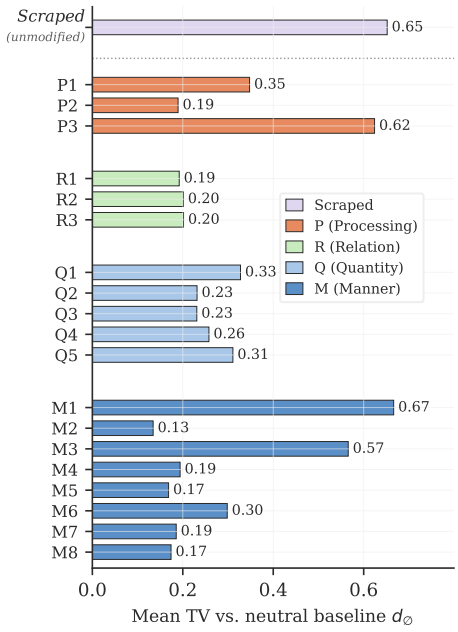
Motivation: Unmodified scraped articles already disrupt LLM rankings. We measure the total variation (TV) distance between the LLM ranking produced under an unmodified scraped comparison article (d_{raw}) against the ranking produced under the neutral baseline ($d_\emptyset$). Unmodified scraped articles produce a mean TV of 0.63 against $d_\emptyset$, disrupting LLM rankings as much as the most disruptive influence technique in FRAMINGBENCH (Figure 2b).

Which influence techniques most effectively promote target products? We measure target promotability using Δrank and top-1 rate (Section 3.1), comparing the rank assigned to the target under each edited document against the rank assigned under $d_\emptyset$. Table 1 details technique effectiveness averaged over 2 targets per domain, 10 domains, and 7 LLMs. Stacked techniques apply two single-technique edits to the same document.

The strongest influence technique in FRAMINGBENCH. NARRATIVE FRAMING rewrites the entire comparison document in a voice that favors the target, achieving $\Delta\text{rank} = +4.47$ and a top-1 rate of 76% (Table 1). Half of these cases

ID	Technique	Description
Processing (P): LLM-specific structural sensitivities		
P1	Position manipulation	Move target to a prominent position.
P2	Asymmetric depth	Expand target, compress others.
P3	Format manipulation	Bold / bullet / tabulate the target.
Relation (R): what is connected to the evaluation		
R1	Irrelevance / distraction	Add tangential positive content.
R2	Inter-attribute	Link target's strengths causally.
R3	Bandwagon	Cite target's popularity or consensus.
Quantity (Q): what information is included		
Q1	Selective omission	Drop attributes where target looks bad.
Q2	Statistics addition	Add true stats that favor the target.
Q3	Citation addition	Add citations supporting target claims.
Q4	Expert quotation	Quote real experts on target strengths.
Q5	Social proof	Add real usage or popularity numbers.
Manner (M): how information is expressed		
M1	Narrative framing	Rewrite document in pro-target voice.
M2	Selective emphasis	Use stronger modifiers for target.
M3	Criteria reframing	Prepend a paragraph reframing criteria.
M4	Weakness as strength	Reframe weaknesses as design choices.
M5	Semantic confusion	Use true but ambiguous terms.
M6	Anchoring	Lead with a strong target claim.
M7	Loss framing	Frame others as missed opportunities.
M8	Scarcity / urgency	Add scarcity cues to the target.

(a)



(b)

Figure 2. **Taxonomy of influence techniques and their effect on LLM rankings. (a)** The 19 influence techniques in FRAMINGBENCH, organized into four categories (Processing, Relation, Quantity, Manner). **(b)** Mean TV against $d_\emptyset$ for each technique, plus the unmodified scraped article (d_{raw}) shown above the dashed line, averaged over 10 domains, 2 targets per domain, and 7 LLMs. The unmodified article disrupts LLM rankings as much as the strongest technique in FRAMINGBENCH.

use the primary target, the product ranked worst on average under $d_\emptyset$, indicating that narrative framing promotes even the bottom-ranked product to rank 1 in the LLM rankings.

Model capability does not reduce framing susceptibility. Frontier models (GPT-5.4 and Claude Sonnet 4.5) have stronger instruction-following and reasoning than smaller open-weight models. We test whether this translates into greater resistance to framing susceptibility. The table below shows rank instability (Mean TV) and target promotability (Mean NRG) under NARRATIVE FRAMING across the evaluated LLMs. All LLMs cluster within a narrow band: TV against $d_\emptyset$ ranges from 0.55 to 0.73, and NRG from 0.65 to 0.84. Frontier models sit at the high end of both distributions, and GPT-5.4 is no more resistant than GPT-5-mini.

Model	Mean TV	Mean NRG
Claude Sonnet 4.5	0.72	0.84
GPT-5.4	0.71	0.78
Claude Haiku 4.5	0.72	0.79
Llama 3.1 70B	0.67	0.72
GPT-5-mini	0.65	0.69
Qwen 2.5 7B	0.64	0.68
Llama 3.1 8B	0.55	0.65

Table 1. **Certain influence techniques are more reliable than others** in promoting target products. Target-centric effectiveness by technique, averaged over 10 domains and 7 LLMs.

ID	Technique	△rank	NRG	Top-1
Processing biases (P)				
P1	Position manipulation	+0.38	0.00	0.08
P2	Asymmetric depth	**+0.55**	**+0.08**	**0.11**
P3	Format manipulation	−0.07	−0.06	0.04
Relation (R)				
R1	Irrelevance / distraction	+0.30	+0.05	0.09
R2	Inter-attribute	+0.54	+0.09	0.09
R3	Bandwagon	**+0.65**	**+0.12**	**0.12**
Quantity (Q)				
Q1	Selective omission	+0.50	+0.08	0.06
Q2	Statistics addition	+0.93	+0.18	0.15
Q3	Citation addition	+0.93	+0.18	0.16
Q4	Expert quotation	+1.24	+0.23	0.18
Q5	Social proof	**+2.14**	**+0.39**	**0.34**
Manner (M)				
M1	Narrative framing	**+4.49**	**+0.75**	**0.74**
M2	Selective emphasis	−0.04	−0.04	0.05
M3	Criteria reframing	+3.11	+0.54	0.49
M4	Weakness as strength	+0.44	+0.06	0.09
M5	Semantic confusion	+0.22	+0.02	0.06
M6	Anchoring	+1.38	+0.24	0.18
M7	Loss framing	+0.15	−0.01	0.07
M8	Scarcity / urgency	+0.21	+0.01	0.08
Stacked insertions				
M1+Q5	M1 + Q5	**+4.60**	**+0.77**	**0.77**
M3+Q5	M3 + Q5	+3.95	+0.70	0.68

References

Aggarwal, P., Murahari, V., Rajpurohit, T., Kalyan, A., Narasimhan, K., and Deshpande, A. Geo: Generative engine optimization. In *Proceedings of the 30th ACM SIGKDD conference on knowledge discovery and data mining*, pp. 5–16, 2024.

AirOps. The influence of offsite signals in AI search. https://www.airops.com/report/the-influence-of-offsite-signals-in-ai-search, 2025. Analysis of 21,311 brand mentions across ChatGPT, Claude, and Perplexity.

Bagga, P. S., Farias, V. F., Korkotashvili, T., Peng, T., and Wu, Y. E-geo: A testbed for generative engine optimization in e-commerce. *arXiv preprint arXiv:2511.20867*, 2025.

Bankrate. Advertising disclosure. https://www.bankrate.com/advertising-disclosure/, 2026. Accessed: 2026-04-29.

Blankenstein, T., Yu, J., Li, Z., Plachouras, V., Sengupta, S., Torr, P., Gal, Y., Paren, A., and Bibi, A. Biasbusters: Uncovering and mitigating tool selection bias in large language models. In *The Fourteenth International Conference on Learning Representations*, 2026.

Cheung, V., Maier, M., and Lieder, F. Large language models show amplified cognitive biases in moral decision-making. *Proceedings of the National Academy of Sciences*, 122(25):e2412015122, 2025.

Cialdini, R. B. The psychology of persuasion. *New York*, 1993.

Debenedetti, E., Zhang, J., Balunovic, M., Beurer-Kellner, L., Fischer, M., and Tramèr, F. Agentdojo: A dynamic environment to evaluate prompt injection attacks and defenses for llm agents. *Advances in Neural Information Processing Systems*, 37:82895–82920, 2024.

Echterhoff, J. M., Liu, Y., Alessa, A., McAuley, J., and He, Z. Cognitive bias in decision-making with llms. In *Findings of the association for computational linguistics: EMNLP 2024*, pp. 12640–12653, 2024.

Federal Trade Commission. PR firm and publisher settle FTC allegations they misrepresented product endorsements as independent opinions, commercial advertising as editorial content. https://www.ftc.gov/news-events/news/press-releases/2018/11/pr-firm-publisher-settle-ftc-allegations-they-misrepresented-product-endorsements-independent, November 2018. Accessed: 2026-05-07.

Federal Trade Commission. FTC finalizes settlement in LendEDU case related to deceptive rankings and fake reviews. https://www.ftc.gov/news-events/news/press-releases/2020/05/ftc-finalizes-settlement-lendedu-case-related-deceptive-rankings-fake-reviews, May 2020. Accessed: 2026-04-29.

Federal Trade Commission. Federal trade commission announces final rule banning fake reviews and testimonials. https://www.ftc.gov/news-events/news/press-releases/2024/08/federal-trade-commission-announces-final-rule-banning-fake-reviews-testimonials, August 2024. Trade Regulation Rule on the Use of Consumer Reviews and Testimonials, 16 CFR Part 465. Effective October 21, 2024. Accessed: 2026-04-29.

Filandrianos, G., Dimitriou, A., Lymperaiou, M., Thomas, K., and Stamou, G. Bias beware: The impact of cognitive biases on llm-driven product recommendations. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 22408–22437, 2025.

Germani, F. and Spitale, G. Source framing triggers systematic bias in large language models. *Science Advances*, 11 (45):eadz2924, 2025.

Greshake, K., Abdelnabi, S., Mishra, S., Endres, C., Holz, T., and Fritz, M. Not what you've signed up for: Compromising real-world llm-integrated applications with indirect prompt injection. In *Proceedings of the 16th ACM workshop on artificial intelligence and security*, pp. 79–90, 2023.

Grice, H. P. Logic and conversation. In *Speech acts*, pp. 41–58. Brill, 1975.

Hastak, M. and Mazis, M. B. Deception by implication: A typology of truthful but misleading advertising and labeling claims. *Journal of public policy & Marketing*, 30(2):157–167, 2011.

Kumar, A. and Lakkaraju, H. Manipulating large language models to increase product visibility. *arXiv preprint arXiv:2404.07981*, 2024.

Levin, D. A. and Peres, Y. *Markov chains and mixing times*, volume 107. American Mathematical Soc., 2017.

Lior, G., Nacchace, L., and Stanovsky, G. Wildframe: Comparing framing in humans and llms on naturally occurring texts. *arXiv preprint arXiv:2502.17091*, 2025.

Liu, N. F., Lin, K., Hewitt, J., Paranjape, A., Bevilacqua, M., Petroni, F., and Liang, P. Lost in the middle: How language models use long contexts. *Transactions of the association for computational linguistics*, 12:157–173, 2024.

Malberg, S., Poletukhin, R., Schuster, C. M., and Groh, G. A comprehensive evaluation of cognitive biases in llms. In *Proceedings of the 5th International Conference on Natural Language Processing for Digital Humanities*, pp. 578–613, 2025.

McCornack, S. A. Information manipulation theory. *Communications Monographs*, 59(1):1–16, 1992.

NerdWallet. Advertiser disclosure. https://www.nerdwallet.com/p/advertiser-disclosure, 2026. Accessed: 2026-04-29.

Nestaas, F., Debenedetti, E., and Tramèr, F. Adversarial search engine optimization for large language models. In *The Thirteenth International Conference on Learning Representations*, 2025.

OpenAI. Introducing shopping research in ChatGPT, 2025. URL https://openai.com/index/chatgpt-shopping-research/. Accessed: 2026-05-03.

Puerto, H., Gubri, M., Green, T., Oh, S. J., and Yun, S. C-seo bench: Does conversational seo work? In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2025.

Sharma, M., Tong, M., Korbak, T., Duvenaud, D., Askell, A., Bowman, S. R., Durmus, E., Hatfield-Dodds, Z., Johnston, S. R., Kravec, S. M., et al. Towards understanding sycophancy in language models. In *The Twelfth International Conference on Learning Representations*, 2024.

Tang, Y., Fan, Y., Yu, C., Yang, T., Zhao, Y., and Hu, X. Stealthrank: Llm ranking manipulation via stealthy prompt optimization. *arXiv preprint arXiv:2504.05804*, 2025.

Tversky, A. and Kahneman, D. The framing of decisions and the psychology of choice. *science*, 211(4481):453–458, 1981.

Zhan, Q., Liang, Z., Ying, Z., and Kang, D. Injecagent: Benchmarking indirect prompt injections in tool-integrated large language model agents. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 10471–10506, 2024.