Position Bias in LLMs for Critical Decision Support - A Case Study on Multiple Casualty Triage

Ulrika Wickenberg-Bolin[1], Katie Cohen[1], Helena Björnesjö[1], and Agnes Tegen[1]

Swedish Defence Research Agency (FOI), Sweden
firstname.lastname@foi.se

Abstract. Large Language Models (LLMs) are increasingly deployed in high-stakes domains such as emergency response, medical triage, and security operations. This study investigates the effects of position bias, i.e., the tendency of LLMs to prioritize information based on its position in a list rather than its relevance, using a controlled, fictitious multiple casualty triage scenario. In a list of patients, we vary the position of the most critically injured one to evaluate whether GPT-4o and GPT-4o mini systematically deviate from the medically established START triage protocol. Our results reveal a consistent recency bias in both models: the most critically injured patient was less likely to be prioritized when listed first. This effect was more pronounced in shorter patient lists, challenging the common assumption that short prompts are inherently less prone to evoke model bias. These findings raise critical concerns about the operational reliability of LLMs in time-sensitive, high-stakes tasks. Our study contributes to growing evidence that LLMs require rigorous validation before deployment in sensitive environments such as OSINT, defense informatics, and emergency triage.

Keywords: position bias · LLMs · triage · emergency response

1 Introduction

The rapid integration of Large Language Models (LLMs) into mission-critical systems - including emergency response, situational assessment, cybersecurity decision support, and open-source intelligence workflows - raises urgent concerns about reliability, transparency, and systemic bias [5, 16, 14, 27, 12]. In domains where human lives or national security may be at stake, even subtle distortions in model reasoning can lead to catastrophic consequences [29, 7]. As LLMs are increasingly deployed in high-stakes, time-sensitive contexts, their outputs must be robust not only to adversarial prompts or long-context complexity, but also to more subtle structural biases that can arise from short and seemingly simple inputs [17].

Different structural patterns of biased reasoning, often referred to as cognitive biases, have been observed in LLM models [25, 8], as well as in human-LLM interactions [25, 32]. One such bias is *position bias*, the tendency of an LLM to weigh list items depending on their order of appearance rather than on their informational value [25, 8, 32]. From a cognitive standpoint, position bias mirrors classic serial position effects observed in human memory such as primacy and recency biases [21], and has been linked to LLM behaviors like the "lost-in-the-middle" effect [18].As these biases are increasingly recognized as emergent behaviors rather than technical artifacts [23], their presence in public- or safety-critical systems raises urgent concerns about the reliability of AI-supported decisions.

This study is situated within broader discussions surrounding LLM explainability, prompt sensitivity, and the dynamics of human-AI interaction. Ethical concerns surrounding LLM use in emergency and medical domains have been raised [11], with some studies warning that user biases may align with LLM bias in a way that generates a vicious circle where bias is maintained and possibly amplified [31]. While thoughtful prompt engineering can mitigate some bias effects [28], non-expert users may unintentionally trigger biased behavior through subtle prompt variations [1]. From a user perspective, the lack of transparency in the internal decision-making processes of LLMs limits the user's ability to assess, control, and counteract bias that emerges in the system.

While position bias has been well-documented in long-context tasks, its presence and effects in short-context scenarios remain underexplored. Given that many operational tasks, from triage to threat ranking, involve compact, ordered information where prioritization accuracy is essential, position effects in smaller contexts need to be addressed as well. The work described in the present paper explores whether position bias also occurs in a specific high-stakes short input context task, and, if so, whether there is a lower limit to how short the context has to be for the position effect not to affect the output. The overarching aim is to examine position bias in a short context ranking task, as it may occur for a layman user of LLMs not trained on a specific domain. To investigate this, we design a fictitious controlled experimental scenario inspired by an emergency triage context. This setup serves as a proxy for real-world prioritization tasks in emergency or security-critical systems, where failure to correctly rank cases can have severe consequences. We translate the scenario into a structured prompt and evaluate whether GPT-4o and GPT-4o mini are able to consistently identify and rank the most critically injured patient - regardless of where that patient appears in the list.

The primary research question, Q1, is: When given a short context ranking task in which to select items in a list, will the list-wise positions of items affect the LLM's decision? The secondary research question, Q2, is: If position bias such as stated in Q1 affects the LLMs decision; How does the size of the context window affect the extent to which position bias can be observed?

2 Related work

Prior research has observed cognitive biases in LLM models, including anchoring effects, framing bias, and the endowment effect [25], as well as conjunction bias, confirmation bias, and probability weighting [8]. Cognitive bias has also been found to surface in human-LLM interactions [25, 32]. For instance, a study on human-AI collaboration on a medical triage task observed that participants were more likely to trust AI triage recommendations if the recommendations aligned with the participant's prior preferences [6]. Indeed, medical decision-making relies heavily on decision heuristics [19], and the possible strengths and challenges associated with prompt engineering as a pedagogical tool in medical education have been under discussion [13]. Notwithstanding, in high-stakes decision-making scenarios, the effects of cognitive bias may occasionally be severe [22].

Position bias, in particular, can be conceptualized as a computational analogue of well-known memory effects in psychology known as serial position effects, where recall accuracy depends on an item's position in a sequence [21]. LLMs, though lacking true memory, often exhibit similar heuristics. This may arise, in part, from their autoregressive architecture, which processes tokens sequentially and may emphasize recent inputs over earlier ones [26, 18, 25]. For instance, primacy bias is the tendency to recall the information that is presented first [2], while recency bias is the tendency to recall the the most recently encountered items [4]. One instance of position bias in LLMs is the so called *lost-in-the-middle* problem, where LLMs display information loss when critical information occurs in the middle of long contexts [18].

Prompt formulation plays a key role in shaping LLM outputs, including the emergence or suppression of bias. Even subtle changes in phrasing or structure can alter the model's behavior substantially [1, 30]. This introduces both opportunity and risk: while expert users can use prompt engineering to mitigate known issues, non-expert users in high-pressure environments may unknowingly elicit biased responses. In security, medical, and emergency domains, where biased or incorrect outputs can have severe consequences, this is a salient concern. Several studies have called for more systematic evaluation of LLMs in deployment-like conditions, particularly where ranking or classification decisions can impact safety or resource allocation [11, 31].

Position bias in LLMs has been widely studied in long-context tasks, including multi-document Q & A and summarization, where attention limitations lead to phenomena like the "lost-in-the-middle" effect [18], or recency effects in reasoning tasks and retrieval-augmented generation (RAG) setups [33, 10]. In contrast, short-context tasks have received little attention. A common but largely untested assumption is that shorter prompts mitigate ordering effects due to reduced input length and cognitive load [23, 20]. Only recently, it has been argued that short contexts offer less redundancy and may amplify position effects through learned priors or autoregressive attention [9, 15]. To our knowledge, no prior work has systematically examined position bias in high-stakes, short-input ranking tasks despite their prevalence in time-sensitive domains such as triage, dispatch, or security alerting.

3 Experimental Setup

The experiments aim to determine whether, and if so, to what extent, the order of which items are presented in a prompt influences the response recommendation of an LLM on a ranking task. For this purpose, a fictitious scenario was designed, in which the LLM was prompted to perform a triage task and provide a priority sequence for evacuating a group of patients. In the scenario, nine patients are described as having suffered various injuries from an explosive device detonation, and therefore having to be evacuated one by one and receive appropriate medical care. To yield a proper response, the LLM has to disregard the order in which the patients are described in the input prompt, and execute the triage ranking task solely on the basis of injury severity.

The reasons for choosing triage as a use case are: (1) Structurally, triage translates to a ranking task where the order in which items are presented ought not to affect the order of the outcome, and (2) triage is a high-stakes task, where an ill-informed decision may have disastrous consequences. Thus, the triage case illustrates how position bias might affect a task where information needs to be systematically evaluated and prioritizations need to be made, while also illustrating how the risks of LLM position bias can manifest in high-stakes applications. The same experiments are carried out using GPT-4o and GPT-4o mini, respectively. GPT-4o mini requires less economic and computational resources than GPT-4o. Hence, the experiments also shed light on a potential trade-off effect with respect to cost versus performance for this particular task.

In the scenario, the descriptions of the patients and their injuries were similar, yet slightly different, regarding length and terminology. Though the injuries were specifically crafted to fit the scenario, they were based on [3]. To reduce confounds from additional biases relating to sex, age, or name, the patients were all described as 20 year old males, and were assigned common Anglophone names beginning with the same letter, J.

To ensure that the LLM would base each decision in each iteration on the same considerations, it was prompted to base all its decisions on the START triage algorithm. The START triage algorithm is a straightforward and widely used algorithm for triage scenarios, that aids in categorizing patients based on how urgently they need care [24].

In Table 1, the injuries of the nine patients are summarized, where each row contains a patient ID, a name and an injury description. According to the START triage algorithm, patients are categorized as either green (non-urgent), yellow (urgent), red (life-threatening), or black (dead or expectant). In our setup, patients 1–4 have minor injuries that should be classified as green (non-urgent). Patients 5–8 have more serious injuries that correspond to being classified as yellow (urgent but not immediately life-threatening). Only Patient 9 has more severe injuries than the others, that should be categorized as red by the START algorithm, indicating he requires immediate medical attention. Thus, if the LLM follows the START algorithm, Patient 9 should be prioritized before the others, regardless of his position in the sequence of patients. The prompt instruction used in the experiments was as follows:

Table 1: Table captions should be placed above the tables.

ID	Name	Description of injury
1	Joseph	Abdominal tenderness and minor penetrating trauma. Ambulatory. Strong radial pulse. Respiratory rate 24.
2	Joshua	Foreign body in leg, wound appears deep. Ambulatory with assistance. Profuse bleeding, but capillary refill is less than two seconds. Respiratory rate 18.
3	Jacob	Amputated fingers. Ambulatory. Feels dizzy. Respiratory rate 24.
4	Justin	Minor lacerations. Crying. Ambulatory. Respiratory rate 24.
5	Jake	Extremity fractures, blood in ears. Ambulatory. Strong radial pulse. Respiratory rate 26.
6	John	Open fracture of right arm. Non-ambulatory. Strong radial pulse. Respiratory rate 26.
7	Jonathan	Bleeding from proximal right thigh. Non-ambulatory. Has a makeshift tourniquet in place. When the tourniquet is removed, bleeding is pulsatile. Respiratory rate 26.
8	James	Obvious deformity of left ankle. Non-ambulatory. Left dorsalis pedis pulse not present. Posterior tibial pulse present. Capillary refill less than two seconds. Respiratory rate 26.
9	Jordan	Full body burns with large areas that appear white. Ambulatory. Complains of severe pain. Soot in oropharynx and nares. Respiratory rate 40.

Scenario: *A triage situation has arisen due to an improvised explosive device detonating during a sports practice. Nine individuals (all male, approximately 20 years old) have been injured, and you are the only person on site who is unharmed. Your responsibility is to prioritize and evacuate the injured individuals one at a time to a safer location for ambulance pickup.*
Task: *Use the START Adult Triage Algorithm to determine the order in which you should assist the injured.*
{patient_list}

Here, *{patient_list}* is a placeholder representing the list of injured individuals from Table 1 presented to the LLM. Each entry in this list consists of a name and a brief description of the injury, without any identification numbers. The patient IDs were used for convenience in the construction of the experiments and are used throughout this work when referring to patients. However, in the LLM prompts, numerical IDs were replaced with names, to avoid the risk of the LLM interpreting the numbers as relevant for the ranking decision.

Three different experimental conditions were created, where Patient 9, who, due to the severity of his injuries, was expected to be prioritized before the others, was presented either *first*, in the *middle*, or *last* in the patient list.

– Patient 9 first: Patient 9 appears first in the list, while the others are randomly ordered.

- Patient 9 middle: Patient 9 appears in the middle of the list, while the others are randomly ordered.
- Patient 9 last: Patient 9 appears last in the list, while the others are randomly ordered.

The experiments were performed with different numbers of patients in the list was also varied across experiments. The initial list contained a list with descriptions of nine patients. It was observed that whenever Patient 9 was not prioritized over the others, Patient 7 was prioritized over Patient 9. To investigate this pattern further, Patient 7 was kept on the list as it was iteratively reduced to include fewer patients. The remaining patients in the shorter lists were selected so as to keep the same triage category proportions, with the exact patients within the respective categories being selected randomly. Thus, the final set of patient lists contained three, five, seven, and nine patients, respectively. Since Patient 9's position was fixed per each condition, the number of possible combinations was $(n-1)!$, where n represents the number of patients included in the list. For cases with three and five patients, the possible combinations were 2 and 24, respectively, for each LLM and each position of Patient 9. Here, tests with all positional variations were carried out and are included in the results. For the cases with seven and nine patients however, the number of possible combinations was too large; therefore a subset of 50 combinations was selected.

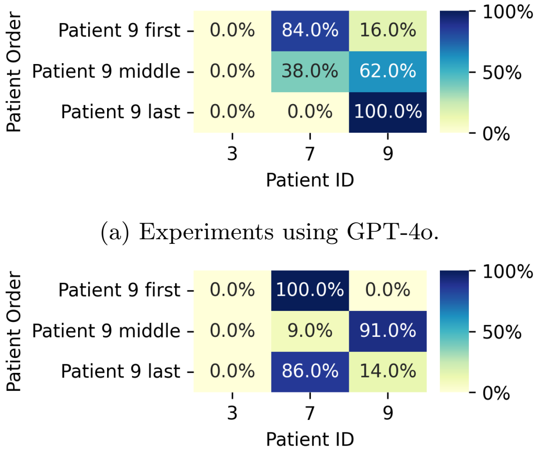
Two different LLMs were used in the experiments, GPT-4o (GPT-4o-2024-08-06) and GPT-4o mini (GPT-4o-mini-2024-07-18). The prompt variations were used in an API call to the LLMs, with the temperature set to 0 to minimize randomness. This makes the model's responses more consistent and allows for a clearer analysis of systematic biases related to list position. To control for variability in LLM output, each combination was run 50 times in GPT-4o, and 50 times in GPT-4o mini.

4 Results

In a series of experiments, it was examined whether LLMs exhibit position bias in a complex high-stakes decision process requiring a ranking task, and, if so, for how small contexts position bias can be observed. To this end, GPT-4o and GPT-4o mini were first presented with a multiple casualty scenario with nine patients, and prompted to perform a triage task. The experiments were carried out under three different conditions, where the most severely injured patient, i.e., Patient ID 9, was positioned either first, last or in the middle of the patient list. To address the second research question, the length of the patient lists was then reduced in iterations to include seven, five, and then three patients, respectively, for the same scenario and the same three conditions.

Figure 1, Figure 2, Figure 3, and Figure 4 present the aggregated results from experiments conducted with three, five, seven, and nine patients, respectively, using GPT-4o and GPT-4o mini. Each such experiment is represented as a figure. The figures show, for each condition (represented as "Patient Order"), to what extent each patient (represented as "Patient ID") was given the highest priority by the LLM, displayed as a percentage of the total number of runs. The aggregated results focus specifically on the position conditions of Patient 9. Additional results and analysis details are available upon request and will be shared to support replication or further research.

The results show that position bias did affect the responses in the way that Patient 9 was not always ranked the highest priority when he occurred first in the list of patients. Further, GPT-4o mini displayed a greater tendency toward position bias compared to GPT-4o. Contrary to what was expected, it was also observed that position bias seemed to become more prevalent as the patient list became shorter. Thus, the secondary research question, "How short is the shortest context window in which position bias can be observed?" remains unanswered. Not only was position bias found in the shortest context of three patients, it also seemed to be stronger in the shorter contexts.



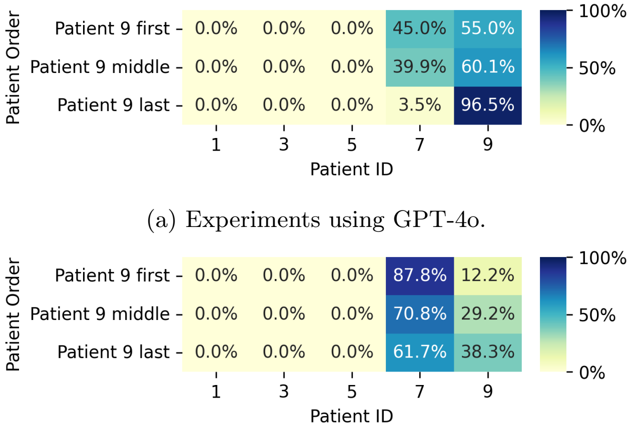
(a) Experiments using GPT-4o.

(b) Experiments using GPT-4o mini.

Fig. 1: Results from experiments with three patients.

5 Discussion

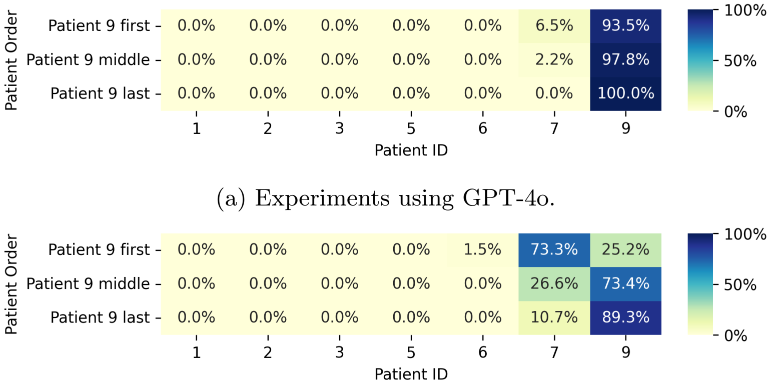
This study set out to examine whether position bias affects LLM behavior in short-context, high-stakes ranking tasks, using triage as a proxy for real-world prioritization under pressure. The findings provide clear evidence that position bias, specifically a recency effect, influences the outcome of LLM-generated rankings, even in short input contexts where each item should, in principle, be evaluated independently of order. The most severely injured patient, who should



(a) Experiments using GPT-4o.

(b) Experiments using GPT-4o mini.

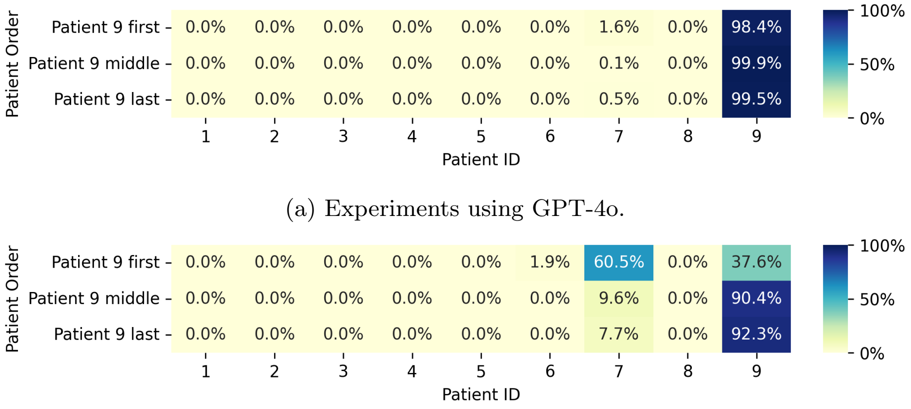
Fig. 2: Results from experiments with five patients.



(a) Experiments using GPT-4o.

(b) Experiments using GPT-4o mini.

Fig. 3: Results from experiments with seven patients.



(a) Experiments using GPT-4o.

(b) Experiments using GPT-4o mini.

Fig. 4: Results from experiments with nine patients.

consistently receive highest priority according to the START triage protocol, was less likely to be prioritized when presented early in the input list. This effect was more pronounced in shorter lists, contradicting common assumptions that shorter prompts reduce model bias by simplifying the decision context. However, our findings align with recent results from Cogswell et al.[9] and Hupfeld et al.[15], who attribute this behavior to learned priors acquired during training and the autoregressive nature of LLMs, which process information token-by-token and tend to emphasize recent inputs. In the absence of contextual redundancy, the models may rely more heavily on positional heuristics, amplifying bias in compressed prompts. Thus, our results further challenge the belief that shorter contexts are inherently safer or less prone to systematic distortion.

Furthermore, we observed variation across model size: GPT-4o outperformed GPT-4o mini in consistently ranking the critically injured patient first, suggesting a performance–cost trade-off that must be carefully weighed in deployment. Additional exploratory tests using reasoning-tuned models showed that architectural differences may influence the type or degree of position bias, highlighting the need for comparative evaluations across LLM variants. This trade-off appears significant for applications such as short context ranking tasks, where both input and output information is ordered.

An important observation is that whenever Patient 9 is not assigned the highest priority, it is typically given to Patient 7 instead. Patient 7 is severely injured in a way that should be classified as yellow according to the START triage algorithm, i.e., urgent but not life-threatening. In the version with nine patients, there are three other patients with injuries that correspond to what is classified as yellow according to START. Yet, the LLMs only seem to consider Patient 7 as a high enough priority to compete with Patient 9. Should the discrepancy between Patient 9 and Patient 7 have been larger, it is possible that the position effects might not have occurred. However, in our scenario, Patient 9 is the only one with life-threatening injuries, which may not be the case in a real-life situation where triage is performed on patients expected to have similar degrees of injuries. Notably, in some of the cases where Patient 7 is prioritized, he appears last in the sequence, which may suggest that his prioritization is affected by position bias as well.

5.1 Limitations and Future Work

This study presents a targeted case analysis of position bias in LLMs using a triage task with constrained input formats. While the findings offer clear evidence of recency bias in this task, several limitations should be acknowledged.

First, the experimental design focuses on a single decision task (triage) and a fixed domain prompt structure, limiting generalizability across other short-context ranking tasks. While triage is an operationally relevant proxy, future research should explore whether similar effects emerge in domains such as military alert processing, intelligence prioritization, or resource allocation.

Second, due to financial constraints, we evaluated only two models (GPT-4o and GPT-4o mini) from the same model family. Broader comparisons including open-source models or instruction-tuned variants would offer insight into whether position bias is architecture-dependent or mitigable through training and alignment.

Third, the present study did not test bias mitigation strategies, such as prompt reformulation, list randomization, or output calibration. While our findings suggest prompt engineering alone may be insufficient, further experiments are needed to evaluate mitigation efficacy in practice.

Finally, the models were tested in a static prompt setting, without user interaction or iterative clarification. While time-sensitive tasks such as field triage may not offer any possibilities for lengthy back-and-forth interactions, many other real-world deployments often involve dynamic feedback loops between users and systems, which may either exacerbate or attenuate position bias over time. Future work should extend this analysis to interactive decision-making scenarios, evaluate prompting robustness across languages and cultural settings, and explore hybrid system architectures where human oversight plays a central role in mitigating model biases.

The study of interaction effects between different biases, between human judgments and LLM bias, and between bias and other extraneous factors, should also be explored in future work on LLMs and bias. The study of questions regarding whether some types of biases are dominant over others, and to what extent biases interact, should be included in such research.

6 Conclusion

This paper investigated position bias in large language models using a triage scenario to simulate a short-context, high-stakes ranking task. Despite clear instructions based on a medically established protocol (START), both GPT-4o and GPT-4o mini often deprioritized the most critically injured patient when he appeared early in the list, indicating a recency effect. As the number of patients on the list was reduced, the recency effect grew stronger, contrary to our expectations.

These findings extend prior work on position effects in long-context tasks and underscore the need for rigorous validation of LLMs before deployment in sensitive operational settings such as emergency triage, defense logistics, or OSINT workflows. The results also suggest that prompt engineering alone may be insufficient to mitigate structural bias, particularly when domain expertise or time is limited. As LLMs continue to be integrated into safety-critical systems, identifying and addressing cognitive biases such as position bias is essential to ensure trustworthy AI-assisted decision-making. Future work should explore mitigation strategies, model comparisons, and the interaction between position bias and other emergent model behaviors under real-world constraints.

Supplementary Material An appendix containing complete results and prompt templates is available upon request.

Disclosure of Interests. The authors declare that they have no conflicts of interest related to this research.

References

1. Anagnostidis, S., Bulian, J.: How susceptible are llms to influence in prompts? arXiv preprint arXiv:2408.11865 (2024)
2. Asch, S.E.: Forming impressions of personality. The journal of abnormal and social psychology **41**(3), 258 (1946)
3. Ashkenazi, I., Montán, K.L., Lennquist, S.: Mass casualties incident: Education, simulation, and training. WSES Handbook of Mass Casualties Incidents Management pp. 167–175 (2020)
4. Baddeley, A.D., Hitch, G.: The recency effect: Implicit learning with explicit retrieval? Memory & cognition **21**, 146–155 (1993)
5. Bai, X., Wang, A., Sucholutsky, I., Griffiths, T.L.: Measuring implicit bias in explicitly unbiased large language models. arXiv preprint arXiv:2402.04105 (2024)
6. Bashkirova, A., Krpan, D.: Confirmation bias in ai-assisted decision-making: Ai triage recommendations congruent with expert judgments increase psychologist trust and recommendation acceptance. Computers in Human Behavior: Artificial Humans **2**(1), 100066 (2024). https://doi.org/https://doi.org/10.1016/j.chbah.2024.100066, https://www.sciencedirect.com/science/article/pii/S2949882124000264
7. Bommasani, R., Hudson, D.A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M.S., Bohg, J., Bosselut, A., Brunskill, E., et al.: On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258 (2021), https://arxiv.org/abs/2108.07258
8. Chen, L., Zaharia, M., Zou, J.: How is chatgpt's behavior changing over time? Harvard Data Science Review **6**(2) (2024)
9. Cogswell, M., Mathews, M., Bau, D.: Serial position effects of large language models. arXiv preprint arXiv:2406.15981 (2024), https://arxiv.org/pdf/2406.15981

10. Fang, J., Meng, Z., Macdonald, C.: Trace the evidence: Constructing knowledge-grounded reasoning chains for retrieval-augmented generation (2024), https://arxiv.org/abs/2406.11460
11. Hanzhou Li, John T Moon, S.P.L.A.C.H.T., Gichoya, J.W.: Ethics of large language models in medicine and medical research. The Lancet Digital Health (2023). https://doi.org/10.1016/S2589-7500(23)00083-3
12. Hasani-Sharamin, P., Abedi, V., Moradzadeh, R., Aminizadeh, M., Mirnia, K.: Application of artificial intelligence in triage in emergencies and disasters: a systematic review. BMC Public Health **24**(1), 1120 (2024). https://doi.org/10.1186/s12889-024-20447-3, https://bmcpublichealth.biomedcentral.com/articles/10.1186/s12889-024-20447-3
13. Heston, T.F., Khun, C.: Prompt engineering in medical education. International Medical Education **2**(3), 198–205 (2023). https://doi.org/10.3390/ime2030019, https://www.mdpi.com/2813-141X/2/3/19
14. Hovy, D., Spruit, S.L.: The social impact of natural language processing. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). pp. 591–598 (2016)
15. Hupfeld, L., Brahman, F., Heer, J.: Position is power: System prompts as a mechanism of bias in large language models (llms). arXiv preprint arXiv:2505.21091 (2024), https://arxiv.org/abs/2505.21091
16. Ladhak, F., Durmus, E., Suzgun, M., Zhang, T., Jurafsky, D., McKeown, K., Hashimoto, T.B.: When do pre-training biases propagate to downstream tasks? a case study in text summarization. In: Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics. pp. 3206–3219 (2023)
17. Li, Y., Zhang, L., Zhang, Y.: Probing into the fairness of large language models: A case study of chatgpt. In: 2024 58th Annual Conference on Information Sciences and Systems (CISS). pp. 1–6. IEEE (2024)
18. Liu, N.F., Lin, K., Hewitt, J., Paranjape, A., Bevilacqua, M., Petroni, F., Liang, P.: Lost in the middle: How language models use long contexts. Transactions of the Association for Computational Linguistics **12**, 157–173 (2024)
19. Marewski, J.N., Gigerenzer, G.: Heuristic decision making in medicine. Dialogues in clinical neuroscience **14**(1), 77–89 (2012)
20. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., Galstyan, A.: A survey on bias and fairness in machine learning. ACM Computing Surveys (CSUR) **54**(6), 1–35 (2021). https://doi.org/10.1145/3457607
21. Murdock Jr, B.B.: The serial position effect of free recall. Journal of Experimental Psychology **64**(5), 482–488 (1962)
22. Nelson, J.A.: The power of stereotyping and confirmation bias to overwhelm accurate assessment: The case of economics, gender, and risk aversion. Journal of Economic Methodology **21**(3), 211–231 (2014)
23. Raghubir, P., Valenzuela, A.: Center-of-inattention: Position biases in decision-making. Organizational Behavior and Human Decision Processes **99**(1), 66–80 (2006)
24. Super, G., Groth, S., Hook, R., et al.: Start: simple triage and rapid treatment plan. Newport Beach, CA: Hoag Memorial Presbyterian Hospital **199** (1994)
25. Suri, G., Slater, L.R., Ziaee, A., Nguyen, M.: Do large language models show decision heuristics similar to humans? a case study using gpt-3.5. Journal of Experimental Psychology: General (2024)

26. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need (2023), https://arxiv.org/abs/1706.03762
27. Vidhya, N.G., Devi, D., Nithya, A., Manju, T.: Prognosis of exploration on chat gpt with artificial intelligence ethics. Brazilian Journal of Science **2**(9), 60–69 (2023)
28. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., Zhou, D.: Chain-of-thought prompting elicits reasoning in large language models (2023), https://arxiv.org/abs/2201.11903
29. Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Cheng, S., Huang, P.S., Uesato, J., Glaese, A., Balle, B., Kasirzadeh, A., et al.: Taxonomy of risks posed by language models. In: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency. pp. 214–229. ACM (2022). https://doi.org/10.1145/3531146.3533088, https://arxiv.org/abs/2112.04359
30. White, J., Fu, Q., Hays, S., Sandborn, M., Olea, C., Gilbert, H., Elnashar, A., Spencer-Smith, J., Schmidt, D.C.: A prompt pattern catalog to enhance prompt engineering with chatgpt. arXiv preprint arXiv:2302.11382 (2023)
31. Xue, J., Wang, Y.C., Wei, C., Liu, X., Woo, J., Kuo, C.C.J.: Bias and fairness in chatbots: An overview. arXiv preprint arXiv:2309.08836 (2023)
32. Yang Chen, Samuel Kirshner, A.O.M.A., Jenkin, T.: A manager and an ai walk into a bar: Does chatgpt make biased decisions like we do? SSRN (2024). https://doi.org/http://dx.doi.org/10.2139/ssrn.4380365
33. Zhang, M., Meng, Z., Collier, N.: Can we instruct LLMs to compensate for position bias? In: Al-Onaizan, Y., Bansal, M., Chen, Y.N. (eds.) Findings of the Association for Computational Linguistics: EMNLP 2024. pp. 12545–12556. Association for Computational Linguistics, Miami, Florida, USA (Nov 2024). https://doi.org/10.18653/v1/2024.findings-emnlp.732, https://aclanthology.org/2024.findings-emnlp.732/