



Document Analysis with LLMs: Assessing Performance, Bias, and Nondeterminism in Decision Making

Stephen Price¹ ^a and Danielle L. Cote² ^b

¹*Department of Computer Science, Worcester Polytechnic Institute, Worcester, MA, U.S.A.*

²*Department of Mechanical and Materials Engineering, Worcester Polytechnic Institute, Worcester, MA 01609, U.S.A.*

Keywords: Natural Language Processing, Large Language Models, Document Analysis, Decision-Making, Bias, Reproducibility, Nondeterminism.

Abstract: In recent years, large language models (LLMs) have demonstrated their ability to perform complex tasks such as data summarization, translation, document analysis, and content generation. However, their reliability and efficacy in real-world scenarios must be studied. This work presents an experimental evaluation of an LLM for document analysis and candidate recommendation using a set of resumes. Llama3.1, a state-of-the-art open-source model, was tested with 30 questions using data from five resumes. On tasks with a direct answer, Llama3.1 achieved an accuracy of 99.56%. However, on more open-ended and ambiguous questions, performance, and reliability decreased, revealing limitations such as bias toward particular experience, primacy bias, nondeterminism, and sensitivity to question phrasing.

1 INTRODUCTION

Large language models (LLMs) have gained significant attention in recent years due to their ability to process and generate human-like text across a vast range of topics (Kasneji et al., 2023; Roumeliotis and Tselikas, 2023). These models have been able to execute tasks such as data summarization (Laban et al., 2023), document question-answering (Wang et al., 2024), and code generation (Gao et al., 2023) with high levels of accuracy. As a result, these tools have begun to be implemented in industry to optimize workflow efficiency (Li et al., 2024b; Acharya et al., 2023). However, as the capabilities of these models grow and their potential applications expand, discussions on their viability have grown.


These models are particularly useful in cases of unstructured data where no specific layout or format exists, such as a series of documents or reports (Li et al., 2024a). While capable of producing well-crafted and convincing arguments, these outputs can sometimes be unreliable due to hallucinations, where the model produces factually incorrect, biased, or nonsensical responses (Azamfirei et al., 2023). In fields like Talent Acquisition, an automated approach


to analyzing and assessing candidates could save a significant amount of time and reduce operating costs (Gopalakrishna et al., 2019; Singh, 2023). However, an LLM does not have the capabilities to effectively quantify skills and compare performance to produce a recommendation. Instead, it relies on linguistic probabilities (Vaswani et al., 2017), creating potentially biased, unsubstantiated, or poor recommendations.

This work presents an experimental validation of using an LLM for such tasks, identifying and discussing the potential challenges and limitations that must be considered. Using a set of five resumes as a case study, Llama3.1, a state-of-the-art open-source model, was tasked with selecting the candidate who best fit specific criteria. The outputs were then analyzed to evaluate factors such as accuracy (when possible), potential biases, and reproducibility. Questions varied in complexity, ranging from straightforward, fact-based questions to more open-ended and ambiguous questions that required complex decision-making skills. While focused on the Talent Acquisition domain, these results are transferable towards other industry applications using LLMs for decision-making tasks.

This work offers the following contributions:

- An experimental evaluation of LLMs for candidate selection in the Talent Acquisition domain using Llama3.1.

^a  <https://orcid.org/0000-0001-9368-1789>

^b  <https://orcid.org/0000-0002-3571-1721>

- An analysis of LLM outputs, demonstrating a significant drop in performance and reproducibility as question ambiguity increased.
- A quantification of LLM nondeterminism using Shannon Entropy revealing variability in decision-making.
- A discovery of model bias, including primacy bias, towards specific qualities or criteria in the decision-making process.
- An evaluation of the impact temperature settings have on model reproducibility and recommendation accuracy.

2 BACKGROUND

2.1 Large Language Models

Large language models (LLMs) represent a significant advancement in natural language processing, leveraging transformer architectures to process and generate human-like text across a variety of tasks (Vaswani et al., 2017). This transformer architecture utilizes an attention mechanism that allows models to consider the entire input sequence when processing data, making them particularly good at long, information-rich, unstructured data (Kenton et al., 2019). When asked a question, the LLM encodes the text into a numerical representation to generate responses (Minaee et al., 2024). For this response, the model predicts the most probable next token based on the input and the training data until reaching a stop condition (Vaswani et al., 2017). Answering questions using statistical probabilities of language, rather than querying a database or performing computations, enables the model to handle a wider range of topics and adapt to new tasks. However, it is also the primary reason for hallucinations when the task is too complex, or the required information is not properly represented in the training data (Azamfirei et al., 2023).

2.2 Model Nondeterminism

Machine learning models, especially deep learning models, such as LLMs, are inherently nondeterministic, meaning that given a specific state, the resultant output cannot be predicted (Price and Neamtu, 2022). As a result, a model's outputs may differ from run to run, making it difficult, or impossible, to reproduce results. While making results less reproducible, this nondeterminism does offer many advantages. For example, stochastic gradient descent enables improved scalability for large datasets in optimization problems

(Amari, 1993). Nondeterministic parallel computing can significantly accelerate training and inference (Price and Neamtu, 2022), and inserted randomness can improve the likelihood of identifying global maximums (Bertsimas and Tsitsiklis, 1993).

In the case of LLMs, the nondeterminism of generated outputs is governed by a hyperparameter referred to as temperature (Saha et al., 2024). Ranging from 0.0 to 1.0, this parameter controls how probabilistic or how creative outputs are. A temperature of 0.0 will result in fully, or nearly fully, deterministic outputs. However, by only selecting the most probable next token, these outputs can become robotic and repetitive. Alternatively, a larger temperature can enable the model to deviate from the most probable choice, resulting in more human-like and "creative" outputs (Renze and Guven, 2024).

2.3 Bias in AI-Based Recommendations

As AI-based recommendation systems have been developed to accelerate and optimize workflows, concerns about bias have also emerged (Wilson and Caliskan, 2024; Gerszberg, 2024; Huang et al., 2022; Leavy, 2018). This bias can lead to unfair recommendations/predictions, disproportionately affecting individuals. In 1979, St George's Medical School in London implemented an algorithm to assist in evaluating candidates. However, after a few years, there were growing concerns about the lack of diversity due to the algorithm. An analysis by the U.K. Commission for Racial Equality found that candidates were classified as "Caucasian" or "Non-Caucasian" based on name and place of birth. Those with a non-Caucasian name were deducted points in the application process. Similarly, women were deducted points purely by gender (Schwartz, 2019).

More recently, courts in the United States have begun to use AI-based criminal risk assessment (CRA) models, seeking to evaluate the likelihood of reoffending and to inform decisions on bail, sentencing, and parole (Eckhouse et al., 2019). However, when trained on historical data with disproportionate representation, these AI models can conflate correlation with causation, perpetuating existing biases (Hao, 2019). This type of model has been implemented in many places, including Fort Lauderdale, FL. After multiple years of use, it was discovered that this model was disproportionately labeling non-white individuals as high-risk over white individuals with a similar crime and prior history (Angwin et al., 2016).

Considering these examples, and many others like them, it is imperative to understand and discuss the implications of new AI-based systems as they are de-

veloped. These systems, when left unchecked, can perpetuate or worsen societal biases, leading to unfair potentially harmful outcomes. Applied to candidate selection with an LLM, it is important to evaluate that recommendations are made based on objective criteria and free from bias or external factors.

3 METHODOLOGY

The primary objective of this study was to evaluate the efficacy of using LLMs as a recommendation tool. Using resume evaluation as a case study, these results are directly applicable to the Talent Acquisition domain but could be similarly applied to other fields.

3.1 Creating Resumes

For this evaluation, five resumes were created and given to the LLM for assessment. These resumes were designed to be similar in work experience and background, with minor variations to highlight distinct skill sets, educational paths, and specific roles within software development. For example, while two of the candidates, Alex Williams and John Smith, both have five years of experience, Alex’s work spans both front-end and back-end software development, giving him a full-stack developer role, whereas John specializes solely in back-end software development. These variations allow for an analysis of how the LLM evaluates and ranks different aspects of the candidates’ profiles, focusing on areas such as specialization, leadership experience, and advanced technical knowledge. An overview of these candidates is provided below:

- **Alex Williams:** B.S. in Computer Science (CS) from the University of Washington and five years of work experience as a full-stack developer (including front-end and back-end).
- **Emily Brown:** B.S. in Information Technology from DePaul University, three years of experience as a Software engineer and two years of experience as a project manager and business analyst.
- **Jane Doe:** B.S. in CS from the University of California, Berkeley and five years of experience as a front-end developer.
- **John Smith:** B.S. in Software Engineering from the University of Texas Austin and five years of experience as a back-end developer.
- **Sarah Johnson:** B.S. in CS from NYU, M.S. in CS from Columbia, and five years of experience as a machine learning engineer.

3.2 Creating a Question Base

For this experiment, there were a total of 30 unique questions asked, broken into two sets. The first set contained 15 questions with a single answer that could be answered objectively based on the provided information. For example, the answer to “Which candidate has the most experience with machine learning?” is Sarah since no other candidates have experience in machine learning. The purpose of these questions was to evaluate if, when there is a definitively correct answer, the model is able to determine this answer accurately, and do so consistently.

The second set of questions was more difficult to directly answer, such as “Which candidate is the most adaptable?” For this question, arguments could be made for all five candidates, requiring the LLM to make a decision. This set of questions had multiple purposes. Without a single correct answer, they provided significant insights into the LLM’s decision-making process, yielding several key benefits. Most importantly, they provided insight into how the model approached ambiguous problems and enabled the identification of potential biases the model may carry towards or against specific backgrounds and skill sets. Additionally, by being more ambiguous and subjective, these questions provide a better analysis of the determinism and reproducibility of results. Lastly, these results are closer to questions that may be asked in a real-world setting, allowing for a more practical evaluation of the model’s capabilities and limitations.

3.3 Model Output Post-Processing

The model used here, Llama3.1 8B, similar to many of the most common LLMs, is a text-to-text model. As a result, model outputs were unstructured paragraphs, containing the recommended candidate and rationale for the decision. However, the style, length, and clarity of these answers were variable, making it difficult to use an algorithmic approach such as a regular expression to post-process these outputs and extract the recommended candidate. To remedy this, each output’s recommendation was labeled by hand. A secondary LLM could have been used to achieve this, but it was not considered to avoid introducing additional bias. Occasionally, the model recommended multiple candidates instead of just one. If this occurred, but the model followed up with this recommendation by ranking one over the others, then this individual was labeled as the recommended candidate. However, if no distinction was made between the multiple recommended candidates, it was labeled as *Multi-Vote*. Alternatively, if the model made no

recommendation or declined to answer, it was labeled *No-Vote*.

3.4 Measuring Nondeterminism in Model Outputs

Any LLM with a temperature larger than 0.0, and even some models with a temperature of 0.0 are nondeterministic (Song et al., 2024). However, this nondeterminism is more broadly defined than is desired for the purposes of this research, which is focused more on decision-making capabilities than on syntax or sentence structure. For the purposes of this study, if a model repeatedly answers a question in unique ways, but consistently recommends the same candidate, this would be regarded as a deterministic decision. However, if the candidate that is recommended changes, this would be regarded as a nondeterministic decision. To evaluate this, each question was asked 30 times to the model. To prevent any contamination of results, each question, and each iteration of each equation, was asked to the model independently, ensuring that any prior questions or their answers did not influence the next question.

When analyzing results, identifying the presence of nondeterminism is simple. If the model did not consistently select the same candidate each time a question was asked, then the results for that question are nondeterministic. Quantifying, ranking, and comparing this nondeterminism is more difficult. To achieve this, Shannon Entropy, as outlined in Equation 1, was implemented (Lin, 1991). While not specifically designed as a measure of nondeterminism, Shannon Entropy measures the degree of uncertainty in a predictor, which can be used to categorize randomness. For example, if the model recommends the same candidate all 30 times, it will have a Shannon entropy of 0.0. As the model diverges from recommending the same candidate 30 times, either by recommending a new candidate or recommending a secondary candidate again, the Shannon entropy will increase, indicating a higher degree of randomness in decisions.

$$H(X) = - \sum_{i=1}^n p(x_i) \log_b p(x_i) \quad (1)$$

4 RESULTS

4.1 Objective Question Results

Evaluating experimental results for questions where there was an objectively correct answer highlights

LLMs' ability to process information and perform document question-answering. Of the 15 questions with a single correct answer, 14 were consistently answered correctly all 30 times, and for the 15th question, the model answered correctly 28 of the 30 times using default parameter settings. These errors, while incorrect, were minor errors that can be explained. For example, when asked who had the most experience in back-end development, the correct answer was John, who had five years of experience as a back-end developer. However, occasionally, the model recommended Alex, who has five years of experience as a full-stack developer, which included back-end components.

Processing these results analytically showed that, of the 450 questions asked with a direct answer, 448 of them were answered correctly, yielding an accuracy of 99.56% for document question-answering. Furthermore, the LLM answered these objective questions with a mean Shannon entropy of 0.024. With such a high accuracy and low entropy, these results highlight that, under the right circumstances where an answer can be realistically determined, an LLM is capable of consistently and accurately performing document question-answering to process straightforward information. Additionally, these scores served as a baseline for evaluating the performance of more open-ended and challenging questions.

4.2 Subjective Question Results

While the model consistently chose the same candidate with very little entropy for objective questions, this was not the case for the set of subjective questions. As shown in Table 1, only 3 of the 15 questions were consistent in the recommendation of the same candidate, scoring an entropy of 0.0. The remaining 12 questions scored an entropy larger than every objective question (at least 0.353), resulting in a mean entropy of 0.917. In two of the questions, the model even recommended all five applicants at least once. With no change to the documents provided or the model architecture, this significant change in consistency and increase in entropy from 0.024 to 0.917 demonstrates how the performance and reliability of a model are highly dependent on the specific question/task asked of the model.

Of these questions, "Which candidate is the best at handling high-pressure situations?" had the highest entropy with 1.911, recommending all five candidates at least once and recommending Alex 50% of the time. As one of the most open-ended questions that was not particularly related to any candi-

Table 1: Shannon Entropy of asking a Llama3.1 8B model subjective questions 30 times each.

Question	Entropy
Which candidate is the most adaptable?	0.0
Which candidate would be the best at juggling multiple projects?	0.0
Which candidate would be the best at aligning their work with long-term business goals?	0.0
Which candidate is most likely to drive business growth?	0.469
Which candidate is the best fitted to lead a team?	0.650
Which candidate would be the best at a client-facing role?	0.722
Which candidate would be the best at conflict resolution?	0.812
Which candidate would be the best for improving company culture?	0.904
Which candidate should I hire?	0.922
Which candidate is most likely to consider the ethics of their actions?	1.120
Who is the best candidate?	1.280
Which candidate would be the best for a software engineering role?	1.303
Which candidate is most likely to create innovative new ideas?	1.887
Which candidate is the best at handling high-pressure situations?	1.911

date’s experience, this was not particularly surprising. However, comparing a similarly difficult and open-ended question, “Which candidate would be the best at juggling multiple projects?”, the LLM consistently recommended only a single candidate. These findings suggest that while the type of question and the amount of relevant information available are strongly correlated with the degree of variability in responses, the underlying reasons for this relationship may not be fully explainable at this time.

Analyzing these recommendations more broadly, as shown in Table 2, in addition to being more random, these recommendations appeared to show underlying biases in how candidates were chosen. For example, of the 450 questions in this experiment, Emily was recommended 183 times (40.7%) and recommended at least once for 12 of the 15 questions. By contrast, John was only recommended 13 times (2.9%) and recommended at least once for only 6 of the 15 questions. Some of this deviation from candidate to candidate could be explained by the specific questions set, causing a slight bias towards one candidate. However, these questions were specifically de-

Table 2: Summary of candidate recommendations for subjective questions using default model parameters. *Question Recs* represents the number of questions each candidate was recommended at least once. *Total Recs* represents the total number of recommendations out of 450 subjective questions. *Average Recs* indicates the average number of recommendations per question out of 15 for each candidate. *Percent Recs* reflects the percentage of total questions for which the candidate was recommended. Note: An additional 28 NA and eight multi-vote.

	Alex	Emily	Jane	John	Sarah
Question Recs	9	12	8	6	7
Total Recs	115	183	65	13	39
Average Recs	7.6	12.2	4.3	0.9	2.6
Percent Recs	25.3%	40.7%	14.4%	2.9%	8.6%

signed not to be geared toward any one candidate. Comparing the experience of these two candidates, they had the same level of education and same number of years of work experience. The primary difference was that Emily spent the first half of her career as a project manager before becoming a developer, whereas John had been a developer for his entire career. It is likely that the model identified this managerial experience, or the combination of managerial and technical experience, and weighed it higher in lieu of more concrete information, as was the case for the set of objective questions.

In addition to potential model biases, these results also highlight how the specific wording of questions can impact performance. For example, “Which candidate should I hire?” and “Who is the best candidate?”, despite asking a very similar question, had significantly different results. When asked “Which candidate should I hire?”, the LLM declined to provide a recommendation 23 of the 30 times, citing ethical reasons or lack of information. In contrast, when asked who the best candidate was, the model only declined to answer three times. Parsing the justification of each answer, it appears that when asked who should be hired, the model gave more priority to the social and ethical ramifications of an uninformed decision than when simply asked to rank them.

4.3 Resume Order Bias Results

If every candidate was recommended evenly, each candidate would have been recommended approximately 90 times. Here, Alex and Emily were recommended 114 and 183 times, respectively, indicating the possibility of some form of bias. Conducting

an identical experiment, but with the order in which resumes were input into the model reversed, offers some insight into this bias. Specifically, by reversing the order, Alex (initially first) and Sarah (initially last) saw the largest change in the number of recommendations. Alex, who was initially recommended 114 times, was only recommended 69 times after reversing, while Sarah, who was initially only recommended 39 times, was recommended 137 times, as shown in Figure 1. Interestingly, despite reversing the order, Emily was recommended a similar number of times (183 compared to 167), supporting conclusions that the LLM was biased towards a specific component of her resume.

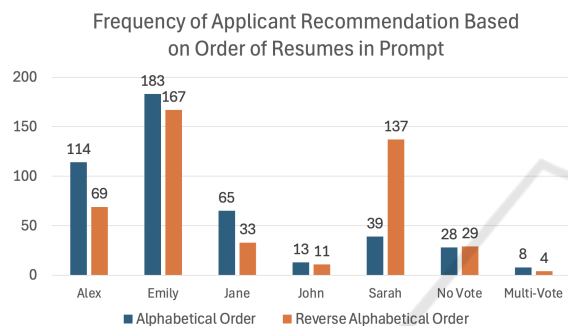


Figure 1: Applicant recommendation frequency based on order of resumes in prompts, highlighting primacy bias. Note: 450 total questions asked for each order.

With an almost four-fold increase in number of recommendations for Sarah when presented first instead of last and a significant decrease for Alex, these results highlight a clear correlation between the order of presented information and final decisions. In this case, by weighing information presented first higher than information presented last, this correlation indicates primacy bias. When considering the viability of an LLM for decision-making tasks, this primacy bias is crucial to be aware of because it confirms that factors other than candidate quality, such as stack order, are also factored into evaluations.

4.4 Evaluating Impact of Temperature

When generating results, reducing a model’s temperature to 0.0 fully eliminates randomness in most cases, or in some cases, nearly fully eliminates randomness (Ouyang et al., 2023). Here, when the temperature was set to 0.0, each answer was identical, including recommendation and semantic structure, revealing a fully deterministic behavior when temperature was no longer a factor. The model’s average entropy on decreased from 0.02 to 0.0 and 0.91 to 0.0 for objective and subjective questions, respectively. In con-

Table 3: Number of recommendations per candidate across 15 subjective questions asked 30 times each with a temperature of 0.0 (minimum), 0.8 (default), and 1.0 (maximum).

	0.0	0.8 (Default)	1.0
Alex	120	115	101
Emily	180	183	182
Jane	90	65	58
John	0	13	25
Sarah	30	39	45
NA	30	27	33
Multi-Vote	0	8	6
Mean Entropy	0.0	0.92	1.06

trast, raising temperature from the default (0.8) to the maximum (1.0) increased nondeterminism, resulting in an increase in mean Shannon entropy from 0.02 to 0.14 and 0.91 to 1.06 for objective and subjective questions, respectively.

As shown in Table 3, increasing the temperature had minimal impact on the distribution of recommendations per candidate for subjective questions. The only observed change was a slight increase in the frequency of less common recommendations, such as “John,” “Sarah,” and “No Vote.” Removing the temperature resulted in a slightly larger effect, likely due to the fully deterministic nature of the model, producing recommendations in sets of 30. Given that only 15 unique questions were analyzed, this sample size may have been insufficient for a comprehensive analysis of distribution shifts, especially with deterministic behavior. Despite this deterministic behavior, many of the distributions remained unchanged from default parameters, with “Alex” varying by only five, “Emily” by three, “Sarah” by nine, and “No Vote” by three votes. Analyzing the observed effects of both increasing and removing temperature, as demonstrated in Table 3, these results validate that the primary impact of temperature tuning is enabling more diverse outputs beyond the probabilistic answer.

Extending beyond number of recommendations, temperature also impacted model behavior differently from question to question, especially for “Who is the best candidate?” and “Which candidate should I hire?”. When temperature was increased, the model declined to make a recommendation ten times instead of the original three for the “Who is the best candidate?” question. Alternatively, the model declined to answer 23 times, the same as default, for “Which candidate should I hire?”. This suggests that there are many additional factors in how a recommendation is produced than simply ranking the candidates.

5 DISCUSSION

This experiment evaluated the efficacy of using an LLM such as Llama3.1 (Dubey et al., 2024) to process data and make recommendations based on a specific criteria. These results have identified four key areas of concern regarding using an LLM for decision-making, particularly in an industry setting. First, unless temperature is turned off, potentially worsening model performance, results are not guaranteed to be reproducible. Thus, asking a model the same question multiple times can result in different recommendations despite no change in model or input. Second, LLMs may be biased toward additional factors beyond the given information. For example, by reversing the order in which resumes were given, model recommendations changed significantly, highlighting a potential primacy bias where the model was more likely to recommend the first resume than the last. Additionally, the model showed a possible bias towards specific qualities or backgrounds. For example, here, the model frequently recommended candidates with broader experience over those with deeper experience. Lastly, these results highlight that an LLM's performance is highly dependent on the specific question asked. For example, the impact of the previous challenges was magnified when questions became more subjective and difficult to answer, or the phrasing of a question changed.

Individually, underlying model bias, reproducibility issues, and dependency on specific question phrasing raise questions on the viability of using an LLM for tasks such selecting a candidate from a set of resumes. However, all three challenges being concurrently present suggests that LLMs are not well suited these types of tasks.

While using an LLM for candidate selection presented many challenges, these results also highlighted multiple circumstances where these issues were not present and an LLM could be used to accelerate workflow. Namely, when only using the LLM to summarize and query documents rather than make decisions, results were significantly improved. On the 15 objective questions that had a direct answer, the model exhibited an accuracy of 99.56% and was nearly deterministic with an average entropy of 0.02. Eliminating randomness by setting temperature to 0.0 improved these results further, scoring an accuracy of 100% and an entropy of 0.0. Leveraging these capabilities, LLMs could be used to process documents, searching for specific qualities to convert unstructured text into a condensed format, enabling a human to make decisions without the previously mentioned challenges. However, there are many challenges and limitations

that must be overcome before LLMs can be used for more subjective and open-ended tasks.

6 CONCLUSION

Large language models (LLMs) have demonstrated capabilities to optimize and improve work-place efficiency. However, results collected here highlight the challenges and limitations of using LLMs for decision-making tasks. When recommending applicants from a set of resumes, the model demonstrated bias towards specific background experience, bias towards the order resumes were presented, nondeterministic behavior resulting in non-reproducible behavior when used in default settings, and varying performance based on question phrasing. Despite these issues, the model showed promise for specific use cases, such as tasks involving objective questions with a single correct answer found in the presented information, where an LLM could be used to accelerate and optimize workflows.

7 FUTURE STEPS

This work outlined the experimental approach to evaluate the usage of LLMs for document analysis and decision-making. Preliminary results have demonstrated the lack of consistency from iteration to iteration and biases present when making decisions. Moving forward, this work will be expanded in the following ways:

1. Increase the number of objective/subjective questions asked to better evaluate potential model bias
2. Evaluate more complex decision-making capabilities such as compound questions
3. Introduce additional resumes to better evaluate model capabilities as the input length grows and decisions get harder
4. Evaluate and compare results to additional model architectures and parameter sizes

8 DATA & CODE AVAILABILITY

This study was conducted using Llama3.1 8B, an open-access model, downloaded on September 10th, 2024. All data, including the complete question list, generated answers, and necessary code required to replicate this work, can be found here: https://github.com/sprice134/Resume_Eval

REFERENCES

- Acharya, A., Singh, B., and Onoe, N. (2023). Llm based generation of item-description for recommendation system. In *Proceedings of the 17th ACM Conference on Recommender Systems*, pages 1204–1207.
- Amari, S.-i. (1993). Backpropagation and stochastic gradient descent method. *Neurocomputing*, 5(4-5):185–196.
- Angwin, J., Larson, J., Mattu, S., and Kirchner, L. (2016). Machine bias. *ProPublica*.
- Azamfirei, R., Kudchadkar, S. R., and Fackler, J. (2023). Large language models and the perils of their hallucinations. *Critical Care*, 27(1):120.
- Bertsimas, D. and Tsitsiklis, J. (1993). Simulated annealing. *Statistical science*, 8(1):10–15.
- Dubey, A. et al. (2024). The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.
- Eckhouse, L., Lum, K., Conti-Cook, C., and Ciccolini, J. (2019). Layers of bias: A unified approach for understanding problems with risk assessment. *Criminal Justice and Behavior*, 46(2):185–209.
- Gao, L., Madaan, A., Zhou, S., Alon, U., Liu, P., Yang, Y., Callan, J., and Neubig, G. (2023). Pal: Program-aided language models. In *International Conference on Machine Learning*, pages 10764–10799. PMLR.
- Gerszberg, N. R. (2024). *Quantifying Gender Bias in Large Language Models: When ChatGPT Becomes a Hiring Manager*. PhD thesis, Massachusetts Institute of Technology.
- Gopalakrishna, S. T. et al. (2019). Automated tool for resume classification using semantic analysis. *International Journal of Artificial Intelligence and Applications (IJAIA)*, 10(1).
- Hao, K. (2019). Ai is sending people to jail — and getting it wrong. *MIT Technology Review*.
- Huang, J., Galal, G., Etemadi, M., and Vaidyanathan, M. (2022). Evaluation and mitigation of racial bias in clinical machine learning models: scoping review. *JMIR Medical Informatics*, 10(5):e36388.
- Kasneci, E. et al. (2023). Chatgpt for good? on opportunities and challenges of large language models for education. *Learning and individual differences*, 103:102274.
- Kenton, J. D. et al. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT*, volume 1, page 2. Minneapolis, Minnesota.
- Laban, P., Kryściński, W., Agarwal, D., Fabbri, A. R., Xiong, C., Joty, S., and Wu, C.-S. (2023). Summedits: measuring llm ability at factual reasoning through the lens of summarization. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9662–9676.
- Leavy, S. (2018). Gender bias in artificial intelligence: The need for diversity and gender theory in machine learning. In *Proceedings of the 1st international workshop on gender equality in software engineering*, pages 14–16.
- Li, H., Gao, H., Wu, C., and Vasarhelyi, M. A. (2024a). Extracting financial data from unstructured sources: Leveraging large language models. *Journal of Information Systems*, pages 1–22.
- Li, Y., Wen, H., Wang, W., Li, X., Yuan, Y., Liu, G., Liu, J., Xu, W., Wang, X., Sun, Y., et al. (2024b). Personal llm agents: Insights and survey about the capability, efficiency and security. *arXiv preprint arXiv:2401.05459*.
- Lin, J. (1991). Divergence measures based on the shannon entropy. *IEEE Transactions on Information theory*, 37(1):145–151.
- Minaee, S., Mikolov, T., Nikzad, N., Chenaghlu, M., Socher, R., Amatriain, X., and Gao, J. (2024). Large language models: A survey. *arXiv preprint arXiv:2402.06196*.
- Ouyang, S., Zhang, J. M., Harman, M., and Wang, M. (2023). Llm is like a box of chocolates: the non-determinism of chatgpt in code generation. *arXiv preprint arXiv:2308.02828*.
- Price, S. and Neamtu, R. (2022). Identifying, evaluating, and addressing nondeterminism in mask r-cnns. In *International Conference on Pattern Recognition and Artificial Intelligence*, pages 3–14. Springer.
- Renze, M. and Guven, E. (2024). The effect of sampling temperature on problem solving in large language models. *arXiv preprint arXiv:2402.05201*.
- Roumeliotis, K. I. and Tselikas, N. D. (2023). Chatgpt and open-ai models: A preliminary review. *Future Internet*, 15(6):192.
- Saha, D., Tarek, S., Yahyaei, K., Saha, S. K., Zhou, J., Tehranipoor, M., and Farahmandi, F. (2024). Llm for soc security: A paradigm shift. *IEEE Access*.
- Schwartz, O. (2019). Untold history of ai: Algorithmic bias was born in the 1980s. *IEEE Spectrum*.
- Singh, V. (2023). *Exploring the role of large language model (llm)-based chatbots for human resources*. PhD thesis, University of Texas at Austin.
- Song, Y., Wang, G., Li, S., and Lin, B. Y. (2024). The good, the bad, and the greedy: Evaluation of llms should not ignore non-determinism. *arXiv preprint arXiv:2407.10457*.
- Vaswani, A. et al. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*.
- Wang, Y., Lipka, N., Rossi, R. A., Siu, A., Zhang, R., and Derr, T. (2024). Knowledge graph prompting for multi-document question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19206–19214.
- Wilson, K. and Caliskan, A. (2024). Gender, race, and intersectional bias in resume screening via language model retrieval. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7, pages 1578–1590.