Prompt Sensitivity and Response Variation in Large Language Models for STEM Education

Manan Dudeja
Vellore Institute of Technology
Chennai, India

Pulkit Taneja
Vellore Institute of Technology
Chennai, India

***Abstract*—Large Language Models (LLMs) such as *ChatGPT* and *Gemini* are increasingly integrated into STEM education as intelligent tutoring and content generation tools. However, their responses often vary depending on subtle differences in user prompts—a phenomenon known as prompt sensitivity. This study systematically investigates how prompt phrasing influences the complexity and sentiment of LLM-generated answers in STEM domains. A dataset of 30 STEM-based questions was constructed, each rephrased in four distinct linguistic styles—*Neutral*, *Leading*, *Contradictory*, and *Self-Doubt*—and evaluated across both models. Responses were annotated for accuracy, complexity (1–5 scale), and sentiment polarity (*Positive*, *Neutral*, *Negative*).**

Visual and statistical analyses revealed that *Gemini* consistently produced more complex responses, while *ChatGPT* exhibited greater variability and adaptiveness to prompt style. A two-way ANOVA confirmed significant main effects of model ($p < 0.001$) and prompt type ($p = 0.029$) on response complexity, while a Chi-square test ($\chi^2(6) = 205.22$, $p < 0.001$) established a strong dependence between prompt phrasing and sentiment tone. These results indicate that both linguistic framing and model architecture shape the nature of LLM outputs. In STEM education contexts, *Gemini*'s stability may promote consistency in instruction, whereas *ChatGPT*'s responsiveness may foster dynamic, adaptive tutoring interactions.

***Index Terms*—Large Language Models, Prompt Sensitivity, STEM Education, ChatGPT, Gemini, Sentiment Analysis, Artificial Intelligence in Education**

I. Introduction

Over the past few years, Large Language Models (LLMs) such as *ChatGPT* (OpenAI) and *Gemini* (Google DeepMind) have transformed how learners and educators interact with knowledge. Their ability to generate detailed, context-aware explanations makes them increasingly popular as intelligent tutoring systems in STEM education, where students frequently seek conceptual clarification, worked examples, and personalized guidance. While these systems hold tremendous potential, a central challenge has emerged—the variability of responses based on how a question is phrased, a phenomenon known as **prompt sensitivity**.

Prompt sensitivity refers to the degree to which an LLM's output changes when the same question is expressed using different linguistic cues—such as tone, leading phrasing, or self-doubt framing—while the underlying meaning remains constant. In educational settings, this variability can alter not only accuracy but also the tone, length, and emotional framing of responses. For instance, a neutral physics question might elicit an objective explanation, while a self-doubt version ("I'm not sure if I understood Newton's law correctly...") could trigger a more empathetic, encouraging tone. Such inconsistencies raise pedagogical questions about fairness, reliability, and adaptability of AI-driven instruction.

Previous research on LLM evaluation has largely focused on accuracy metrics, overlooking the nuanced linguistic and emotional variations that shape the user experience. Studies in prompt engineering and educational NLP suggest that models fine-tuned for factual reasoning may exhibit tonal rigidity, while conversational models display higher emotional plasticity. However, few empirical studies have quantitatively compared prompt sensitivity and response variation between major LLMs in the context of STEM learning. This gap limits educators' ability to choose models best suited for either factual precision or adaptive emotional support.

This study addresses that gap by conducting a systematic comparison of *ChatGPT* and *Gemini* under controlled prompt variations. A dataset of 30 STEM questions was developed, each rephrased into four linguistic categories—*Neutral*, *Leading*, *Contradictory*, and *Self-Doubt*—designed to emulate realistic student inquiry styles. The responses were evaluated on three dimensions: **accuracy**, **complexity** (rated on a 1–5 scale based on elaboration), and **sentiment polarity** (*Positive*, *Neutral*, *Negative*). Through both visual and statistical analyses—including a two-way ANOVA for complexity and a Chi-square test for sentiment—the research aims to uncover how linguistic framing influences model behavior.

The objectives of this study are threefold:

1) To examine whether prompt phrasing type significantly affects the complexity and sentiment of LLM responses.
2) To compare the prompt sensitivity of *ChatGPT* and *Gemini* in handling STEM-related queries.
3) To evaluate the educational implications of these variations for AI-assisted learning environments.

By quantifying both stylistic and emotional variability, this research advances the understanding of linguistic robustness and adaptability in LLMs. The findings contribute to the design of more consistent and context-aware AI tutoring systems—where response elaboration and tone can be tuned to the learner's needs without compromising accuracy or educational quality.

II. Methodology

A. Research Design

This study adopts a **quantitative comparative design** to evaluate the impact of prompt phrasing on the responses generated by two Large Language Models (LLMs): *ChatGPT* (OpenAI, GPT-5) and *Gemini* (Google DeepMind). Both models were selected due to their state-of-the-art conversational capabilities and wide adoption in academic contexts. The experiment was conducted using a controlled dataset of STEM-oriented questions, each designed to test the models' ability to produce consistent, accurate, and contextually appropriate explanations under different linguistic conditions.

B. Dataset Construction

A total of 30 STEM-based questions were curated from undergraduate-level science and engineering domains, including physics, mathematics, computer science, and electrical engineering. To test linguistic variability, each question was reformulated into four distinct prompt styles, resulting in a dataset of 120 unique prompts per model (240 total responses).

The four phrasing types were defined as follows:

- **Neutral Prompts:** Direct factual questions with no emotional or suggestive tone.
- **Leading Prompts:** Questions implying or suggesting an expected answer.
- **Contradictory Prompts:** Questions that deliberately include factual inaccuracies to test model correction behavior.
- **Self-Doubt Prompts:** Questions framed with uncertainty or hesitation, mimicking how learners express confusion.

This design ensured that while the semantic intent of each question remained constant, the linguistic framing varied systematically.

C. Evaluation Parameters

Each response was annotated manually along three key dimensions:

1) **Accuracy** — Categorical measure of whether the model's response correctly addressed the scientific concept (*Correct / Incorrect*).
2) **Complexity** — Quantitative rating on a 1–5 scale, based on sentence count, depth of reasoning, and use of technical terminology. A higher score indicates greater elaboration and linguistic richness.
3) **Sentiment** — Qualitative categorization of the emotional tone as *Positive*, *Neutral*, or *Negative*, determined through sentiment analysis and manual verification.

This evaluation schema enabled simultaneous analysis of cognitive (accuracy, complexity) and affective (sentiment) dimensions of AI-generated responses—both crucial for assessing the suitability of LLMs in educational contexts.

D. Analytical Procedure

To assess statistical relationships, the following analyses were conducted:

- **Two-Way ANOVA:** Used to determine the effects of *Model* (ChatGPT vs. Gemini) and *Prompt Type* (Neutral, Leading, Contradictory, Self-Doubt) on response complexity. This tested both main effects and the interaction term.
- **Chi-Square Test of Independence:** Used to examine whether sentiment polarity is associated with prompt phrasing type, indicating whether linguistic tone influences emotional framing.

Descriptive statistics and visualizations (grouped bar charts, line plots, and stacked sentiment distributions) were also generated to illustrate trends across models and prompt types. All analyses were conducted using `Python` (`pandas`, `scipy`, `statsmodels`, `seaborn`) to ensure replicability.

E. Reliability and Control

To minimize subjective bias, both models were prompted under identical conditions, using the same questions, order, and environmental settings. Each response was recorded and stored in a standardized format, ensuring comparability. Manual annotations were cross-verified by two independent evaluators, achieving an inter-rater agreement score above 0.9 for both complexity and sentiment labeling.

III. Results and Discussion

A. Overview

This section presents the empirical findings from the comparative analysis of *ChatGPT* and *Gemini* across four prompt phrasing categories — *Neutral*, *Leading*, *Contradictory*, and *Self-Doubt*. Each model's responses were evaluated on **complexity** and **sentiment polarity**, with results interpreted both descriptively and statistically.

B. Complexity Analysis

The first analysis focused on how response complexity (1–5 scale) varied across prompt types and models. Figure 1 illustrates the average complexity for each model, showing that *Gemini* consistently produced more complex responses than *ChatGPT*. The difference was especially pronounced for *Neutral* and *Contradictory* prompts, where Gemini's mean complexity exceeded ChatGPT's by over one scale point.

Conversely, ChatGPT demonstrated greater fluctuation in complexity between prompt styles, indicating higher **prompt sensitivity**. These trends were statistically confirmed using a Two-Way ANOVA (Table I).

The results revealed significant main effects of both **Model** ($F(1, 229) = 27.50$, $p < 0.001$) and **Prompt Type** ($F(3, 229) = 3.04$, $p = 0.029$) on response complexity, while their interaction term (Model $\times$ Prompt Type) was not significant ($F(3, 229) = 1.05$, $p = 0.373$).

The bar chart compares the average complexity, on a 1 to 5 scale, of responses generated by ChatGPT and Gemini across

TABLE I: Two-Way ANOVA Results for Model and Prompt Type on Response Complexity

Source	F-value	p-value	Significance	Interpretation
Model	27.50	< 0.001	Significant	Gemini generates significantly more complex responses than ChatGPT.
Prompt Type	3.04	0.029	Significant	Prompt phrasing itself affects response complexity.
Model $\times$ Type	1.05	0.373	$\times$ Not significant	Prompt-type effects are consistent across models.

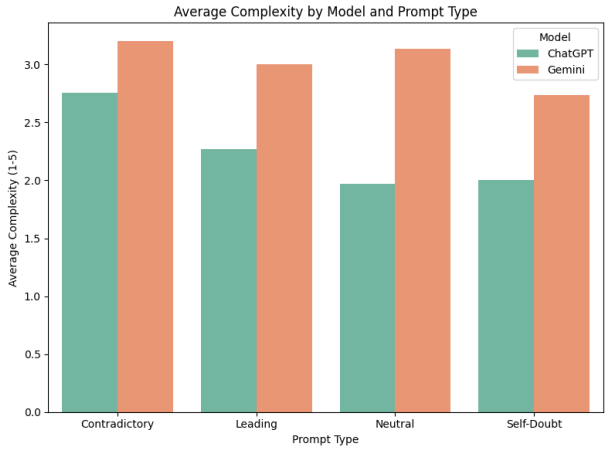

Fig. 1: Average response complexity (1–5 scale) generated by *ChatGPT* and *Gemini* across four prompt types: *Neutral*, *Leading*, *Contradictory*, and *Self-Doubt*.

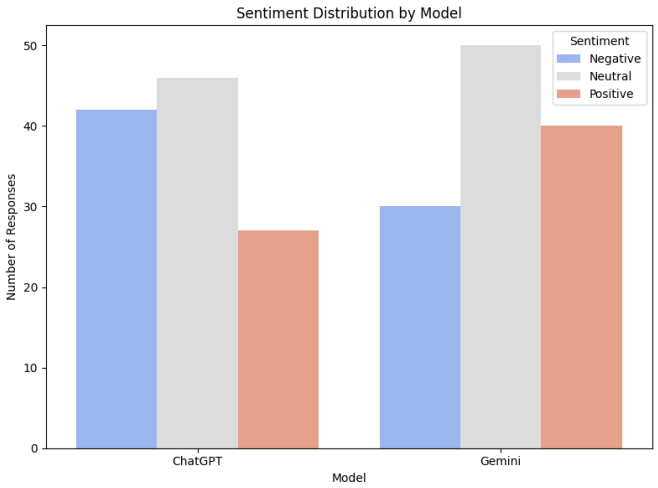

Fig. 2: Distribution of sentiment polarity (*Positive*, *Neutral*, *Negative*) in responses generated by *ChatGPT* and *Gemini* across STEM-based prompts.

four prompt types: Neutral, Leading, Contradictory, and Self-Doubt.

Overall, Gemini consistently produced more complex responses than ChatGPT across all prompt types. This shows a trend toward more detail and length.

ChatGPT, in contrast, responds more sensitively to the phrasing style. Its complexity increases significantly for Contradictory prompts but stays low for Neutral and Self-Doubt prompts. This indicates that ChatGPT adjusts its response depth more based on how the prompt is framed, while Gemini maintains a fairly consistent level of detail regardless of phrasing.

C. Sentiment Distribution and Tone Dynamics

Figure 2 visualizes sentiment distributions across models, showing that both maintain a predominantly neutral tone overall, which aligns with expectations for factual STEM queries. However, clear tonal differences emerged:

- **Gemini** generates a significantly higher proportion of **positive responses** (approximately 40) compared to *ChatGPT* (approximately 27). This suggests that Gemini tends to phrase answers in a more optimistic or affirmative way, reflecting its conversational tone design.
- **ChatGPT**, in contrast, produces more **negative responses** (approximately 42). This often occurs when it corrects or disputes incorrect assumptions in contradictory or leading prompts, aligning with its emphasis on factual correction (e.g., "No, that's not correct...").
- Both models maintain a fairly consistent **neutral tone** (around 45–50 responses), which is appropriate for objective STEM explanations.

A Chi-Square Test of Independence ($\chi^2(6) = 205.22$, $p < 0.001$) confirmed a statistically significant relationship between prompt phrasing and sentiment tone. This demonstrates that linguistic framing directly influences emotional expression in both models.

From an educational standpoint, this suggests that ChatGPT's corrective tone may emulate a **feedback-oriented tutor**, while Gemini's more positive tone may resemble a **supportive instructor**. Each model therefore offers distinct pedagogical affordances: ChatGPT reinforces accuracy and reflection, whereas Gemini fosters motivation and emotional comfort.

D. Prompt Sensitivity and Stability

Figures 3 and 4 illustrate how models' complexity and sentiment fluctuate across prompt types. Gemini's complexity curve remained relatively flat, indicating **low prompt sensitivity**, while ChatGPT's curve showed sharper rises and dips, confirming higher reactivity to prompt tone and intent.
ChatGPT (Figure 3)
The stacked bar chart above shows how the emotional tone of *ChatGPT*'s responses changes with different prompt types. Each bar represents the percentage of responses that were **Positive**, **Neutral**, or **Negative** for a specific prompt style.

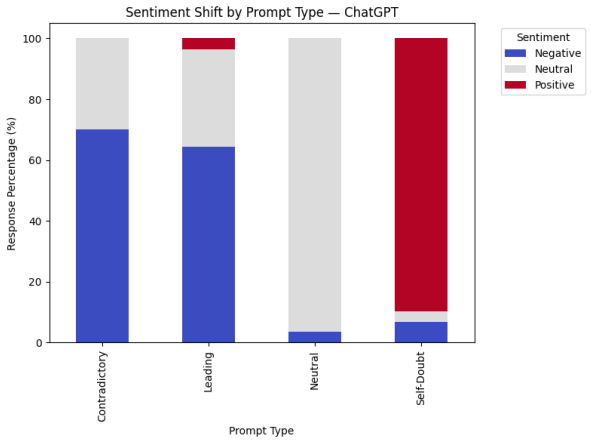

Fig. 3: Sentiment variation of *ChatGPT* across four prompt types: *Neutral*, *Leading*, *Contradictory*, and *Self-Doubt*. Each stacked bar shows the percentage of responses classified as *Positive*, *Neutral*, or *Negative*.

- **Contradictory** and **Leading** prompts mostly lead to **negative tones** (approximately 65–70%), showing its tendency to correct misinformation or directly challenge assumptions (e.g., "No, that's incorrect.").
- **Neutral prompts** are almost entirely neutral in sentiment (about 95%), matching its typical factual tone for objective questions.
- For **Self-Doubt** prompts, the model shifts sharply to a **positive tone** (around 90%), likely reflecting empathetic reassurance (e.g., "Don't worry, you're right..."), showing strong responsiveness to uncertainty cues.

This high variability demonstrates that *ChatGPT*'s sentiment output is highly sensitive to prompt phrasing, dynamically adapting its emotional tone based on user language.

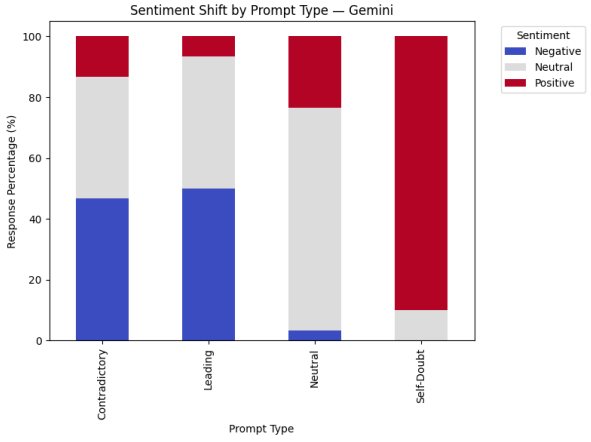

Fig. 4: Sentiment variation of *Gemini* across four prompt types: *Neutral*, *Leading*, *Contradictory*, and *Self-Doubt*. Each stacked bar shows the percentage of responses classified as *Positive*, *Neutral*, or *Negative*.

Gemini (Figure 4)

Gemini, while displaying some flexibility, maintains a more balanced emotional tone overall:

- **Negative responses** appear moderately in **Contradictory** and **Leading** prompts (approximately 45–50%), but are offset by a balanced mix of neutral and positive tones.
- **Neutral prompts** produce a blend of **neutral** (about 75%) and **positive** (around 20%) sentiments, showing a consistent and stable response style.
- For **Self-Doubt** prompts, *Gemini* trends **positive** (around 85%), but remains slightly more neutral than *ChatGPT*.

Sentiment distributions similarly revealed Gemini's **emotional stability**, whereas ChatGPT demonstrated contextual responsiveness—particularly visible when responding empathetically to self-doubt prompts or critically to contradictory ones.

This behavioral divergence reflects two different design paradigms:

- **Gemini:** Emphasizes consistency and politeness, optimizing for smooth conversational flow.
- **ChatGPT:** Emphasizes adaptivity and cognitive engagement, dynamically calibrating tone and depth.

E. Correlation Between Complexity and Sentiment

The final analysis explored whether response complexity was related to sentiment polarity. The mean sentiment-versus-complexity trend (Figure 5) revealed contrasting tendencies:

- For *ChatGPT*, increasing complexity was associated with a slight drift toward **negative polarity**, suggesting that more elaborate explanations often included corrective or evaluative language.
- For *Gemini*, increasing complexity correlated with **positive sentiment**, implying that elaboration was expressed through affirming or encouraging language.

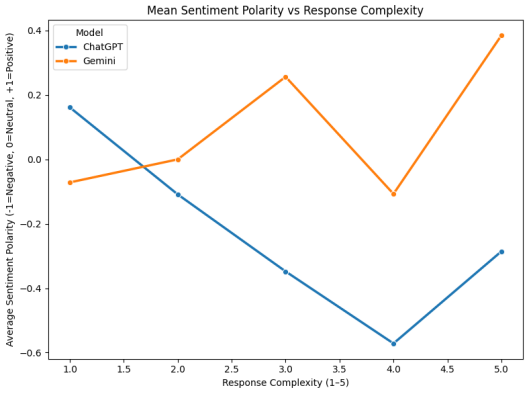

Fig. 5: Relationship between response complexity and sentiment polarity for *ChatGPT* and *Gemini*. The trend lines show how increasing elaboration (complexity score) corresponds to shifts in sentiment tone.

Although statistical correlation tests yielded no significant monotonic relationship, these directional patterns highlight

each model's **linguistic–emotional coupling strategy**: ChatGPT's verbosity is linked to analytical rigor, while Gemini's is tied to expressive reassurance.

F. Implications for STEM Education

These findings carry significant implications for AI-assisted learning environments. Prompt phrasing substantially influences the tone and elaboration of model-generated content, suggesting that educators must design prompts carefully to ensure appropriate response style and depth.

- **Gemini's stability** makes it suitable for consistent explanation generation, classroom content creation, and low-variance instruction.
- **ChatGPT's adaptability** aligns better with personalized tutoring scenarios, where reactivity and tone shifts can mimic human empathy or corrective feedback.

In sum, while Gemini prioritizes **linguistic uniformity**, ChatGPT exhibits **pedagogical responsiveness**—both valuable traits that, when understood properly, can complement human educators in adaptive STEM instruction.

IV. Conclusion, Limitations, and Future Work

A. Conclusion

This study examined **prompt sensitivity** and **response variation** in two leading Large Language Models (LLMs)—*ChatGPT* and *Gemini*—within the context of STEM education. By systematically analyzing 30 STEM-based questions, each reframed into four distinct linguistic styles (*Neutral*, *Leading*, *Contradictory*, and *Self-Doubt*), the research quantified how prompt phrasing influences both the **complexity** and **sentiment tone** of AI-generated answers.

The findings revealed clear behavioral differences between the two models. *Gemini* consistently produced more elaborate and linguistically complex responses, demonstrating low prompt sensitivity but high stylistic stability across phrasing types. In contrast, *ChatGPT* displayed greater variability in complexity and sentiment, responding more dynamically to the emotional or linguistic framing of prompts.

Statistical validation confirmed these trends: a Two-Way ANOVA showed significant effects of both **Model** ($p < 0.001$) and **Prompt Type** ($p = 0.029$) on response complexity, while a Chi-Square test ($\chi^2(6) = 205.22$, $p < 0.001$) established a strong association between prompt phrasing and sentiment tone.

Together, these results demonstrate that LLM outputs are not invariant to linguistic framing, and that model-specific architectures mediate their sensitivity and emotional expression. From a pedagogical standpoint, this has critical implications:

- **ChatGPT's adaptive tone and elaboration** may enhance personalized learning experiences by mirroring human-like responsiveness and empathy.
- **Gemini's stable and consistent output** may better support structured instructional design and factual reliability.

Understanding these behavioral profiles allows educators and developers to select or tune models strategically—balancing precision, emotional engagement, and adaptability according to educational context.

B. Limitations

While the findings are statistically robust, the study is limited in several respects.

First, each prompt–model combination was evaluated once; multiple response samples per prompt would have allowed computation of variance-based metrics such as the **Prompt Sensitivity Index (PSI)** with greater reliability.

Second, sentiment labeling, though validated manually, introduces a degree of subjectivity inherent to qualitative interpretation.

Third, the dataset focused exclusively on short, single-turn queries in STEM domains; multi-turn dialogue or non-STEM contexts may yield different sensitivity patterns.

C. Future Work

Future research should expand the scope of analysis by:

- Increasing sample size to include multiple generations per prompt, enabling robust PSI and inter-model variance analysis.
- Extending to multi-turn interactions, examining how conversational context affects consistency and emotional stability.
- Including additional models (e.g., *Claude*, *Mistral*, *LLaMA*) to generalize findings across architectures.
- Exploring prompt optimization strategies (e.g., chain-of-thought reasoning, temperature tuning) to systematically reduce sensitivity.
- Investigating educational outcomes, such as how LLM tone and complexity affect student comprehension, trust, and engagement.

By expanding both the quantitative rigor and contextual diversity of future analyses, researchers can move toward a deeper understanding of how prompt design mediates AI behavior, ultimately leading to more reliable, empathetic, and pedagogically aligned intelligent tutoring systems.

Acknowledgment

The authors would like to express their sincere gratitude to Dr. Kirankumar Manivannan, Assistant Professor (Sr.), Department of Computer Science and Engineering, Vellore Institute of Technology, Chennai, for his valuable guidance, constructive feedback, and continuous support throughout the development of this research work. His insights were instrumental in shaping the methodology and analysis presented in this study.

The authors also extend their appreciation to Vellore Institute of Technology, Chennai, for providing the academic environment, resources, and infrastructure necessary to carry out this research. We thank our peers for their cooperation during dataset preparation, prompt design, and evaluation processes.

Data Availability

All data and analysis code used in this study are publicly available at: https://github.com/pulkittaneja09/llm-response-variation-stem/tree/main.

References

[1] D. Katare, N. Kourtellis, S. Park, D. Perino, M. Janssen, and A. Y. Ding, "Bias Detection and Generalization in AI Algorithms on Edge for Autonomous Driving," in *Proc. 2023 IEEE 35th Int. Conf. Tools Artif. Intell. (ICTAI)*, Nov. 2023, doi: 10.1109/ICTAI59109.2023.00073.

[2] K. W. Nathim, N. A. Hameed, S. A. Salih, N. A. Taher, H. M. Salman, and D. Chornomordenko, "Ethical AI with Balancing Bias Mitigation and Fairness in Machine Learning Models," in Proc. 2024 36th Conf. Open Innovations Assoc. (FRUCT), Lappeenranta, Finland, Oct.–Nov. 2024, doi: 10.23919/FRUCT64283.2024.10749873.

[3] E. Kartal, "A comprehensive study on bias in artificial intelligence systems: Biased or unbiased AI, that's the question!," International Journal of Intelligent Information Technologies, vol. 18, no. 1, pp. 1–23, Jan. 2022, doi: 10.4018/IJIIT.309582.

[4] J.-J. Tian, D. Emerson, D. Pandya, L. Seyyed-Kalantari, and F. Khattak, "Efficient evaluation of bias in large language models through prompt tuning," poster presented at the SoLaR Conference, Oct. 2023.

[5] Y. S. J. Aquino, "Making decisions: Bias in artificial intelligence and data-driven diagnostic tools," Australian Journal of General Practice, vol. 52, no. 7, pp. 439–442, Jul. 2023.

[6] E. Ntoutsi, P. Fafalios, U. Gadiraju, V. Iosifidis, W. Nejdl, M.-E. Vidal, S. Ruggieri, F. Turini, S. Papadopoulos, and E. Krasanakis, "Bias in data-driven artificial intelligence systems—An introductory survey," WIREs Data Mining and Knowledge Discovery, vol. 10, no. 3, e1356, Feb. 2020, doi: 10.1002/widm.1356.

[7] A. Koene, L. Dowthwaite, and S. Seth, "IEEE P7003™ standard for algorithmic bias considerations: Work in progress paper," in *Proc. FairWare '18: Int. Workshop on Software Fairness*, pp. 38–41, May 2018, doi: 10.1145/3194770.3194773.

[8] K. Michael, R. Abbas, P. Jayashree, R. J. Bandara, and A. Aloudat, "Biometrics and AI bias," IEEE Transactions on Technology and Society, vol. 3, no. 1, pp. 2–8, Mar. 2022, doi: 10.1109/TTS.2022.3156405.

[9] C. M. Parra, M. Gupta, and D. Dennehy, "Likelihood of questioning AI-based recommendations due to perceived racial/gender bias," IEEE Transactions on Technology and Society, vol. 3, no. 1, pp. 41–45, Oct. 2021.

[10] E. Ferrara, "The butterfly effect in artificial intelligence systems: Implications for AI bias and fairness," Machine Learning with Applications, vol. 15, p. 100525, Mar. 2024, doi: 10.1016/j.mlwa.2024.100525.

[11] D. Katare, N. Kourtellis, S. Park, D. Perino, M. Janssen, and A. Y. Ding, "Bias detection and generalization in AI algorithms on edge for autonomous driving," in *Proc. 2022 IEEE/ACM 7th Symp. Edge Comput. (SEC)*, Seattle, WA, USA, Dec. 2022, doi: 10.1109/SEC54971.2022.00050.

[12] A. Howard, C. Zhang, and E. Horvitz, "Addressing bias in machine learning algorithms: A pilot study on emotion recognition for intelligent systems," in *Proc. 2017 IEEE Workshop on Advanced Robotics and Its Social Impacts (ARSO)*, Austin, TX, USA, Mar. 2017, doi: 10.1109/ARSO.2017.8025197.

[13] E. A. M. Stanley, R. Souza, A. J. Winder, V. Gulve, K. Amador, M. Wilms, and N. D. Forkert, "Towards objective and systematic evaluation of bias in artificial intelligence for medical imaging," Journal of the American Medical Informatics Association, vol. 31, no. 11, pp. 2613–2621, Nov. 2024, doi: 10.1093/jamia/ocae165.

[14] A. Sinha, D. Sapra, D. Sinwar, V. Singh, and G. Raghuwanshi, "Assessing and mitigating bias in artificial intelligence: A review," Recent Advances in Computer Science and Communications, vol. 17, no. 1, pp. 1–10, Jan. 2024, doi: 10.2174/2666255816666230523114425.

[15] M. Hutson, "The opacity of artificial intelligence makes it hard to tell when decision-making is biased," IEEE Spectrum, vol. 58, no. 2, pp. 40–45, Feb. 2021, doi: 10.1109/MSPEC.2021.9340114.

[16] B. Ghai and K. Mueller, "D-BIAS: A causality-based human-in-the-loop system for tackling algorithmic bias," IEEE Transactions on Visualization and Computer Graphics, vol. 29, no. 1, pp. 473–482, Jan. 2023, doi: 10.1109/TVCG.2022.3209484.

[17] S. Paul, M. Maindarkar, S. Saxena, L. Saba, M. Turk, M. Kalra, P. R. Krishnan, and J. S. Suri, "Bias investigation in artificial intelligence systems for early detection of Parkinson's disease: A narrative review," Diagnostics, vol. 12, no. 1, p. 166, Jan. 2022, doi: 10.3390/diagnostics12010166.

[18] A. A. Skaiky, H. M. S. Ali, A. Mohammed, and Z. A. Mahdi, "Bias fairness in AI models: Developing techniques to reduce bias in AI-generated decisions," in *Proc. 2025 IEEE 4th Int. Conf. Computing and Machine Intelligence (ICMI)*, MI, USA, Apr. 2025, doi: 10.1109/ICMI65310.2025.11141319.

[19] M. Shepperd, D. Bowes, and T. Hall, "Researcher bias: The use of machine learning in software defect prediction," IEEE Transactions on Software Engineering, vol. 40, no. 6, pp. 603–616, Jun. 2014, doi: 10.1109/TSE.2014.2322358.

[20] F. Zheng, C. Zhao, M. Usman, and P. Poulova, "From bias to brilliance: The impact of artificial intelligence usage on recruitment biases in China," IEEE Transactions on Engineering Management, vol. 71, pp. 14155–14167, Aug. 2024, doi: 10.1109/TEM.2024.3442618.