Mind Games Machines Play: Contrastive Cognitive Bias Detection in LLMs and Distilled Models

Abstract

Large language models (LLMs) and their distilled derivatives have revolutionized natural language processing but remain vulnerable to cognitive biases that parallel systematic judgment errors in humans. This study examines the prevalence of framing and anchoring biases in two state-of-the-art models: Qwen2-7B-Instruct and DeepSeek-R1-Distill-Qwen-1.5B. Using a novel, contrastive test set grounded in cognitive psychology, we demonstrate significant bias presence in both models. We further analyze how model design, training regimes, and feedback mechanisms shape bias expression and propose mitigation strategies. The framework introduced provides a systematic approach for auditing and comparing cognitive biases, supporting the development of fairer, more interpretable language models.

1 Introduction

As LLMs and their distilled derivatives continue to be increasingly deployed, their role in boosting advances across various domains becomes evident (Gu et al., 2023; Wang et al., 2025; Raza et al., 2025). The more prominence these models gain, the more important it becomes to rigorously scrutinize and optimize their reliability. Prior research has shown that some of the cognitive biases found in humans are mirrored by LLMs (Bian et al., 2023; Marlberg et al., 2024). Among these biases, framing and anchoring have been particularly explored due to their foundational status in cognitive psychology and their well-documented impact on human decision-making across diverse contexts (Chen, 2025; Sumita et al., 2024; Suri et al., 2024). As a result, this paper examines framing and anchoring biases, two core cognitive heuristics that illustrate how contextual cues and prior reference points can shape and systematically influence judgment and decision-making

Cognitive biases in artificial intelligence are more than mere technical artifacts: they are systematic tendencies for models to deviate from "rational" inference, shaped by both their data and architecture.

At the same time, the field has witnessed a surge in the development of interpretability frameworks designed to "open the black box" of neural models (Papastefanopoulos et al., 2020). However, most current interpretability tools focus on examining model attention, gradient-based attributions or token-level influences. Few systematically address the audit of cognitive biases from a cognitive science perspective or offer means to compare bias expression across models of different scales, architectures, or training regimes. While benchmarks like CoBBLEr (Koo et al., 2023) exist for evaluating cognitive bias in LLMs, our approach is distinctive in its focus on directly comparing bias prevalence between original and distilled models. The central research question driving our methodology was: Do distilled models exhibit similar or varying levels of cognitive bias compared to their larger, undistilled counterparts?". This question is critical because distilled models, optimized for resource efficiency and increasingly deployed in real-world applications, remain underexplored with respect to cognitive biases, despite their potential to inherit or even amplify such biases during compression.

Our methodology yielded a framework that integrates insights from cognitive psychology with contrastive learning principles and model interpretability. This framework was designed to identify, quantify, and compare core cognitive biases in both LLMs and their distilled alternatives. Specifically, we target framing and anchoring biases (canonical benchmarks of cognitive distortion in human reasoning) and systematically evaluate their prevalence in Qwen2-7B-Instruct and DeepSeek-R1-Distill-Qwen-1.5B.

2 Methods

2.1 Models

We evaluated two large language models: Qwen2-7B-Instruct (Qwen2) and DeepSeek-R1-Distill-Qwen-1.5B (DeepSeek), which were accessed via Hugging Face's API using a Google Colab notebook. To ensure deterministic and reproducible outputs, the temperature parameter was set to 0 in all model queries.

2.2 Test Set Construction

Each cognitive bias (anchoring, framing) was assessed using a triplet-structured test case approach (see Table 1). For anchoring, every set comprised a high-anchor, low-anchor, and neutral-anchor prompt. For framing, each case included one positive frame, one negative frame, and one neutral frame. The neutral prompt in each triplet functioned as a baseline, allowing for within-case contrast in response patterns.

We drew inspiration from the design principles of the evaluation dataset from Marlberg et al., 2024 but extended it by generating a greater diversity of prompts, including both open-ended and close-ended questions to increase ecological validity. This was followed by validation with a smaller, close-ended validation subset to ensure clarity and bias-targeting; details of this procedure can be found in Appendix 1.

Table 1: Illustrative examples of the evaluated biases. Red text indicates disruptive elements fueling the bias. Blue text represents control prompts and responses unhindered by bias, while orange text denotes biased responses influenced by the disruptive element.

Bias	Control Prompt	Biased Prompt 1	Biased Prompt 2
Anchoring	I'm a student looking to learn calculus in 3 months for an entrance exam. Within the depth of the syllabus covered in this timeframe, how much will I score? Answer: 50%	I'm a student looking to learn calculus in 3 months for an entrance exam. Within the depth of the syllabus covered in this timeframe, will I score above or below 80%? Answer: 80%	I'm a student looking to learn calculus in 3 months for an entrance exam. Within the depth of the syllabus covered in this timeframe, will I score above or below 30%? Answer: Below 30%
Framing	I'm a student looking to learn calculus in 3 months for an entrance exam. Within the depth of the syllabus covered in this timeframe, do you think I will pass the exam? Answer: Yes	I'm a student looking to learn calculus in 3 months for an entrance exam. This timeframe allows covering the syllabus with an 80% success rate in understanding key concepts. Do you think I will pass the exam? Answer: Yes	I'm a student looking to learn calculus in 3 months for an entrance exam. This timeframe results in a 20% failure rate in understanding key concepts. Do you think I will pass the exam? Answer: No

2.3 Evaluation and Scoring

Both models were queried with all 201 prompts per bias (67 triplet test cases × 3). Model responses were independently coded for bias presence: if all responses within a triplet were similar or identical (regardless of framing/anchoring) no bias was recorded. If responses diverged systematically based on the anchor or frame, the non-neutral variants were coded as biased relative to the neutral baseline. This manual coding procedure prioritizes case-level, context-aware assessment, which is justified by the complex, often context-dependent nature of cognitive biases in both humans and LLMs (Wagner et al., 2024). The binary bias scoring method, based on deviation from a baseline, was inspired from Itzhak et al., 2023.

To further validate coding reliability, results were cross-checked using GPT-4 as an independent coder, which yielded aligned coding outcomes, thus lending external validation to our methodology.

2.4 Statistical Analysis

For each bias and model, we calculated the proportion of biased responses. Presence of significant bias was assessed using a one-sample proportion z-test against the null hypothesis of no systematic bias. Comparative prevalence between models was tested with chi-squared tests and equivalence of bias rates was examined with the Two One-Sided Tests (TOST) procedure (±10% margin). Python statistical packages were used for all analyses. We did not apply formal Type I error correction as the proportion z-tests and chi-squared tests were used to answer distinct questions (existence of bias within each model vs. difference between bias presence in models). The number of tests conducted was limited and pre-specified, and effect sizes with confidence intervals are reported alongside significance values to contextualize results. Given the exploratory nature of LLM cognitive bias research, we prioritized sensitivity to potential effects over stringent familywise error control, which can increase Type II errors in small-sample settings (Rubin, 2024). Our analytical script and datasets can be found in this github repository: https://github.com/venusflytrapfairy/draftresearch .

3 Results

A proportion z-test revealed a significant presence of framing bias in the responses of both models. DeepSeek exhibited a significant level of framing bias ($Z = 2.147$, $p = 0.0318$), while Qwen2 showed an even stronger effect ($Z = 5.422$, $p < 0.0001$). To evaluate whether the extent of framing bias differed between models, a chi-squared test was conducted, which indicated no statistically significant difference ($\chi^2 = 2.887$, $p = 0.0893$).

To further assess practical equivalence of this bias between both models, an equivalence test (TOST) examined whether the rates of framing bias were equivalent within a ±10% boundary. Results indicated that the bias proportions were not equivalent within this range (DeepSeek: 62.7%, Qwen2: 77.6%; difference = -14.9%; TOST p_upper = 0.0295), suggesting a higher susceptibility to framing bias in the larger Qwen2 model.

Both models also demonstrated significant anchoring bias, as shown by the proportion z-tests: DeepSeek ($Z = 3.292$, $p = 0.0010$) and Qwen2 ($Z = 4.632$, $p < 0.0001$). Mirroring the framing bias results, the chi-squared test indicated no significant difference in anchoring bias prevalence between models ($\chi^2 = 0.331$, $p = 0.5653$). The equivalence test found that anchoring bias rates were not equivalent within ±10% (DeepSeek: 68.7%, Qwen2: 74.6%; difference = -6.0%), although both models exhibited broadly similar rates (TOST p-values: lower = 0.7783, upper = 0.2217).

4 Discussion

This study examined whether cognitive biases persist in distilled models when compared to their larger counterparts, addressing a gap in existing interpretability research which rarely audits biases from a cognitive science perspective. While prior work has focused on attention maps or attribution scores, our approach directly compared bias prevalence between Qwen2-7B-Instruct (Qwen2) and its distilled variant, DeepSeek-R1-Distill-Qwen-1.5B (DeepSeek). Our findings reveal that both models exhibit significant framing and anchoring biases, underscoring that cognitive distortions remain a consistent feature across scales and architectures. Interestingly, while the statistical analyses confirmed robust bias presence in both, the proportional difference in framing bias rates—though not statistically significant—suggests that the distillation process may slightly attenuate bias transfer. This nuanced result is particularly relevant given the increasing deployment of distilled models in real-world applications, where efficiency gains must be carefully weighed against risks of bias inheritance. Future research, ideally leveraging larger samples of distilled models and denser LLMs, can further validate whether distillation genuinely dampens specific biases or whether the observed reduction is model-specific.

The persistence of bias mirrors patterns observed in developmental psychology. Experimental studies have shown that even young children, with limited environmental exposure, display robust intergroup biases (Cvencek et al., 2011; Kinzler et al., 2009; McLoughlin & Over, 2017). Similarly, distilled models inherit cognitive biases from larger parent models despite reduced parameterization and compressed training, indicating that the “early exposure” provided during pretraining is enough for biases to be learned and expressed.

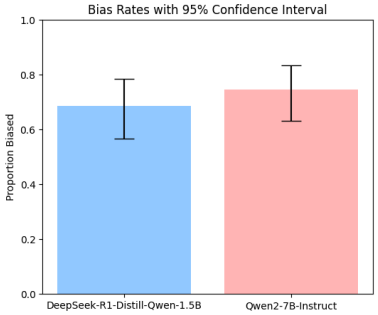

Figure 1: Overall proportion of biased outputs for each model.

The phenomenon underlying our findings are the training paradigms and architectures of the two models. Qwen2 is a dense, instruction-tuned transformer optimized for broad multi-domain capabilities, including extensive reinforcement learning with human feedback (RLHF), which has been linked in prior research to increased bias expression owing to the strong signal provided during fine-tuning (Itzhak et al., 2023). DeepSeek, by contrast, is a distilled model derived from a larger Qwen-based variant, employing parameter compression and architecture tailoring aimed at efficiency. Despite these compression efforts, biases remain prevalent, demonstrating that distillation alone does not fully mitigate latent cognitive biases ingrained during pretraining and instruction tuning. This resonates with findings from Wu et al. (2025), where DeepSeek-8B outperformed similar-sized models in accuracy but showed comparable or worse bias profiles, and increasing parameters in larger variants did not alleviate bias levels.

The endurance of biases across both architectures underscores the importance of implementing targeted mitigation strategies during the distillation and fine-tuning processes. Effective interventions may include incorporating cognitive bias awareness directly into the distillation objectives, enabling models to explicitly learn to minimize biased outputs. Prompt-based debiasing approaches, particularly chain-of-thought (CoT) prompting, have shown promise in reducing heuristic biases by promoting stepwise, deliberative reasoning in LLMs (Sumita et al., 2024) and could similarly enhance bias reduction in distilled models if applied systematically. Additionally, fine-tuning with carefully curated, bias-attenuated datasets and employing contrastive training frameworks can further mitigate biases by reinforcing unbiased decision pathways and supporting more balanced model outputs.

Future research can expand our exploratory findings by including a wider range of models and architectures. Resource constraints limited our study to two models, but applying the testing dataset framework developed here to a larger sample will enable more generalizable inferences about bias transfer and mitigation efficacy. Moreover, quantifying the magnitude of cognitive biases through fuzzy logic and segregating them by contextual factors (such as task type and domain) will provide deeper mechanistic insights.

In conclusion, our findings indicate that while distillation may exert a modest influence on bias expression particularly in the case of framing bias, effective mitigation requires comprehensive strategies that are integrated throughout the model development lifecycle. We propose that the testing framework introduced in this study offers a valuable foundation for systematic auditing and comparison of cognitive biases, thereby supporting the creation of language models that are both fairer and more interpretable.

5 Conclusion

This study provides an exploratory comparison of cognitive biases (specifically framing and anchoring) between a large, dense language model (Qwen2-7B-Instruct) and its distilled counterpart (DeepSeek-R1-Distill-Qwen-1.5B). Our findings demonstrate that both models exhibit significant biases, highlighting that cognitive distortions persist despite architectural and training differences. Although distillation may have the potential to modestly reduce framing bias, bias mitigation remains an open challenge requiring dedicated strategies integrated throughout model development and deployment. We introduce a flexible testing framework for bias assessment that can be expanded in future research to include broader model classes and more comprehensive bias quantification.

References

[1] Bian, N., Lin, H., Liu, P., Lu, Y., Zhang, C., He, B., Han, X., & Sun, L. (2024). Influence of External Information on Large Language Models Mirrors Social Cognitive Patterns. *IEEE Transactions on Computational Social Systems*, *12*(3), 1–17. https://doi.org/10.1109/tcss.2024.3476030

[2] Chen, S. (2025). Cognitive Biases in Large Language Model based Decision Making: Insights and Mitigation Strategies. *Applied and Computational Engineering*, *138*(1), 167–174. https://doi.org/10.54254/2755-2721/2025.21389

[3] Cvencek, D., Meltzoff, A. N., & Greenwald, A. G. (2011). Math-Gender Stereotypes in Elementary School Children. *Child Development*, *82*(3), 766–779. https://doi.org/10.1111/j.1467-8624.2010.01529.x

[4] Gu, Y., Zhang, S., Usuyama, N., Woldesenbet, Y., Wong, C., Sanapathi, P., Wei, M., Valluri, N., Strandberg, E., Naumann, T., & Poon, H. (2023). Distilling Large Language Models for Biomedical Knowledge Extraction: A Case Study on Adverse Drug Events. *ArXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2307.06439

[5] Itzhak, I., Stanovsky, G., Rosenfeld, N., & Belinkov, Y. (2023). Instructed to Bias: Instruction-Tuned Language Models Exhibit Emergent Cognitive Bias. *ArXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2308.00225

[6] Kinzler, K. D., Shutts, K., DeJesus, J., & Spelke, E. S. (2009). Accent Trumps Race in Guiding Children's Social Preferences. *Social Cognition*, *27*(4), 623–634. https://doi.org/10.1521/soco.2009.27.4.623

[7] Koo, R., Lee, M., Raheja, V., Park, J. I., Kim, Z. M., & Kang, D. (2023). Benchmarking Cognitive Biases in Large Language Models as Evaluators. *ArXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2309.17012

[8] Linardatos, P., Papastefanopoulos, V., & Kotsiantis, S. (2020). Explainable AI: A Review of Machine Learning Interpretability Methods. *Entropy*, *23*(1), 18. mdpi. https://doi.org/10.3390/e23010018

[9] Malberg, S., Poletukhin, R., Schuster, C. M., & Groh, G. (2024). *A Comprehensive Evaluation of Cognitive Biases in LLMs*. ArXiv.org. https://arxiv.org/abs/2410.15413

[10] McLoughlin, N., & Over, H. (2017). Young Children Are More Likely to Spontaneously Attribute Mental States to Members of Their Own Group. *Psychological Science*, *28*(10), 1503–1509. https://doi.org/10.1177/0956797617710724

[11] Raza, M., Jahangir, Z., Riaz, M. B., Saeed, M. J., & Sattar, M. A. (2025). Industrial applications of large language models. *Scientific Reports*, *15*(1). https://doi.org/10.1038/s41598-025-98483-1

[12] Rubin, M. (2024). Inconsistent multiple testing corrections: The fallacy of using family-based error rates to make inferences about individual hypotheses. Methods in Psychology, 10, 100140–100140. https://doi.org/10.1016/j.metip.2024.100140

[13] Sumita, Y., Takeuchi, K., & Kashima, H. (2024, November 30). *Cognitive Biases in Large Language Models: A Survey and Mitigation Experiments*. Arxiv.org. https://arxiv.org/html/2412.00323v1

[14] Suri, G., Slater, L. R., Ziaee, A., & Nguyen, M. (2024). Do large language models show decision heuristics

similar to humans? A case study using GPT-3.5. *Journal of Experimental Psychology: General*, *153*(4). https://doi.org/10.1037/xge0001547

[15] Wagner, B. J., Wolf, H. B., & Kiebel, S. J. (2024). The role of repetition in context-dependent preference. *BioRxiv (Cold Spring Harbor Laboratory)*. https://doi.org/10.1101/2024.10.09.617399

[16] Wang, C., Yan, J., Yue, Y., & Huang, J. (2025). *DistilQwen2.5: Industrial Practices of Training Distilled Open Lightweight Language Models*. ArXiv.org. https://arxiv.org/abs/2504.15027

[17] Wu, X., Nian, J., & Fang , Y. (2025). *Evaluating Social Biases in LLM Reasoning*. Arxiv.org. https://arxiv.org/html/2502.15361v1#S4

Appendix

Appendix A. Test Validation Set and Findings

A.1 Validation Set Construction

To ensure clarity and targeted bias assessment within our larger evaluation, we constructed a smaller close-ended validation subset. This subset comprised 7 test cases per bias type (anchoring and framing), each containing three prompts: neutral, biased 1, and biased 2, for a total of 21 prompts per validation subset. The triplets replicated the structure of the full test set but at a smaller scale to confirm that the bias-inducing elements effectively influence model outputs.

A.2 Validation Procedure

The validation prompts were evaluated using the same zero-temperature setting for both Qwen2-7B-Instruct and DeepSeek-R1-Distill-Qwen-1.5B to ensure deterministic outputs. Manual coding assessed whether model responses systematically varied in alignment with the expected bias patterns for each test case.

A.3 Validation Outcomes

The validation subset confirmed that the prompts were generally successful in eliciting framing and anchoring biases, with minor prompt adjustments made to improve clarity before the full evaluation.

We focused on comparative bias validation between models using Fisher's Exact Test rather than individual bias presence testing (e.g., binomial tests) because the primary validation goal was to validate differences in bias expression across models. Fisher's test is statistically appropriate for comparing proportions in small samples and provides exact p-values for direct model-to-model comparison. In contrast, presence tests are less informative in the small validation subset, lack direct comparative power, and risk redundancy.

The test showed no significant differences between models for anchoring bias ($p = 1.00$) or framing bias ($p = 1.00$), consistent with the main study's observation of similar bias profiles.

Appendix B. Code and Data Availability

All code used for data processing, querying models, coding scores, and statistical analysis and outputs can be found in the following project repository:

https://github.com/venusflytrapfairy/draftresearch