Evaluating the Capabilities of LLMs in Mimicking Human Learning Behaviors: An Interdisciplinary Approach

Wendy Liang *Advised by Professor Soroush Vosoughi, Daniel Rockmore*
Dartmouth College, Department of Mathematics and Computer Science

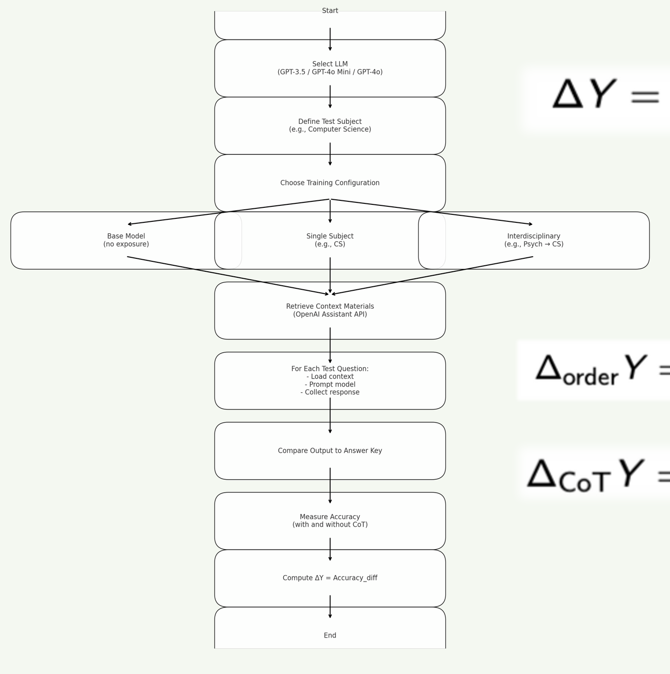
INTRODUCTION

This study presents a novel framework to evaluate how large language models (LLMs) respond to interdisciplinary learning—mirroring human-like knowledge transfer across domains. Using a series of controlled experiments, we assess whether prior exposure to one subject (e.g., Economics) affects model performance on another (e.g., Computer Science), and whether the sequence of exposure or reasoning strategies like Chain-of-Thought (CoT) modulate this effect.

Our analysis, centered on GPT-4o and evaluated via AP-style multiple-choice questions, finds that interdisciplinary exposure can enhance or hinder performance depending on subject pairing and order. CoT generally improves accuracy and reduces sensitivity to ordering, especially in structured domains. These results highlight LLMs' potential to model aspects of cognitive flexibility and inform the design of AI training and educational tools that better support cross-domain reasoning.

DATA & METHODS

Training and testing data were prepared using OCR and web scraping from AP-style materials across seven subjects. Each experiment tested LLMs on 60 multiple-choice questions per subject pairing. The AP framework ensures consistent rigor, and MCQs allow objective, reproducible assessment. The experimental process is shown below, covering three designs: subject pairing, order effects, and Chain-of-Thought prompting.



$$\Delta Y = f_{M_{s_1 \to s_2}}(X_{\text{CS}}) - f_{M_{\text{CS}}}(X_{\text{CS}})$$

$$\Delta_{\text{order}} Y = f_{M_{s_1 \to s_2}}(X_{\text{CS}}) - f_{M_{s_2 \to s_1}}(X_{\text{CS}})$$

$$\Delta_{\text{CoT}} Y = f_M^{\text{CoT}}(X_{\text{CS}}) - f_M^{\text{noCoT}}(X_{\text{CS}})$$

RESULTS

Figure 1. Subject-Specific Transfer Effects
Computer Science (CS) shows the greatest variation across interdisciplinary pairings, making it the most informative test case. Some pairings (e.g., Psych–CS) boost performance, while others (e.g., Latin–CS) introduce interference.

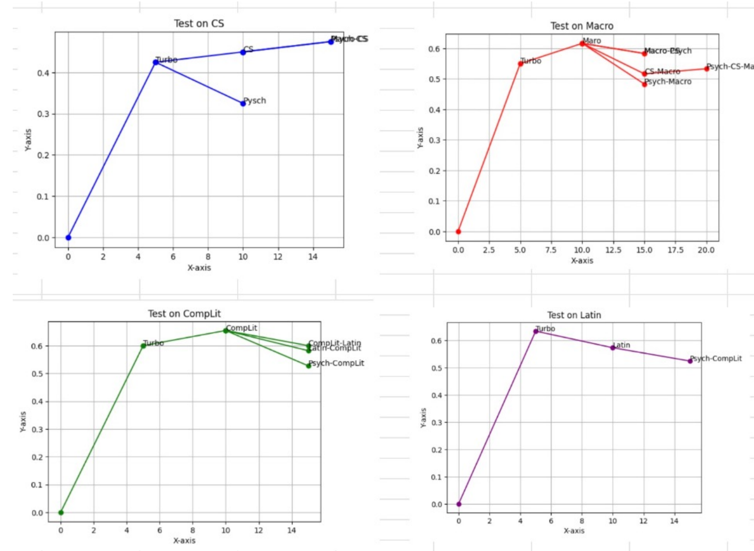


Figure 2. Performance Trajectories by Model and Context
Model performance varies by architecture and training condition. GPT-4o benefits most from interdisciplinary input; OpenAI GPT-3.5 drops under CS-only and partially recovers. LangChain GPT-3.5 shows steady gains and minimal variation. Stronger models generalize better and respond more to structured context.

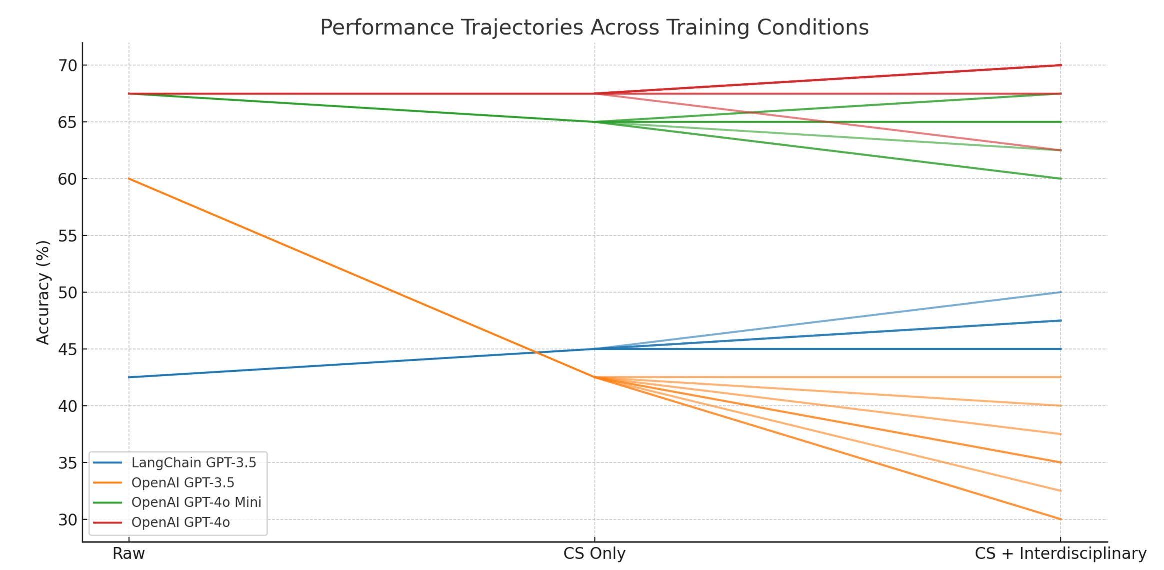


Figure 3. Subject-Specific Transfer Effects

- CS-only training (67.5%) underperforms Raw (70.5%), suggesting narrow exposure reduces generalization.
- Interdisciplinary pairings improve CS performance by +1.36 points on average ($p < 0.001$).
- Top gains over CS-only: Psych–Econ: +5.0, Stats–Psych: +4.0, Psych–Latin: +4.0
- Largest drops (vs. Raw): CS–Psych, Latin–CS, CS–CompLit (all –4.5)

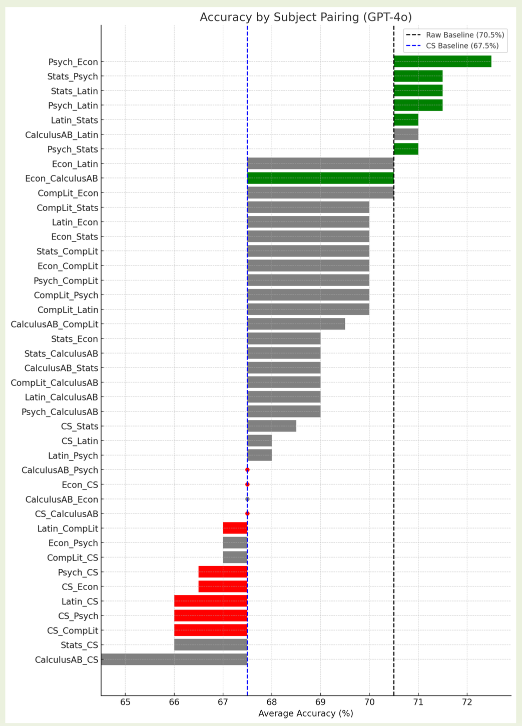


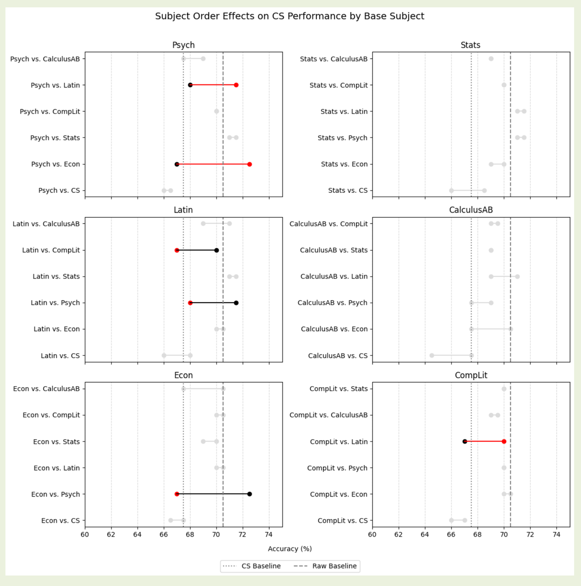
Figure 4. Subject Order Effects on CS Performance

- No overall effect: Mean difference across reversed subject pairs is +0.14 pts ($p = 0.76$), not statistically significant.
- Significant asymmetries in specific pairs:

Psych–Econ outperforms Econ–Psych: +5.5 pts ($p = 0.025$)
Psych–Latin outperforms Latin–Psych: +3.5 pts ($p = 0.044$)
CompLit–Latin outperforms Latin–CompLit: +3.0 pts ($p = 0.041$)

Directional trend:

- Psychology helps more when introduced first
- Latin helps more when introduced second



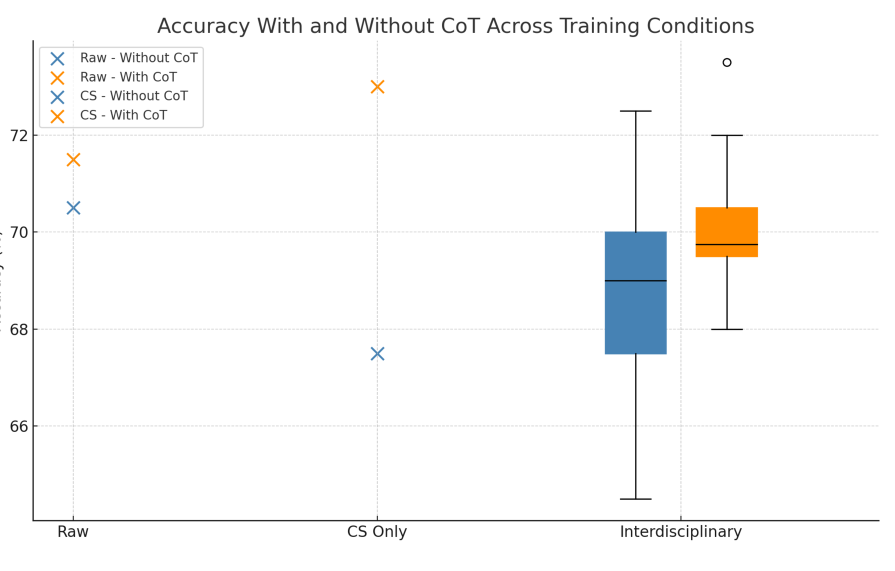
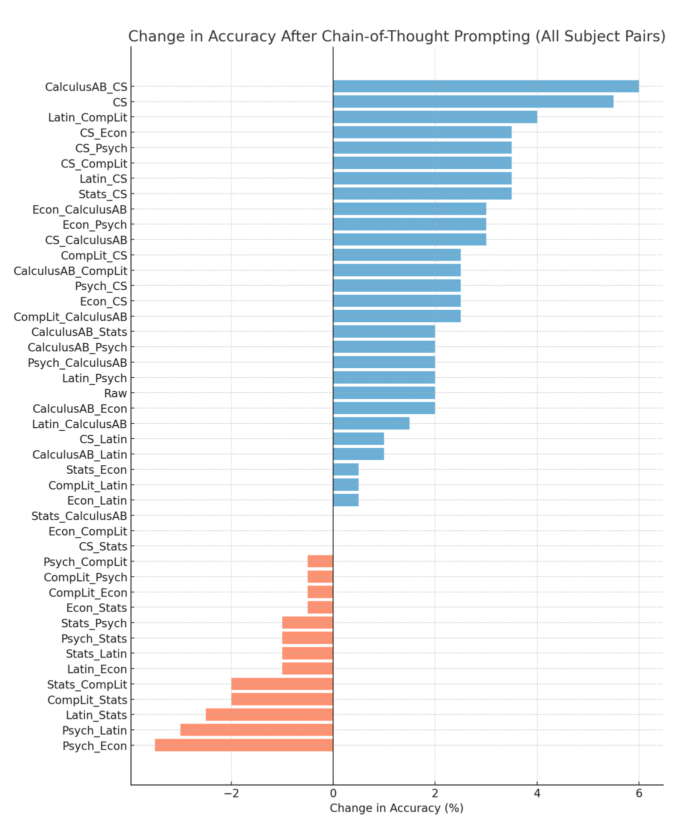
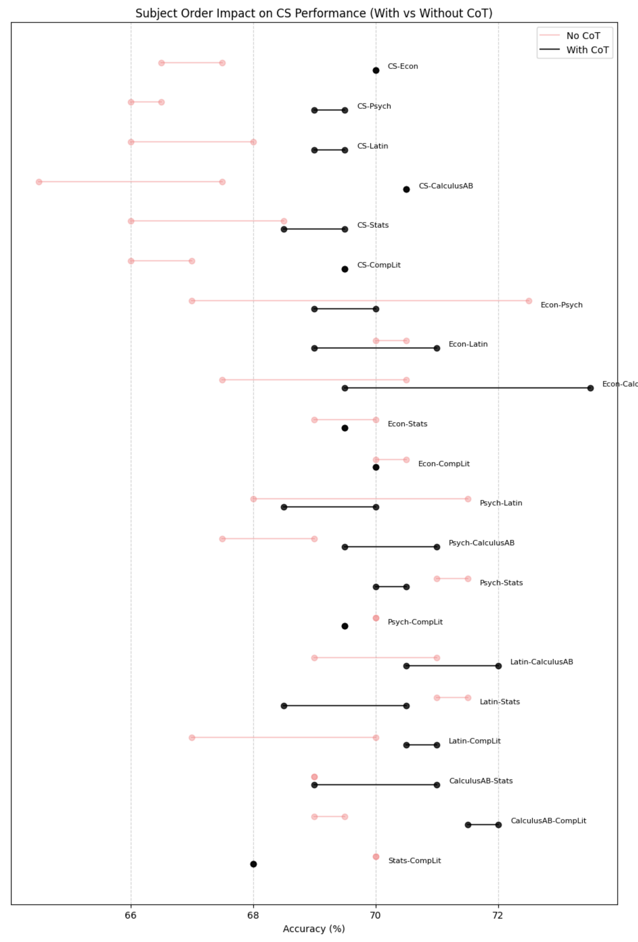


Figure 5. Chain-of-Thought (CoT) Prompting

- CoT improves overall performance: Average gain: +1.23 points, $p = 0.0004$
- CS-only improves from 67.5% to 73.0%, exceeding Raw with CoT (72.5%)
- Top improvements: CalculusAB–CS (+6.0), CS–Econ (+3.5), Latin–CS (+3.5)
- Largest declines: Psych–Latin (–3.0), Latin–Stats (–2.5), Psych–Econ (–2.0)
- Order effects with CoT:
  - No significant change in direction: +0.048 pts, $p = 0.930$
  - Slight reduction in order sensitivity: –0.619 pts in absolute effect, $p \approx 0.085$





CONCLUSIONS

While CS-only training underperforms the raw model—suggesting that narrow exposure may limit generalization—interdisciplinary input helps mitigate this rigidity. On average, subject pairings improve performance by +1.36 points over the CS-only baseline, indicating that cross-domain context can enhance reasoning. However, the benefits are not universal; some combinations improve accuracy, while others introduce interference, reflecting the nuanced nature of transfer in LLMs.

This complexity extends to the order of exposure. Although no consistent directional effect emerged, certain asymmetries suggest that the sequence in which subjects are introduced matters. For example, Psychology boosts downstream performance when presented first, while Latin benefits from being second—hinting at subject-specific roles in priming or absorbing knowledge.

CoT prompting further shapes these dynamics. By encouraging step-by-step reasoning, CoT increases accuracy by 1.23 points, reverses the CS-only performance drop, and modestly reduces sensitivity to ordering. This suggests that structured reasoning may help LLMs integrate interdisciplinary input more effectively.

Together, these findings reveal that LLMs do not passively absorb new information—they exhibit structured, context-sensitive learning behaviors. Subject pairing, order, and reasoning strategy all interact to influence generalization, offering insights for educational AI, curriculum sequencing, and model interpretability.

REFERENCES

Carmichael, T., & LaPierre, Y. (2014). *Interdisciplinary learning works: The results of a comprehensive assessment of students and student learning outcomes in an integrative learning community*. *Issues in Interdisciplinary Studies*, (32), 53–78. Retrieved from https://eric.ed.gov/?id=EJ1117882
Chan, C.-M., Chen, W., Su, Y., Yu, J., Xue, W., Zhang, S., Fu, J., & Liu, Z. (2023). *ChatEval: Towards better LLM-based evaluators through multi-agent debate*. arXiv preprint arXiv:2308.07201. https://doi.org/10.48550/arXiv.2308.07201
Gao, C., Lan, X., Lu, Z., Mao, J., Piao, J., Wang, H., Jin, D., & Li, Y. (2023). *S3: Social-network simulation system with large language model-empowered agents*. arXiv preprint arXiv:2307.14984. http://arxiv.org/abs/2307.14984
Leng, Y., & Yuan, Y. (2024). *Do LLM agents exhibit social behavior?* arXiv preprint arXiv:2312.15198. http://arxiv.org/abs/2312.15198
Li, Y., Zhang, Y., & Sun, L. (2023). *MetaAgents: Simulating interactions of human behaviors for LLM-based task-oriented coordination via collaborative generative agents*. arXiv preprint arXiv:2310.06500. http://arxiv.org/abs/2310.06500
Liu, H.-Y., Hsu, D.-Y., Han, H.-M., Wang, I.-T., Chen, N.-H., Han, C.-Y., Wu, S.-M., Chen, H.-F., & Huang, D.-H. (2022). Effectiveness of interdisciplinary teaching on creativity: A quasi-experimental study. *International Journal of Environmental Research and Public Health*, *19*(10), 5875. https://doi.org/10.3390/ijerph19105875
Rockmore, D. N. (Ed.). (2017). *What are the arts and sciences? A guide for the curious*. Hanover, NH: Dartmouth College Press.
Webb, T., Holyoak, K. J., & Lu, H. (2023). *Emergent analogical reasoning in large language models*. arXiv preprint arXiv:2212.09196. http://arxiv.org/abs/2212.09196
Xu, C., Wu, C.-F., Xu, D.-D., Lu, W.-Q., & Wang, K.-Y. (2022). Challenges to student interdisciplinary learning effectiveness: An empirical case study. *Journal of Intelligence*, *10*(4), 88. https://doi.org/10.3390/jintelligence10040088
Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E. H., Le, Q. V., & Zhou, D. (2022). *Chain-of-thought prompting elicits reasoning in large language models*. arXiv preprint arXiv:2201.11903. https://arxiv.org/abs/2201.11903
Sprague, Z., Yin, F., Rodriguez, J. D., Jiang, D., Wadhwa, M., Singhal, P., Zhao, X., Ye, X., Mahowald, K., & Durrett, G. (2024). *To CoT or not to CoT? Chain-of-thought helps mainly on math and symbolic reasoning*. arXiv preprint arXiv:2409.12183.

ACKNOWLEDGEMENTS

I am deeply grateful to my advisors, Professors Soroush Vosoughi, Daniel Rockmore, and Peter Mucha, for their guidance, feedback, and encouragement throughout this project. I also thank the Dartmouth College Department of Mathematics and the Computer Science program for their academic support. Finally, I'm thankful to my friends and family for always being there for me.