Toward Controllable and Trustworthy LLM Reasoning: From Failure Mapping to Cognition-inspired Control and Real-world Impact

Ben Zhou

Arizona State University
benzhou@asu.edu

Large Language Models (LLMs) have advanced rapidly and raised the bar for what AI is expected to do. However, accompanied with such progress is a stronger consensus that these models consistently fail in out-of-distribution reasoning, especially on tasks that require abstraction, transfer, or long-horizon planning. While acceptable for most consumer use, these issues prevent AI from being safely deployed in high-stakes settings (e.g., healthcare), where stakeholders cannot trust AI models that exhibit uncontrollable and unpredictable failures. In this talk, I will discuss our work and insights on how to make LLM reasoning controllable and trustworthy, by 1) understanding the mechanisms of LLM reasoning and predicting when LLM will fail; 2) improving model reasoning and generalization based on such insights; and 3) moving towards trustworthy AI applications through such improvements, and identifying new problems to form a healthy positive-feedback loop.

I will first discuss my analytical work on understanding LLM reasoning and identifying LLM failure modes. I will begin with our work on conceptual reasoning (Zhou et al. 2024), demonstrating that LLMs can appear competent yet break under abstraction, spurious cues, and long-horizon dependencies. I will then explain that a potential reason is that models latch onto deceptive semantic biases and spurious correlations (Li et al. 2024b), and how these spurious reasoning patterns pose trustworthiness concerns, such as low reasoning consistency (Yu et al. 2024) and safety rule violations (Xu et al. 2024). I will also share our methodologies to effectively predict potential failures of LLM reasoning as a control mechanism (Li et al. 2024a).

To address these issues, I will discuss our work to improve controllability and trust in LLM reasoning. In the first set of approaches, we aim to "steer" model reasoning towards human reasoning: the only known reasoning system humans can trust. I will talk about our work on injecting human-cognition-inspired priors into learning schemes and optimization processes, such as problem decomposition (Zhou et al. 2022; Chen et al. 2024); internalizing fine-grained cognitive feedback (Dineen et al. 2025), and steering self-reflection (RRV et al. 2025). I will also discuss our work to improve reasoning abstraction and consistency (Ye et al. 2025b; Zhou et al. 2025), as well as augmenting next word prediction with fine-grained human explanations (Xu et al. 2025). The second set of approaches aims to derive trust from the model itself. On this front, I will introduce our work on mitigating overthinking in reinforcement learning (RL) (Li et al. 2025) and on applying RL and soft outcome matching for next-word prediction (Shen et al. 2025). I will also discuss our work towards understanding the potential of multi-model evolution (Liu et al. 2025).

Finally, I will discuss our work on translating our insights into real-world impact and trustworthy applications. I will start with demonstrating that some analytical metrics introduced above can be used to improve model performance in downstream tasks (Jung et al. 2025). I will then introduce our work on improving complex decision-making under high uncertainty with Bayesian-infused inference (Feng et al. 2025) and later applying such insights to build patient-facing chatbots that adapt to patients' subtle emotional needs (Srinivasan et al. 2025). I will close the loop with a roadmap outlining how these three directions (i.e., identifying reasoning issues, resolving them, and applying solutions in real-world applications) will form a healthy, self-evolving loop. To illustrate this, I will use the high-stakes domain of healthcare AI as an example to showcase the potential for developing controllable and trustworthy AI applications that account for LLM reasoning risks and solutions (Ye et al. 2025a).

References

Chen, S.; Zhang, H.; Chen, T.; Zhou, B.; Yu, W.; Yu, D.; Peng, B.; Wang, H.; Roth, D.; and Yu, D. 2024. Sub-Sentence Encoder: Contrastive Learning of Propositional Semantic Representations. In *NAACL 2024*, 1596–1609.

Dineen, J.; RRV, A.; Liu, Q.; Xu, Z.; Ye, X.; Shen, M.; Li, Z.; Lu, S.; Baral, C.; Chen, M.; et al. 2025. QA-LIGN: Aligning LLMs through Constitutionally Decomposed QA. In *EMNLP 2025 (Findings)*.

Feng, Y.; Zhou, B.; Lin, W.; and Roth, D. 2025. BIRD: A Trustworthy Bayesian Inference Framework for Large Language Models. In *ICLR*.

Jung, D.; Liu, Q.; Huang, T.; Zhou, B.; and Chen, M. 2025. Familiarity-aware Evidence Compression for Retrieval Augmented Generation. In *EMNLP 2025 (Findings)*.

Li, B.; Zhou, B.; Fu, X.; Wang, F.; Roth, D.; and Chen, M. 2024a. FamiCom: Further Demystifying Prompts for Language Models with Task-Agnostic Performance Estimation. *arXiv preprint arXiv:2406.11243*.

Li, B.; Zhou, B.; Wang, F.; Fu, X.; Roth, D.; and Chen, M. 2024b. Deceptive Semantic Shortcuts on Reasoning Chains: How Far Can Models Go without Hallucination? In *NAACL 2024*.

Li, R.; Luo, Z.; Zhang, Q.; Li, R.; Zhou, B.; Payani, A.; and Du, X. 2025. AALC: Large Language Model Efficient Reasoning via Adaptive Accuracy-Length Control. *arXiv preprint arXiv:2506.20160*.

Liu, Q.; Dineen, J.; Huang, Y.; Zhang, S.; Poon, H.; Zhou, B.; and Chen, M. 2025. ArenaBencher: Automatic Benchmark Evolution via Multi-Model Competitive Evaluation. *arXiv preprint arXiv:2510.08569*.

RRV, A.; Dineen, J.; Handa, D.; Uddin, M. N.; Parmar, M.; Baral, C.; and Zhou, B. 2025. ThinkTuning: Instilling Cognitive Reflections without Distillation. In *EMNLP 2025*.

Shen, M.; Xu, Z.; Ye, X.; Dineen, J.; and Zhou, B. 2025. BOW: Bottlenecked Next Word Exploration. *arXiv preprint arXiv:2506.13502*.

Srinivasan, A.; Dineen, J.; Afzal, M. U.; Sarfraz, M. U.; Riaz, I. B.; and Zhou, B. 2025. RECAP: Transparent Inference-Time Emotion Alignment for Medical Dialogue Systems. *arXiv preprint arXiv:2509.10746*.

Xu, N.; Wang, F.; Zhou, B.; Li, B.; Xiao, C.; and Chen, M. 2024. Cognitive Overload: Jailbreaking Large Language Models with Overloaded Logical Thinking. In *NAACL 2024 (Findings)*.

Xu, Z.; Shen, M.; Dineen, J.; Li, Z.; Ye, X.; Lu, S.; Rrv, A.; Baral, C.; and Zhou, B. 2025. ToW: Thoughts of Words Improve Reasoning in Large Language Models. In *NAACL 2025*.

Ye, X.; Dineen, J.; Li, Z.; Xu, Z.; Chen, W.; Lu, S.; Huang, Y.; Shen, M.; Tran, P.; Yum, J.-E. I.; et al. 2025a. Evaluating Medical LLMs by Levels of Autonomy: A Survey Moving from Benchmarks to Applications. *arXiv preprint arXiv:2510.17764*.

Ye, X.; Shrivastava, S.; Li, Z.; Dineen, J.; Lu, S.; Ahuja, A.; Shen, M.; Xu, Z.; and Zhou, B. 2025b. CC-LEARN: Cohort-based Consistency Learning. *arXiv preprint arXiv:2506.15662*.

Yu, X.; Zhou, B.; Cheng, H.; and Roth, D. 2024. Reasonagain: Using extractable symbolic programs to evaluate mathematical reasoning. *arXiv preprint arXiv:2410.19056*.

Zhou, B.; Jain, S.; Zhang, Y.; Ning, Q.; Wang, S.; Benajiba, Y.; and Roth, D. 2025. Self-supervised analogical learning using language models. *arXiv preprint arXiv:2502.00996*.

Zhou, B.; Richardson, K.; Yu, X.; and Roth, D. 2022. Learning to Decompose: Hypothetical Question Decomposition Based on Comparable Texts. In *EMNLP 2022*.

Zhou, B.; Zhang, H.; Chen, S.; Yu, D.; Wang, H.; Peng, B.; Roth, D.; and Yu, D. 2024. Conceptual and unbiased reasoning in language models. *arXiv preprint arXiv:2404.00205*.