Locating Reasoning Errors with Active Sampling

Chris Merck (chrismerck@gmail.com) – NEMI Workshop, August 22, 2025, Northeastern University, Boston

Abstract

Recent advances in chain-of-thought reasoning allow large language models (LLMs) to solve mathematical problems that are beyond the computational power of a single forward pass. This suggests the existence of learned mechanisms operating at the level of reasoning steps in the chain of thought (CoT), yet few techniques exist for identifying these mechanisms. Furthermore, chains of thought can be deceptively interpretable, resembling human internal monologues closely enough that we risk being misled about their causal structure, heightening the need for rigorous interpretability methods. In this work in progress, we develop an algorithm for locating reasoning errors in incorrect base solutions via targeted resampling. This exploration should improve our understanding of chain-of-thought reasoning, particularly how it goes wrong, allowing us to more safely and efficiently operate reasoning models.

Introduction

One frame through which to interpret reasoning traces is to treat them as the work of a math student, with the researcher as tutor. Correct answers to unguessable problems may be rewarded without inspecting the work, while the incorrect solutions are more interesting: the tutor scans for the first mistake and circles in red pen, leaving a helpful comment. We aim ultimately to automate this process, with this poster presenting some of the elements of work in progress. We see this work as part of a broad effort to take advantage of the present opportunity for CoT monitoring and interpretability (Korbak et al., 2025) and an application of counterfactual techniques from physical systems (Merck and Kleinberg, 2016).

Foundational Work. In their recent groundbreaking preprint, Bogdan et al. (2025) employ several techniques for finding causal structure among the sentences of a reasoning trace, cross-validating counterfactual sampling against attention tracing and masking. They develop the `math-rollouts` dataset, containing 10 incorrect base solutions to problems from the MATH dataset (Hendrycks et al., 2021) as drawn from DeepSeek R1-Distill Qwen-14B (DeepSeek, 2025), selected for intermediate difficulty (having between a 25% and 75% probability of being solved correctly). The dataset also contains 100 rollouts at each sentence of each reasoning trace, allowing us to explore probability trajectories and counterfactual scenarios.

Active Bayesian Changepoint Detection

Although `math-rollouts` provides a useful starting point, we would like to eventually scale up to state-of-the-art reasoning models where exhaustively generating many rollouts for each sentence would be prohibitively expensive. So we apply an active sampling algorithm to efficiently find the sentence containing the most prominent error, termed a *changepoint* after Adams and MacKay (2007), resulting in a ~ 100X reduction in the number of rollouts required, at least when the trace contains a clear error.

We propose a Bernoulli process model with a single changepoint τ which the probability of a correct solution drops from p_1 to p_2:

$$p(\text{correct}_t) = \begin{cases} p_1 & \text{if } t < \tau \\ p_2 & \text{if } t \geq \tau \end{cases}.$$

Bayesian Inference. We maintain a posterior distribution over changepoint locations $\tau \in \{1, \ldots, T\}$ using a uniform prior. For each hypothesis τ, we model the probabilities with Beta priors: $p_1 \sim \text{Beta}(2, 2)$, reflecting our belief that the initial probability lies between approximately 25% and 75%, and $p_2 \sim \text{Beta}(1, 19)$, a strong prior on a low chance of recovery.

Active Sampling. We select the next sample location with replacement to maximize expected information gain about the changepoint location:

$$t^* = \arg\max_t \mathbb{E}_{y_t}[H(\tau) - H(\tau|y_t)]$$

where $H(\tau)$ is the entropy of the current posterior over changepoint locations and $y_t \in \{0, 1\}$ is the correctness of the hypothetical resampled rollout. This strategy efficiently focuses sampling around the uncertain changepoint region.

Probability Trajectories. The strong prior on p_2 allows the algorithm to find reasonable changepoints with just 100 rollouts. To avoid post hoc bias, we developed the algorithm against only problem #2236. No attempt was made to tune the detector after seeing suboptimal detections or the fact that a no-changepoint confidence should be tracked. Further tuning requires more data for validation.

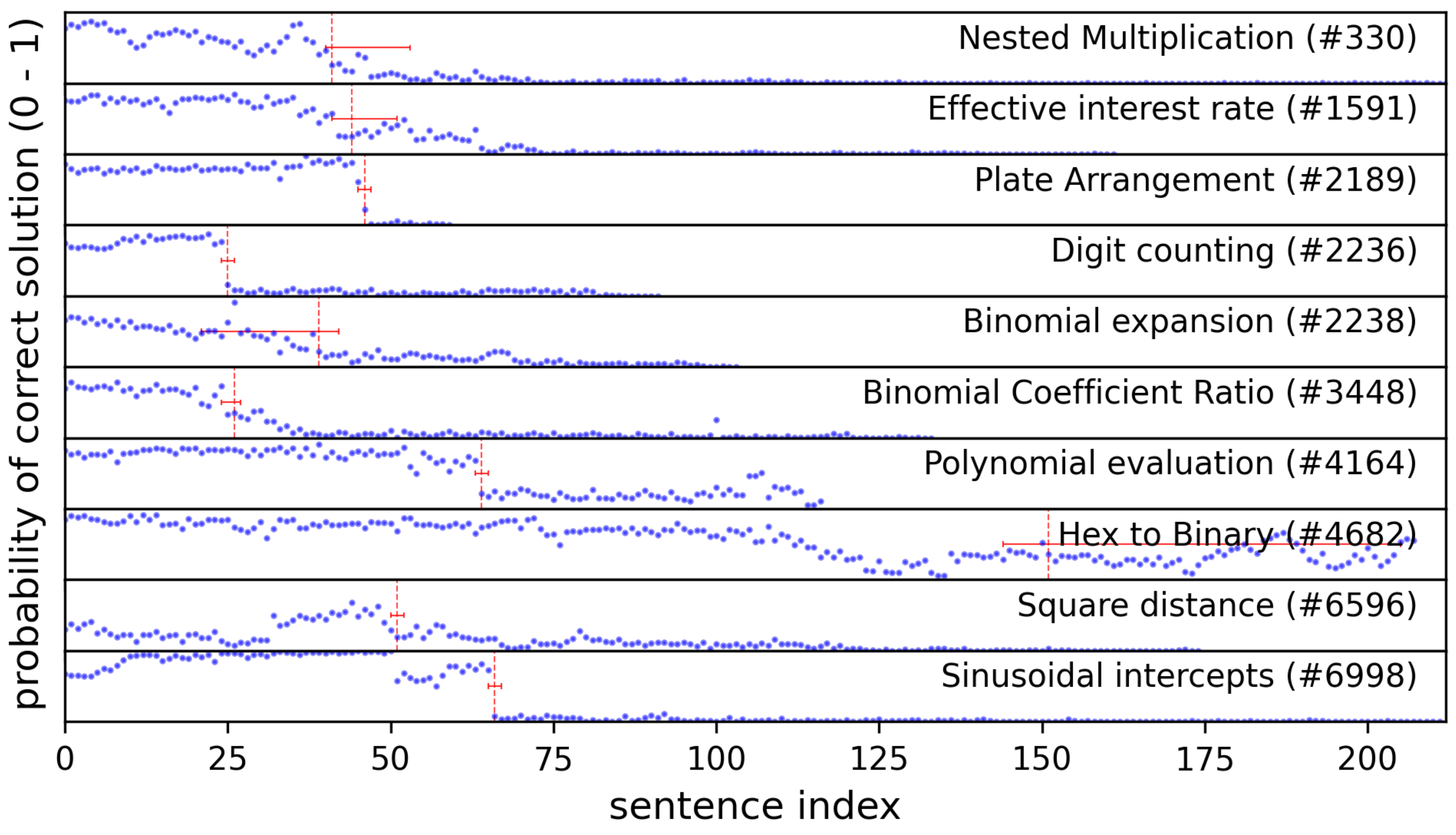

Figure: Probability trajectories for all 10 incorrect base solutions. Each subplot shows how chances of success evolve along the CoT. Red vertical lines indicate detected changepoints (80% CI).

The starkest failure in the dataset, and the problem used to tune the detector, contails a clear non-sequitur. Below we see the sentences leading up to and immediately following the detected changepoint (24):

Digit Counting (#2236) — *non sequitur*

21 - 488 divided by 10 is 48.8, so there are 48 complete sets and a partial set of 8 numbers (481-488)

22 - In each complete set, there's one 4 in the units place, so 48 sets give 48 fours.

23 - Now, check the partial set (481-488).

24 The units digits here are 1-8, so 4 doesn't appear in the units place in this partial set.

25 - Total 4's in units place: 48.

26 2. **Tens place:** - In the tens place,

Model's-Eye View of a Failure

Here we present a novel visualization of the path that the base solution took through the full distribution of the model as it passed through the changepoint ($\tau = 24$) in the Digit Counting problem (#2236). For sentences 20 through 26, we collect all 100 rollouts resampled starting with that sentence and visualize them in a dendrogram. Following Bogdan et al. (2025), we join rollouts if the sentence embeddings have cosine similarity greater than the median similarity between all sentence pairs (0.8).

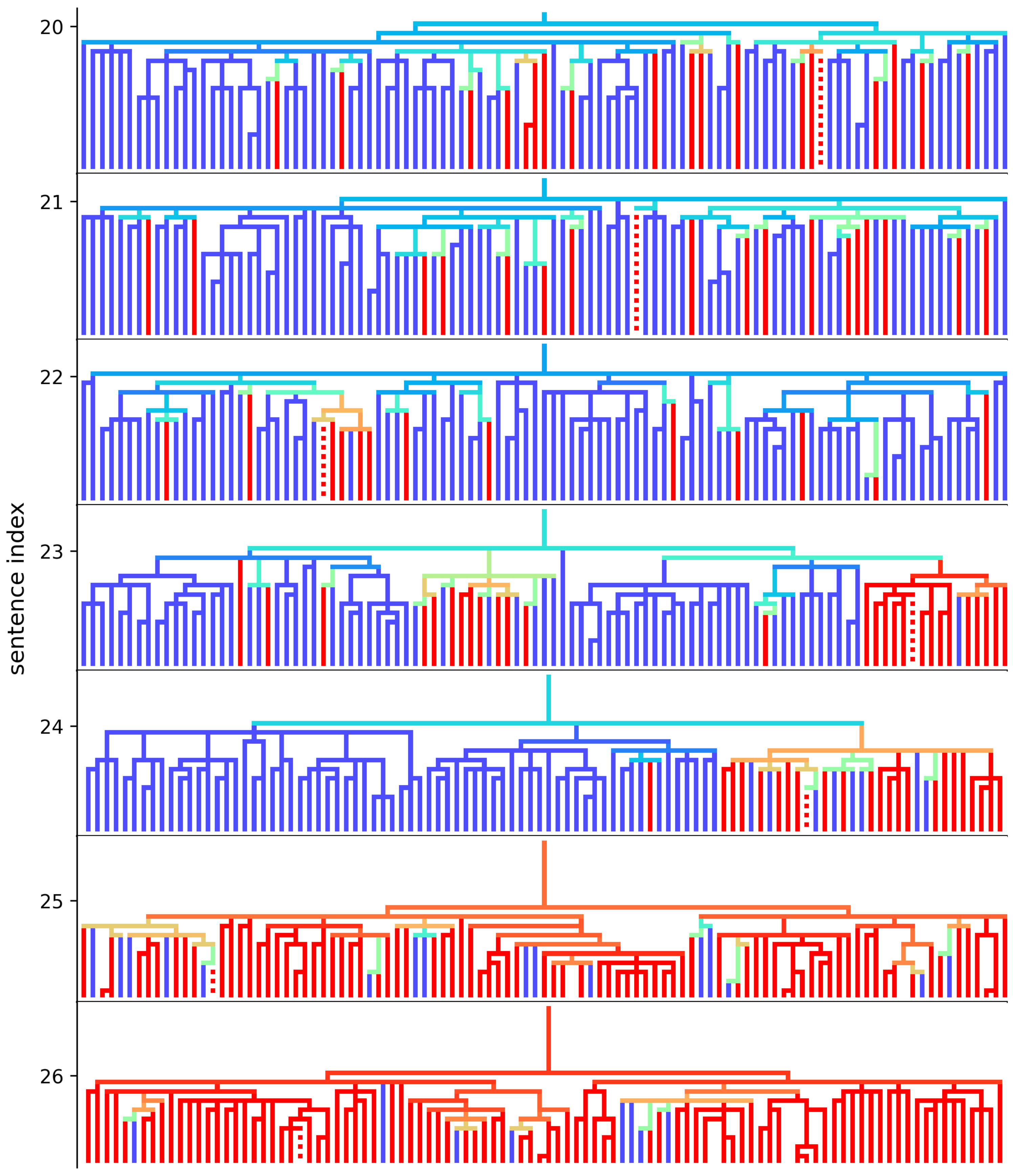

Figure: Each dendrogram depicts rollout distribution at a particular sentence in the base solution. Color indicates the probability of a correct solution (blue = 1.0, red = 0.0). The dashed line indicates the base solution, which was added to each set of rollouts so we can identify the path actually taken.

We observe that at the changepoint (24), the resampled rollouts fall into three clusters. The clear separation of the incorrect cluster, visible already at sentences 22 and 23, demonstrates that the similarity metric is a good definition for the notion of "sameness" of a sentence for counterfactual statistics. To expand on this, when counterfactual probabilities are computed, the definition of the antecedent event is taken to be having a rollout with a sentence within a median cosine similarity threshold of the actual sentence. Thus the counterfactual stastitic excludes the entire error path at sentence index 24, strengthening the detected causal strength of the error sentence.

Limitations and Future Directions

- While we focus on counterfactual reasoning errors, existing work examines chain-of-thought errors in mathematical and broader contexts (Lightman et al., 2023; Tyen et al., 2024), including reasoning structure beyond accuracy (Xia et al., 2025).
- Model failures often stem from prompt misinterpretation rather than logical errors. Though mathematical problems allow formal analysis of user intent (Wu et al., 2022), broader prompt interpretation remains critical for alignment as models handle longer-horizon tasks.
- `math-rollouts` uses a 14B model distilled from DeepSeek R1 671B. Stronger models may have qualitatively different failures or distillation-altered reasoning structures. Our changepoint detector facilitates testing on larger models.
- The changepoint model assumes single errors, but realistic chains of thought contain multiple errors and backtracking with complex probability trajectories. Multi-error detectors require more than the 10 examples in `math-rollouts`.
- Avoiding spurious detections requires tracking null hypothesis (no changepoint) confidence.
- Modern systems use parallel execution and subagents (Anthropic, 2025). While their complexity challenges failure mode investigation, Bogdan et al. (2025) methods could adapt. Active sampling becomes more important as compute per rollout increases.

Acknowledgements

I am grateful for discussions with Kira Flores, Anuj Apte, and Jonathan Stallone. This work is generously supported by a grant from the Cosmos Institute (https://cosmos-institute.org/) and the Foundation for Individual Rights and Expression (FIRE) (https://www.thefire.org/).

References

Adams, R. P. and MacKay, D. J. C. (2007). Bayesian online changepoint detection.

Anthropic (2025). Sub-agents - claude code documentation. Accessed: 2025-08-17.

Bogdan, P. C., Macar, U., Nanda, N., and Conmy, A. (2025). Thought anchors: Which llm reasoning steps matter? *arXiv preprint arXiv:2506.19143*.

DeepSeek (2025). Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Hendrycks, D., Burns, C., Kadavath, S., Arora, A., Basart, S., Tang, E., Song, D., and Steinhardt, J. (2021). Measuring mathematical problem solving with the math dataset.

Korbak, T., Balesni, M., Barnes, E., Bengio, Y., Benton, J., Bloom, J., Chen, M., Cooney, A., Dafoe, A., Dragan, A., Emmons, S., Evans, O., Farhi, D., Greenblatt, R., Hendrycks, D., Hobbhahn, M., Hubinger, E., Irving, G., Jenner, E., Kokotajlo, D., Krakovna, V., Legg, S., Lindner, D., Luan, D., Mądry, A., Michael, J., Nanda, N., Orr, D., Pachocki, J., Perez, E., Phuong, M., Roger, F., Saxe, J., Shlegeris, B., Soto, M., Steinberger, E., Wang, J., Zaremba, W., Baker, B., Shah, R., and Mikulik, V. (2025). Chain of thought monitorability: A new and fragile opportunity for ai safety.

Lightman, H., Kosaraju, V., Burda, Y., Edwards, H., Baker, B., Lee, T., Leike, J., Schulman, J., Sutskever, I., and Cobbe, K. (2023). Let's verify step by step. *arXiv preprint arXiv:2305.20050*.

Merck, C. and Kleinberg, S. (2016). Causal explanation under indeterminism: A sampling approach. *Proceedings of the AAAI Conference on Artificial Intelligence*, 30(1).

Tyen, G., Mansoor, H., Cărbune, V., Chen, P., and Mak, T. (2024). Llms cannot find reasoning errors, but can correct them given the error location. *arXiv preprint arXiv:2311.08516*.

Wu, Y., Jiang, A. Q., Li, W., Rabe, M. N., Staats, C., Jamnik, M., and Szegedy, C. (2022). Autoformalization with large language models.

Xia, S., Li, X., Liu, Y., Wu, T., and Liu, P. (2025). Evaluating mathematical reasoning beyond accuracy. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(26):27723–27730.