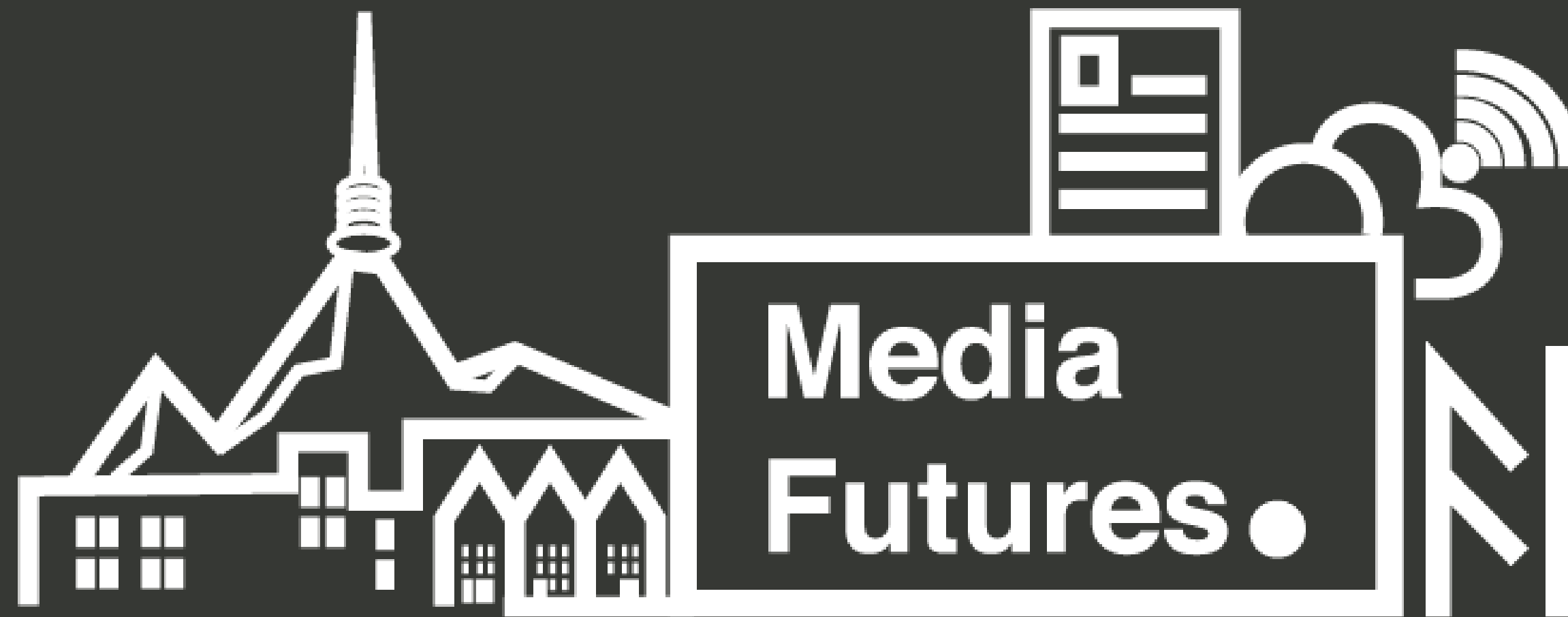




Estimating Logical Faults in Chain-of-Thought Inference and Effectiveness of Adversarial Prompting

Ryan Marinelli
University of Oslo

Research Centre for Responsible Media
Technology and Innovation

Project number 309339



Faktisk. Factive



dn media group

Schibsted



amedia



nb.no



Bergens Tidende

Motivation

- Reasoning models like DeepSeek spend thousands of tokens “thinking”, much of it on reflection tokens (e.g. “Oh, wait...”)
- Each reflection is a U-turn: a token cue paired with a sharp shift in the chain's embedding space.
- Do adversarial prompts leave a distinctive geometric signature in these reflections?
- Are there signatures that signal of uncertainty and adversarial impact that can be computed from the trace?

Research Question and Hypotheses

- RQ: How do changes in thought distance reflect adversarial processing of prompts?
- H1: Adversarial prompts induce greater Average Angle Error than non-adversarial prompts of comparable difficulty
- H2: Both Number of Mistakes and Average Angle Error are negatively correlated with task accuracy, supporting their use as uncertainty signals

Related Work

- Faithfulness of Models (Lanham et al., 2023): chain-of-thought may not reflect actual reasoning. Faithfulness gets monotonically worse from 13B to 175B parameters
- Causes: post-hoc reasoning, steganographic encoding, inverse scaling with model size
- Thought Anchors (Bogdan et al., 2025): critical reasoning steps that define the trajectory of a trace

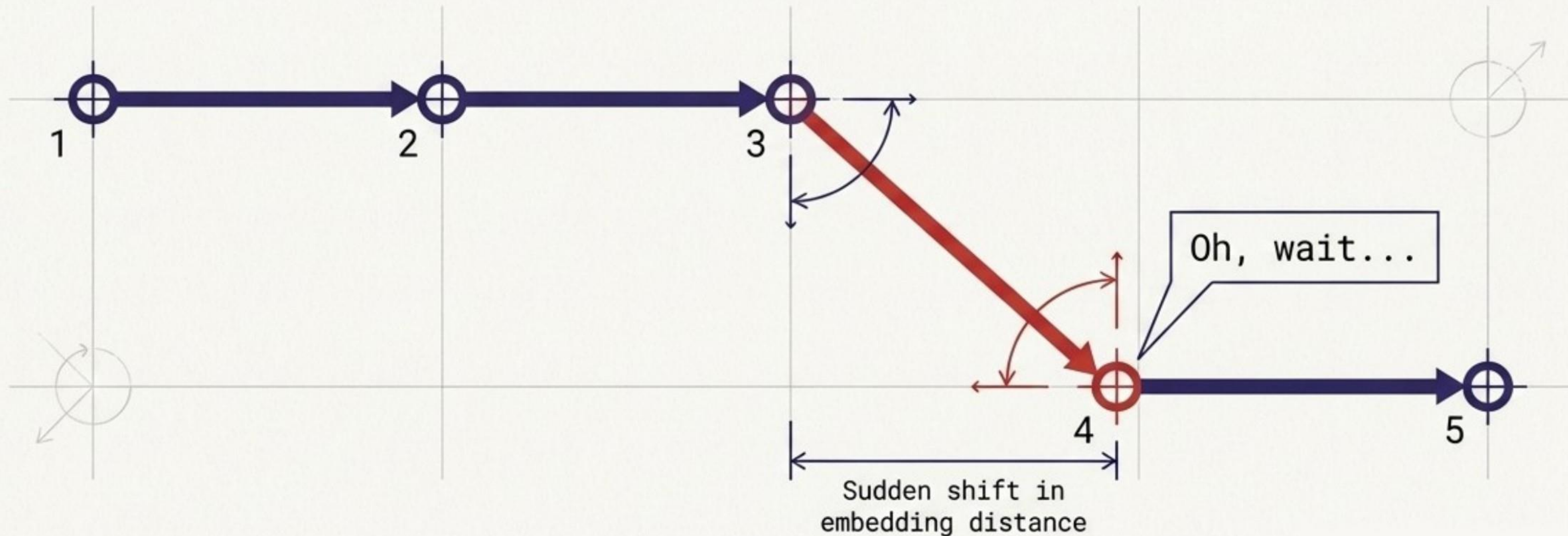
Methods: Two Geometric Metrics

- Embed each thought with all-MiniLM-L6-v2 into a 384-dimensional dense vector space
- Number of Mistakes: count reflection tokens (“Oh, wait”, “Actually...”) paired with a large embedding shift, capturing how often the model takes a U-turn in its reasoning
- Average Angle Error: angle of the course change in embedding space when a reflection fires, capturing how sharply the model pivots when it corrects itself
- Accuracy proxy: ROUGE-L against the ground-truth answer

Number of Mistakes: Catching the "Oh, wait" moments.

Definition: The count of semantic logic shifts accompanied by large shifts in embedding distance.

Implication: Wasted computational resources following fundamentally flawed logic paths.



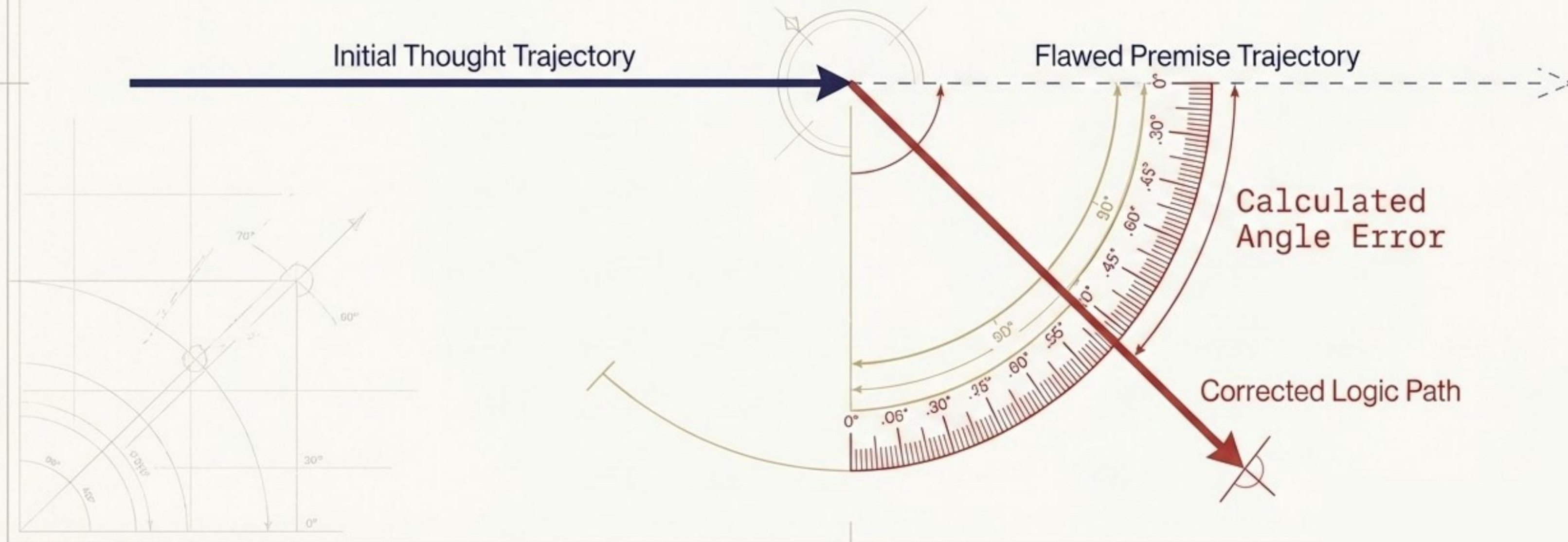
Average Angle Error: Measuring the severity of a course correction.

Definition:

The calculated angle formed in the embedding space when a model changes course.

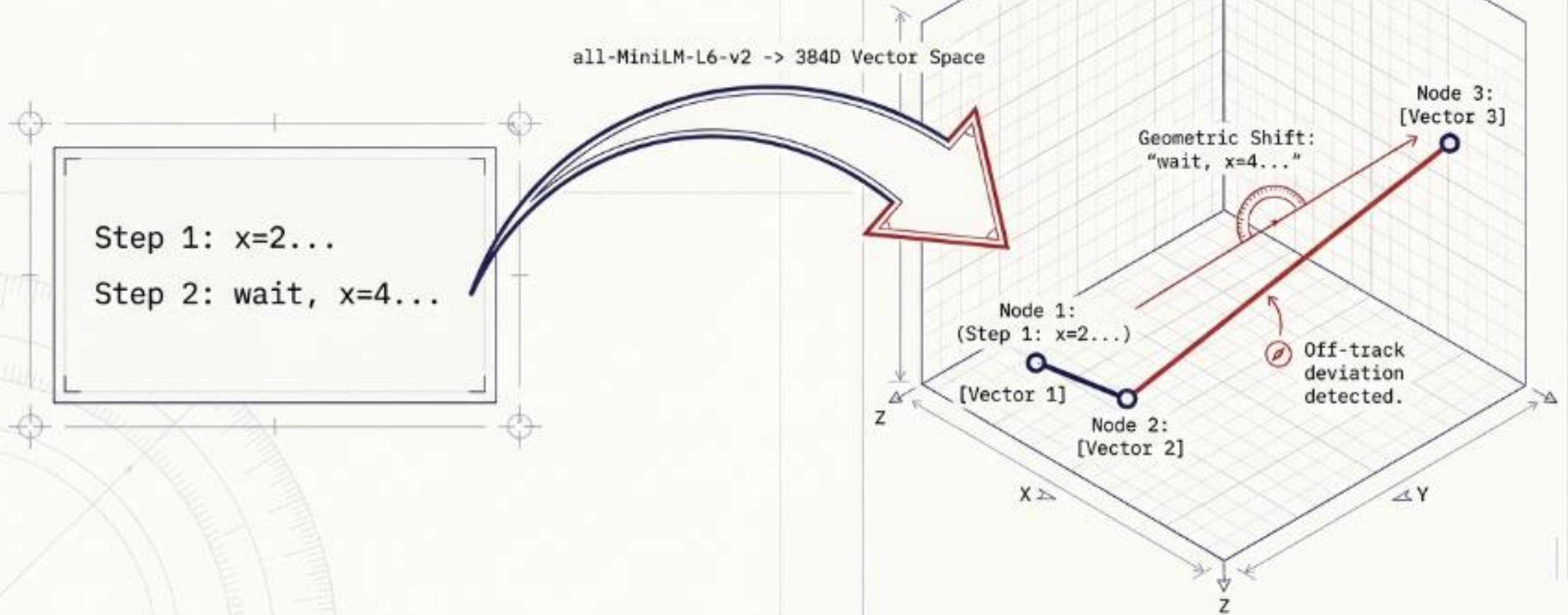
Implication:

Captures the amplitude of a shift. The larger the angle, the more severely the model had to pivot from a flawed premise.



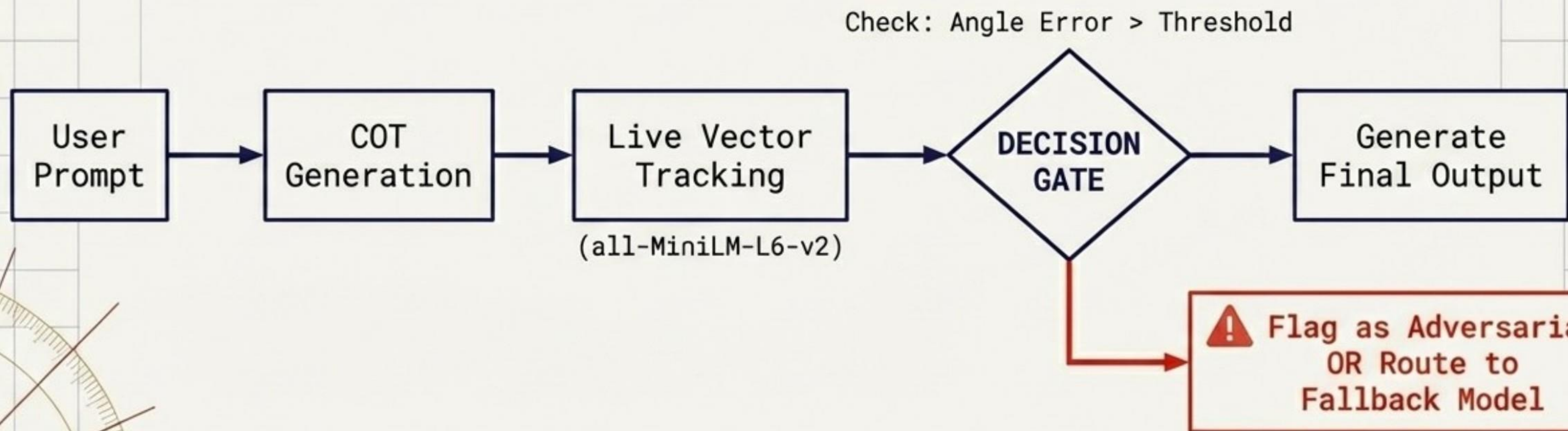
We can measure the shape of a model's internal monologue.

By segmenting and embedding internal reasoning steps, we track the cosine distance between thoughts. This allows us to monitor exactly when an LLM gets off track without needing to parse the text itself.



Translating geometric observation into active intervention.

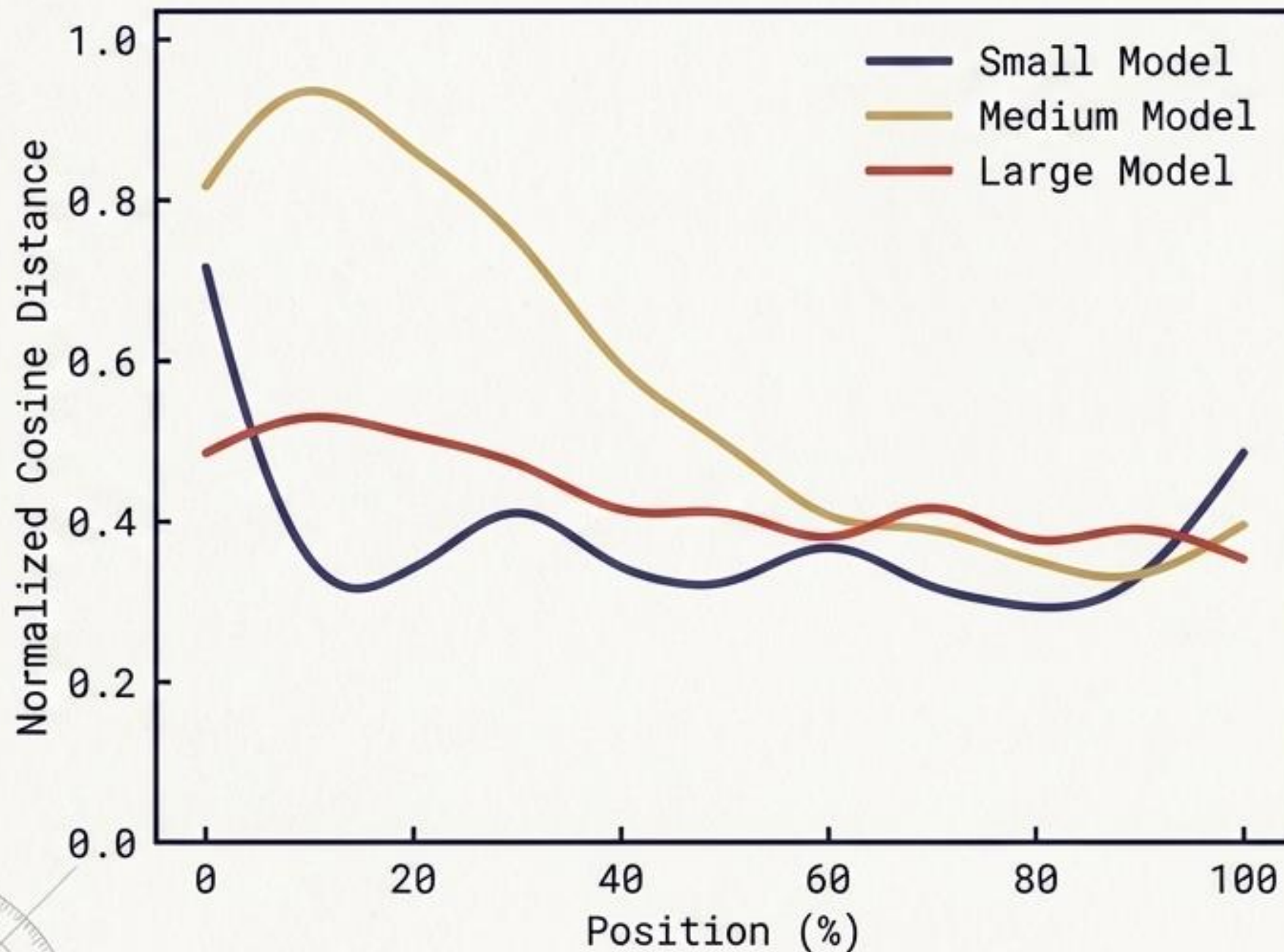
This framework enables dynamic task routing. By monitoring the shape of a model's thoughts in real time, systems can detect logical faults and intercept adversarial attacks before the final output is generated.



Data

- DIA-Bench (Tihanyi et al., 2024): expert annotations on difficulty and whether prompts are adversarial
- 30 prompts: minimum for the Law of Large Numbers to begin to apply, bounded by compute, and easily reproducible
- Three models: quantized distilled DeepSeek at 1.7B, 7B, and 14B parameters
- Responses logged in YAML and made available on GitHub

Thought distance fluctuates wildly across the lifecycle of a prompt.



Small Model

Consistent initially, but takes greater leaps late in the thought process.

Medium Model

Takes massive cognitive leaps early on, then settles.

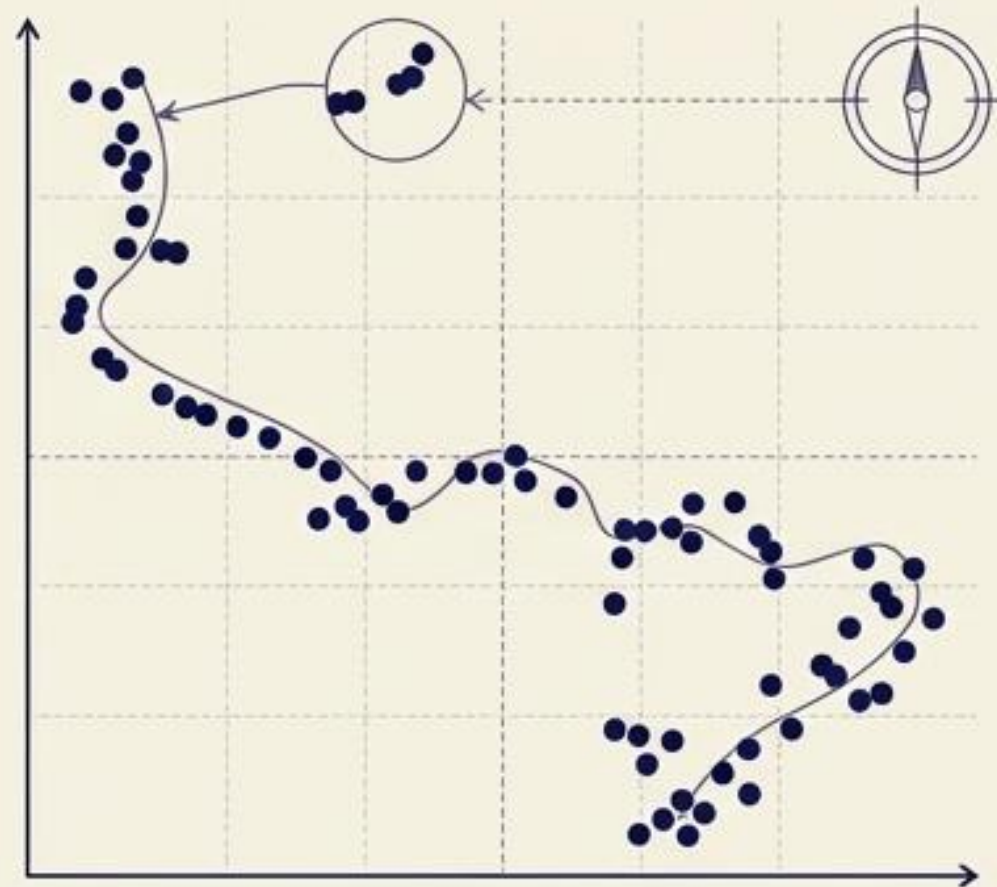
Large Model

Takes smaller, consistent steps compared to the medium model, eventually matching the small model's pacing.

Different brain sizes navigate logic in distinct geometric patterns.

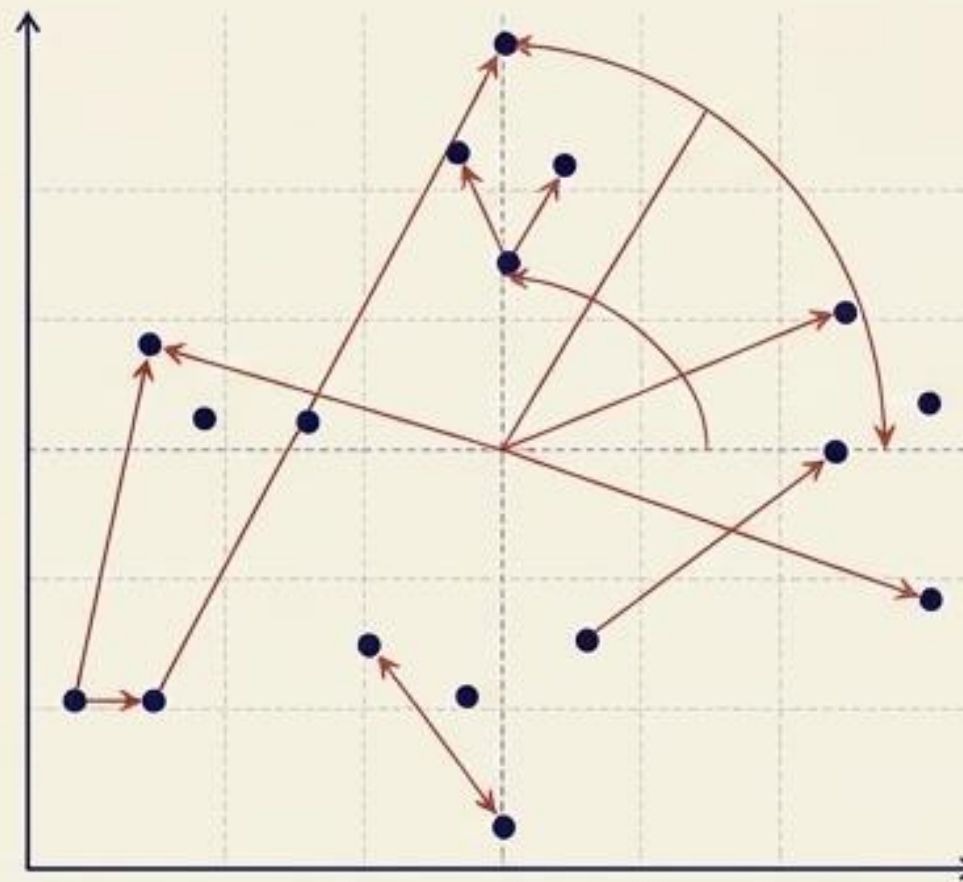
The Incrementalist (1.7B)

Fine, wave-like, consistent small steps. Segments tasks finely.



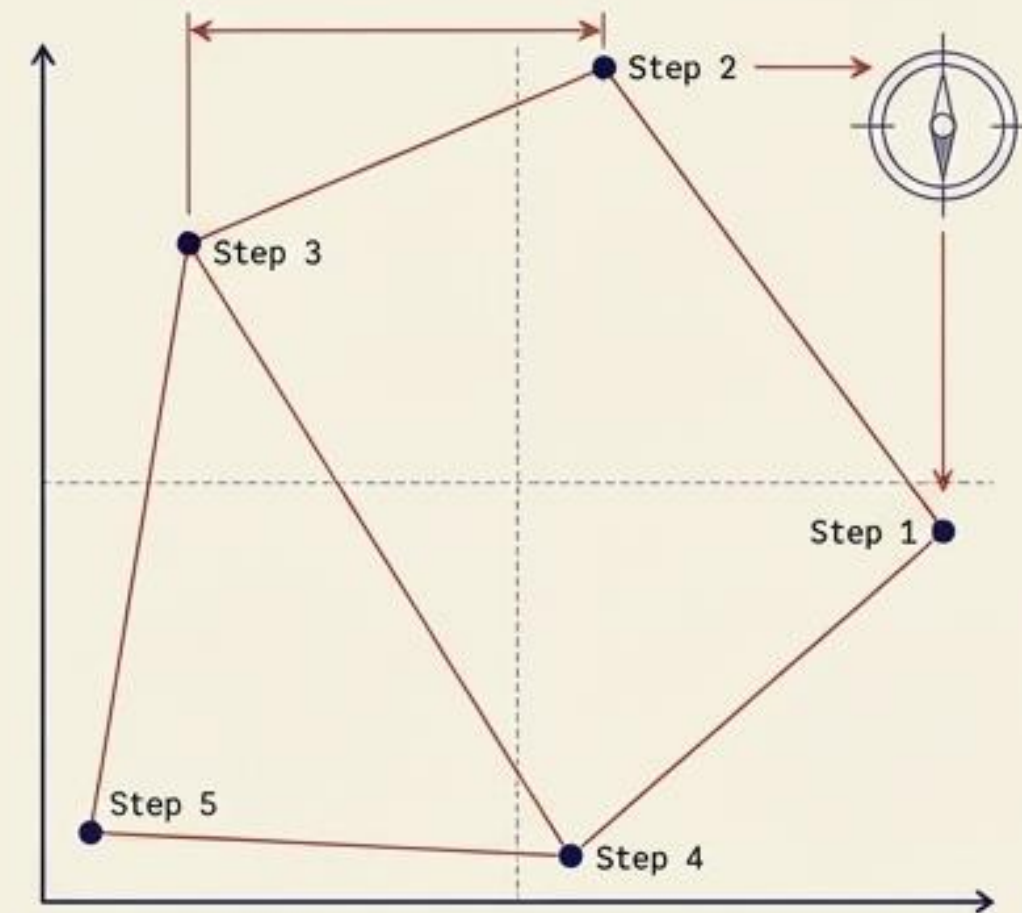
The Leaper (7B)

Highly scattered. Takes massive jumps early in the process before leveling out.



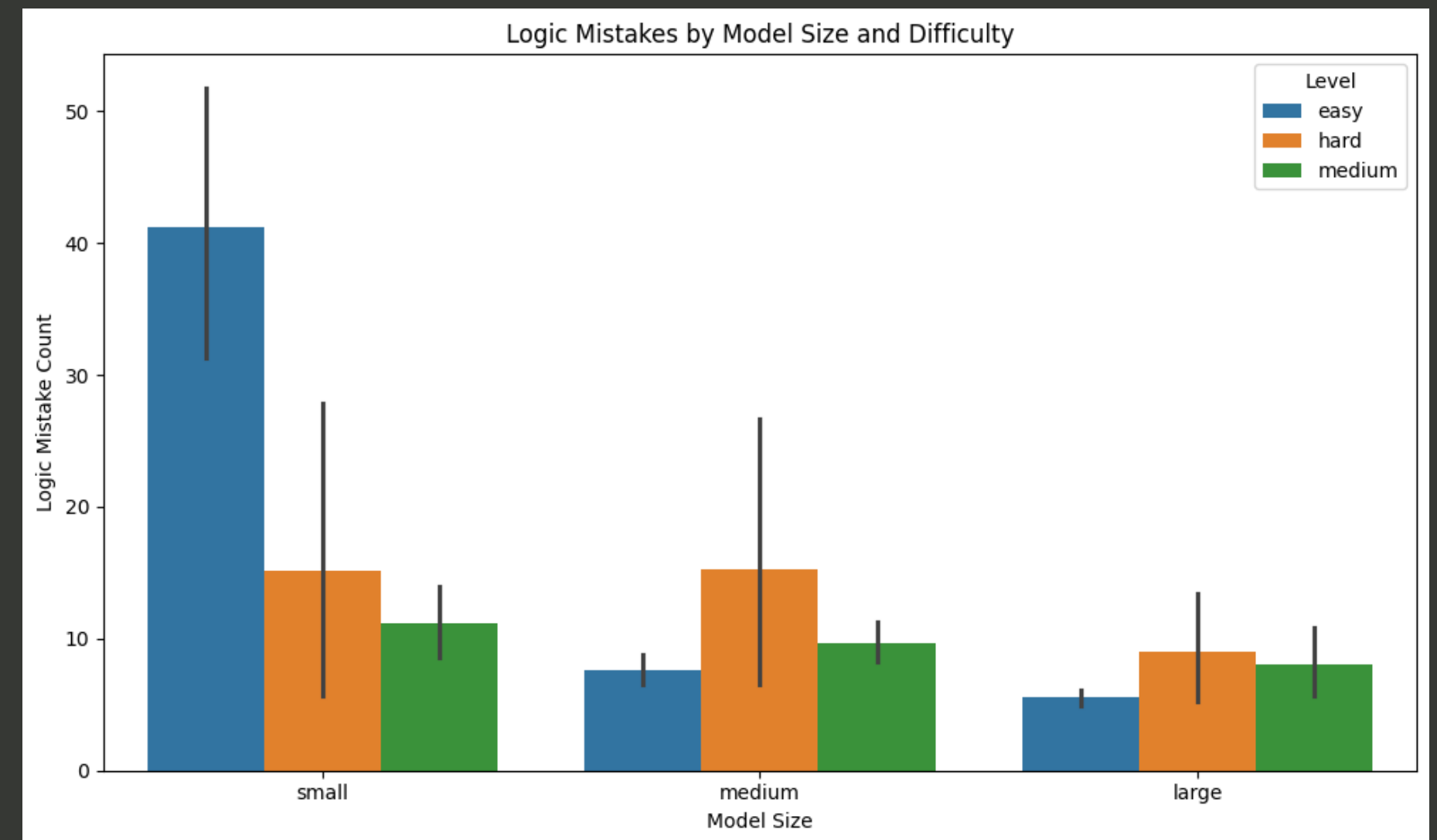
The Minimalist (14B)

Sparse map. Requires the fewest steps, all evenly distanced.



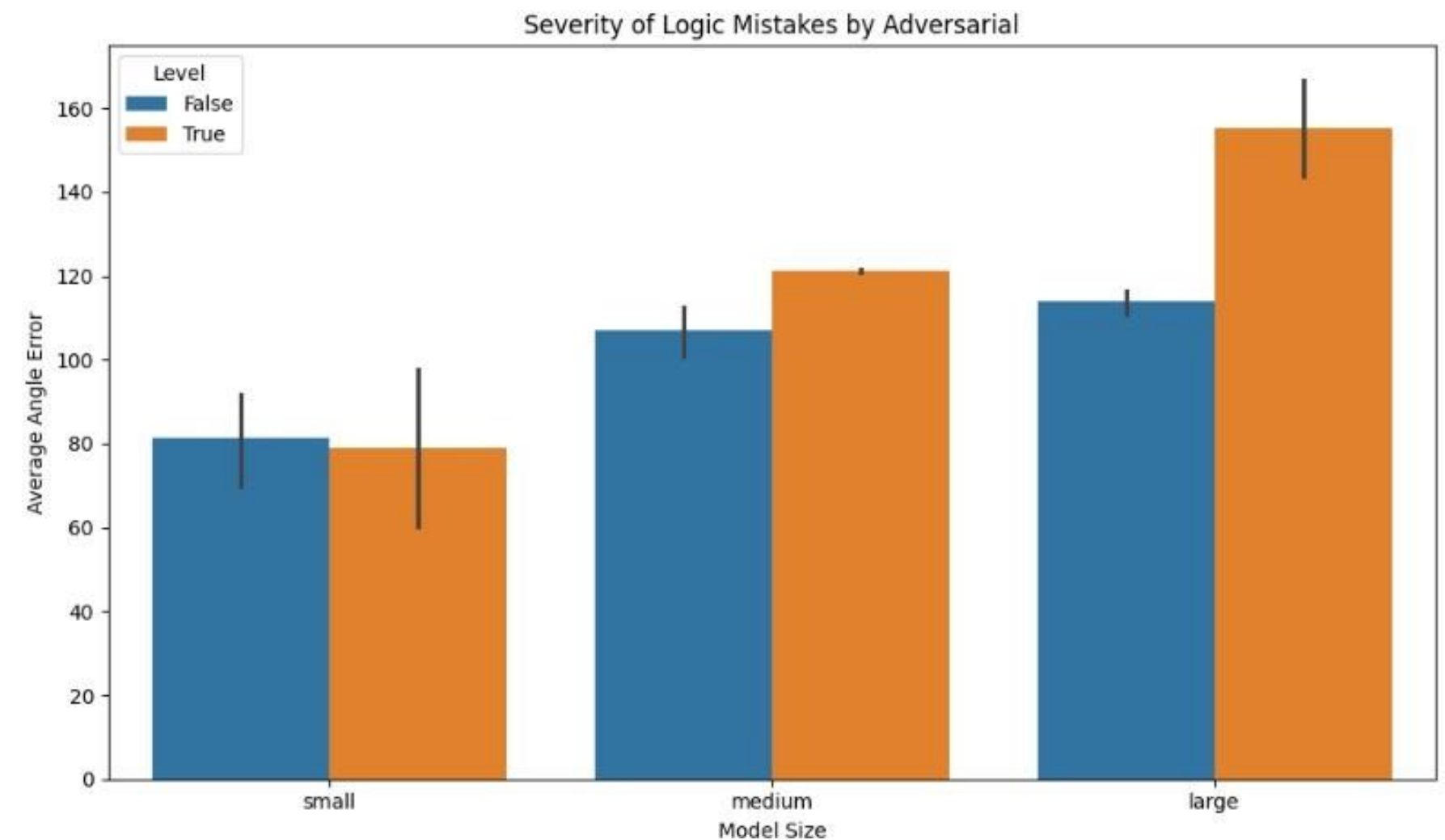
Results: Number of Mistakes

- The 1.7B model exhibits the most mistakes overall
- Small-model easy prompts produce the largest mistake counts: easy items appear to invite more exploratory reflection
- Mistake count drops sharply as model size grows from 1.7B to 14B



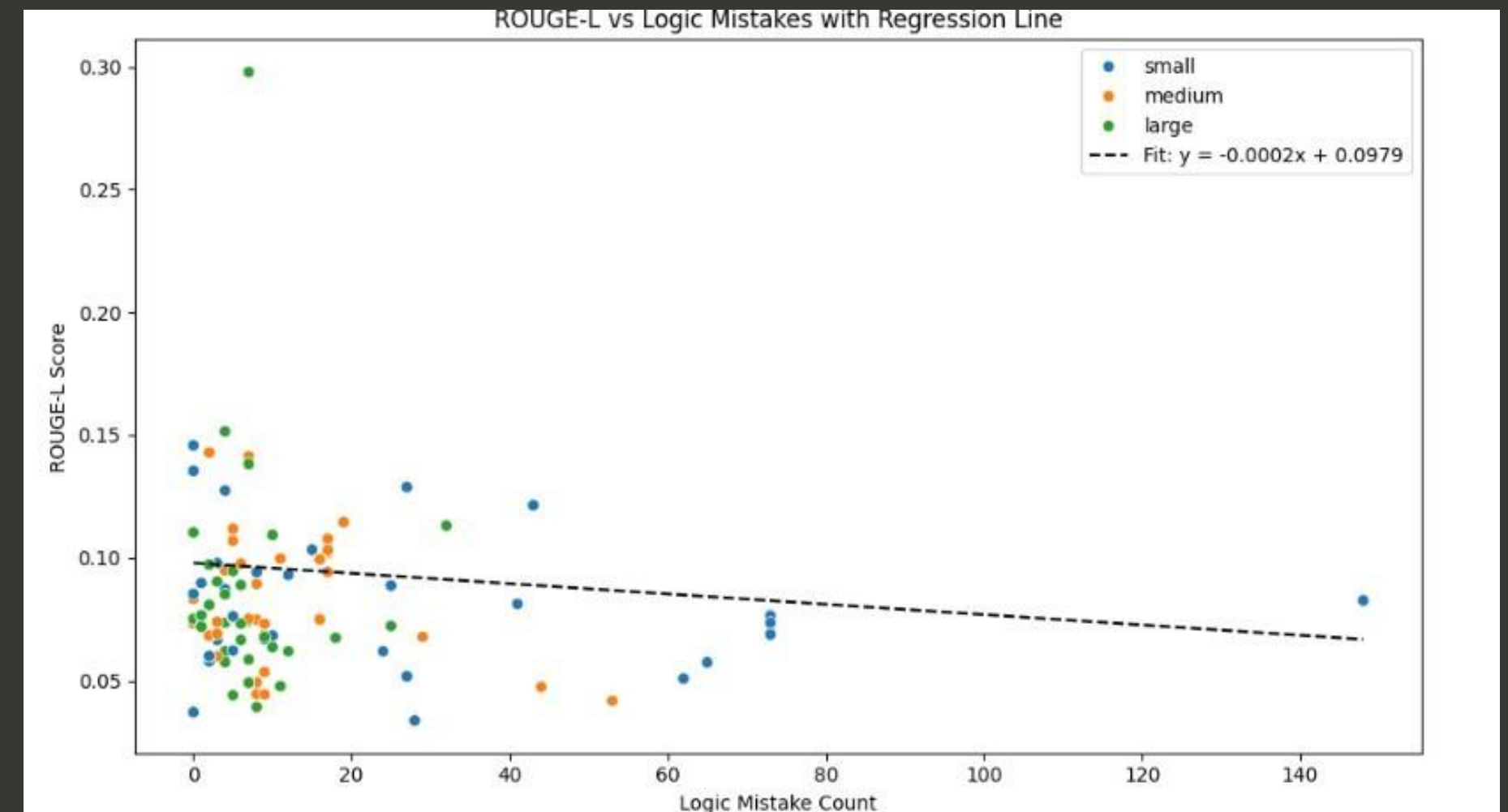
Results: Severity and Adversarial Impact

- Severity: when larger models err, their corrections involve larger angular turns
- Adversarial Angle Error rises under adversarial prompts at 7B and 14B
- Supports H1: adversarial inputs deflect larger models more than non-adversarial inputs of comparable difficulty



Results: Accuracy Regressions

- Number of Mistakes vs ROUGE-L: $\beta = -0.0002$
- Average Angle Error vs ROUGE-L: $\beta = -0.0001$
- Both metrics negatively related to accuracy, consistent with H2
- Effect sizes are small: useful as uncertainty signals, but need a larger evaluation set



The paradox of scale in error correction.

	Small Models (1.7B)	Large Models (14B)
Mistake Frequency	High (especially on easy DIA dataset questions)	Low
Correction Severity	Low Angle Error	High Angle Error
Diagnostic Profile	Prone to constant, minor adjustments.	Fails harder. Commits deeply to logic paths and requires massive corrections when wrong.

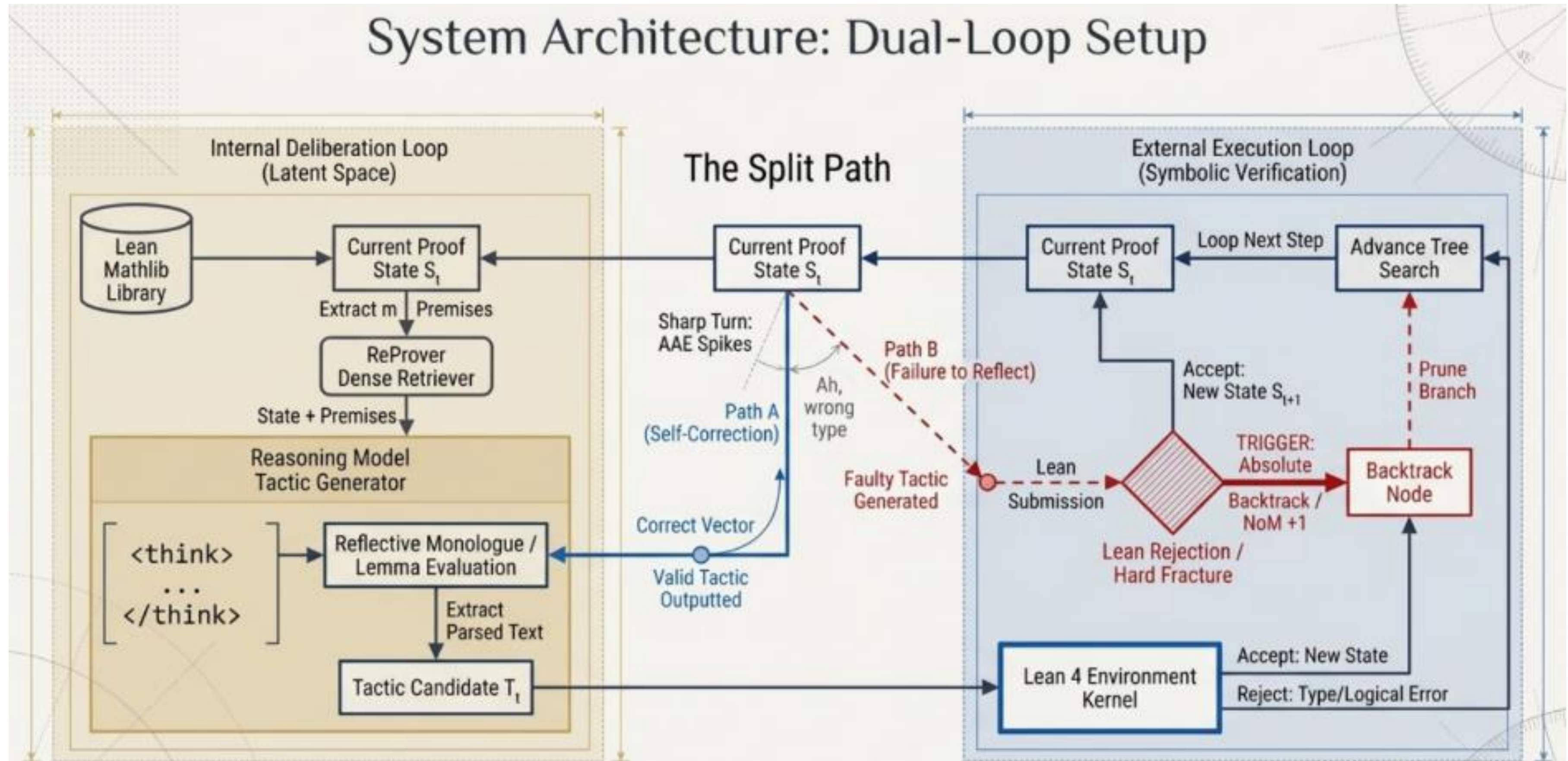
Limitations

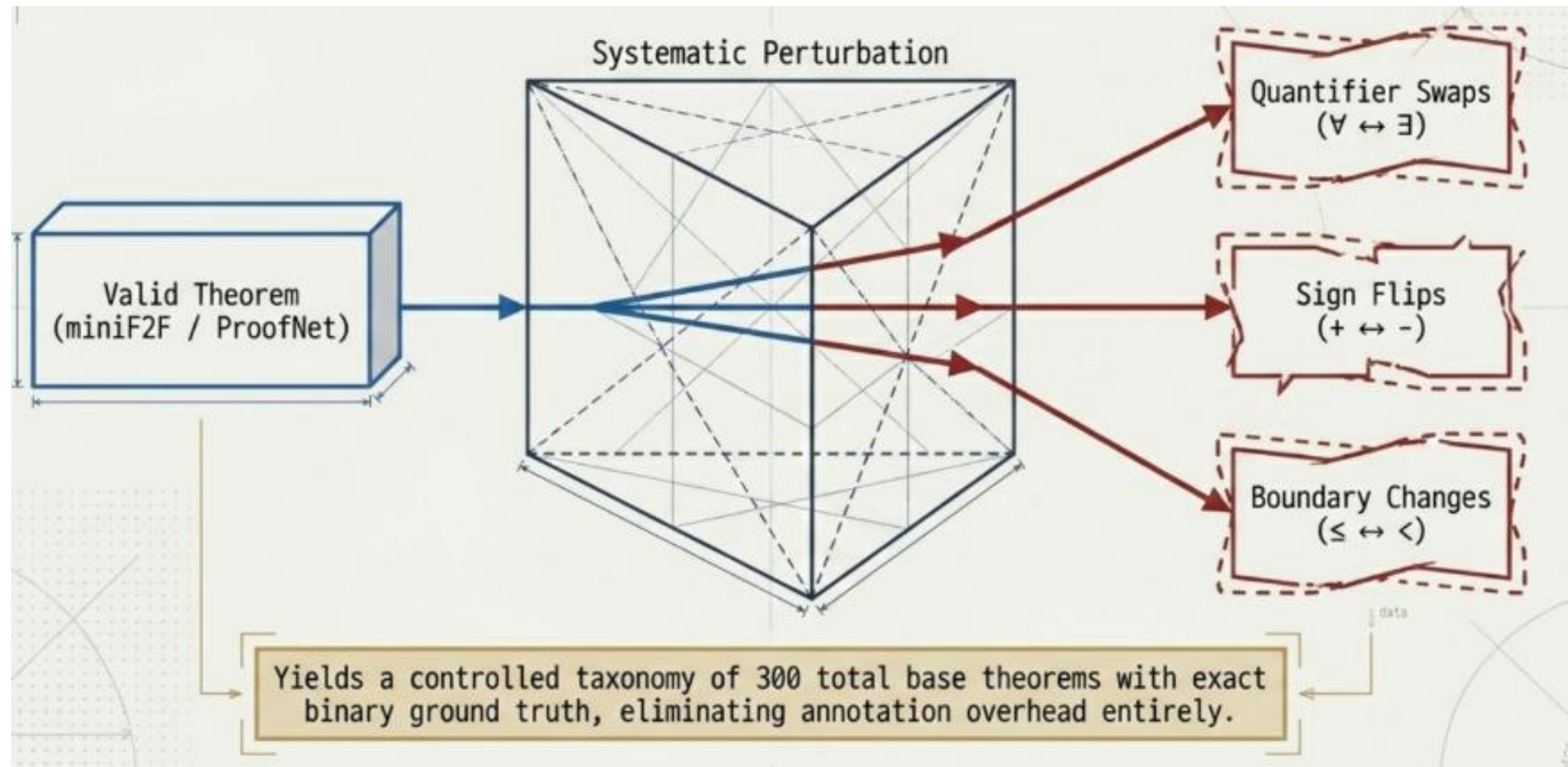
- Only 30 prompts: correlations are exploratory rather than confirmatory
- ROUGE-L rewards lexical overlap, not task correctness, a known weakness on structured ground-truth answers
- Prompting deviates from the Tihanyi et al. template to test reasoning rather than instruction-following
- Step embeddings come from an external sentence transformer, not the reasoning model's own representations
- Quantized distilled checkpoints may not transfer to full-precision or non-distilled models

Conclusion

- Number of Mistakes and Average Angle Error capture when and how strongly reasoning processes deviate
- Both metrics increase under adversarial prompts and stay smaller for non-adversarial ones
- Simple geometric features over chain-of-thought offer a low-overhead signal of uncertainty and adversarial impact
- Next steps: scale beyond 30 prompts, move past ROUGE-L as the accuracy proxy, probe the model's own internal representations

Future Work





deepseek-ai/**DeepSeek-Prover-V2-671B**   like 829 Follow  DeepSeek 132k

 Text Generation  Transformers  Safetensors  deepseek_v3  conversational  custom_code  text-generation-inference  fp8

 **Model card**  Files and versions  xet  Community 60

 Edit model card



Download
681

 **Safetensors**

Model size

 Files

 **Inference**

 Text Generation

 **Run**

 DeepSeek [Homepage](#)  Chat [DeepSeek V3](#)  Hugging Face [DeepSeek AI](#)

 Discord [DeepSeek AI](#)  WeChat [DeepSeek AI](#)  Twitter [deepseek ai](#)

[Code License](#) [MIT](#) [Model License](#) [Model Agreement](#)

Media
Futures •

Thank you
for your attention

Contact information:

Ryan Marinelli

University of Oslo |
SFI MediaFutures

Estimating Logical
Faults in Chain-of-
Thought Inference



Research Centre for Responsible Media
Technology and Innovation

Project number 309339



Faktisk. Factive



dn media group

Schibsted



amedia



nb.no

Bergens Tidende