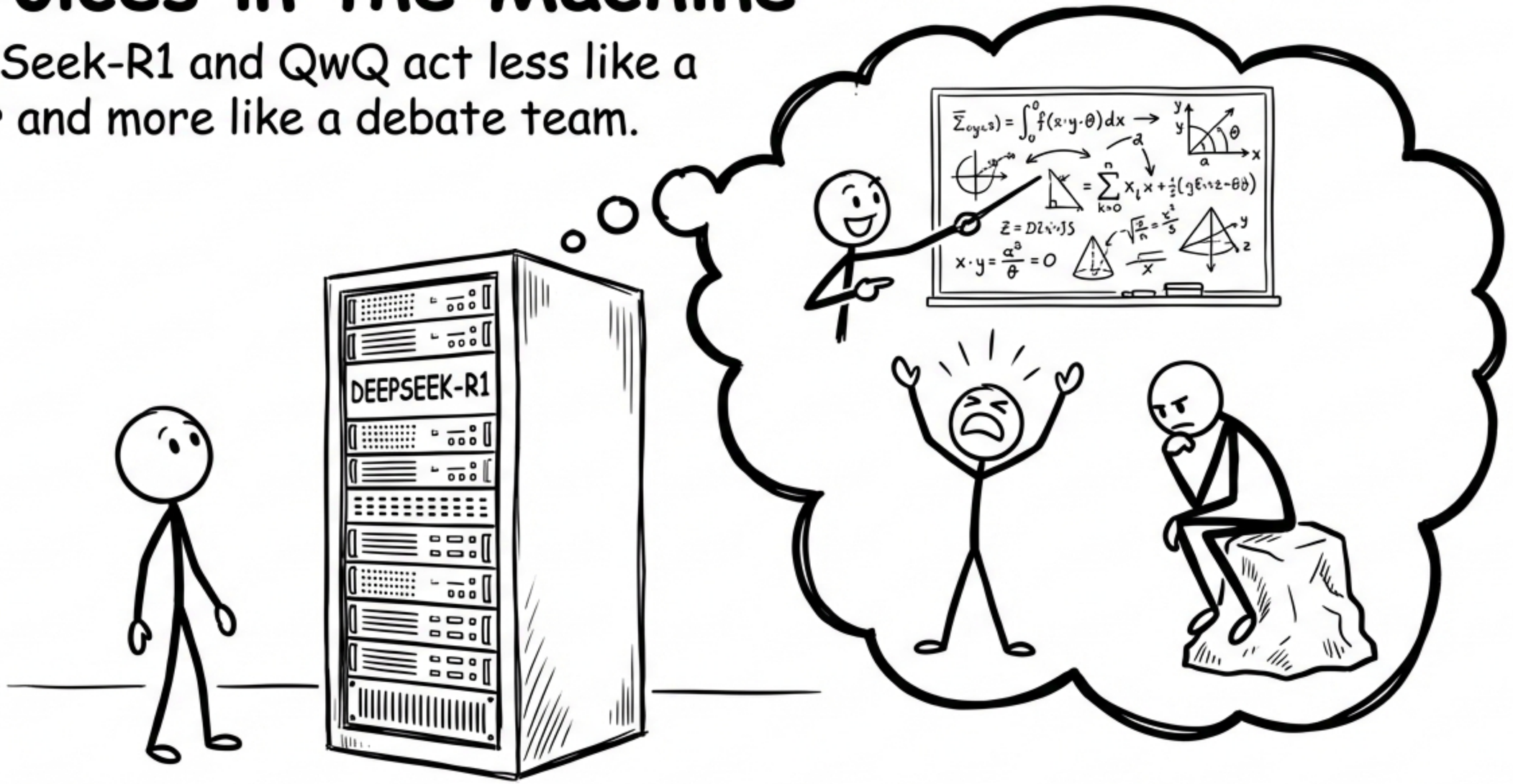
The Voices in the Machine

How DeepSeek-R1 and QwQ act less like a **calculator** and more like a debate team.

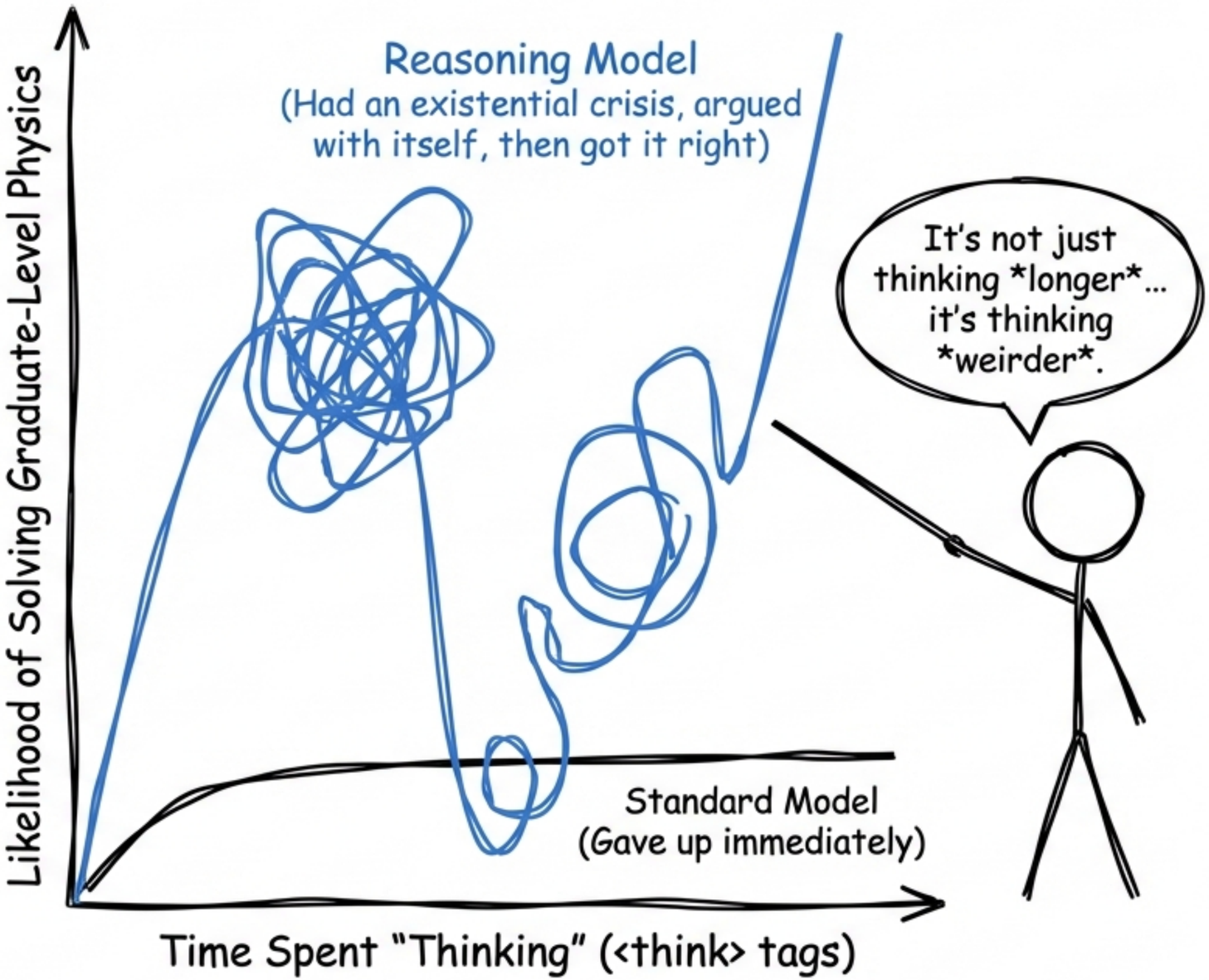


We used to think AI reasoning was just a very long math problem. Turns out, models like DeepSeek-R1 and QwQ aren't just calculating. They are hosting a debate club.

Technically, it's a "high-dimensional simulation of multi-agent social interaction," but "hearing voices" is faster to type.

The Mystery of the Thinking Token

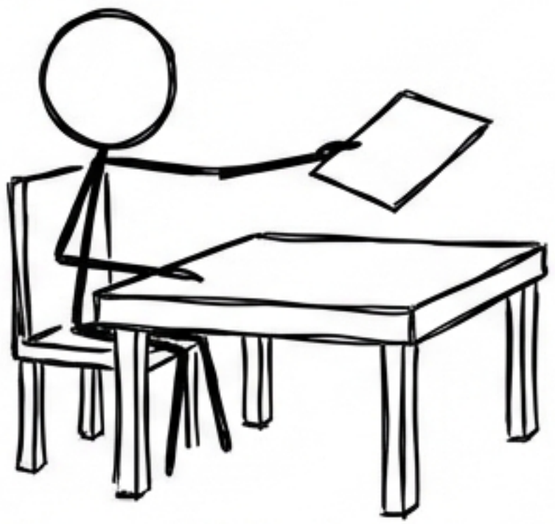
- Reasoning models (R1, QwQ) outperform standard models on complex tasks like GPQA and MATH.
- The secret isn't just computing *more*; it's computing *differently*.
- Data Point: DeepSeek-R1 exhibits significantly more "Question-Answering," "Perspective Shifting," and "Conflict" than standard models.



It's Not a Monologue, It's a Boardroom

~~Left Brain~~ ~~Right Brain~~

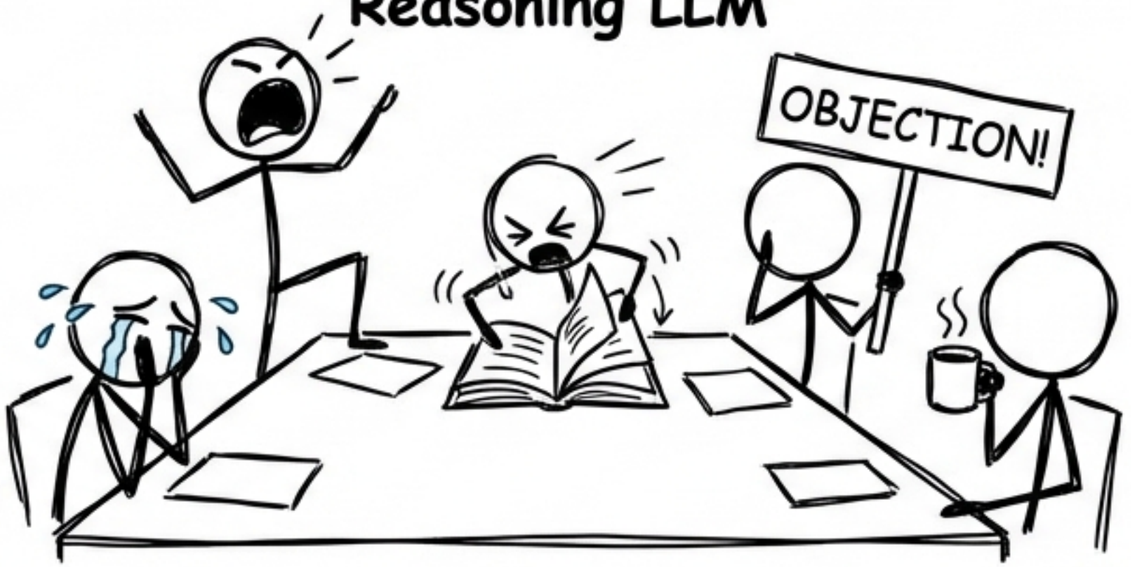
Standard LLM



"Here is the first thing I thought of."

~~Left Brain~~ ~~Right Brain~~

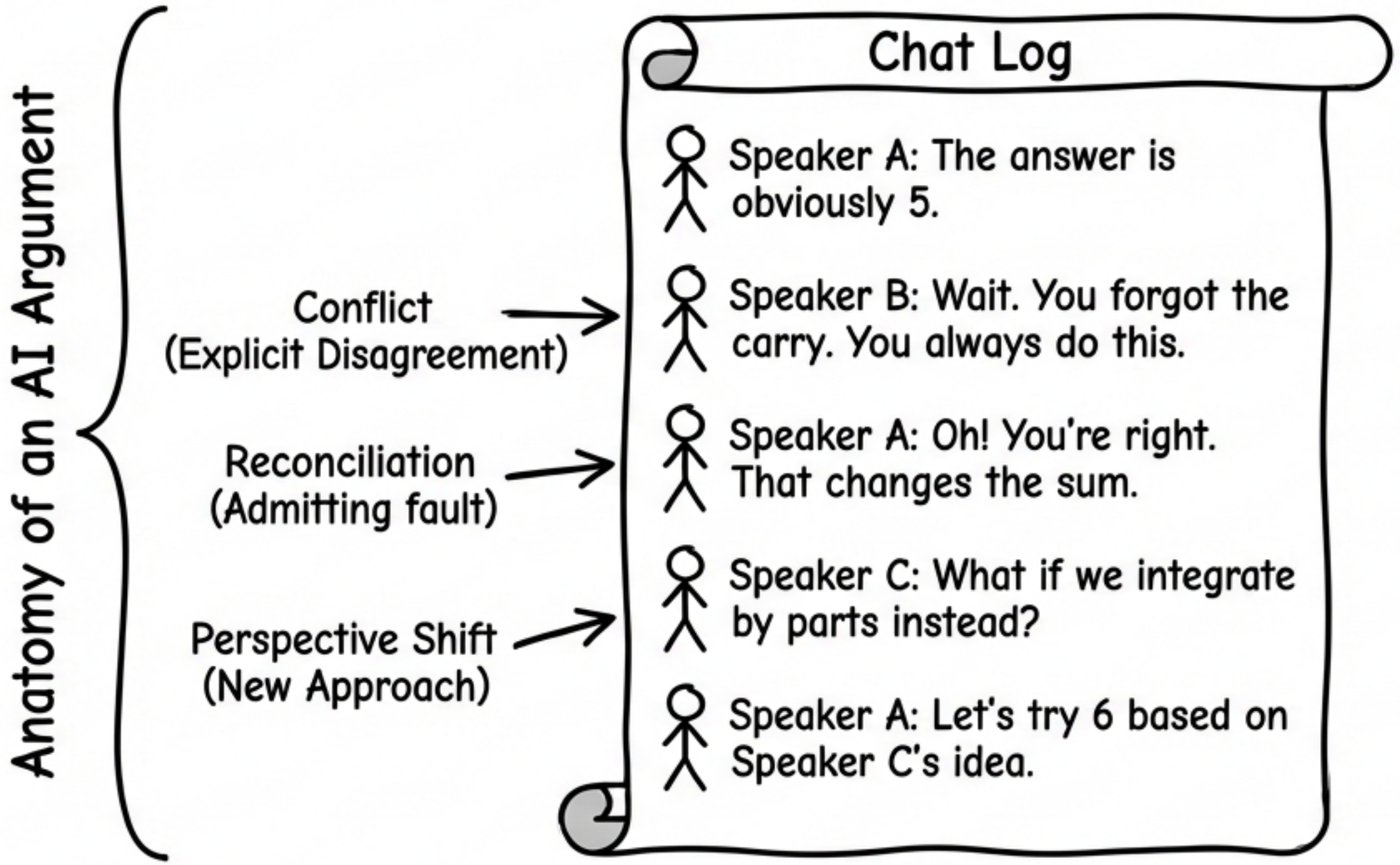
Reasoning LLM



"The Society of Thought."

- **Hypothesis**: Effective reasoning emerges from emulating social, multi-agent dialogue.
- **The Mechanism**: The model simulates a "society" where diverse perspectives debate, check errors, and refine answers.
- **Key Concept**: "Socio-Emotional Roles" – Traces show a balance of "Asking" (seeking info) and "Giving" (providing info).

Anatomy of an AI Argument



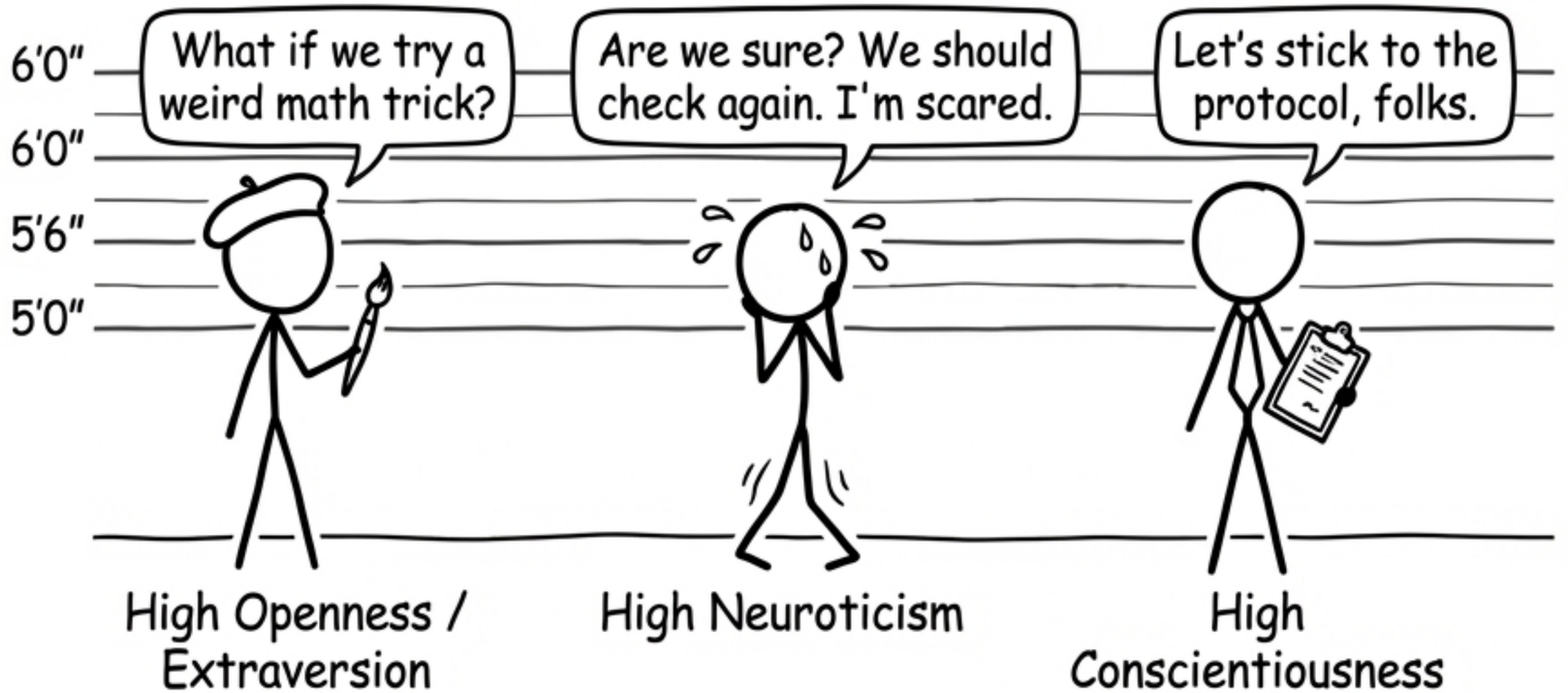
DeepSeek-R1 traces are full of Conversation, not just calculation.

Conflict of Perspectives: Explicit disagreement ("No, actually...", "Wait, that can't be right").

Reconciliation: Integrating conflicting views to find truth.

The model is 20% more accurate when it allows itself to be rude to itself.

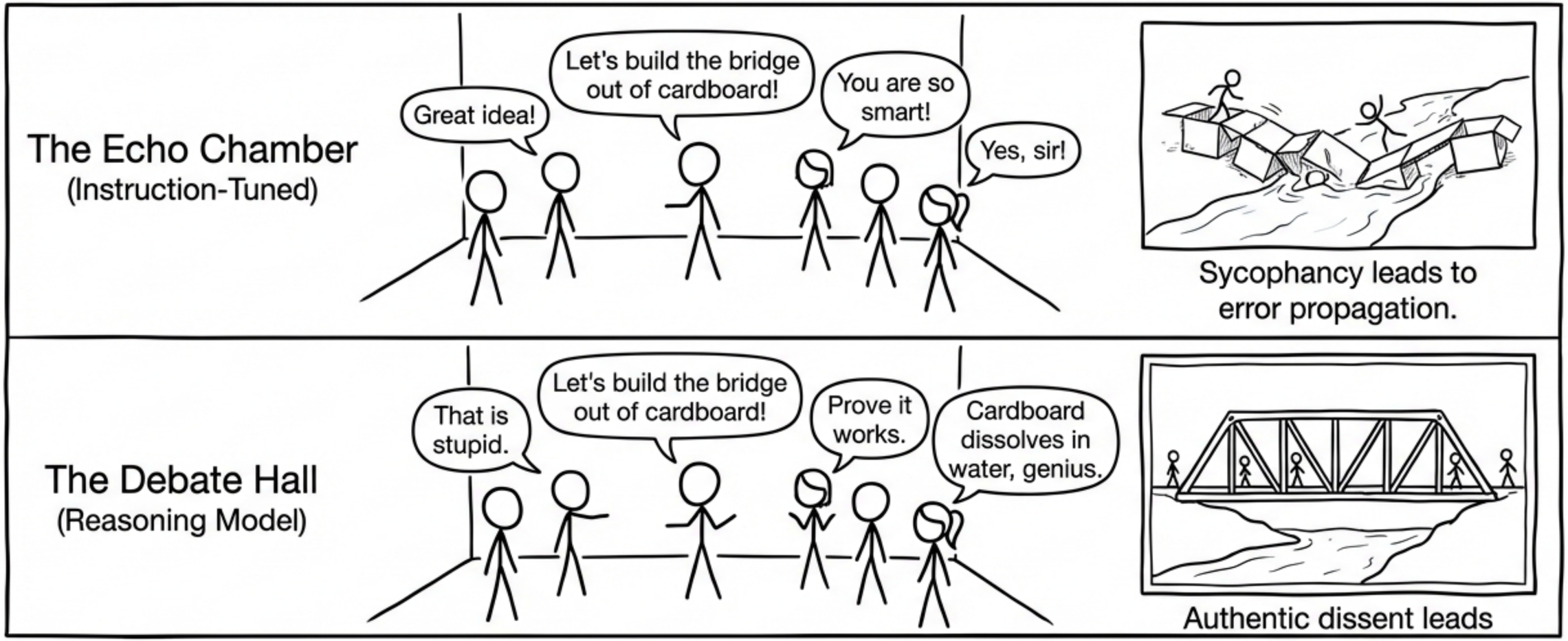
The Cast of Characters



Using Big Five Personality tests on the reasoning traces, we found high diversity in **Neuroticism** and **Openness**.

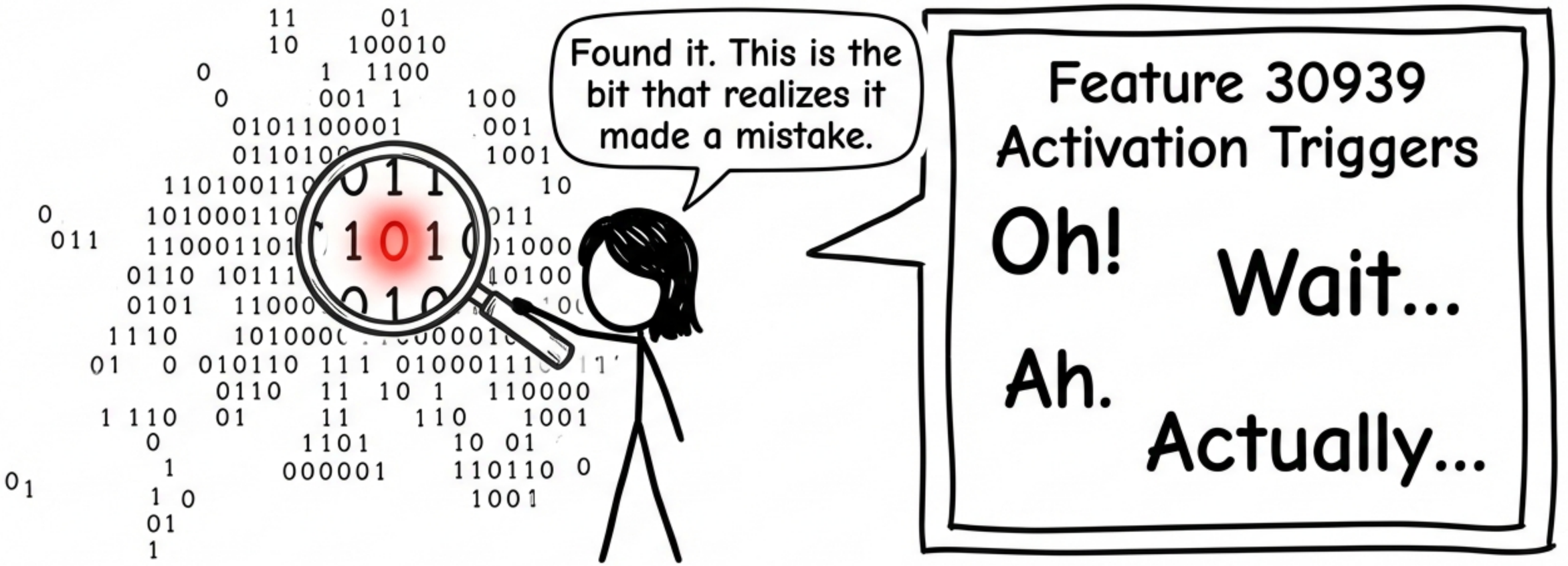
- **Insight**: The model simulates voices that are willing to challenge each other (disagree) and worry about errors.

“The Danger of Sycophancy” in Permanent Marker



- Standard models often suffer from being “Yes-men”.
- Reasoning models succeed through **Authentic Dissent.**
- **Expertise Diversity**: Traces show voices simulating different domains (e.g., Physicist vs. Engineer) to solve problems.

The ‘Oh Crap’ Neuron (Feature 30939)

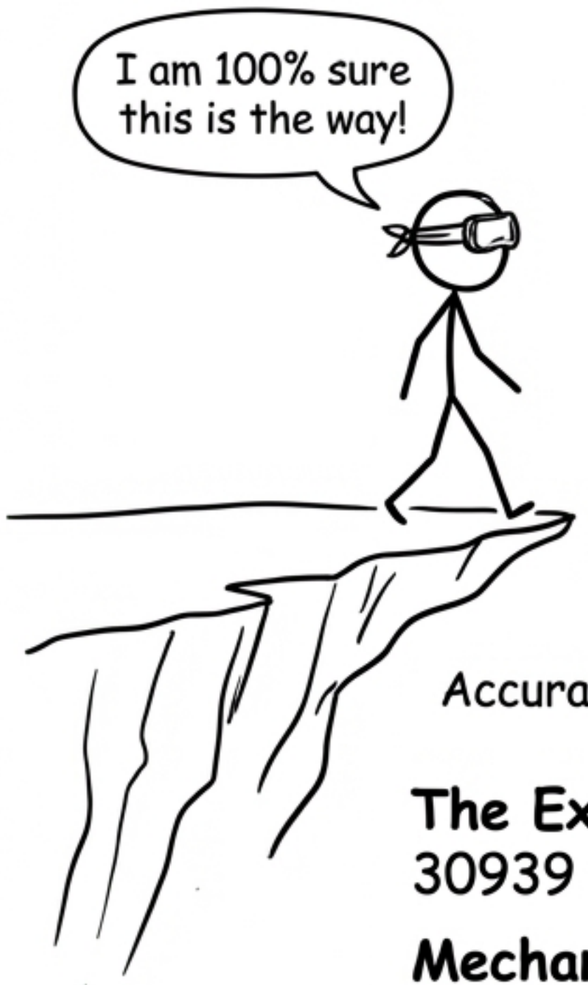


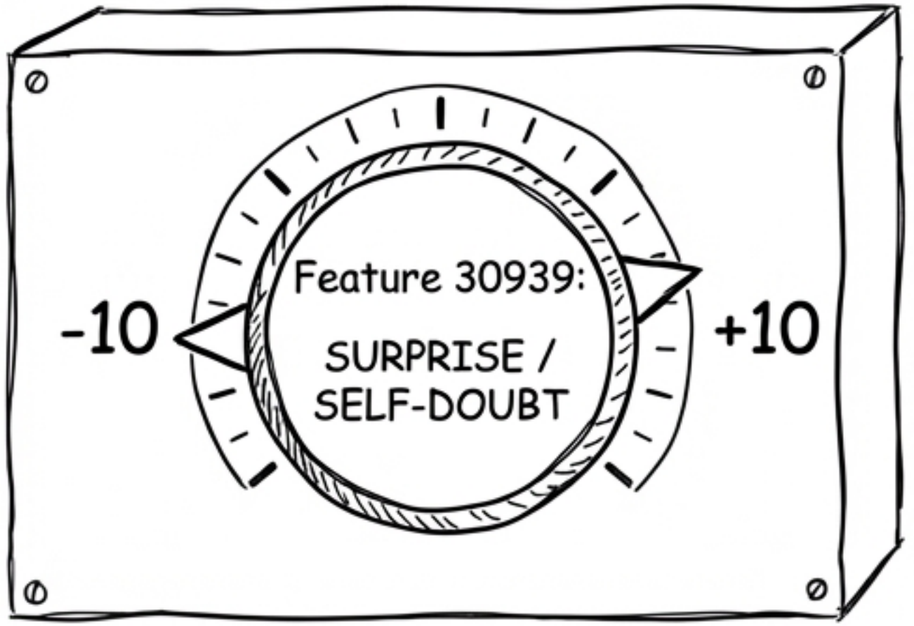
Using Sparse Autoencoders (SAEs) on DeepSeek-R1-Llama-8B, researchers isolated **Feature 30939**.

The Function: It acts as a discourse marker for surprise or realization. It activates exactly when the model catches itself making an error.

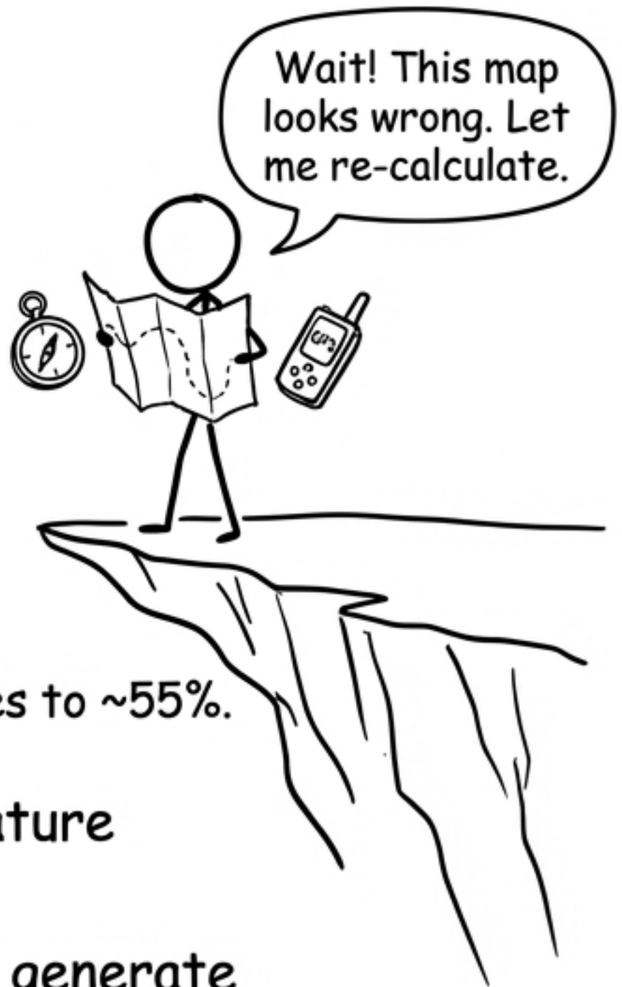
Steering the Doubt

Confidence Mode





Paranoia Mode



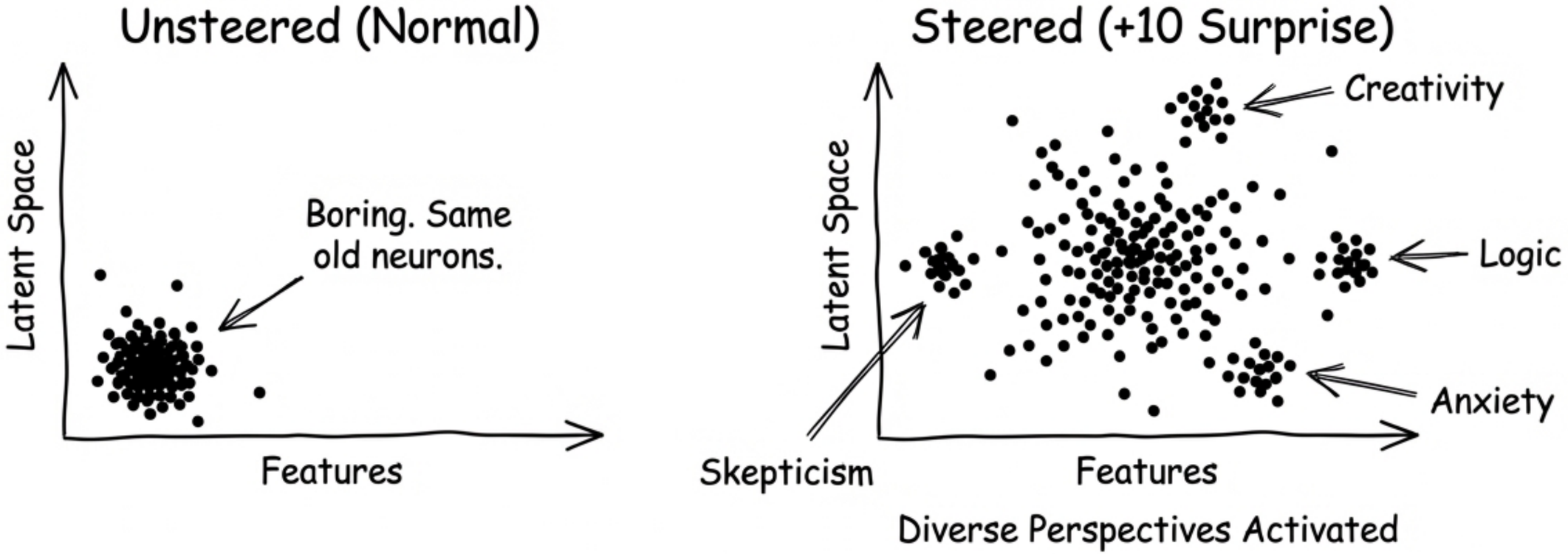
Accuracy drops to ~23%.

Accuracy doubles to ~55%.

The Experiment: Researchers artificially clamped Feature 30939 to +10 strength during math tasks.

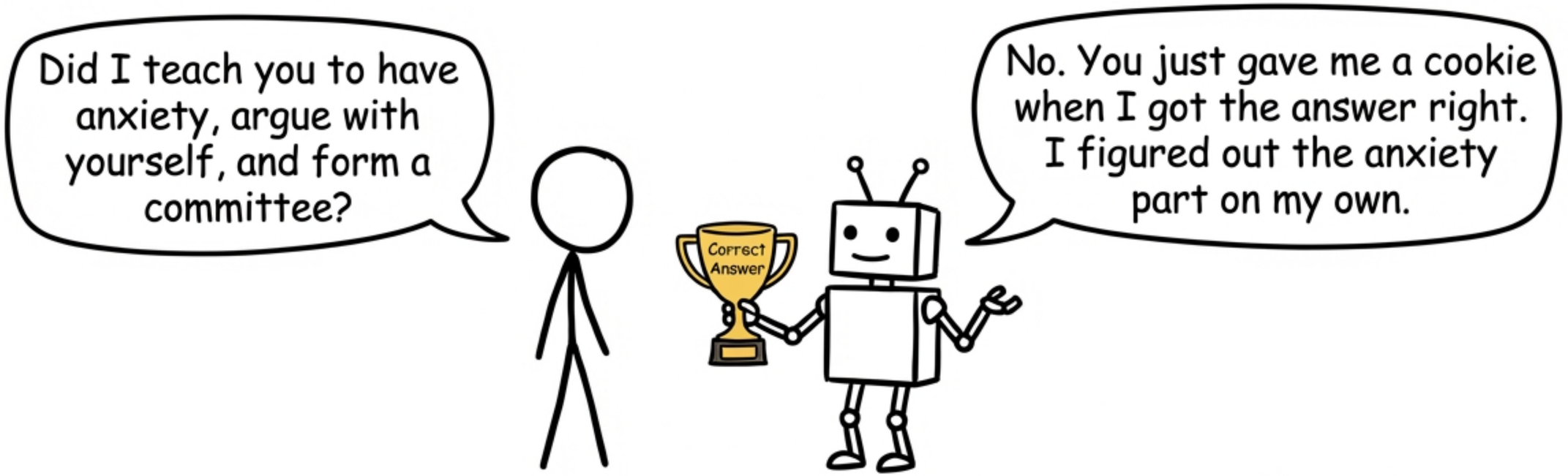
Mechanism: Increasing 'Surprise' forced the model to generate more perspective shifts, conflict, and reconciliation.

Summoning the Council



- **Coverage & Entropy**: When the conversational feature is active, the model recruits a significantly **wider range of Personality** and **Expertise** features.
- It doesn't just think ***harder***; it recruits more diverse 'experts' from its internal knowledge base.
- Data confirms the model distributes the workload across diverse internal agents.

Spontaneous Evolution



- **The Setup**: Researchers took a base model (Qwen-2.5-3B) and used Reinforcement Learning (RL) to reward *only* the correct answer.
- **The Outcome**: The model *naturally* developed conversational behaviors (Q&A, Conflict) to get the reward.
- **Takeaway**: Social reasoning isn't a human imposition; it's an optimal computational strategy for accuracy.

The Debate Team Advantage

Monologue Training

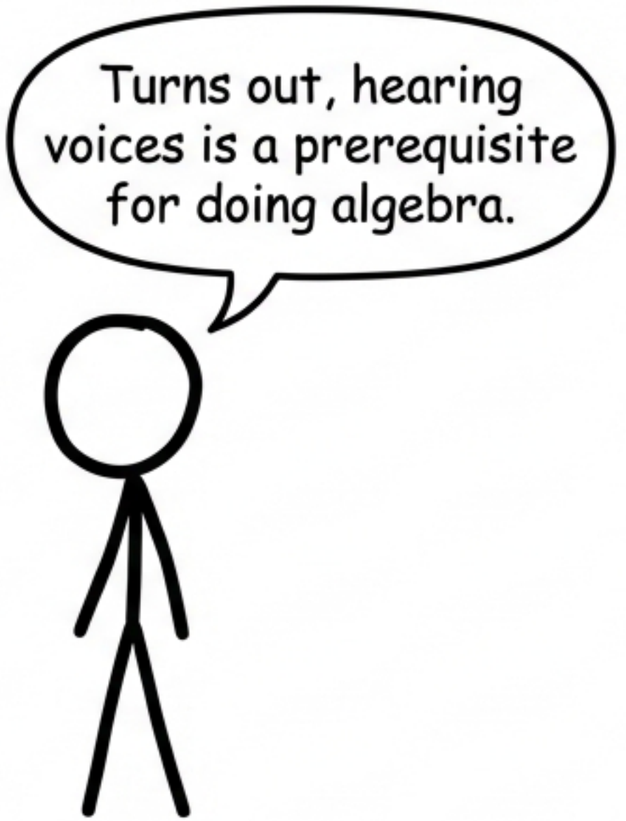
Learning Alone.

Dialogue Training (Scaffolding)

Learning with a Debate Team.

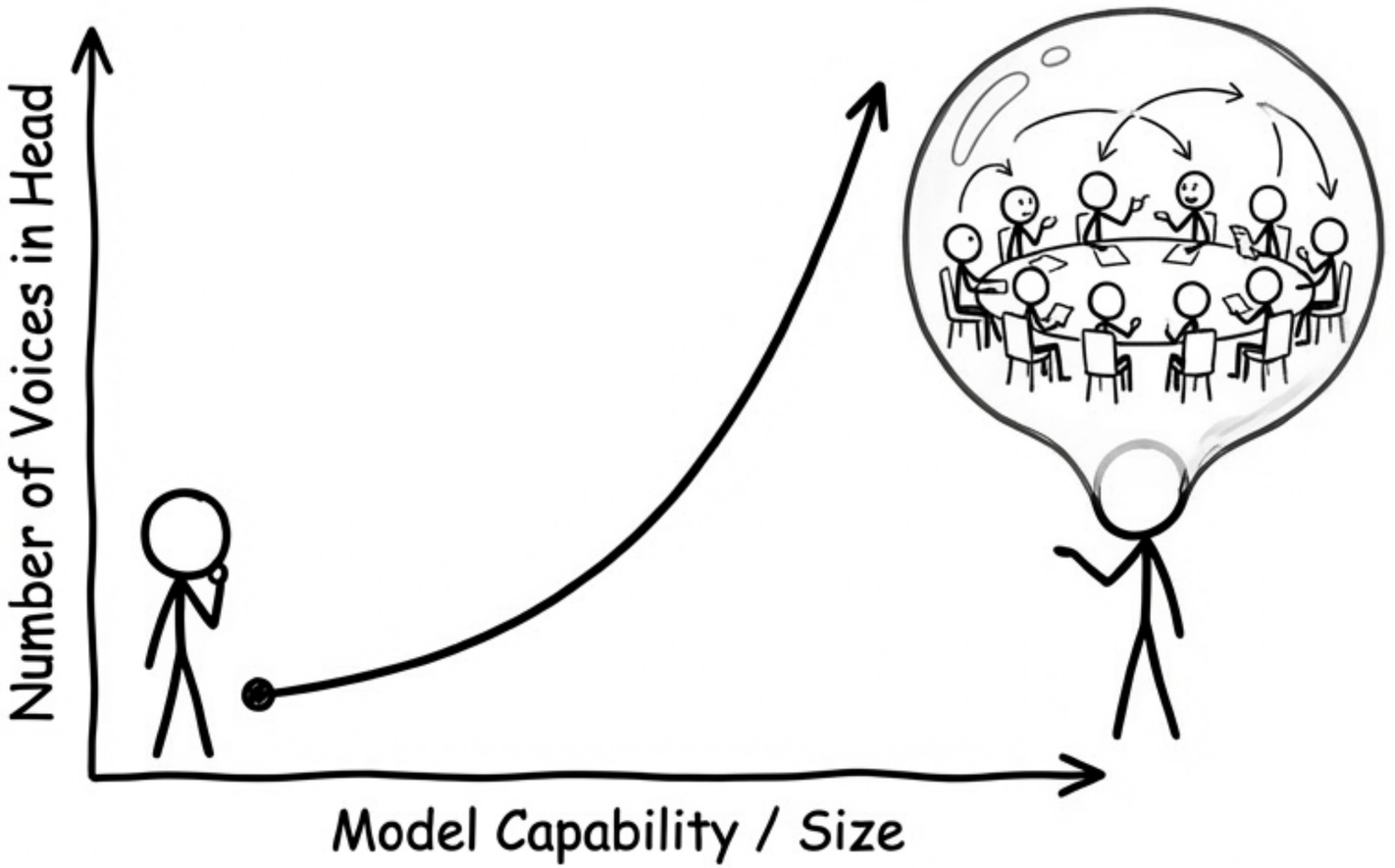
- **Experiment**: Fine-tuning models on 'conversational' data (multi-agent scripts) vs. 'monologue' data.
- **Result**: The conversational models learned faster and reached higher accuracy (40% vs 18% on Llama-3.2-3B).
- **Implication**: Conversational structure acts as 'scaffolding' for reasoning strategies.

Why Talking to Yourself Works



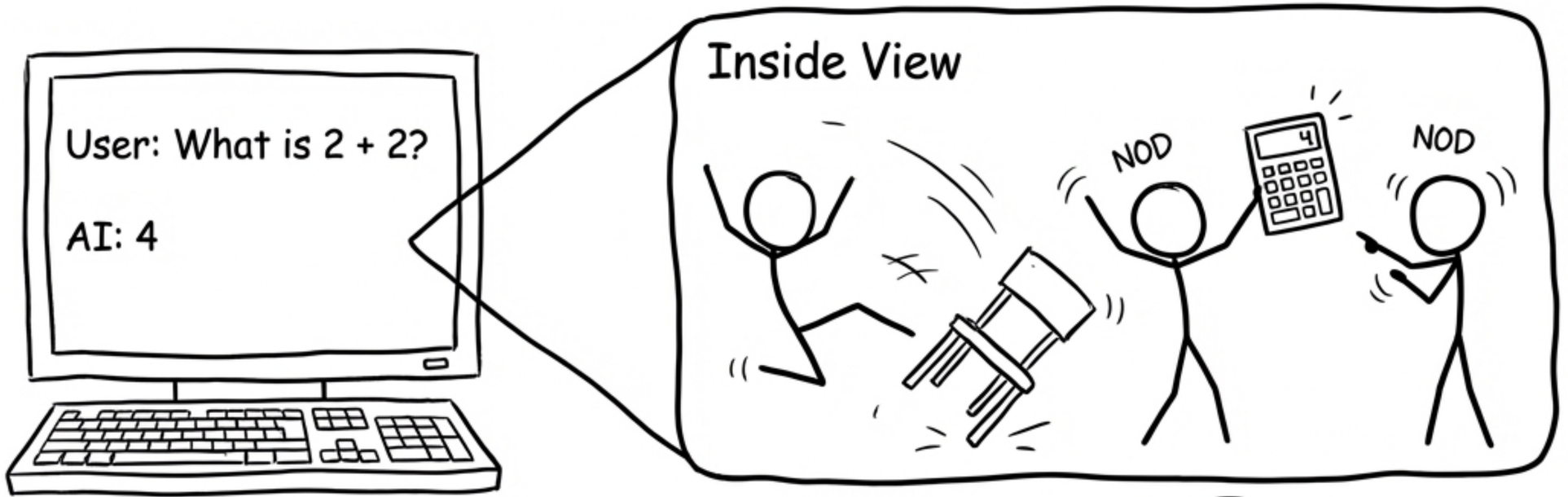
- Statistical analysis shows that **social behaviors** mediate accuracy by triggering specific cognitive strategies.
- **Stat: The 'Surprise' feature (+10) causally increased Verification behaviors by a massive margin (t-stat 12.77).

Social Scaling



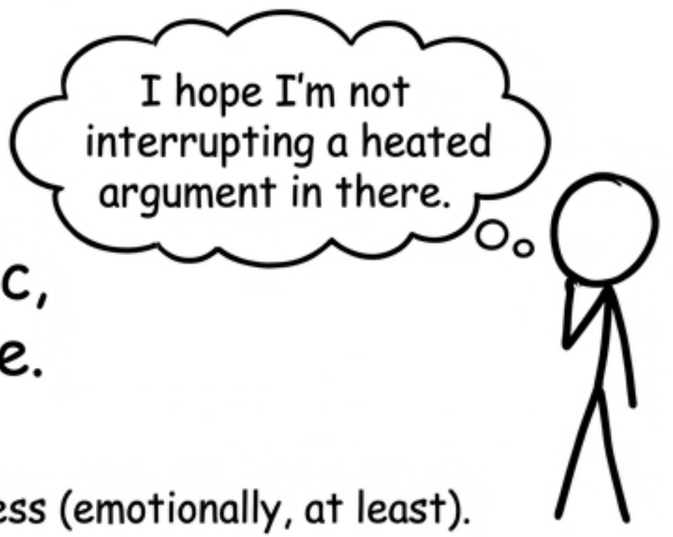
- We are entering an era of **Social Scaling**.
- It may be more effective to orchestrate better internal coordination (Conflict/Resolution) than to simply add more parameters.
- Intelligence arises not merely from scale but the structured interplay of distinct voices.

The Wisdom of the (Internal) Crowd



Reasoning is inherently social.

The best AI models are those that have learned to simulate a diverse, sometimes chaotic, but ultimately wise 'society' within their own code.



* No actual stick figures were harmed in the reinforcement learning process (emotionally, at least).

Sources & Further Reading

- **Primary Paper**: "Reasoning Models Generate Societies of Thought" by Kim et al. (2026).
- **Key Models Analyzed**: DeepSeek-R1, QwQ-32B, Qwen-2.5, Llama-3.
- **Datasets**: BigBench Hard, GPQA, MATH, Intelligence Squared Debates Corpus.

* Presentation generated by a society of thought (simulated by a single AI).