http://ww.equationfarm.comP

Exploration of the Equation Space Utilizing RAG Based LLM's in a Opposition Format.

As of 4th Quarter 2025, AI agents are solving significantly more portions of advanced mathematics, approaching Phd level.

Currently as reported by Grok 4 (The Users Subscription Model)

Available Evidence on Grok 4's Performance

1. **Humanity's Last Exam (HLE) Benchmark**:
    - **Description**: HLE is a challenging benchmark with 2,500 PhD-level questions across multiple disciplines, including mathematics, physics, and engineering. It's designed to test frontier-level knowledge and reasoning.
    - **Performance**:
        - Grok 4 (base model, without tools) scored 25.4% accuracy on HLE.
        - Grok 4 Heavy (multi-agent model with tools like code execution and web search) scored 44.4% to 50.7% on HLE, significantly outperforming competitors like Google's Gemini 2.5 Pro (~21.6%) and OpenAI's GPT models (~20–21%).
        - These scores suggest Grok 4 solves PhD-level math problems (a subset of HLE) with 25–50% accuracy, depending on whether tools are used.
2. **FrontierMath Benchmark**:
    - **Description**: Conducted by Epoch AI, FrontierMath tests advanced undergraduate to early-career research-level math problems, which align with PhD-level difficulty. It focuses on background knowledge, creative problem-solving, and complex computations.
    - **Performance**: Grok 4 achieved 12% and 14% (±2%) across two evaluations, placing it second behind OpenAI's o4-mini (19%). This indicates a lower success rate (12–14%) on highly specialized PhD-level math problems, though the benchmark is noted for its extreme difficulty.
3. **AIME 2025 Benchmark**:
    - **Description**: The American Invitational Mathematics Examination (AIME) is a high school competition, but its problems approach advanced undergraduate or early graduate-level complexity, relevant to PhD-level skills.
    - **Performance**: Grok 4 scored 95% to 100% on AIME 2025, demonstrating near-perfect performance on structured math problems. While not strictly PhD-level, this suggests strong capability on problems requiring advanced reasoning.

By this extension we can then theorize the following cyclic process:

1. AI Agent is asked to discover or 'role-play' a scientist looking for a new theory that would explain some unknown observation in the Universe. It is to create these new theories just as any research scientist would do so.

2. AI Agent then is instructed to verify and see if their new theory is accurate or wrong, and to test it multiple times.

3. If the theory is proved wrong, it is instructed to continue this cycle repeatedly.

4. Once the AI Agent has found something that it believes is a new theory advancing known knowledge it is compiled to paper format for human introspection.

It should be noted that the process of creating new theories by the AI systems can be done in literally minutes along with its proof and eliminations. It was found that in only a few iterations, the AI models were suggesting, auditing and checking their own proofs, at a rate *far faster than humans can verify.* RAG (Retrieval- Augmented Generation type AI's) and model variants will look to be the strong contenders in this development.

It is key that one does not presume that everything that an AI model will produce is wrong, nor to assume that it is correct, but that by having the AI model audit it's own work – it can save researchers months.