Researching with Thinking Models: AI-Assisted Paths in Mathematics

Jonas Henkel

Philipps-Universität Marburg

Graduate Training Workshop for the Mathematical Sciences
Isaac Newton Institute for Mathematical Sciences, Cambridge
5 – 9 January 2026

Outline

AI in Mathematical Research

From Frontier Models to Daily Practice

Frontier Performance vs. Practical Reliability

The Frontier: High-End Autonomous Performance

- `Gemini Deep Think` wins IMO Gold in 2025, officially permitted to compete autonomously in the 4.5 hour time frame.
- `AlphaEvolve` autonomously discovers a faster matrix multiplication, beating the 56-year-old Strassen Algorithm.

Frontier Performance vs. Practical Reliability

The Frontier: High-End Autonomous Performance

- `Gemini Deep Think` wins IMO Gold in 2025, officially permitted to compete autonomously in the 4.5 hour time frame.
- `AlphaEvolve` autonomously discovers a faster matrix multiplication, beating the 56-year-old Strassen Algorithm.

The Reality: A Significant Gap

- Under standard direct prompting, accessible models like `Gemini 3 Pro` fail to secure even a Bronze medal at the IMO.
- This isn't just about scores. We must understand the nature of their errors.

Frontier Performance vs. Practical Reliability

The Frontier: High-End Autonomous Performance

- `Gemini Deep Think` wins IMO Gold in 2025, officially permitted to compete autonomously in the 4.5 hour time frame.
- `AlphaEvolve` autonomously discovers a faster matrix multiplication, beating the 56-year-old Strassen Algorithm.

The Reality: A Significant Gap

- Under standard direct prompting, accessible models like `Gemini 3 Pro` fail to secure even a Bronze medal at the IMO.
- This isn't just about scores. We must understand the nature of their errors.

This creates the central tension:

How do we harness the power without falling victim to the flaws?

- Scientific analysis (like the **Open Proof Corpus** (2025)) reveals systematic flaws:
  - *Self-critique blindness*
  - *Reluctance to admit failure* (leads to hallucination)
  - *Overgeneralization* from simple cases

Case Study: The "Illusion of Rigour"

1. The Complex Query
Asking for a proof or derivation of a difficult question.

2. The plausible Invention
The AI generates a proof with perfect structure and standard phrasing ("It is trivial to see...").

3. The Logical Gap
A critical condition (e.g., compactness) is silently assumed or hallucinated to bridge a gap.

Trigger: Reluctance to admit failure & Gap filling

Result: High confidence masking low competence

Case Study: The "Illusion of Rigour"

1. The Complex Query
Asking for a proof or derivation of a difficult question.

↓

2. The plausible Invention
The AI generates a proof with perfect structure and standard phrasing ("It is trivial to see...").

Trigger: Reluctance to admit failure & Gap filling

↓

3. The Logical Gap
A critical condition (e.g., compactness) is silently assumed or hallucinated to bridge a gap.

Result: High confidence masking low competence

The Danger: We trust the *form* and overlook the *flaw*.

The Core Challenge: From Flaw to Methodology

This duality of power and fallibility is our new reality.

The Core Challenge: From Flaw to Methodology

This duality of power and fallibility is our new reality.

The challenge is therefore *methodological*: How do we work *with* these flaws to achieve reliable results?

The Core Challenge: From Flaw to Methodology

This duality of power and fallibility is our new reality.

The challenge is therefore *methodological*: How do we work *with* these flaws to achieve reliable results?

> *"[...] I expect, say, 2026-level AI, when used properly, will be a trustworthy co-author in mathematical research [...]"*
>
> — *Terence Tao (2023)*

Status Jan 2026: Evidence supports this.
Huang et al. (Oct 2025) showed that models fail solo ($\approx 30\%$), but reach Gold level ($> 85\%$) at the IMO within a *verification loop*.
This necessitates the shift from consumer to critical navigator.

The Copilot Model: A Structured Approach

To fulfill this role of the *critical navigator* in daily research, we propose a structured workflow:

The Copilot Model

The Copilot Model: A Structured Approach

To fulfill this role of the *critical navigator* in daily research, we propose a structured workflow:

The Copilot Model

This framework defines specific interaction patterns for mathematicians and is detailed in:

H. (2025). *The mathematician's assistant: integrating AI into research practice.*
Mathematische Semesterberichte.
DOI: 10.1007/s00591-025-00400-0

The Copilot Model: 5 Guiding Principles

- This model is built on the following five core guiding principles:
  1. **The Copilot, Not the Pilot**
     Copilot assists, suggests, compute. Pilot: direction, judgment, verification. Decides: When to use copilot.

The Copilot Model: 5 Guiding Principles

- This model is built on the following five core guiding principles:
  1. **The Copilot, Not the Pilot**
     Copilot assists, suggests, compute. Pilot: direction, judgment, verification. Decides: When to use copilot.
  2. **The Principle of Critical Verification**
     First, check for plausibility. Then, verify with literature, code, or another model.

The Copilot Model: 5 Guiding Principles

- This model is built on the following five core guiding principles:
  1. **The Copilot, Not the Pilot**
     Copilot assists, suggests, compute. Pilot: direction, judgment, verification. Decides: When to use copilot.
  2. **The Principle of Critical Verification**
     First, check for plausibility. Then, verify with literature, code, or another model.
  3. **Understand the Non-Human Nature of AI**
     Understand how the AI's mind works. Be aware of flaws like persistent memory.

The Copilot Model: 5 Guiding Principles

- This model is built on the following five core guiding principles:
  1. **The Copilot, Not the Pilot**
     Copilot assists, suggests, compute. Pilot: direction, judgment, verification. Decides: When to use copilot.
  2. **The Principle of Critical Verification**
     First, check for plausibility. Then, verify with literature, code, or another model.
  3. **Understand the Non-Human Nature of AI**
     Understand how the AI's mind works. Be aware of flaws like persistent memory.
  4. **The Art of Prompting & Model Selection**
     Effective use requires iterative practice. Crucial heuristic: Quality correlates with time. Favor models that "think" or "research" for minutes, not seconds.

The Copilot Model: 5 Guiding Principles

- This model is built on the following five core guiding principles:
  1. **The Copilot, Not the Pilot**
     Copilot assists, suggests, compute. Pilot: direction, judgment, verification. Decides: When to use copilot.
  2. **The Principle of Critical Verification**
     First, check for plausibility. Then, verify with literature, code, or another model.
  3. **Understand the Non-Human Nature of AI**
     Understand how the AI's mind works. Be aware of flaws like persistent memory.
  4. **The Art of Prompting & Model Selection**
     Effective use requires iterative practice. Crucial heuristic: Quality correlates with time. Favor models that "think" or "research" for minutes, not seconds.
  5. **The Experimental Mindset**
     Discover the specific strengths and weaknesses for your own field of research.

The Toolkit

The current top & affordable models on math benchmarks

Our Toolbox: The All-Rounders (Google & OpenAI)

Google: Gemini 3 Pro

- **Key Strength:** Huge context window (>1M tokens) for analyzing large documents, strong reasoning
- **Recommendation:** Use it for free via Google AI Studio to get full control over creativity (temperature).
- *Weakness:* "Persistent Memory" issues, literature search.

OpenAI: GPT-5.2 Thinking

- **Key Strength:** Excellent for web searching for 5-10 minutes (!), reasoning.
- Integrates well with more complex tasks (`Agent mode`).
- *Weakness:* Smaller context window than Gemini.
- Available for $\approx$ 20 \$/month.

Our Toolbox: The All-Rounders (Google & OpenAI)

Google: Gemini 3 Pro

- **Key Strength:** Huge context window (>1M tokens) for analyzing large documents, strong reasoning
- **Recommendation:** Use it for free via Google AI Studio to get full control over creativity (temperature).
- *Weakness:* "Persistent Memory" issues, literature search.

OpenAI: GPT-5.2 Thinking

- **Key Strength:** Excellent for web searching for 5-10 minutes (!), reasoning.
- Integrates well with more complex tasks (`Agent mode`).
- *Weakness:* Smaller context window than Gemini.
- Available for $\approx$ 20 \$/month.

Synergy: Use GPT to *find* literature (Search), then upload the PDFs to Gemini to *analyze* them (Context).

***Note:* Both offer "Frontier" tiers (`Deep Think`, etc.) at a much higher price point ($\approx$ 275\$/month).**

Our Toolbox: The Specialists (Speed, Search & Style)

- **For Real-Time Info & Deep Reasoning: Grok 4 (xAI)**
  - *Pro:* Best integration with live web search.
  - *Observation:* Capable of extreme inference-time compute.
    - We observed it "thinking" for **30 minutes** (!) on an inequality, resulting in a remarkably elegant proof.
  - *Con:* Web search cannot be disabled (can sometimes interfere).
  - Available for 30 $/month.

Our Toolbox: The Specialists (Speed, Search & Style)

- **For Real-Time Info & Deep Reasoning: Grok 4 (xAI)**
  - *Pro:* Best integration with live web search.
  - *Observation:* Capable of extreme inference-time compute.
    - We observed it "thinking" for **30 minutes** (!) on an inequality, resulting in a remarkably elegant proof.
  - *Con:* Web search cannot be disabled (can sometimes interfere).
  - Available for 30 $/month.
- **For Automated Literature Search: Deep Research Tools**
  - Available in both Google's and OpenAI's ecosystems.
  - They perform an intensive, automated search for you over 5-15 minutes.

Our Toolbox: The Specialists (Speed, Search & Style)

- **For Real-Time Info & Deep Reasoning: `Grok 4` (xAI)**
  - *Pro:* Best integration with live web search.
  - *Observation:* Capable of extreme inference-time compute.
    - We observed it "thinking" for **30 minutes** (!) on an inequality, resulting in a remarkably elegant proof.
  - *Con:* Web search cannot be disabled (can sometimes interfere).
  - Available for 30 $/month.
- **For Automated Literature Search: `Deep Research` Tools**
  - Available in both Google's and OpenAI's ecosystems.
  - They perform an intensive, automated search for you over 5-15 minutes.
- **For Polishing Your Writing: `DeepL Write`**
  - Excellent for refining language, rephrasing, and improving style.

A Rapidly Evolving Landscape

- The AI field is developing at an extraordinary pace:
- Benchmarks and studies on specific models are usually outdated within 6 months.
- Therefore, the *ways* of usage are more durable than knowledge of any single model.

The Workflow

7 Ways of AI Usage

The 7 Ways: A Practical Framework (Part 1)

1. Creativity & Ideation

- Formulate novel conjectures or research questions.
- Construct non-obvious counterexamples.
- Design new exercises for your students.
- *Tools: `Gemini 3 Pro` (creative mode), `GPT-5.2 Thinking`.*

The 7 Ways: A Practical Framework (Part 1)

1. **Creativity & Ideation**
   - Formulate novel conjectures or research questions.
   - Construct non-obvious counterexamples.
   - Design new exercises for your students.
   - *Tools: `Gemini 3 Pro` (creative mode), `GPT-5.2 Thinking`.*
2. **Literature Search**
   - Identify foundational papers in a new field.
   - Track the evolution of a specific concept over time.
   - *Tools: `Deep Research` tools, `GPT-5.2 Thinking` (with search).*

The 7 Ways: A Practical Framework (Part 1)

1. **Creativity & Ideation**
   - Formulate novel conjectures or research questions.
   - Construct non-obvious counterexamples.
   - Design new exercises for your students.
   - *Tools: `Gemini 3 Pro` (creative mode), `GPT-5.2 Thinking`.*
2. **Literature Search**
   - Identify foundational papers in a new field.
   - Track the evolution of a specific concept over time.
   - *Tools: `Deep Research` tools, `GPT-5.2 Thinking` (with search).*
3. **Literature Analysis**
   - Extract and compare definitions across multiple sources.
   - Identify implicit assumptions in an author's argument.
   - *Tool: `Gemini 3 Pro`.*

     For reliable analysis, only use documents you upload into the session!

The 7 Ways: A Practical Framework (Part 2)

4. **Interdisciplinarity**
   - Act as a "universal translator" between mathematical subfields.
   - Facilitate collaboration with other disciplines (e.g., mathematical physics, computer science, engineering) by bridging gaps in technical language.
   - *Tools: All major LLMs.*

The 7 Ways: A Practical Framework (Part 2)

4. **Interdisciplinarity**
   - Act as a "universal translator" between mathematical subfields.
   - Facilitate collaboration with other disciplines (e.g., mathematical physics, computer science, engineering) by bridging gaps in technical language.
   - *Tools: All major LLMs.*
5. **Mathematical Reasoning**
   - Explore alternative proof strategies for a known result.
   - Diagnose the specific flaw in a broken proof attempt.
   - *Tools: `Gemini 3 Pro`, `GPT-5.2 Thinking`.*

     Safety Strategy: Use a multi-model approach to counter "self-critique blindness" and to verify stated formulas.

The 7 Ways: A Practical Framework (Part 2)

4. **Interdisciplinarity**
   - Act as a "universal translator" between mathematical subfields.
   - Facilitate collaboration with other disciplines (e.g., mathematical physics, computer science, engineering) by bridging gaps in technical language.
   - *Tools: All major LLMs.*
5. **Mathematical Reasoning**
   - Explore alternative proof strategies for a known result.
   - Diagnose the specific flaw in a broken proof attempt.
   - *Tools: `Gemini 3 Pro`, `GPT-5.2 Thinking`.*

     Safety Strategy: Use a multi-model approach to counter "self-critique blindness" and to verify stated formulas.
6. **Social Aspect**
   - Reduce the barrier to asking "dumb questions."
   - Force clarification of one's own thoughts before presenting to colleagues.
   - *Tools: Any LLM as a 24/7 sparring partner.*

The 7 Ways: A Practical Framework (Part 3)

7. Writing

- Rephrase complex arguments for different audiences (e.g., introduction vs. technical section).
- Generate compelling abstracts and summaries from the full text.
- *Tools: `Gemini 3 Pro` (low creativity), `DeepL Write`.*

Conclusion

Final Thoughts

Conclusion: From Hype to Reality

The Reality

Not Automation, but **Augmentation**.

Conclusion: From Hype to Reality

The Reality

Not Automation, but **Augmentation**.

The Method

The key to reliability is the **Copilot Model**.

Conclusion: From Hype to Reality

The Reality

Not Automation, but **Augmentation**.

The Method

The key to reliability is the **Copilot Model**.

The Role

This demands a new core competency:
The researcher as the Critical Navigator.

Thank You

Questions & Discussion

Jonas Henkel

henkelj@mathematik.uni-marburg.de

Figure 1: Full article: *The mathematician's assistant: integrating AI into research practice* doi.org/10.1007/s00591-025-00400-0

This talk provided a high-level overview of a complex topic. For further discussion, feel free to ask now or contact me via email.