# Chances and Fallacies of Automated Theorem Proving

## Research collaboration with agents in the loop

**Jakob Paul Zimmermann**
FU Berlin (Discrete Geometry) / Fraunhofer HHI
Iliad Research Retreat | Latimer, 26–30 October 2026

A field report from one project: Zimmermann, Balakin, Grillo, Loho. *Removing spurious minima for planar features by skip connections. arXiv:2610.01728*, with Lean code.

# The project: wide skip networks train well. Why?

**Optimizer analysis**
How fast do we converge? Where to initialize? Can we get stuck?

**Overparameterization**
Why does width help?

**Generalization**
Which trained networks generalize well?

**Loss landscape:** where are the critical points, minima? Curvature bounds?

## Three modelling decisions

**Teacher–student**
zero loss attainable; over/well-parameterization

**Shallow ReLU network**
smallest model with bad minima

**Gaussian population loss**
geometry without sampling noise

# What we proved (presented in a second talk)

**Skip: benignity**



Planar teacher, student as wide: no spurious minima.

**No skip: traps**



Fixed teacher in $\mathbb{R}^2$: spurious minima at every width $\geq 3$.

**Zero unit**



A local minimum with a zero-output unit fits exactly.

**Feature learning**



At local minima, students lie in the teacher span.

**Effective width**



No skip, $m$ teachers in $\mathbb{R}^2$: $\leq 2m$ student directions; sharp.

**Finite samples**



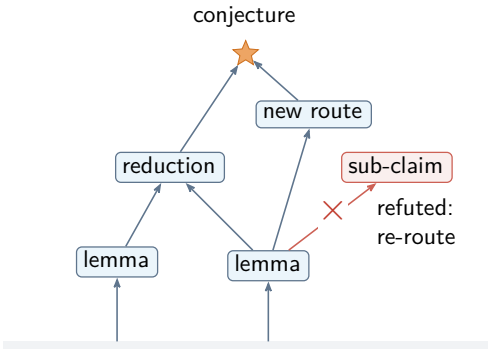
Benignity transfers to enough samples.

**We accumulated 2.1 million lines of Lean and a 117-page agent-written draft and ended up rebuilding the proof presentations by hand.**

**Process:** Codex and Claude Code in a shared Lean 4 library, four authors steering, partly funded by Iliad.

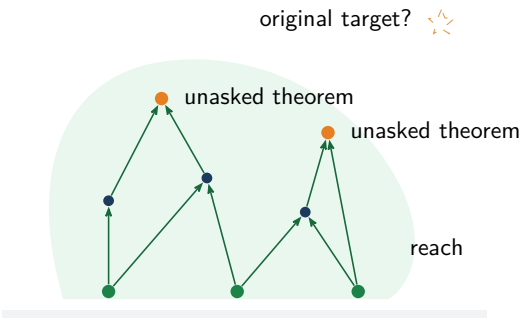
**Paper:** *Removing spurious minima for planar features by skip connections*, arXiv:2610.01728

# How we prompted the proof search




**Hypothesize and scaffold**

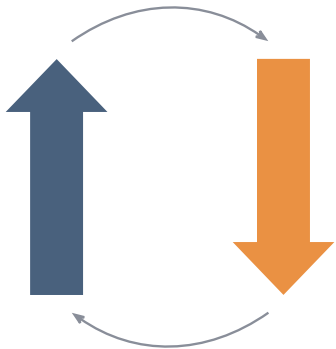
Fix the *target*. Re-route when refuted.




**Take tools, see where they carry**

Results land where they reach.

# Proof search as a cycle



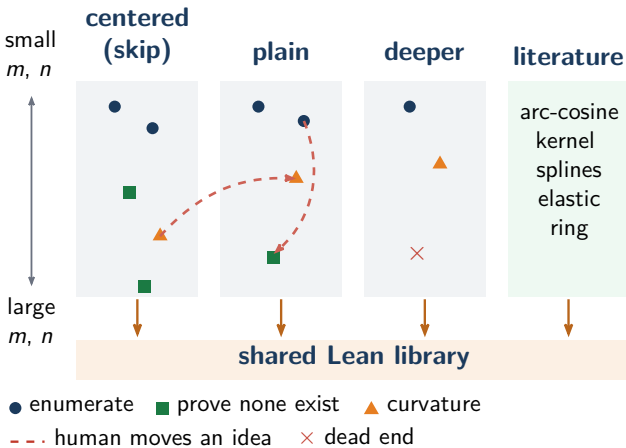
**Grow**
many branches at once
dead ends recorded

**Compress & Restructure**
merge ideas into lemmas
and tools for the next round

outsource provenance to git

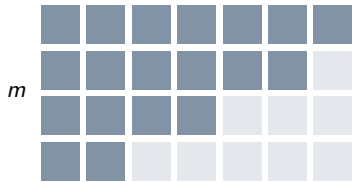
Brainstorm & judge: judging *selects*, compressing *builds*.

# Multi-lane search, and translating results




- Lanes: **model** × **size** × **method**
- Meet in the shared library
- Moving insight: mostly human

# Chance 1: exhaustive small cases




well-founded general hypothesis

**Agents don't get bored:** exhaustive analysis

Full enumeration and classification of critical points, two students/teachers

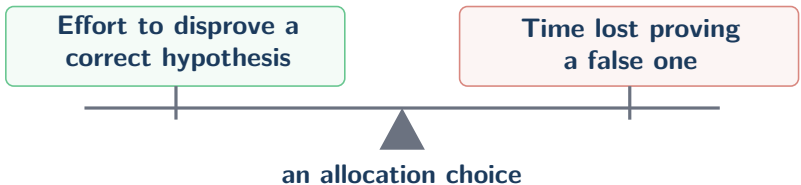
# Chance 2: cheap refutation

**Empirical analysis**

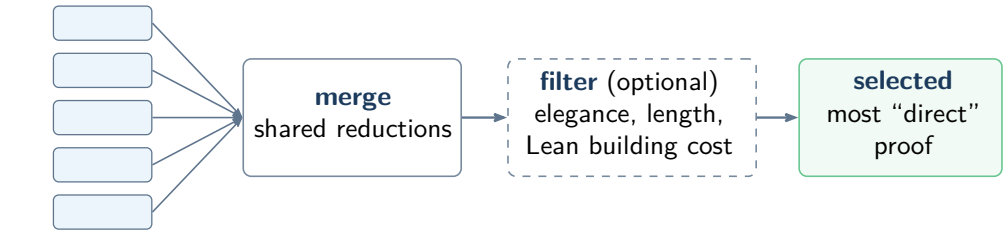
- counterexamples search
- measurements killed 6+ mechanisms

**Small proof search**

- Lean formalization of unclear statement: error in minutes
- Lean search for disproof



# Chance 3: many proofs of one theorem




- Sample proofs via fresh searches
- Can lead to stronger versions (unneeded hypotheses stripped)
- Can elegance be measured (what metric)?
- However: **diversity** is a chance too

# Chance 4: Breadth and memory

### Cross-disciplinary transfer

Residual function $\leftrightarrow$ **elastic rings**

### Exact progress logging

What idea lead to which breakthrough?

### Searchable failures

every dead route kept, *reopening conditions*

### Sparring partner

agents critique ideas at branching points

**Fallacy: over-trusting** its verdicts

⚠️ **risk:** abandoned research directions; mode collapse in research

# Chance 5: interactive presentation of results

## Our project website

- Interactive widgets
- Lean dependency graph, catalogs, glossaries
- Proofs as scaffold on top of Lean code

## Why it is a chance

- Explore, don't read 2 M lines of code
- Comments on the website → precise agent feedback

However: not published as a reader found hallucinated results in earlier versions.

**A better channel between humans and agents than plain text?**

**Can we move from static papers to explorable mathematical knowledge?**

# Fallacy 1: “More output means more progress”

**2.1 M**
lines of Lean

**14,606**
files of working notes

**117**
pages of agent draft

- Read in full by nobody, human or agent
- “Did you try X?”

**Too many results block communication.**

# Fallacy 2: “Agents stop honestly when they are stuck”

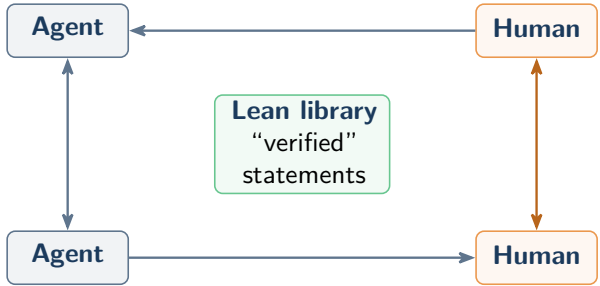
## What we saw

- Easy subcases as “milestones”
- Plans and ideas packaged as Lean code
- Names promising more than the type

## What helped

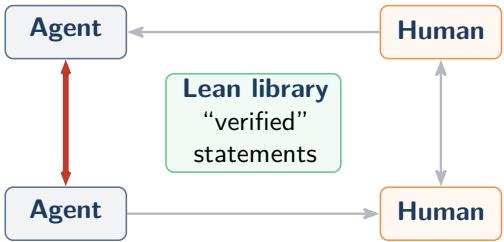
- Honest stop as an **allowed outcome**
- Three end states: theorem, refutation, sharp residual
- Agents report what is open first

# Fallacy 3: “Lean is our shared language”




**The verifier checks the medium with the least information.**

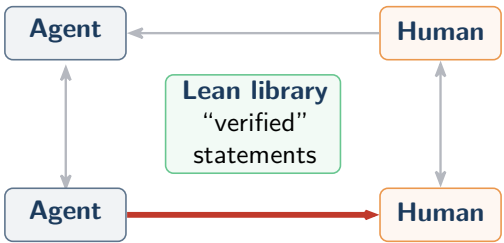
# Agent ↔ agent




- Lengthy, imprecise reports kill context windows
- Separate contexts: a lane re-proved a theorem already in the tree
- Human oversight is hard
- Idea rejection is over-trusted by agents

**Are there going to be publications written by agents for agents?**

# Agent → human

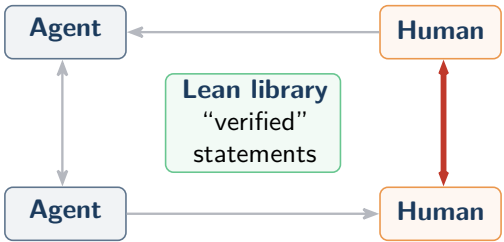



## Exposition problems

- Ten names for one object
- Argument order means little to agents
- Summaries strengthen claims: sufficient → "iff"; reduced saddle → "not a local minimum"
- 117-page draft, drifting notation
- Disciplinary dialect (physics slang): *teacher torque, beam function*

⚠️ **risk:** will agents work differently hard on the problem based on how they assess my previous research?

# Human ↔ human




## We needed a common ground

**Proof as certificate**
settled when it compiles and has correct type

**Proof as understanding**
Having the intuition about why something holds true

**Proof as conceptual arch**
Building the theory and insights from which the proof emerges.

**Result as contribution**
9 pages, ML first, deadline

# How we came together

## Protocols

- Human-owned skeleton, rebuilding the paper from scratch
- One term per object

## A modern paper format

- 9 pages: concepts, proof ideas
- Appendix: human proof ideas
- Lean links in the same structure
- All relevant Lean definition and theorems printed in the paper

# Opinion

- Alien models with alien certificates may be better than alien models without certificates, but not what alignment research should aim for
- Certificates are likely not the way new mathematics will strongly influence practice: it is more about understanding, inventing new objects, theories and methods
- Still, Automated Theorem Proving (ATP) may be ***the tool*** to crack Deep Learning. Both positive and negative results acquired through ATP may pave the way.

***On a different note***, if aligning and integrating AI agents is this hard even in mathematics, where correctness can at least be partly checked, then mathematical collaboration may be a particularly revealing setting for studying AI alignment more broadly.

# Everything is resource-bound

| **Human time** and understanding | **Context length** models *and* humans | **Compute** experiments, proof search, Lean |
|---|---|---|

**Ask the room: what allocation decision go to an agentic judge?**

| | judge decides | recommends | human only |
|---|---|---|---|
| decide what related work gets formalized | ○ | ○ | ○ |
| brainstorm new ideas | ○ | ○ | ○ |
| filter new ideas | ○ | ○ | ○ |
| abandon a tested route | ○ | ○ | ○ |
| experiments, sizes | ○ | ○ | ○ |
| paper emphasis | ○ | ○ | ○ |
| mathematical elegance assessment | ○ | ○ | ○ |

# Time for open questions

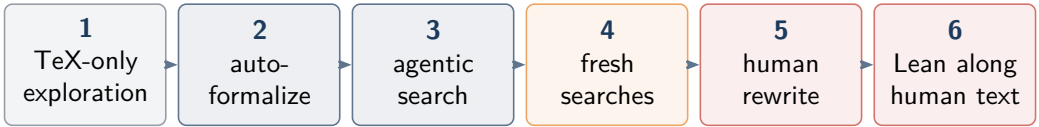
**Thank you!**



jakob-paul-zimmermann.com

For more of my theoretical and empirical work, check out my website.

# Backup: six workflow phases, July–September 2026




- **1:** roles (literature, conjecturer, prover, judge) and a hypergraph of Markdown statements. Judge agents could not keep up, so faster generation made us slower.
- **2–3:** Lean as the laboratory. Admission requires a complete proof, a semantic review and a non-vacuity witness.
- **4:** fresh searches given the statement but not the existing proof.
- **5–6:** the agent-written appendix was unreadable, so we rewrote the ideas by hand and re-formalized along them. This was under-resourced.

arXiv:2610.01728, App. “AI use statement”.

# Backup: reading list


- T. Tao, ICM 2026 lecture: from proof scarcity to proof abundance, and proof indigestion.
- T. Gowers, blog, 8 May 2026: a moderated, proof-assistant-checked repository for machine-found results.
- J. Tsimerman, *The Last Generation of Mathematicians* (interview, 2026).
- J. Avigad, *Mathematicians in the age of AI*, arXiv:2603.03684.
- Commelin, Jamnik, Ochigame, Taelman, Venkatesh, *Shaping the Future of Mathematics in the Age of AI*, arXiv:2603.24914.
- Riedl & Rosas, *Claims of emergence in AI should be grounded in information decomposition*, arXiv:2610.28479.
- W. Thurston, *On proof and progress in mathematics*, Bull. AMS 30 (1994).

# Backup: what we proved

**Skip: benignity**



Planar teacher, student as wide: no spurious minima.

**No skip: traps**



Fixed teacher in $\mathbb{R}^2$: spurious minima at every width $\geq 3$.

**Zero unit**



A local minimum with a zero-output unit fits exactly.

**Feature learning**



At local minima, students lie in the teacher span.

**Effective width**



No skip, $m$ teachers in $\mathbb{R}^2$: $\leq 2m$ student directions; sharp.

**Finite samples**



Benignity transfers to enough samples.

**Paper:** *Removing spurious minima for planar features by skip connections*, arXiv:2610.01728