

Playing with Frankl's conjecture using LLM agents

what works and mostly, what doesn't

2026-09-03

Eric P. Hanson

Contents



1	Introduction	3
1.a	Motivation	4
1.b	Frankl's conjecture	5
1.c	History and progress	6
1.d	2016: a Polymath project	7
2	The project	8
2.a	Setup	9
2.b	Shared tooling	10
2.c	Methodology: very "unsupervised"	11
2.d	Workstreams (or lines of attack)	12
2.e	Many generalizations refuted	13
2.f	Results: a lot of stuff... ..	14
3	Observations	15
3.a	I chose a bad problem to observe progress	16
3.b	Agent collaboration	17
3.c	Specific failure modes	19
3.d	Outlook	20
3.e	References	21

1 Introduction

Motivation



- LLMs are good at math now
 - ▶ On July 19, Levent Alpöge found a counterexample to the Jacobian conjecture, then on August 19 a counterexample for the Carathéodory conjecture.
 - ▶ LLMs are proving successful at formalizing math in lean, with LLM-driven contributions to mathlib starting to become more common (918 closed PRs mentioning Claude, Codex, or LLMs)
- Around a year ago, I had gotten interested in Frankl's conjecture because a few years ago there was progress in it due to an entropic bound, similar to ones I considered in my PhD
- On Saturday August 22, I decided why not see if I can setup some LLM agents and have a go at Frankl's conjecture
- I am not an expert on this!
 - ▶ At my day job, I work on problems like “how does narcolepsy present in overnight sleep” (and training classifiers) and “how can we get junior engineers to learn more from production incidents” – not many agents, and no pure math.

Frankl's conjecture



Conjecture (Frankl)

Any finite family of sets which is closed under union and contains a non-empty set has an element which belongs to at least half the sets.

A family on $\{1, 2, 3\}$, closed under union:

$$\mathcal{F} = \{\emptyset, \{1\}, \{2, 3\}, \{1, 2, 3\}\}$$

- each union of two members is a member, for example $\{1\} \cup \{2, 3\} = \{1, 2, 3\}$;
- call an element *abundant* if it is in at least half the sets. Here every element is in 2 of the 4 sets, so all of them are abundant.

	1	2	3
$\emptyset$			
$\{1\}$			
$\{2, 3\}$			
$\{1, 2, 3\}$			

all three elements are abundant

History and progress



Let n be the number of base elements, and m the number of sets in the family.

- **1979.** Frankl states the conjecture. It first appears in print in 1984 [BS15].
- **1989–2017.** Special cases. The conjecture is proven for $n \leq 12$ elements [VŽ17], for $m \leq 50$ sets [Lo 94], and for lower semimodular lattices [Rei00], in the lattice form of the problem. See the survey [BS15].
- **2017, Karpas** [Kar17]. Proved the conjecture for $m \geq (\frac{1}{2} - c)2^n$ for some unspecified $c > 0$.
 - ▶ Used a reformulated version of the problem: consider a single set as a boolean vector, where the i th entry corresponds to element i belonging to a set. Karpas used Fourier analysis over $\mathbb{F}_2^n$ on the indicator function for the family of sets to obtain his bound.
- **2022, Gilmer** [Gil22]. Proved some element is in at least 1% of the sets.
 - ▶ Used an entropic argument. Let A and B be independent and uniform on the family $\mathcal{F}$. Union closure puts $A \cup B$ in $\mathcal{F}$, so $H(A \cup B) \leq H(A)$ since uniform A has maximal entropy. If every element has frequency at most 0.01, one can show $H(A \cup B) \geq 1.26H(A)$. This is a contradiction, so some element has frequency above 0.01.
 - ▶ Immediate followups [AHS22, CL22, Peb22, Saw22] sharpened to $\psi = (3 - \sqrt{5})/2 \approx 0.382$ fraction of the sets. Pretty close to 0.5!

2016: a Polymath project



Tim Gowers proposed the conjecture as a Polymath project [Gow16a]. It ran for two months in the open, as the blog posts FUNC1 to FUNC4 [Gow16b, Gow16c, Gow16d, Gow16e], with a wiki and hundreds of comments [Pol16].

- The working method was to propose a statement which implies the conjecture, and then to attack that statement.
- Many strengthenings and variants were proposed. Most were refuted within days, often by a small explicit family.
- The project did not prove the conjecture. It stopped in the spring of 2016.
- Many human “agents” working together to try to crack it – but ultimately did not do so.

2 The project

Setup



- Funnel everything into a Typst¹ website to ensure agent's work was visible/legible and I could check on things remotely
- All agents work in a single directory so agents could collaborate “live” instead of e.g. in PR reviews.
 - This mostly worked but had some conflicts.
- Each agent was instructed to end with a journal entry in a “work log” and update the site with their findings (new pages, update review pages, etc)
- They maintained a “map” that outlines all the lines of work, named statements, statuses
- Used pre-push git hooks to check lots of things
 - all experiments have lockfiles (uv.lock or Manifest.toml), any new pages must be mentioned in the work log entry, any new counterexamples must be syntactically valid, work log must be append-only, no files over 1 MB, and the whole site builds correctly
- Got a \$200 Claude max plan. Not cheap but wanted to give it a go.

¹A modern LaTeX alternative

Shared tooling



- Setup a Lean² proof directory with CI checks
- Zotero³ integration with bibliography
- a “job slots” system (file lock per slot) for cross-process cooperative load-balancing
 - provided `tmap`, `tforeach` etc primitives to use threading in a way that used the slot system (lose slots as tasks finish if another job is waiting)
- experiment convention with reproducible code, checked in JSON results (aggregated)
- Main Julia library. Constructions, symmetry, checkpointing, and the search tools. When a piece of code outlives its experiment, had the agents move it into the package.
- Checked-in zoo of counterexamples
- Refutation pipeline. Check against the examples in the zoo, then the complete corpus of all 29 327 union-closed families at $n = 5$, then a SAT solver, then simulated annealing / heuristics. Also used integer programming (HiGHS) where appropriate.

²programming language/proof checker

³Citation manager

Methodology: very “unsupervised”



First prompts:

ok we have a whole framework setup here. we're gonna prove frankl!! it will take some time but we can do it. Please decide on something to do and do it, following our workflow :)

can you think of a new direction to take on for frankl?

can you review the graph formulation of frankl and write up a lit review on it, following our workflow? (note other agents are working on different pages, should be fine)

are there next steps that help us push beyond small n numerics?

I saw my role as directing the infrastructure setup then letting it play out to some extent.

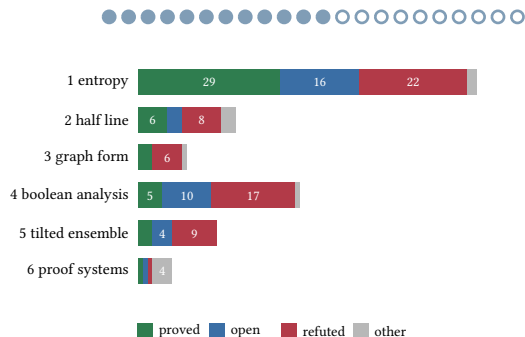
Workstreams (or lines of attack)

The conjecture has five equivalent phrasings.

1. **Entropy**, the method behind every modern record. The deepest line.
2. **The half line**, on Poonen's classification of the tight families.
3. **The graph form**, on bipartite graphs.
4. **Boolean analysis**, following Karpas.
5. **The tilted ensemble**, which reweights the members by size using a thermodynamic analogy.

We also considered:

6. **Proof systems**, which attempts to constrain what proofs are possible.
7. **Alternate conjectures**. Maybe we can find a counterexample to something else using the machinery built here. (Not yet!)

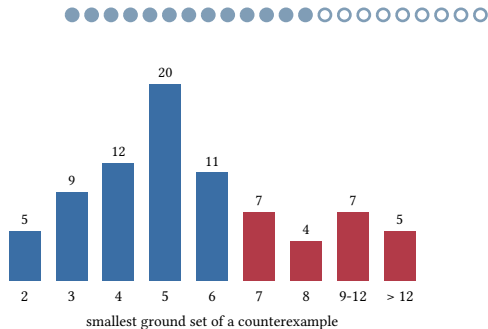


164 named statements, 165 experiments, 250 sessions. The conjecture is still open.

Many generalizations refuted

Like Polymath, the agents proposed generalizations of Frankl, special cases, and related statements. Then they tried to refute them. **They refuted 80 statements, with 105 explicit families.**

- **Families with six elements or fewer refute 57 of the 80.** A corpus sweep or a SAT solver finds such a family in minutes. The other 23 need a constructed family, with up to 80 elements.
- **One line of work made four attempts at one inequality.** Each attempt asked for less than the one before. The agents refuted all four: at $n = 6$ (corpus, minutes), at $n = 40$, at $n = 80$, and at $n = 240$. The $n = 80$ attempt fails by $6 \cdot 10^{-7}$. At $n = 240$, a *pair* of families shows that no inequality of this shape is true. The parent statement stays open. It is the proof strategy that fails.
- The agents also kept 21 families that almost refute a statement. Each entry records the objective and the failure threshold.



The 80 refuted statements, by the smallest family that refutes each one. Blue shows the range where the corpus and the solver give an answer, at $n \leq 6$.

Results: a lot of stuff...



Pages	140
experiment / note / review / meta	67 / 52 / 9 / 12
journal entries, one file each	250
words of prose	696 970
Experiments	165
committed result files	1 468
plots	41
Zoo entries	125

Julia	121 832 lines
of which the shared library	10 959
Python	11 407 lines
Lean	2 112 lines, 99 theorems
tools and scripts	4 199 lines
Named statements	164
proved / open / refuted / other	47 / 34 / 63 / 20
references to them	3 932
Commits	860

...but no proof.

One likely “real” result is that the agents say they have proved $c = 3.3 \cdot 10^{-7}$ holds in Karpas’s theorem (he left it unspecified). I am considering writing it up but it would take a fair amount of time for me to fully onboard onto the proof and build my mental model for me to trust it, and/or formally verify it all.

There are a lot of deadends and some claimed negative results about what arguments cannot work.

3 Observations

I chose a bad problem to observe progress



How do you know you are making progress when proving a theorem?

- if programming is theory building [Nau85], then math certainly is too
- human world: you build a mental model, you test it, find holes, and chase them
- You can sometimes *feel* that model solidify or shift. You know you learned something new when it contradicts your model

How do you know you are making progress when your LLM agents are proving a theorem?

- they prove special cases? May be trivial or very, very hard
- they refute generalizations? Not necessarily interesting...
- if we had a *shared* mental model, i.e. I took the time to follow exactly what they were doing, and manage them closely, then it could work – basically an accelerated version of doing math normally. But I wanted to try hands-off (had other things to do too!).

If we had a “blueprint”, a map of all the intermediate results/statements (ideally all true), then we could trace progress along formalizing each result.

But we don't even know if Frankl's conjecture is true! No blueprint available.

Agent collaboration



I tried a few models of collaboration.

Parallel work, shared workspace.

- Here, I set each agent off on its own task, usually a different line of attack on Frankl, but they are sometimes encouraged to read the work log and other ongoing work.
- This sometimes led to conflicts over shared files and rarely led to helpful simultaneous collaboration
- Having clear records of past work did make future work easier

Two “peer” agents exchanging direct messages

- while debugging the job slot system (which agent was hogging the slots and why wasn't the cooperative load balancing working) I found learned each agent has an ID in `~/claude/sessions` and they have a send-message tool to send messages knowing the ID.
- I set up two “peers” and asked them to work together on a task, sending messages to collaborate, each choosing complementary approaches
- One worked on a proof, the other on numeric/computational checks
- Kind of worked, but not clear it's better than a subagent reviewer

Agent collaboration (ii)



One “manager” and multiple independent contributors (ICs)

- This worked OK! Used manager primarily to avoid failure modes and be able to be more hands-off
- Told manager to not worry about the math/details and instead keep pushing the ICs to keep going, avoid tunnel vision, avoid big computational sweeps and instead think analytically
- Manager could be a much cheaper model (Sonnet) than the ICs (Fable)
- Again, used the send-message tool for the manager to communicate with the agents
- Could help them avoid converging while sending relevant new info from one to the other

Specific failure modes



These may be more specific to my prompting and/or the specific models being used (a mix of Fable 5 and Opus 5).

- **Imprecise writing: incomplete searches described as exhaustive searches.** A sweep found no counterexample up to a size limit and reported that none exists. A later run found one at twice that size.
 - Fix: accurately/completely record what was done. Review agents help here.
- **Stale claims: a statement is refuted in one place and left unchanged in others.**
 - Fix: git pre-commit hook warns when a line you deleted is still on another page.
- **Self-checks that contain the same error as the code they check.** A search reported a known counterexample as clean. The check had the same off-by-one error as the search.
 - Fix: use different code path for checks.
- **Use of more compute in place of a new idea.** An agent that is stuck asks for one more element, ten thousand more starts, or a longer run. This looks like progress and almost never changed a conclusion.
 - Fix: estimate the cost in CPU-hours first, and state which result the run can settle.

Outlook



I think for a problem as hard as Frankl's conjecture, it needs something more:

1. Expert direction. I think a mathematician pairing/directing agents could get further, building their mental map in parallel.
2. Potentially, brute scaling. I could have at most 2-3 Fable agents running for a day or two before running out of tokens. I could imagine getting further with more agents and more time (and more compute than my laptop).

I think other problems though are likely more amenable! I can see better why LLM program evolution is a good fit: you start feasible, you can stay feasible each iteration, and you have a clear metric to measure (runtime or performance or whatever), and you don't need to reach perfection.

Proofs are binary, either you have proved it or not. Very hard to know when you are 50% there. You might be 0% there then 90% in one move! And you might never find that move.

References



- [AHS22] Alweiss, R., Huang, B., and Sellke, M., *Improved Lower Bound for Frankl's Union-Closed Sets Conjecture*, 2022, <https://arxiv.org/abs/2211.11731>.
- [BS15] Bruhn, H. and Schaudt, O., *The Journey of the Union-Closed Sets Conjecture*, *Graphs and Combinatorics* **31** (2015, no. 6) 2043–2074, <https://doi.org/10.1007/s00373-014-1515-0>.
- [CL22] Chase, Z. and Lovett, S., *Approximate Union Closed Conjecture*, 2022, <http://arxiv.org/abs/2211.11689>.
- [Gil22] Gilmer, J., *A Constant Lower Bound for the Union-Closed Sets Conjecture*, 2022, <http://arxiv.org/abs/2211.09055>.
- [Gow16a] Gowers, W. T., *Frankl's Union-Closed Conjecture — a Possible Polymath Project?*, Gowers's Weblog 2016, <https://gowers.wordpress.com/2016/01/21/frankls-union-closed-conjecture-a-possible-polymath-project/>.
- [Gow16b] Gowers, W. T., *FUNC1 — Strengthenings, Variants, Potential Counterexamples*, Gowers's Weblog 2016, <https://gowers.wordpress.com/2016/01/29/func1-strengthenings-variants-potential-counterexamples/>.
- [Gow16c] Gowers, W. T., *FUNC2 — More Examples*, Gowers's Weblog 2016, <https://gowers.wordpress.com/2016/02/08/func2-more-examples/>.
- [Gow16d] Gowers, W. T., *FUNC3 — Further Strengthenings and Variants*, Gowers's Weblog 2016, <https://gowers.wordpress.com/2016/02/13/func3-further-strengthenings-and-variants/>.
- [Gow16e] Gowers, W. T., *FUNC4 — Further Variants*, Gowers's Weblog 2016, <https://gowers.wordpress.com/2016/02/22/func4-further-variants/>.

References (ii)



- [Kar17] Karpas, I., *Two Results on Union-Closed Families*, 2017, <http://arxiv.org/abs/1708.01434>.
- [Lo 94] Lo Faro, G., *Union-Closed Sets Conjecture: Improved Bounds*, Journal of Combinatorial Mathematics and Combinatorial Computing **16** (1994) 97–102.
- [Nau85] Naur, P., *Programming as Theory Building*, Microprocessing and Microprogramming **15** (1985, no. 5) 253–261, <https://www.sciencedirect.com/science/article/pii/0165607485900328>.
- [Peb22] Pebody, L., *Extension of a Method of Gilmer*, 2022, <https://arxiv.org/abs/2211.13139>.
- [Pol16] Polymath, *Polymath 11 Is Now Open*, The polymath blog 2016, <https://polymathprojects.org/2016/02/07/polymath-11-is-now-open/>.
- [Rei00] Reinhold, J., *Frankl's Conjecture Is True for Lower Semimodular Lattices*, Graphs and Combinatorics **16** (2000, no. 1) 115–116, <https://doi.org/10.1007/s003730050008>.
- [Saw22] Sawin, W., *An Improved Lower Bound for the Union-Closed Set Conjecture*, 2022, <https://arxiv.org/abs/2211.11504>.
- [VŽ17] Vučković, B. and Živković, M., *The 12-Element Case of Frankl's Conjecture*, IPSI Transactions on Internet Research **13** (2017, no. 1) 65–71, <http://ipsitransactions.org/journals/papers/tir/2017jan/p9.pdf>.