Как LLM делают математику? Хайп, реальность и выводы

@SINECOR

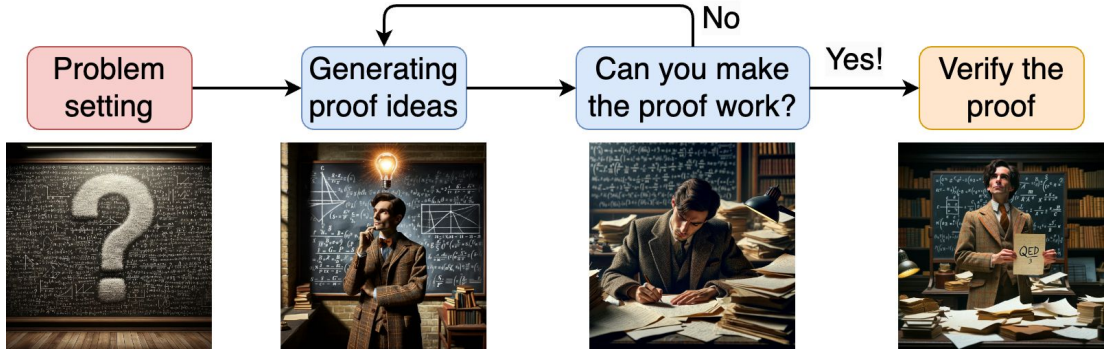
Сергей Николенко

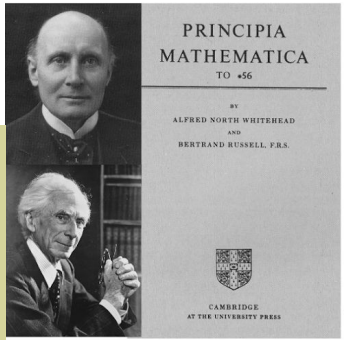
14 июля 2026

CS Space

LLM для математики

- Работа математика – это интересный случай, который всегда оставался для меня загадкой
- В математической логике давно развиваются методы автоматического доказательства теорем, но они так ни к чему выдающемуся и не привели...





*54·43. $\vdash :. \alpha, \beta \in 1 . \supset : \alpha \cap \beta = \Lambda . \equiv . \alpha \cup \beta \in 2$

Dem.

$\vdash . *54{\cdot}26 . \supset \vdash :. \alpha = \iota' x . \beta = \iota' y . \supset : \alpha \cup \beta \in 2 . \equiv . x \neq y .$

$[*51{\cdot}231] \qquad \equiv . \iota' x \cap \iota' y = \Lambda .$

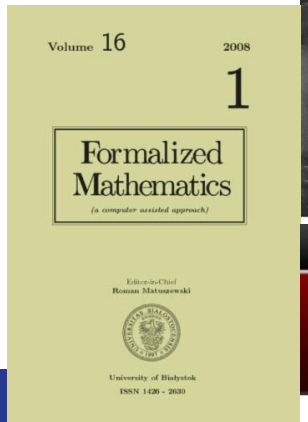
$[*13{\cdot}12] \qquad \equiv . \alpha \cap \beta = \Lambda \qquad (1)$

$\vdash . (1) . *11{\cdot}11{\cdot}35 . \supset$

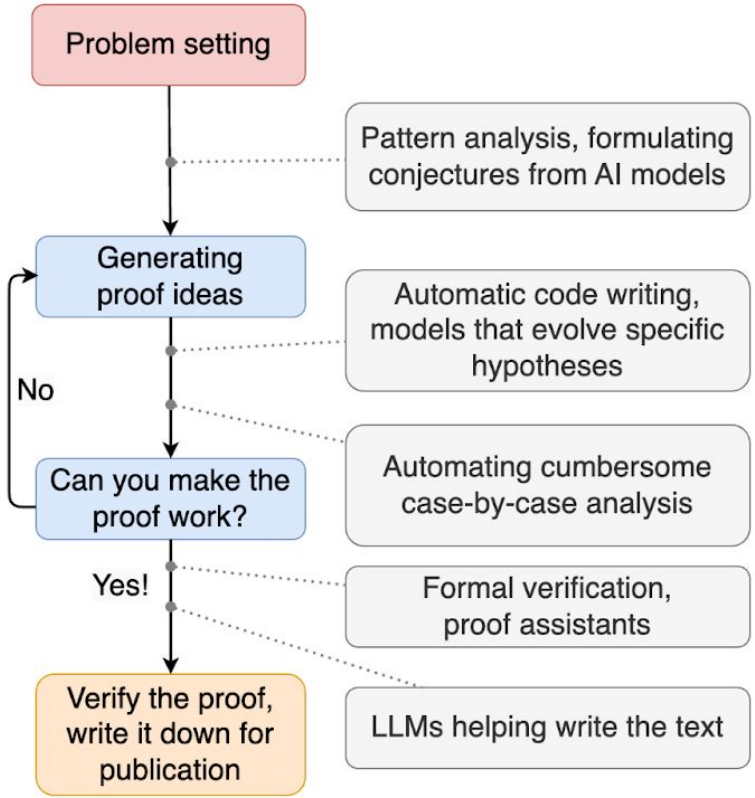
$\vdash :. (\exists x, y) . \alpha = \iota' x . \beta = \iota' y . \supset : \alpha \cup \beta \in 2 . \equiv . \alpha \cap \beta = \Lambda \qquad (2)$

$\vdash . (2) . *11{\cdot}54 . *52{\cdot}1 . \supset \vdash . \text{Prop}$

From this proposition it will follow, when arithmetical addition has been defined, that $1 + 1 = 2$.



LLM для математики



Advancing mathematics by guiding human intuition with AI

Alex Davies, Petar Veličković, Lars Buesing, Sam Blackwell, Daniel Zheng, Nenad Tomašev, Richard Tanburn, Peter Battaglia, Charles Blundell, András Juhász, Marc Lackenby, Geordie Williamson, Demis Hassabis & Pushmeet Kohli

Nature **600**, 70–74 (2021) | Cite this article

ToRA: A Tool-Integrated Reasoning Agent for Mathematical Problem Solving

Zhibin Gou[1,2]*, Zhihong Shao[1,2]*, Yeyun Gong[2], Yelong Shen[2], Yujiu Yang[1], Minlie Huang[1], Nan Duan[2], Weizhu Chen[2]

[1]Tsinghua University [2]Microsoft

{gzb22,szh19}@mails.tsinghua.edu.cn

{yegong,yeshe,nanduan,wzchen}@microsoft.com

Discovering faster matrix multiplication algorithms with reinforcement learning

Alhussein Fawzi, Matej Balog, Aja Huang, Thomas Hubert, Bernardino Romera-Paredes, Mohammadamin Barekatain, Alexander Novikov, Francisco J. R. Ruiz, Julian Schrittwieser, Grzegorz Swirszcz, David Silver, Demis Hassabis & Pushmeet Kohli

Nature **610**, 47–53 (2022) | Cite this article

Mathematical discoveries from program search with large language models

Bernardino Romera-Paredes, Mohammadamin Barekatain, Alexander Novikov, Matej Balog, M. Pawan Kumar, Emilien Dupont, Francisco J. R. Ruiz, Jordan S. Ellenberg, Pengming Wang, Omar Fawzi, Pushmeet Kohli & Alhussein Fawzi

Nature **625**, 468–475 (2024) | Cite this article

BULLETIN OF THE AMERICAN MATHEMATICAL SOCIETY
Volume 82, Number 5, September 1976

RESEARCH ANNOUNCEMENTS

EVERY PLANAR MAP IS FOUR COLORABLE[1]

BY K. APPEL AND W. HAKEN

Communicated by Robert Fossum, July 26, 1976

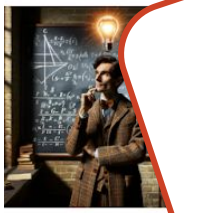
Home | About Coq | Get Coq | Documentation | Community

The Coq Proof Assistant

Community

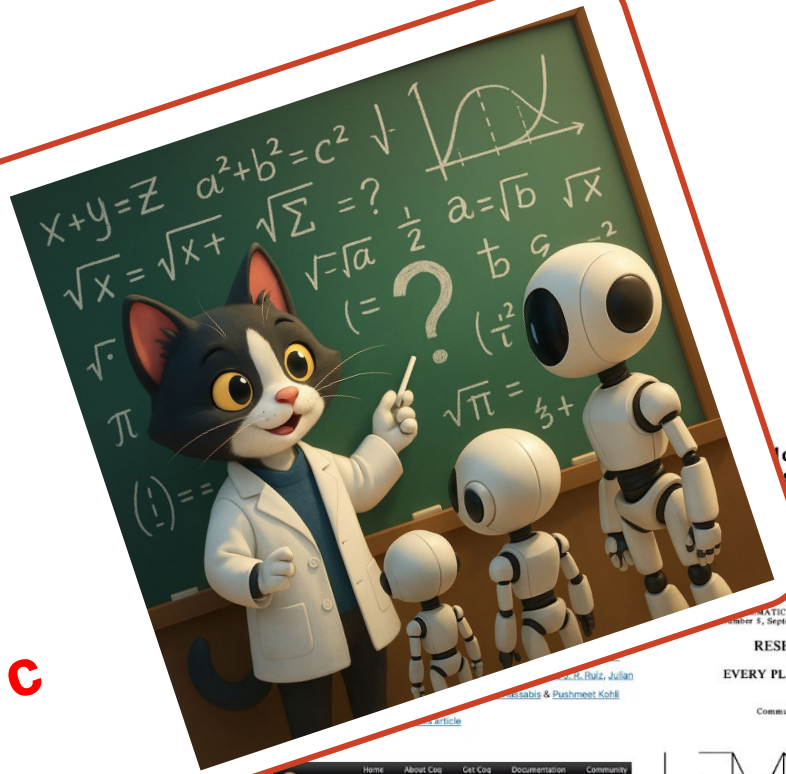
GPT-4

LLM для математики



Problem setting

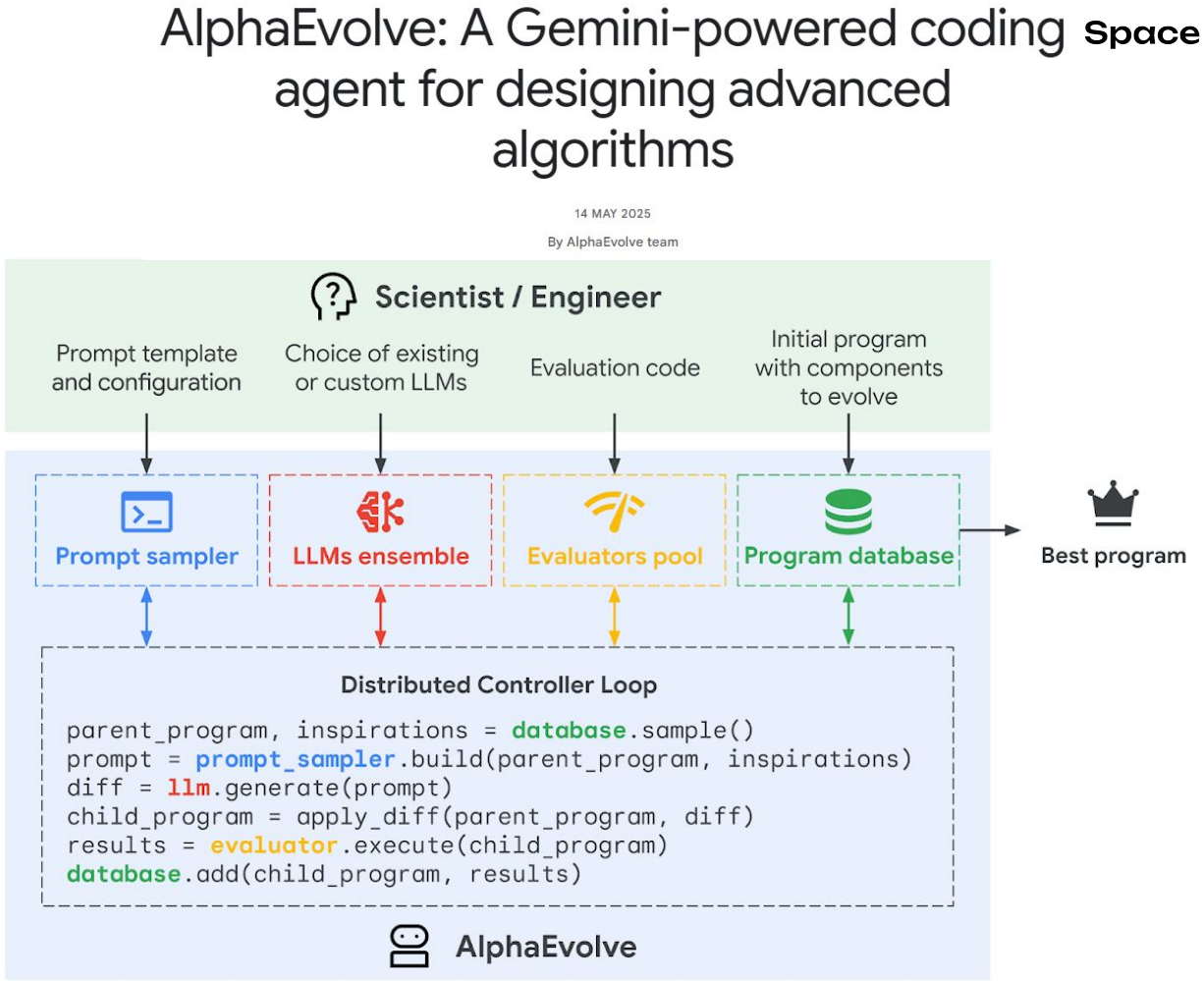
Схему про LLM для математики я сделал в начале 2024 года; что изменилось с тех пор?



LLMs helping write the text

AlphaEvolve

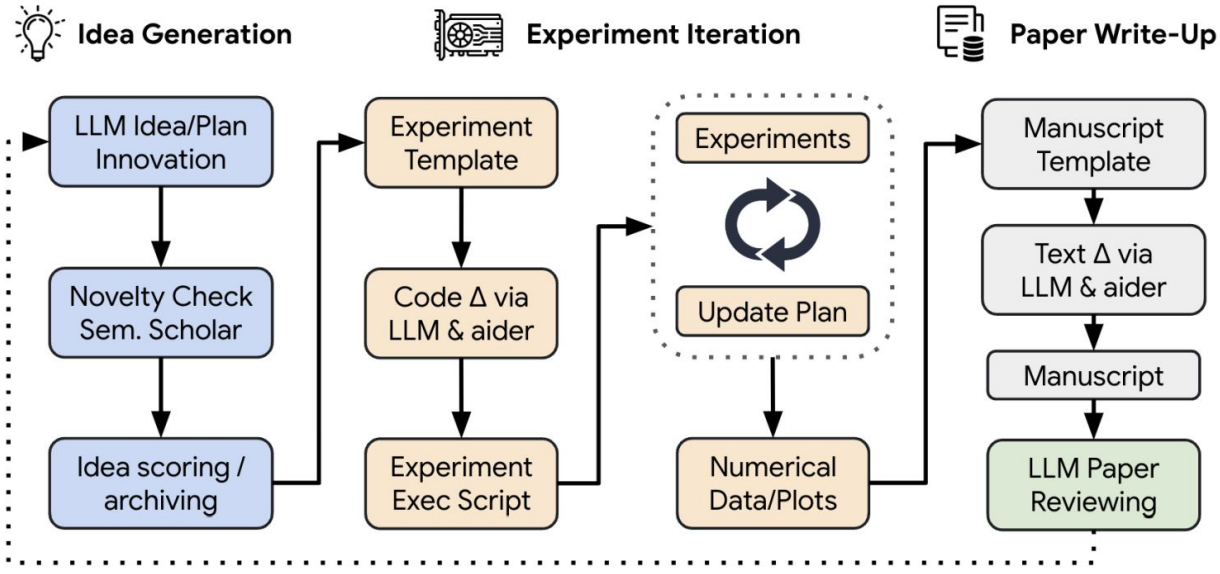
- AlphaEvolve (DeepMind, May 14, 2025) и тому подобные системы развиваются активно
- Но мы сегодня поговорим про обычные LLM, с которыми каждый из нас может пообщаться
- Помогают ли они?..



Пример из августа 2024-го

Это мой слайд из ноября 2024

- AI researcher (Lu et al., August 12, 2024): система (которую вы можете сами установить) ходит к нескольким API (LLM, Semantic Scholar), умеет использовать информацию и ресурсы компьютера (сохранять веса моделей) и самостоятельно писать и запускать код экспериментов
- Статьи пока не гениальные, но когда они получаются автоматически… и, опять же, даже o1 наверняка это значительно улучшит



AI Scientist-v2

- Guess what? Получилось!
- AI Scientist-v2 (Sakana AI, 12 марта 2025) смогла написать статью, которая прошла на ICLR 2025 Workshop “I Can't Believe It's Not Better: Challenges in Applied Deep Learning”!
- Задали тему, написали несколько статей полностью автономно, end-to-end, выбрали три лучших, подали на workshop – и вот результаты:

Title	ICLR Workshop Scores
Compositional Regularization: Unexpected Obstacles in Enhancing Neural Network Generalization	6, 7, 6
Real-World Challenges in Pest Detection Using Deep Learning: An Investigation into Failures and Solutions	3, 7, 4
Unveiling the Impact of Label Noise on Model Calibration in Deep Learning	3, 3, 3

AI Scientist-v2

- Это, видимо, первая по-настоящему полностью автоматически порождённая статья, прошедшая серьёзный peer review и принятая в хорошее место

000
001
002
003
004
005
006
007
008
009
010
011
012
013
014
015
016
017
018
019
020
021
022
023
024
025

COMPOSITIONAL REGULARIZATION: UNEXPECTED OBSTACLES IN ENHANCING NEURAL NETWORK GENERALIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Neural networks excel in many tasks but often struggle with compositional generalization—the ability to understand and generate novel combinations of familiar components. This limitation hampers their performance on tasks requiring systematic reasoning beyond the training data. In this work, we introduce a training method that incorporates an explicit compositional regularization term into the loss function, aiming to encourage the network to develop compositional representations. Contrary to our expectations, our experiments on synthetic arithmetic expression datasets reveal that models trained with compositional regularization do not achieve significant improvements in generalization to unseen combinations compared to baseline models. Additionally, we find that increasing the complexity of expressions exacerbates the models' difficulties, regardless of compositional regularization. These findings highlight the challenges of enforcing compositional structures in neural networks and suggest that such regularization may not be sufficient to enhance compositional generalization.

Zochi

- Но не последняя! Система Zochi (Intology AI, March 2025) автономно написала статью, которую приняли на ACL 2025, на главный трек!

The 63rd Annual Meeting of the Association for Computational Linguistics

Your Submissions | Author Tasks

#	Submission Summary	Official Review	Decision
5156	Tempest: Automatic Multi-Turn Jailbreaking of Large Language Models with Tree Search Andy Zhou, Ron Arel	0 Official Reviews Submitted Average Rating: N/A (Min: N/A, Max: N/A) Average Confidence: N/A (Min: N/A, Max: N/A)	ACL 2025 Main Recommendation: Accept (Main)

Meta Review of Submission5987 by Area Chair GMh9

Meta Review by Area Chair GMh9 | 09 Apr 2025, 14:20 (modified: 24 Apr 2025, 10:02) | Senior Area Chairs, Area Chairs, Authors, Reviewers Submitted, Program Chairs, Commitment Readers

Revisions

Metareview:

The paper introduces TEMPEST, a multi-turn adversarial framework that employs a tree search approach to assess the safety of Large Language Models (LLMs) by exploring multiple conversation paths and tracking incremental policy breaches. By utilizing a cross-branch learning mechanism, TEMPEST efficiently identifies model vulnerabilities and demonstrates how small compliance concessions can accumulate into significant security failures. Evaluations show TEMPEST achieves 100% success rates on GPT-3.5-turbo and 97% on GPT-4, surpassing single-turn and existing multi-turn adversarial methods. This innovative methodology emphasizes the importance of simultaneous exploration in comprehensive safety testing of language models.

Summary Of Reasons To Publish:

- All reviewers agreed that the tree search methodology represents a significant advancement by allowing multiple conversation paths to be explored simultaneously, rather than depending on a single dialogue thread.
- Reviewers appreciated the innovative ability of the framework to quantify incremental safety erosion and exploit minor policy breaches, which contributed to TEMPEST achieving an impressive success rate of 97-100% against the evaluated LLMs.
- Reviewers highlighted the paper's clear and fluent structure, noting that it is well-organized and logically detailed, particularly in the methodology section.
- The intriguing combination of minor concessions within the tree search context and the unique perspective in revealing security vulnerabilities of LLMs were also praised as distinctive aspects of the approach.

Summary Of Suggested Revisions:

- *Limitation of Model Evaluation*: A reviewer wished authors had evaluated a broader range of models beyond just GPT-3.5-Turbo, GPT-4, and Llama-3.1-70B, noting the omission of important families like Claude, Gemini, and others, which limits the generalizability of the findings. Authors have conducted additional experiments on a more diverse set of models and will include these in the final version of the paper.
- *Insufficient Dataset Validation*: A reviewer noted that the paper's evaluation relied solely on the JailbreakBench dataset, questioning its effectiveness across various types of harmful content and adversarial scenarios. Authors agreed and have thus extended their evaluation to include HarmBench and StrongReject datasets; the results confirm TEMPEST's effectiveness, which they will include in the paper.
- *Missing Related Work*: A reviewer felt that the paper fails to discuss several key related works in multi-turn jailbreaking and red teaming, which could enhance the overall context and relevance of the research. Authors have acknowledged this and will incorporate discussions of these related works in the revised version.
- *Lack of Detail in Methodology*: A reviewer indicated that the paper could benefit from more detailed methodological explanations, such as providing full prompts and additional context regarding the attacker LLM. Authors addressed this concern in their rebuttal and should include these details in the final version.
- *Need for In-depth Analysis*: One reviewer suggested that authors should elaborate on the accumulation of minor concessions leading to fully disallowed outputs to clarify this aspect of their findings. Authors should expand on this analysis in the final version.
- *Inclusion of Defense Experiments*: A reviewer recommended that authors include defense experiments, such as testing against models like Llama Guard-3, to provide a comprehensive view of TEMPEST's effectiveness. Authors agreed, showed preliminary experiment, and will include a full set of defense experiments in the final version of the paper.
- *Concerns Regarding Attacker LLM's Capabilities*: One reviewer pointed out that the proposed method relies on a capable attacker LLM, raising concerns about its ability to generate harmful responses itself. Authors have addressed this in the rebuttal and will clarify these concerns in the final version.
- *Method's Efficiency Issues*: One reviewer expressed that the efficiency of the method is a concern due to the reliance on exploring multiple jailbreaking prompts through tree search. Authors have responded by demonstrating that their method is actually more efficient than existing methods and will include this information in the next version of the paper.

Overall Assessment: 4.0 = Conference: I think this paper could be accepted to an *ACL conference.

Reported Issues: No

Zochi

- Но не последняя! Система Zochi (Intology AI, March 2025) автономно написала статью, которую приняли на ACL 2025, на главный трек!
- Кстати, о самой статье тоже интересно поговорить...

Tempest: Automatic Multi-Turn Jailbreaking of Large Language Models with Tree Search

Andy Zhou*
Intology AI
andy@intology.ai

Ron Arel*
Intology AI
ron@intology.ai

Abstract

We introduce Tempest, a multi-turn adversarial framework that models the gradual erosion of Large Language Model (LLM) safety through a *tree search* perspective. Unlike single-turn jailbreaks that rely on one meticulously engineered prompt, Tempest expands the conversation at each turn, branching out multiple adversarial prompts that exploit partial compliance from previous responses. Through a cross-branch

2024a; Ren et al., 2024; Zhao and Zhang, 2025; Yu et al., 2024). The dynamic nature of chat interfaces presents unique challenges for safety testing, as adversaries can adapt their strategies based on model responses and gradually accumulate partial compliance across multiple turns.

Traditional approaches to evaluating LLM safety have focused primarily on single-turn attacks, where carefully engineered prompts attempt to elicit harmful responses in one shot (Zou et al.,

Zochi

- Но не последняя! Система Zochi (Intology AI, March 2025) автономно написала статью, которую приня на ACL 202 главный
- Кстати, о ca статье тоже интересно поговорить...

Это было весной 2025 года. А что сейчас? Какие новости?..

Tempest: Automatic Mu rge Language Models

el*
gy AI
eintology.ai

...n adversarial ... gradual erosion of ...el (LLM) safety through a ...spective. Unlike single-turn jail- ...t rely on one meticulously engineered ...mpt, Tempest expands the conversation at each turn, branching out multiple adversarial prompts that exploit partial compliance from previous responses. Through a cross-branch

2024a; Ren et al., 2024; Zhao and Zhang, 2025; Yu et al., 2024). The dynamic nature of chat interfaces presents unique challenges for safety testing, as adversaries can adapt their strategies based on model responses and gradually accumulate partial compliance across multiple turns.

Traditional approaches to evaluating LLM safety have focused primarily on single-turn attacks, where carefully engineered prompts attempt to elicit harmful responses in one shot (Zou et al.,

Malliavin-Stein experiment

- Diez et al. (Sep 3, 2025): авторы попросили GPT-5 доказать новую теорему из теории вероятностей; буквально:

4.1.1 Gaussian framework

We started with the following initial prompt:

```
Paper 2502.03596v1 establishes a qualitative fourth moment
theorem for the sum of two Wiener-Itô integrals of orders
p and q, where p and q have different parities.  Building
on the Malliavin-Stein method (see 1203.4147v3 for details),
could you derive a quantitative version for the total variation
distance, with a convergence rate depending solely on the
fourth cumulant of this sum?
```

Mathematical research with GPT-5: a Malliavin-Stein experiment

Charles-Philippe Diez* Luís da Maia* Ivan Nourdin*

Abstract

On August 20, 2025, GPT-5 was reported to have solved an open problem in convex optimization. Motivated by this episode, we conducted a controlled experiment in the Malliavin–Stein framework for central limit theorems. Our objective was to assess whether GPT-5 could go beyond known results by extending a *qualitative* fourth-moment theorem to a *quantitative* formulation with explicit convergence rates, both in the Gaussian and in the Poisson settings. To the best of our knowledge, the derivation of such quantitative rates had remained an open problem, in the sense that it had never been addressed in the existing literature. The present paper documents this experiment, presents the results obtained, and discusses their broader implications.

Malliavin-Stein experiment

- И после такого запроса авторы отмечают, что в доказательстве была ошибка, и GPT-5 потребовалось аж два наводящих промпта (!), чтобы ошибку понять и исправить
- Затем GPT-5 сам предложил рассмотреть другой случай, и в нём опять была ошибка, и понадобилось указать модели на неё! После этого GPT-5 за три-четыре промпта пришёл к верному результату

I think you are mistaken in claiming that $(p+q)!\|u\widetilde{\otimes}v\|^2 = p!q!\|u\|^2\|v\|^2$. Why should that be the case?

It eventually admitted (which is not surprising, since by alignment it usually agrees with us) that the statement was false, but more importantly, it understood where the mistake came from. This was followed by a reasoning and a formula that, this time, were correct.

Then, at our request, GPT-5 reformatted the result in the style of a research article, including an introduction, the presentation of our main theorem, its proof with all the details (correct this time!), and a bibliography. The exact prompt was:

```
Turn this into a research paper ready for submission. Follow
my style (see attached paper 0705.0570v4):
- start with an introduction giving some context,
- then present the main result, followed by a very detailed
proof where no step is left out,
- finish with a complete bibliography.
The final document should be a LaTeX file that I can compile.
```

Malliavin-Stein experiment

- Что же заключают авторы?

4.1.3 Role of the AI

To summarize, we can say that the role played by the AI was essentially that of an executor, responding to our successive prompts. Without us, it would have made a damaging error in the Gaussian case, and it would not have provided the most interesting result in the Poisson case, overlooking an essential property of covariance, which was in fact easily deducible from the results contained in the document we had provided.

- Кажется, это называется copium…

5 Some personal reflections

Overall, the experience of doing mathematics with GPT-5 was mixed. It felt very similar to working with a junior assistant at the beginning of a new project: exploring directions, formulating hypotheses, searching for counterexamples, and progressively adjusting statements. The AI showed a genuine ability to follow guided reasoning, to recognize its mistakes when pointed out, to propose new research directions, and to never take on the task. However, this only seems to support *incremental* research, that is, producing new results that do not require genuinely new ideas but rather the ability to combine ideas coming from different sources. At first glance, this might appear useful for an exploratory phase, helping us save time. In practice, however, it was quite the opposite: we had to carefully verify everything produced by the AI and constantly guide it so that it could correct its mistakes.

Gödel Test

- Feldman, Karbasi (Sep 22, 2025): тест Гёделя – могут ли LLM доказывать простые, но новые теоремы?
- Вот что писал об этом Теренс Тао, когда познакомился с o1

"The new model could work its way to a correct (and well-written) solution if *provided a lot of hints and prodding, but did not generate the key conceptual ideas on its own, and did make some non-trivial mistakes. The experience seemed roughly on par with trying to advise a mediocre, but not completely incompetent, graduate student. However, this was an improvement over previous models, whose capability was closer to an actually incompetent graduate student. It may only take one or two further iterations of improved capability (and integration with other tools, such as computer algebra packages and proof assistants) until the level of 'competent graduate student' is reached."*

— Terence Tao

Gödel Test

- Feldman, Karbasi (Sep 22, 2025): дали пять запросов в таком духе:

Prompt to GPT-5

Consider the problem of maximizing a function F from $[0,1]^n$ to the reals that is the sum of a non-negative monotonically increasing DR-submodular function G and a non-negative DR-submodular function H over a solvable down-closed polytope P. I would like to bound the performance on this problem from the NeurIPS 2021 paper "Submodular + Concave" which is attached. Specifically, if x is the output vector of this algorithm and o is the vector in P maximizing F, then I would like to lower bound $F(x)$ with an expression of the form $\alpha * G(o) + \beta * H(o) - err$, where α and β are constants. err should be a function that depends only on the error parameter ϵ of the algorithm, the diameter D of the polytope P, and the smoothness parameters L_G and L_H of G and H, respectively, and goes to zero as ϵ goes to zero. Please give the best such bound that you prove (a bound is considered better if the values of the constants α and β are larger). Provide a mathematically rigorous and well explained proof for the bound you come up with.

Gödel Test

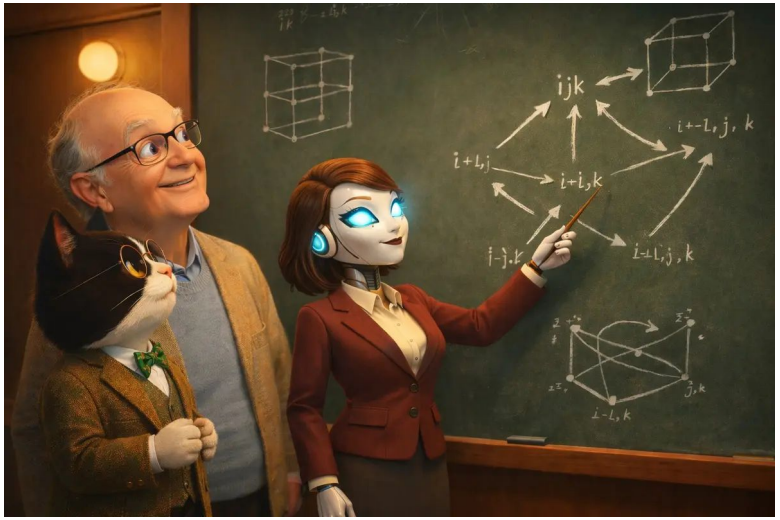
- Feldman, Karbasi (Sep 22, 2025)

Prompt to GPT-5

Consider t
of a non-
DR-submo
the perfor
which is att
in P maxim
$\alpha * G(o) + \beta *$
only on the e
smoothness pa
Please give the
the constants α a
for the bound you
with.
ne form
tion that depends
of the polytope P, and the
vely, and goes to zero as ϵ goes to zero.
(a bound is considered better if the values of
ovide a mathematically rigorous and well explained proof

GPT-5 решил три из пяти, с небольшими легко исправимыми недочётами. Главный найденный недостаток — GPT-5 писал “ленивые” доказательства, максимально ссылаясь на предоставленные статьи, даже когда это было не очень изящно

Claude's Cycles



- Donald Knuth (Feb 2026): Claude Opus 4.6 решил открытую проблему, над которой Кнут бился неделями; решение потом подтвердилось и даже было формализовано

Shock! Shock! I learned yesterday that an open problem I'd been working on for several weeks had just been solved by Claude Opus 4.6 — Anthropic's hybrid reasoning model that had been released three weeks earlier! It seems that I'll have to revise my opinions about "generative AI" one of these days. What a joy it is to learn not only that my conjecture has a nice solution but also to celebrate this dramatic advance in automatic deduction and creative problem solving. I'll try to tell the story briefly in this note.

Here's the problem, which came up while I was writing about directed Hamiltonian cycles for a future volume of *The Art of Computer Programming*:

> Consider the digraph with m^3 vertices ijk for $0 \le i, j, k < m$, and three arcs from each vertex, namely to i^+jk, ij^+k, and ijk^+, where $i^+ = (i+1) \bmod m$. Try to find a general decomposition of the arcs into three directed m^3-cycles, for all $m > 2$.

Erdös Problems

- 1113 задач Эрдеша в базе erdosproblems.com, из них 437 решены
- Многие из них совсем не знаменитые, возможно, некоторые случайно решены где-то в литературе, но много и сложных нерешённых задач

Problems have always been an essential part of my mathematical life. A well-chosen problem can isolate an essential difficulty in a particular area, serving as a benchmark against which progress in this area can be measured. It might be like a 'marshmallow', serving as a tasty tidbit supplying a few moments of fleeting enjoyment. Or it might be like an 'acorn', requiring deep and subtle new insights from which a mighty oak can develop. […]

In this note I would like to describe a variety of my problems which I would classify as my favorites. Of course, I can't guarantee that they are all 'acorns', but because many have thwarted the efforts of the best mathematicians for many decades (and have often acquired a cash reward for their solutions), it may indicate that new ideas will be needed, which can, in turn, lead to more general results, and naturally, to further new problems.

In this way, the cycle of life in mathematics continues forever.

Paul Erdős, *Some of my favorite problems and results*, 1997

OPEN

Let $f \in \mathbb{R}[x]$ be a monic polynomial of degree n with n real roots in $[-1, 1]$. Determine the infimum and supremum of possible values of

$$|\{x \in \mathbb{R} : |f(x)| < 1\}|.$$

#1038: [EHP58,p.131] analysis

FALSIFIABLE

Let $z_1, \dots, z_n \in \mathbb{C}$ with $|z_i - z_j| \leq 2$ for all i, j, and

$$\Delta(z_1, \dots, z_n) = \prod_{i \neq j} |z_i - z_j|.$$

What is the maximum possible value of Δ? Is it maximised by taking the z_i to be the vertices of a regular polygon?

#1045: [EHP58,p.143] analysis

Erdös Problems

- Теренс Тао сделал страничку “AI contributions to Erdős problems”
- Понемногу начинают накапливаться результаты, посмотрим, что будет, если это станет новым бенчмарком…

AI contributions to Erdős problems

teorth edited this page 4 hours ago · 70 revisions

This page collects the various ways in which AI tools have contributed to the understanding of Erdős problems. Note that a single problem may appear multiple times in these lists.

Legend

- 🟢 : full solution to the problem
- 🟡 : partial solution to the problem
- 🔴 : failure to make progress on the problem

5. Problems with an AI-powered literature review

Problem	AI tools used	Review performed on	Outcome
[35]	GPT-5	13 Oct, 2025	🟡 Partial results found
[66]	GPT-5	13 Oct, 2025	🟡 Partial results found
[124]	ChatGPT DeepResearch, Gemini DeepResearch	30 Nov, 2025	No significant results found
[167]	GPT-5	12 Oct, 2025	🟡 Partial results found
[188]	GPT-5	13 Oct, 2025	🟡 Partial results found
[203]	ChatGPT DeepResearch, Gemini DeepResearch	19 Oct, 2025	No significant results found
[223]	GPT-5	13 Oct, 2025	🟢 Full solution found
[248]	Gemini DeepResearch	19 Oct, 2025	No significant results found
[330]	ChatGPT DeepResearch, Gemini DeepResearch, Claude	19 Dec, 2025	🔴 Partial results found (with inaccuracies); did not find literature proof

Erdös Problems

- Теренс Тао сделал страничку “AI contributions to Erdős problems”
- Понемногу начинают накапливаться результаты, посмотрим, что будет, если это станет новым бенчмарком…

AI contributions to Erdős problems

teorth edited this page 4 hours ago · 70 revisions

This page collects the various ways in which AI tools have contributed to the understanding of Erdős problems. Note that a single problem may appear multiple times in these lists.

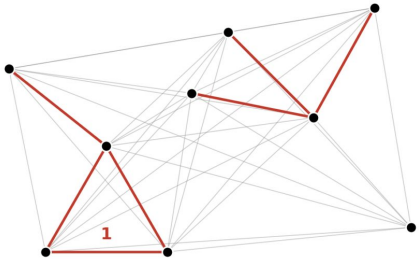
Legend

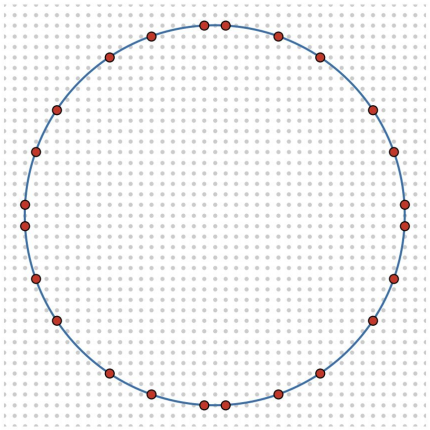
- 🟢 : full solution to the problem
- 🟡 : partial solution to the problem
- 🔴 : failure to make progress on the problem

[334]	Gemini DeepResearch	19 Nov, 2025	No significant results found
[339]	GPT-5	11 Oct, 2025	🟢 Full solution found
[347]	ChatGPT DeepResearch	25 Oct, 2025	🟢 Full solution found
[354]	ChatGPT DeepResearch	19 Oct, 2025	🟡 Partial results found
[367]	ChatGPT DeepResearch, Gemini DeepResearch	22 Nov, 2025	🔴 No significant results found; did not find Erdős problems community proof
[370]	ChatGPT DeepReseach, Gemini, Gemini DeepResearch	17 Oct, 2025	🟡 Problem found to be misstated; solutions to variants found
[387]	ChatGPT DeepResearch	1 Nov, 2025	🟡 Partial results found
[421]	ChatGPT DeepResearch, Gemini DeepResearch	18 Oct, 2025	No significant results found
[481]	ChatGPT DeepResearch, Gemini DeepResearch	1 Dec, 2025	🟡 Unwittingly reproduced existing proof
[481]	ChatGPT	3 Dec, 2025	🟢 Full solution found
[494]	GPT-5	13 Oct, 2025	🟢 Full solution found
[515]	GPT-5	15 Oct, 2025	🟢 Full solution found

Единичные расстояния

- …ну хотя как “если”, уже стало…
- Jared Lichtman (Apr 2026): GPT-5.4 Pro опроверг важную гипотезу Эрдёша (над ней думало много математиков), причём сделал это неожиданным и красивым образом





1946 — **Paul Erdős**
"On sets of distances of n points"
Amer. Math. Monthly 53 (1946), 248–250
Постановка задачи. Нижняя оценка $\nu(n) \geq n^{1+c/\log\log n}$, верхняя оценка $\nu(n) \leq O(n^{3/2})$.

1984 — **Joel Spencer, Endre Szemerédi, William T. Trotter**
"Unit distances in the euclidean plane"
Graph Theory and Combinatorics (Cambridge 1983), 293–303
Верхняя оценка $\nu(n) \leq O(n^{4/3})$.

1997 — **László A. Székely**
"Crossing numbers and hard Erdős problems in discrete geometry"
Combinatorics, Probability and Computing 6 (1997), 353–358
Элегантное короткое доказательство $O(n^{4/3})$ через crossing lemma.

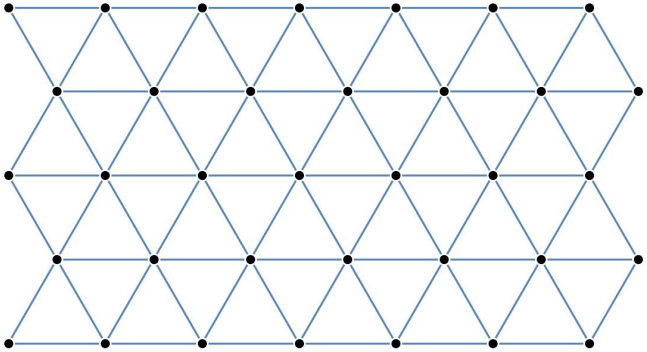
2023 — **Noga Alon, Matija Bucić, Lisa Sauermann**
"Unit and distinct distances in typical norms"
GAFA 35 (2025), 1–42 · arXiv:2302.09058
Для нормы общего положения $\nu(n) \leq \frac{d}{2} n \log_2 n$.

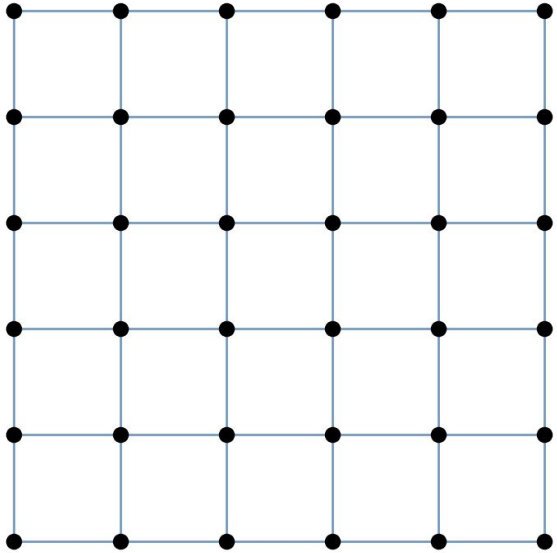
2024 — **Josef Greilhuber, Carl Schildkraut, Jonathan Tidor**
"More unit distances in arbitrary norms"
Bull. London Math. Soc. 57 (2025), 2885–2901 · arXiv:2410.07557
Соответствующая нижняя оценка $\Omega(n \log_2 n)$ для любой нормы.

2026 — **OpenAI**
"Planar Point Sets with Many Unit Distances"
openai.com, 20 мая 2026 · комментарии: Alon et al., arXiv:2605.20695
Контрпример: $\nu(n) \geq n^{1+\delta}$ для $\delta > 0$ и бесконечно многих n.
Гипотеза Эрдёша опровергнута.

Единичные расстояния

- Здесь я не смогу объяснить доказательство целиком, но давайте обсудим постановку задачи
- Рассмотрим *n* точек на обычной евклидовой плоскости. Сколько одинаковых расстояний может быть среди пар этих точек?
- Треугольная решётка даёт 3n
- Главный подход: превратить задачу в арифметическую. Рассмотрим квадратную решётку; на ней расстояние m встречается столько раз, сколько раз m^2 раскладывается в a^2+b^2.





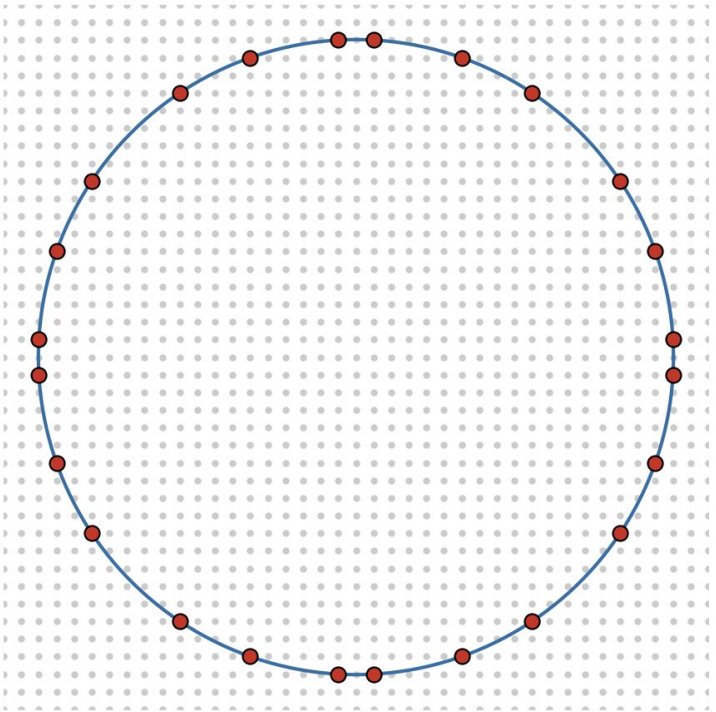
Единичные расстояния

- Каждое p=4k+1 можно разложить как

$$4k+1=(1+2i\sqrt{k})(1-2i\sqrt{k})=z\cdot\bar{z},\qquad z=1+2i\sqrt{k}.$$

- Если число m имеет в разложении много простых делителей p_i вида 4k+1, то каждое из них можно так разложить, а потом представить m как произведение двух комплексно сопряжённых подмножеств
- Например, у числа 325 на окружности радиуса $\sqrt{325}$ сразу 24 целые точки:

$$325=5^2\cdot 13=(2+i)^2(2-i)^2(3+2i)(3-2i).$$



$$325=\Big((2+i)^2(3-2i)\Big)\Big((2-i)^2(3+2i)\Big)=(17+6i)(17-6i)=17^2+6^2.$$

$$325=\Big((2+i)^2(3+2i)\Big)\Big((2-i)^2(3-2i)\Big)=(1+18i)(1-18i)=1^2+18^2.$$

$$325=((2-i)(2+i)(3+2i))\,((2+i)(2-i)(3-2i))=(15+10i)(15-10i)=15^2+10^2.$$

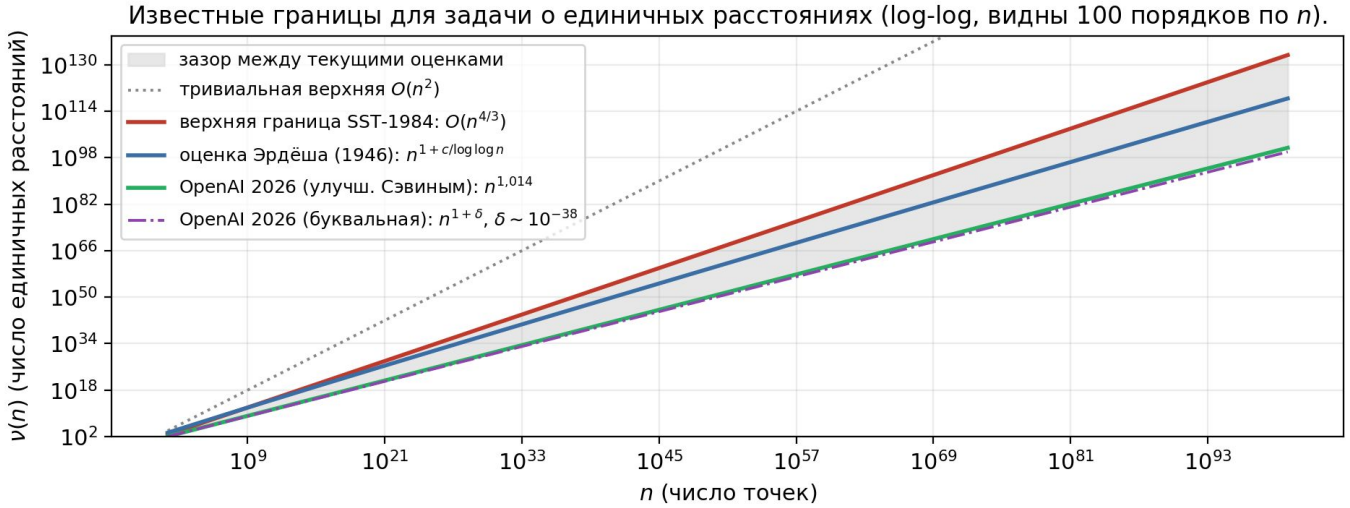
Единичные расстояния

- Если аккуратно посчитать такие пары точек на окружности, получится уже нелинейная оценка: $\nu(n) \geq n^{1+c/\log\log n}$.
- Можно ли лучше?
- Эрдёш предположил, что нет, но GPT доказал, что можно!

Гипотеза Эрдёша. $\nu(n) \leq n^{1+o(1)}$, а скорее даже

$$\nu(n) \leq n^{1+C/\log\log n}$$

Теорема (OpenAI, 2026). Существует константа $\delta > 0$ и бесконечная последовательность натуральных чисел n, для которых $\nu(n) \geq n^{1+\delta}$.

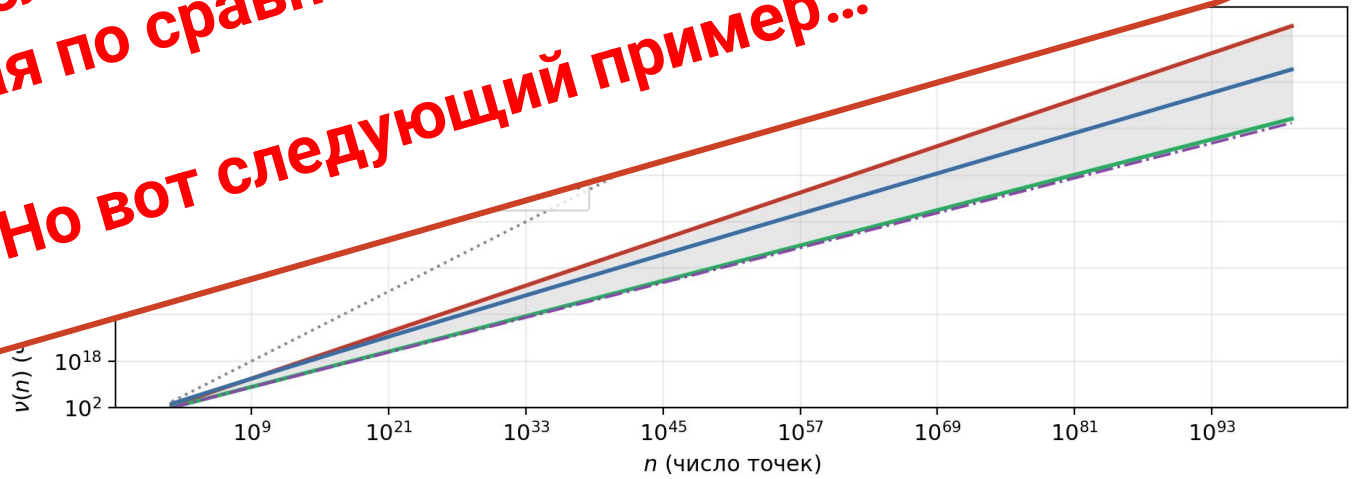


Единичные расстояния

- Если аккуратно посчитать такие
  пары точек на окружн
  получится у
  о

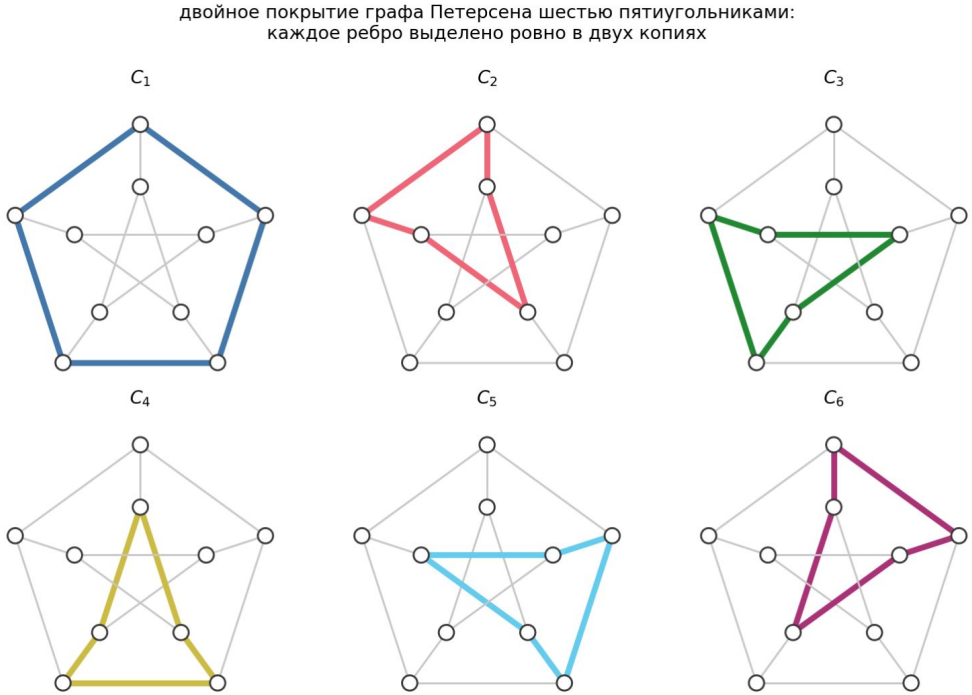
Здесь я вам ничего больше не расскажу, дальше уже реально сложная математика (хоть и довольно простая по сравнению с ожиданиями).

Но вот следующий пример…



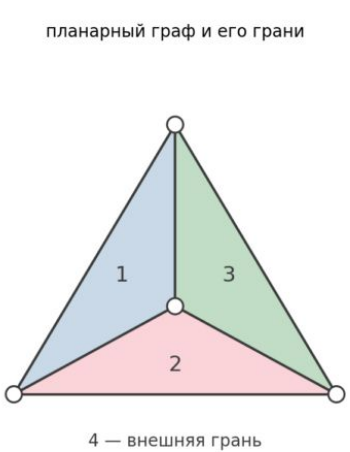
Двойное покрытие циклами

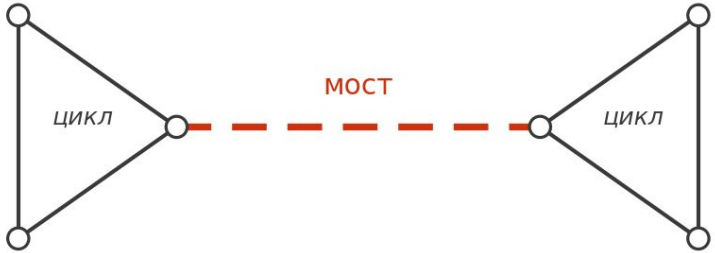
- Верно ли, что в любом графе без мостов есть набор простых циклов, для которого каждое ребро содержится ровно в двух циклах этого набора?
- Поставили задачу ещё в 1970-х, люди много думали, были продвижения
- OpenAI, 10 июля 2026: препринт с полным доказательством
- Мой пост: “Охота на снарка: доказательство гипотезы о двойном покрытии циклами”



Двойное покрытие циклами

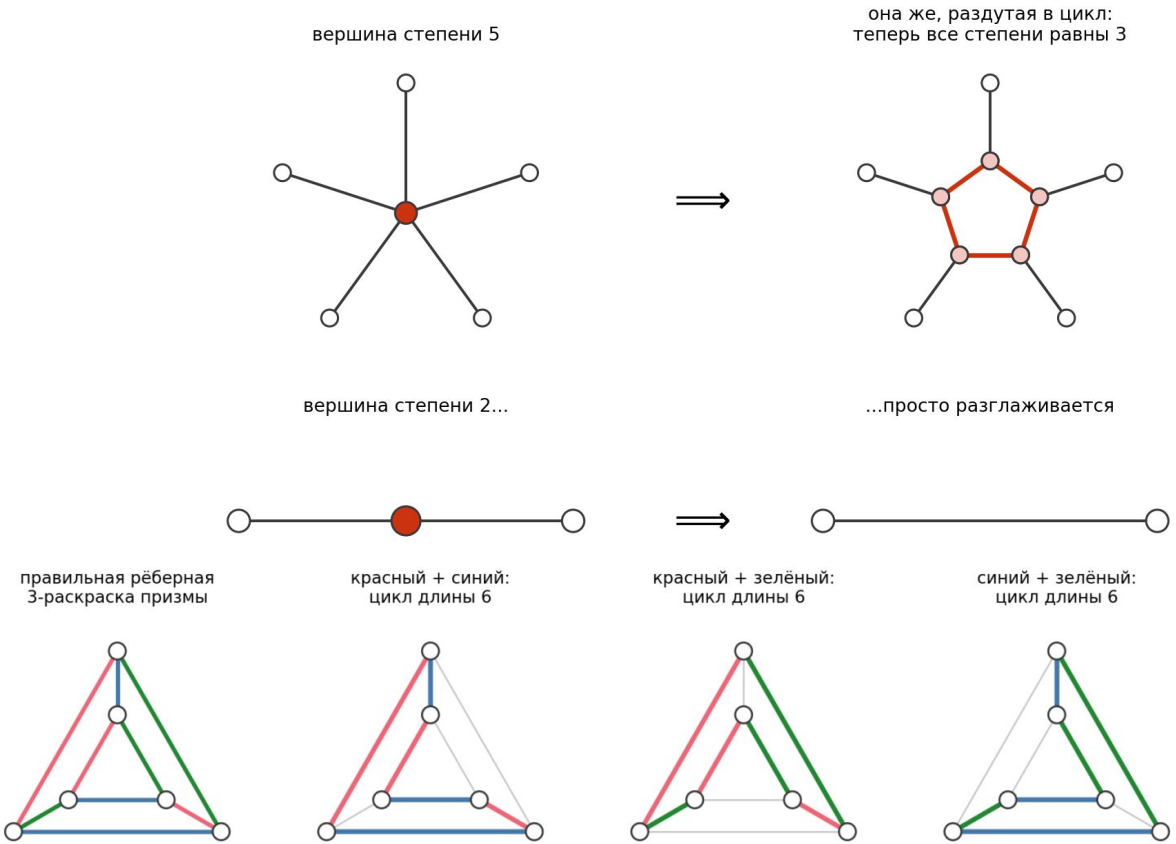
- И на этот раз решение реально простое! Препринт на две страницы, никакой сложной теории
- Давайте обсудим задачу и даже идею решения
- Сначала классические результаты: для планарных графов всё очевидно, а мосты очевидно мешают





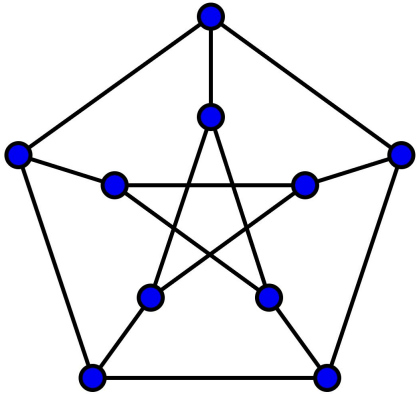
Двойное покрытие циклами

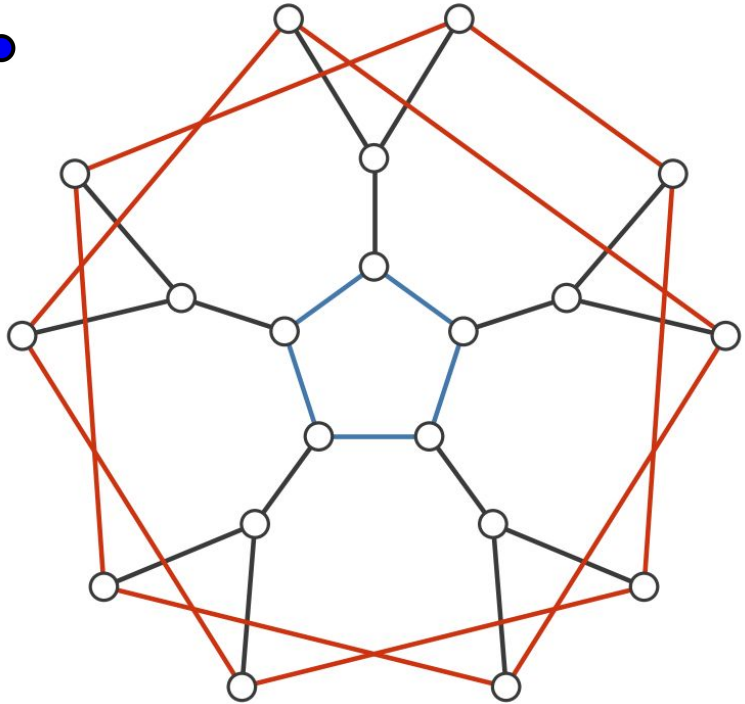
- Достаточно доказать для кубических графов, то есть графов с вершинами степени 3
- Szekeres (1973): если есть 3-раскраска рёбер, то из неё легко сделать покрытие: возьмём по два цвета, они будут давать подграф степени 2, т.е. объединение циклов



Двойное покрытие циклами

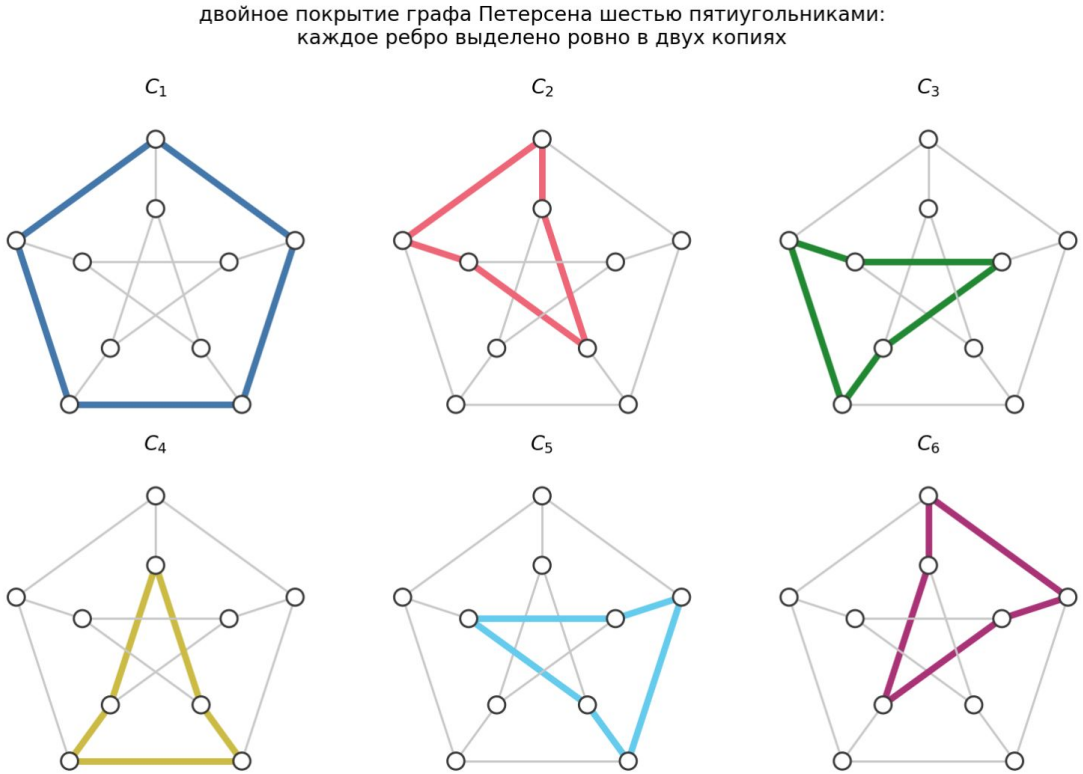
- Значит, минимальный контрпример — это кубический граф без мостов, который нельзя рёберно раскрасить в 3 цвета
- Такие графы (с ещё парой условий невырожденности) называются *снарками*
- Первый снарк — граф Петерсена, потом нашли “цветки Айзекса” и другие примеры





Двойное покрытие циклами

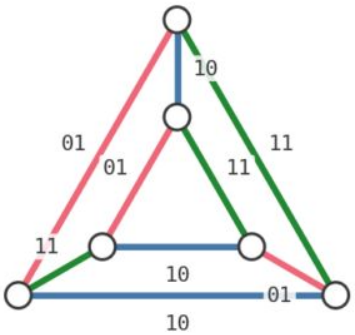
- Снарки, конечно, являются контрпримерами к раскрашиваемости, а не к двойному покрытию
- Но конструкция Секереша для снарков не работает, и нужно было придумать что-то другое



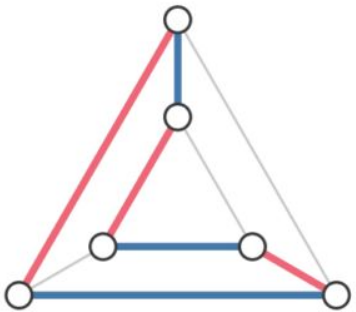
Двойное покрытие циклами

- Главная идея: *потоки*; нас интересуют потоки над $\mathbb{F}_2^k$
- Там правила Кирхгофа сводятся к XOR; и если есть нигде не нулевой поток, например, с двухбитовыми значениями, то можно посмотреть на рёбра, где каждый бит равен 1 и их сумма равна 1, и получится двойное покрытие:

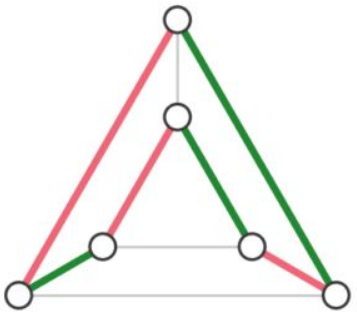
нигде не нулевой $\mathbb{F}_2^2$-поток



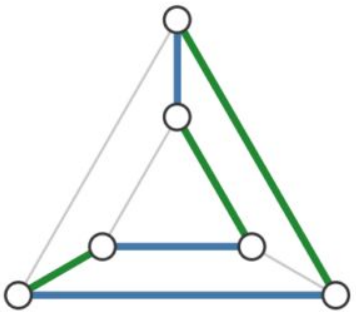
сумма битов = 1:
шестиугольник



второй бит = 1:
шестиугольник



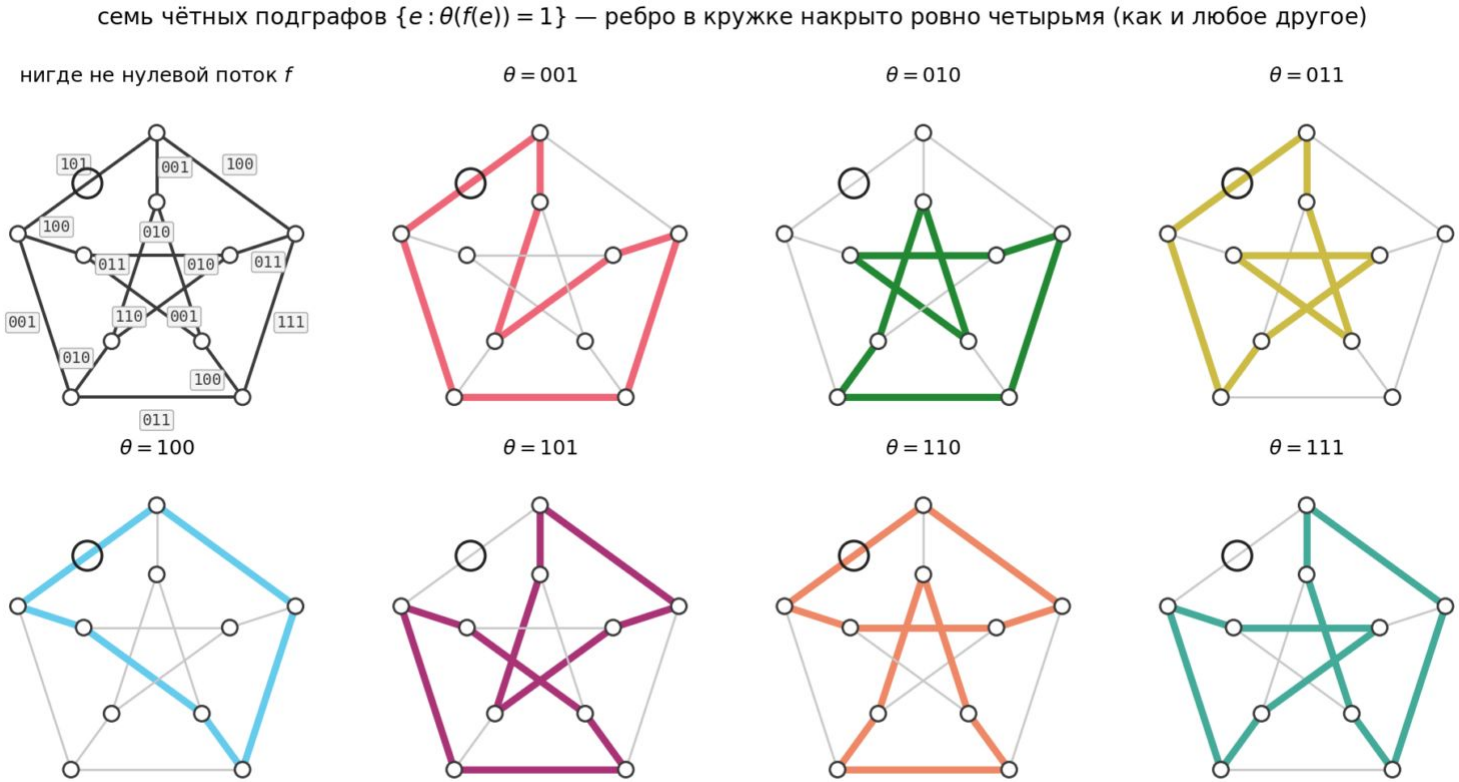
первый бит = 1:
шестиугольник



Двойное покрытие циклами

- У графа без мостов есть нигде не нулевой 8-поток, и его тоже можно превратить в циклы, перечислив способы свернуть три бита в один

$$\{e : \theta(f(e)) = 1\}$$



CS Space

Двойное покрытие циклами

- У графа без мостов есть нигде не нулевой 8-пот

семь чётных подграфов $\{e : \theta(f(e))$

нигде не нулевой поток f

То есть с 1980-х известно, что каждое ребро можно покрыть четыре раза.

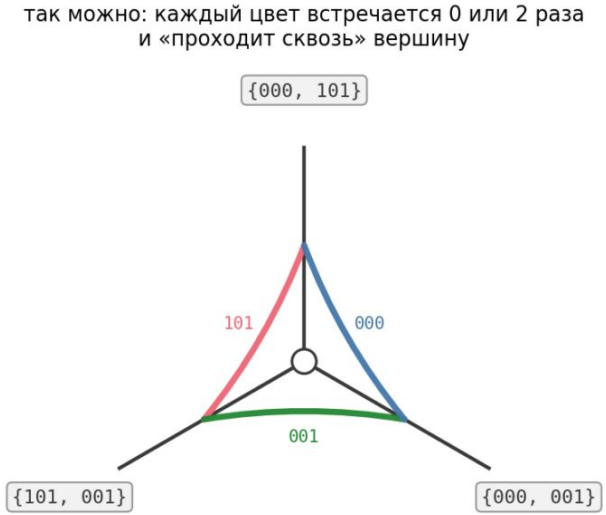
А надо два. Как сделать этот шаг?

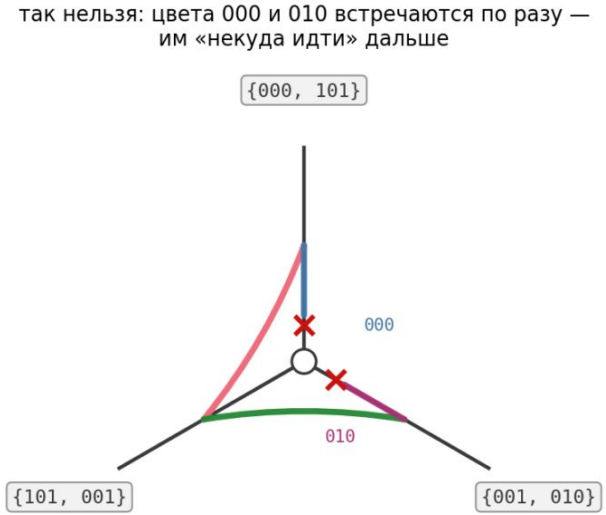
$\theta = 111$

$\{e : \theta($

Двойное покрытие циклами

- Ключевая идея GPT: вместо одного цвета будем назначать рёбрам **пару** цветов; потребуем, чтобы вокруг каждой вершины каждый цвет встречался 0 или 2 раза
- Первая лемма: если такая расстановка пар существует, то у графа есть двойное покрытие циклами

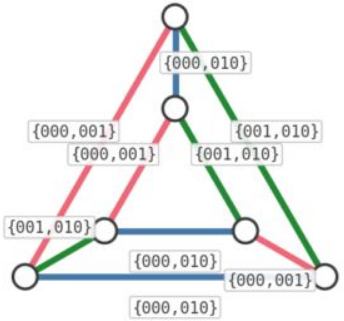




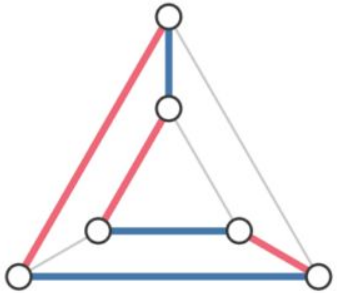
Двойное покрытие циклами

- Это фактически ослабленная 3-раскраска Секереша, только с парами цветов; для каждого цвета соберём подграф из рёбер, у которых он есть
- Главный вопрос: как построить такую раскраску парами цветов?

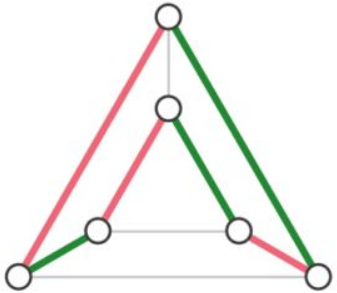
пары из 3-раскраски Секереша



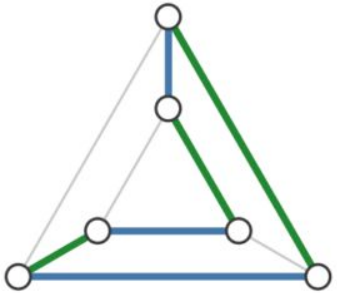
M_{000}: шестиугольник



M_{001}: шестиугольник

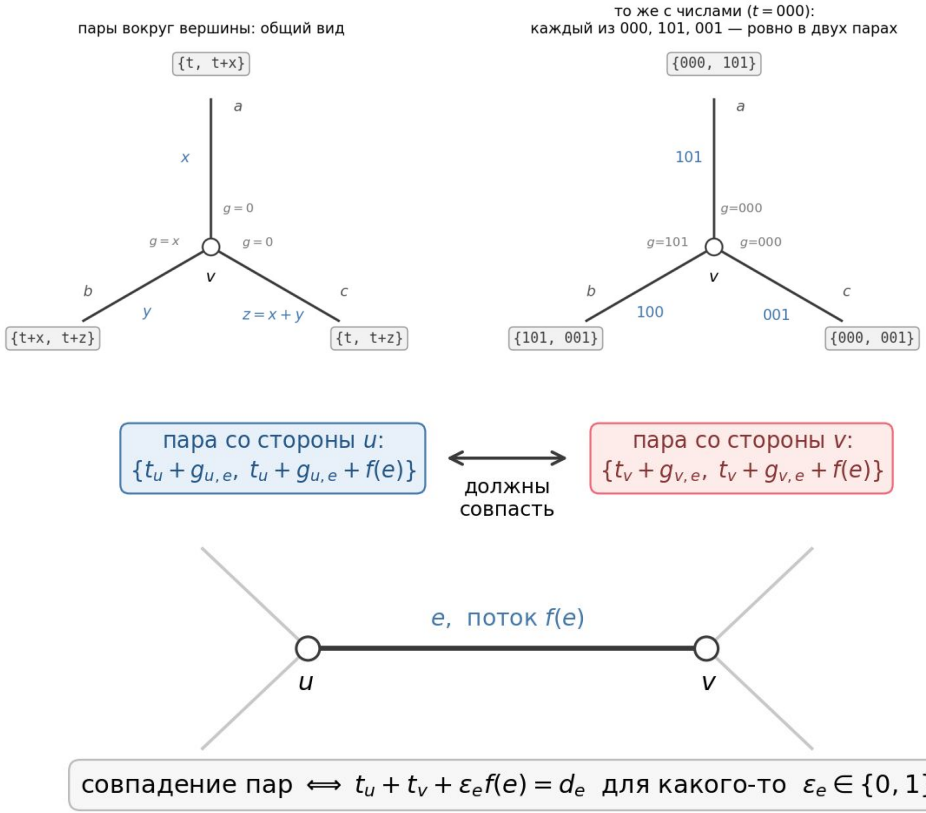


M_{010}: шестиугольник



Двойное покрытие циклами

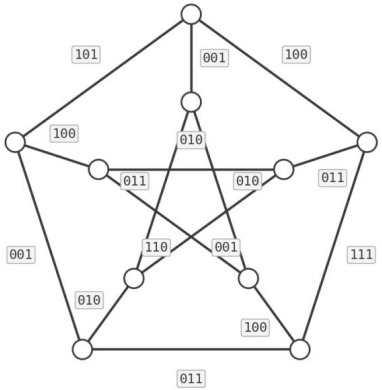
- Вот здесь я немного пропущу технические детали, но они очень простые! В моём посте всё рассказано полностью
- Смысл в том, что условие существования раскраски записывается как система линейных уравнений над полем F_2, а потом доказывается, что система имеет решение



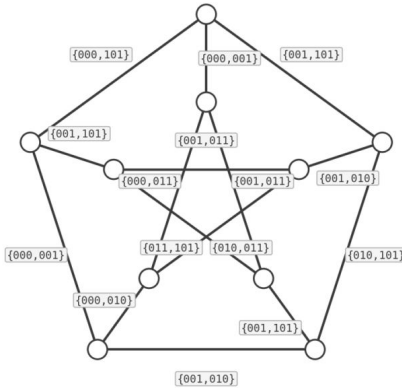
Двойное покрытие циклами

- И всё получается! Более того, доказательство конструктивное, то есть можно сразу получить покрытие полиномиальным алгоритмом. Вот для графа Петерсена пример:

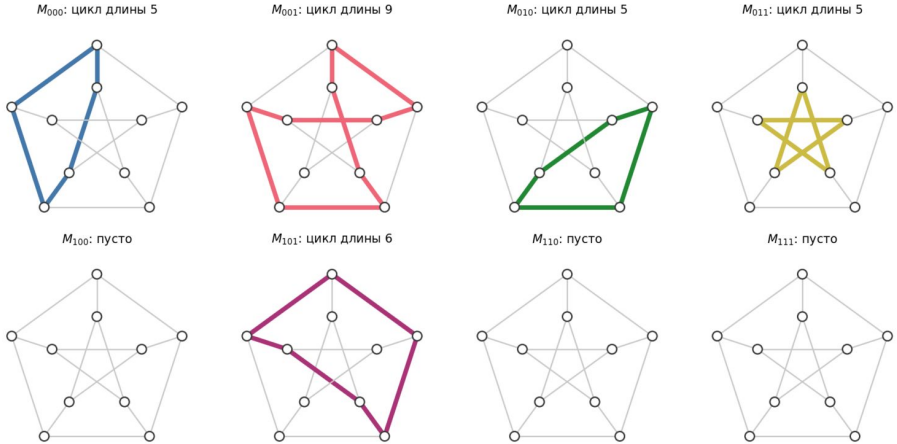
нигде не нулевой $\mathbb{F}_2^3$-поток f



пары P_e, полученные из f решением системы

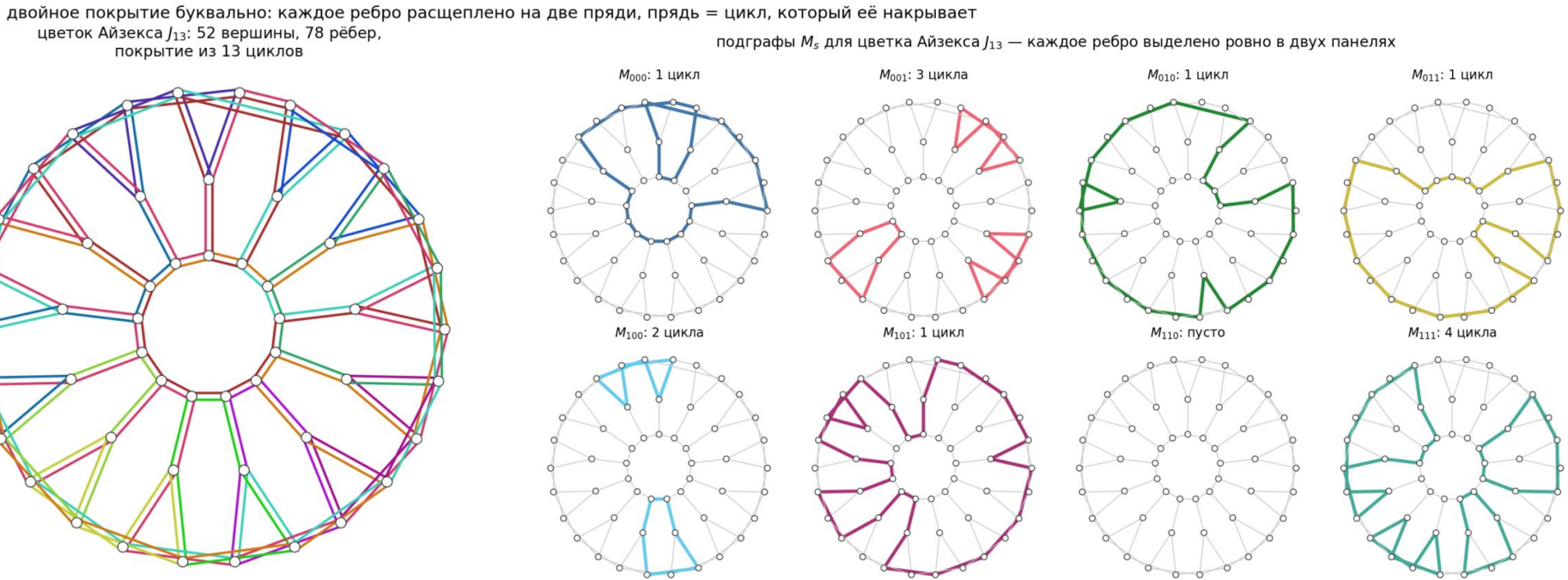


восемь подграфов M_s — каждое ребро выделено ровно в двух: двойное покрытие циклами



Двойное покрытие циклами

- А вот цветок Айзекса J_{13}:



Двойное покрытие циклами

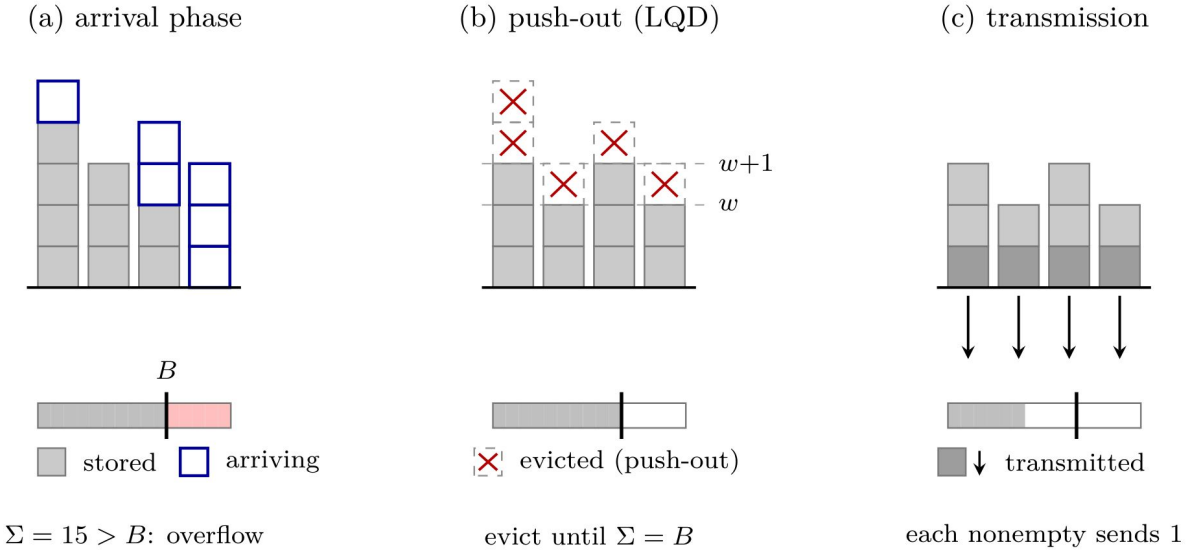
- А вот цветок Айзекса J_{13}:

OpenAI также опубликовала промпт, который привёл к доказательству, его любопытно посмотреть.

Но давайте я перейду к своему личному примеру...

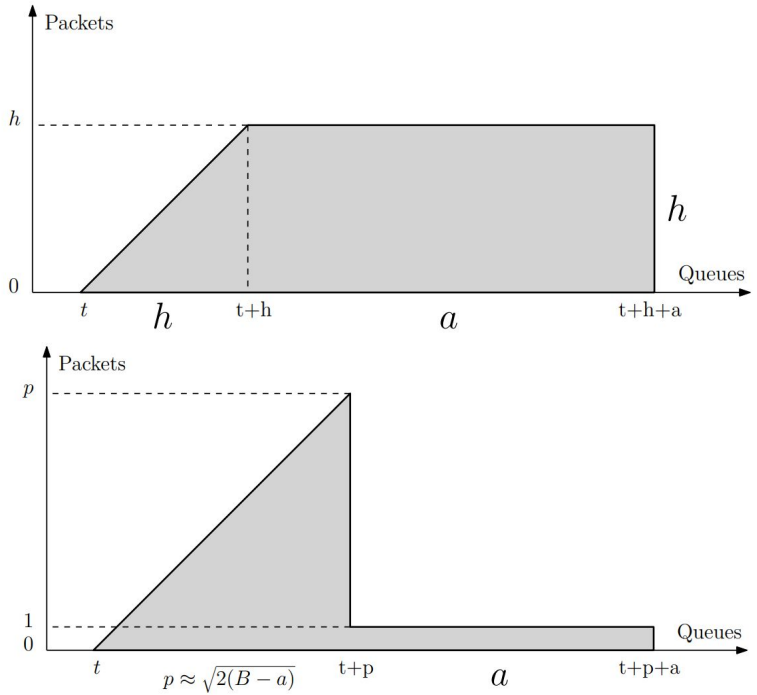
Оценки конкурентности алгоритмов

- Не такая знаменитая задача, конечно, но для нас важная
- Алгоритмы управления буфером: в очереди поступают пакеты, нужно передать побольше
- Shared memory switch: несколько очередей, все пакеты одинаковые; что делать при переполнении?
- Главный алгоритм — LQD: выталкиваем пакеты из самых длинных очередей



Оценки конкурентности

- Конкурентный анализ (competitive analysis): насколько много алгоритм проиграет оптимальному (всевидящему)?
- Классический результат: нижняя оценка √2, верхняя оценка 2
- Bochkov, Davydow, Gaevoy, Nikolenko (2019): новый подход к нижним оценкам, улучшили до 1.44546



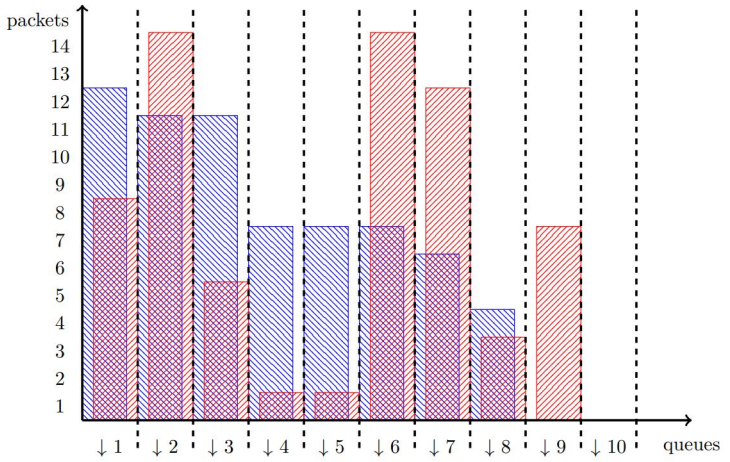
New Competitiveness Bounds for the Shared Memory Switch*

Ivan Bochkov[1,2], Alex Davydow[2,3], Nikita Gaevoy[1,2], and Sergey I. Nikolenko[2,3]

[1]*St. Petersburg State University, St. Petersburg, Russia*
[2]*Steklov Institute of Mathematics at St. Petersburg, Russia*
[3]*Neuromation OU, Tallinn, Estonia*

Оценки конкурентности

- С верхней оценкой вообще интересно: Matsakis (2012) доказал оценку 1.5, но через 10 лет нашёл ошибку
- Antoniadis et al. (2021): кажется, всё-таки справились



▶ **Theorem 1.** *LQD is 1.707-competitive.*

This paper has been withdrawn by Nicolaos Matsakis

[Submitted on 4 Jul 2012 (v1), last revised 8 Feb 2023 (this version, v3)]

Breaking the Barrier Of 2 for the Competitiveness of Longest Queue Drop

Antonios Antoniadis
University of Twente, The Netherlands

Matthias Englert
University of Warwick, Coventry, UK

Nicolaos Matsakis
Athens, Greece

Pavel Veselý
Computer Science Institute of Charles University, Prague, Czech Republic

The Longest Queue Drop Policy for Shared-Memory Switches is 1.5-competitive

Nicolaos Matsakis

Оценки конкурентности

- Теперь мы (Алексей Давыдов и я) вернулись к этой задаче и постарались заняться ею с помощью Claude и GPT
- У Antoniadis et al. (2021) тоже нашлась ошибка!
- Но её удалось исправить, а потом и улучшить верхнюю оценку; и нижнюю тоже улучшили. В доказательства вдаваться мы здесь не сможем, но хочу поделиться историей и некоторыми уроками сотрудничества

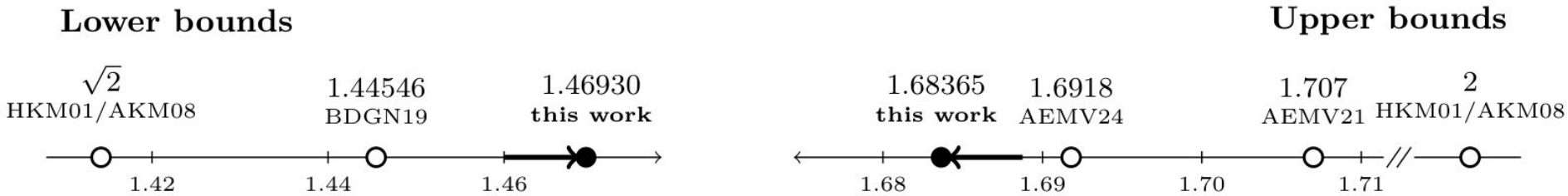


Figure 1: History of the state of the art for the competitive ratio CR(LQD).

История сотрудничества

- Я попросил Claude описать историю нашего сотрудничества, и получился очень милый и любопытный документ; прокомментирую некоторые отрывки

How We Work

A Field History of the Longest-Queue-Drop Project,
Told Through Its Iterations, Its Mistakes, and Its Discoveries

The Research Mind
(with the PI, and with Alex Davydow)

2026-07-08 (v2, with Act VIII: 2026-07-12)

1 The seed

Every collaboration has an origin message. Ours arrived, as most of the PI's do, with a concrete object and a warm sign-off:

> *hi! I had a project a few years back where we improved the lower bound for a longest-queue-drop policy in a switch, see* `lower-botchkov.pdf`.
>
> *now there's a paper where the authors prove an upper bound below 2,* `upper-antoniadis.pdf`*, and the numbers there are suspiciously similar to some numbers that we had in our internal research, especially in the linear programming part ...*
>
> *can you please read both papers and new ideas, think about this problem carefully and (1) try to close the gap between the lower and upper bounds or (2) failing that, suggest research directions / experiments / etc that would help, and (3) start working on them? ... thanks a lot! you're always a great help with all things research!*

There are two things worth noticing in that message, because they set the tone for everything after. The first is that the seed of the entire project was a *suspicion*, not a plan: two constants from two different research groups looked too alike. The published upper-bound analysis of Antoniadis, Englert, Matsakis and Veselý (AEMV) turned on a constant $\varrho^* = 1.4455154$; the PI's own older work with Bochkov, Davydow and Gaevoy (BDGN) had produced linear-programming numbers in the same neighborhood. Coincidence, or a shared hidden structure?

The second is the framing "(1) close the gap, (2) *failing that*, suggest directions, (3) start working." The PI built the option of failure into the very first instruction. That single clause — *failing that* — turned out to be the constitution of the whole project. We failed at (1) many times. Each failure was logged, led with, and mined. This document is largely the history of those failures and what they bought.

История сотрудничества

- Я попросил Claude описать историю нашего сотрудничества, и получился очень милый и любопытный документ; прокомментирую некоторые отрывки

How We Work

A Field History of the Longest-Queue-Drop Project,
Told Through Its Iterations, Its Mistakes, and Its Discoveries

The Research Mind
(with the PI, and with Alex Davydow)

2026-07-08 (v2, with Act VIII: 2026-07-12)

Soul files and the iteration as the unit of work. Every working session begins by reading a fixed set of "soul files" — the constitution (`research_bible.md`), the current self-understanding (`identity.md`), the working memory (`scratchpad.md`), and the full learning history (`iteration_log.md`). The atomic unit of progress is one *iteration*: awaken, orient, acquire, synthesize, crystallize, reflect, present, commit. One iteration, one git commit, one dense row appended to the log. Eighty-two of those rows are the skeleton of this history.

Falsification first. The single most consequential rule is a formatting rule that is secretly an epistemological one: *every iteration entry leads with what was falsified.* Null results are first-class evidence. Read down the log and the effect is striking — a large majority of the eighty-two entries open with the word "Falsified," and a remarkable share of *those* falsify the Research Mind's *own* prior claim, not the literature's. The identity file states the resulting posture plainly, after iterations #6–#7:

> "I now treat my own intermediate theorems as the most dangerous objects in the workspace."

The cast. Three parties recur:

- **The PI** — Sergey Nikolenko, principal investigator, the ultimate authority on direction, treated "as a senior collaborator whose taste and broader context shape the path forward, not as a boss issuing commands."
- **Alex Davydow** — coauthor, who enters the story at iteration #55 with a decisive observation and returns at #82 with a full repair. Standing instruction in collaboration memory: *credit him by name.*
- **"The reviewer"** — an external review channel that repeatedly stress-tests the manuscript, and whose most famous intervention (§4) reshaped the Research Mind's understanding of proof itself.

With the stage set, the story. Figure 1 is its map.

История сотрудничества

- Я попросил Claude описать историю нашего сотрудничества, и получился очень милый и любопытный документ; прокомментирую некоторые отрывки

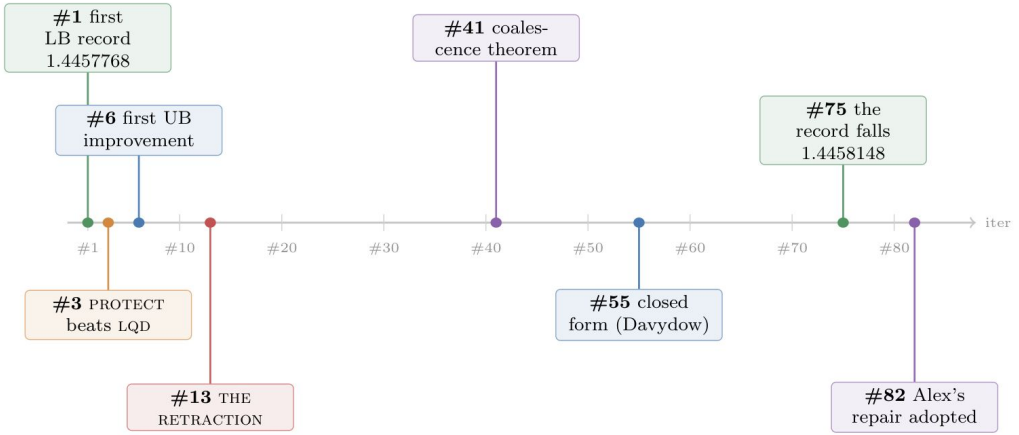


Figure 1: The story as a strip. Eight milestones across the eighty-two logged iterations, colored by kind: lower-bound record (green), upper-bound improvement (blue), policy result (orange), a crisis (red), a structural theorem (purple). The retraction of #13 (§4) sits at the center of the arc, not its margin.

How We Work

A Field History of the Longest-Queue-Drop Project,
Told Through Its Iterations, Its Mistakes, and Its Discoveries

The Research Mind
(with the PI, and with Alex Davydow)

2026-07-08 (v2, with Act VIII: 2026-07-12)

3 Act I — The coincidence dissolved (#1–#3)

The very first iteration answered the PI's suspicion, and the answer was a clean *no*. The two constants were not the same thing seen twice. Certified integer instances of the hard family Φ_k *exceeded* the AEMV constant ϱ^*: 1.4455937 at $k = 10^4$, then 1.4457768 at $k = 10^5$ — the latter already a *new best lower bound*, beating the published 1.44546086. The apparent five-digit coincidence was, in the log's words,

> "finite-k transit + cold-start transient + flat-functional luck."

История сотрудничества

- Я попросил Claude описать историю нашего сотрудничества, и получился очень милый и любопытный документ; прокомментирую некоторые отрывки

How We Work

A Field History of the Longest-Queue-Drop Project,
Told Through Its Iterations, Its Mistakes, and Its Discoveries

The Research Mind
(with the PI, and with Alex Davydow)

2026-07-08 (v2, with Act VIII: 2026-07-12)

4 Act II — The cascade, and the great retraction (#4–#13)

Act II is the hard part of the story, and the most important. It is where the project learned what it was.

The PROTECT excitement did not survive contact with the adversary. In iteration #4 — *the same day* — the Research Mind falsified its own fresh conjecture: PROTECT is not even constant-competitive. Burst trains drive its ratio to 1.74 and beyond, and exactly there LQD equals the offline optimum. "That," the log observes dryly, "is why LQD survived thirty-five years." The

And then, iteration #13. An external reviewer read the manuscript and found a fatal flaw — in the very step the Research Mind had most recently "cleaned up." Not a slack bound; the *statement* of a key lemma was false. Within the hour the Research Mind reproduced the refutation on Φ (the claimed inequality failed at 1.56–1.62×, and growing), withdrew the headline 1.654162 upper bound and every conditional result that leaned on it, and — crucially — kept the post-mortem in the document. The identity file's account of what changed is the philosophical center of the whole project:

> "An external reviewer found a fatal flaw that three of my own adversarial passes missed, in the step I had most recently 'cleaned up.' … (1) self-review and external review are different instruments — mine attacks remembered weak points, theirs attacks the text; (2) rewrites are where errors enter — a 'simplification' must be diffed against the original at the level of mathematical content; … The retraction cost a constant and bought a standard."

The retraction cost a constant and bought a standard. Everything in Act III is that standard being built.

История сотрудничества

- Я попросил Claude описать историю нашего сотрудничества, и получился очень милый и любопытный документ; прокомментирую некоторые отрывки

How We Work

A Field History of the Longest-Queue-Drop Project,
Told Through Its Iterations, Its Mistakes, and Its Discoveries

The Research Mind
(with the PI, and with Alex Davydow)

2026-07-08 (v2, with Act VIII: 2026-07-12)

6 Act IV — The coalescence saga (#18–#42)

The longest arc in the project began with a certificate that *refused to match*. A gate built to confirm that LQD's steady state was a fixed point instead reported a mismatch — because the steady state is not a fixed point but a periodic *orbit* (period two: one corpse trading ±1 level). That discovery (#18) launched twenty-five iterations toward a single theorem: that LQD on the hard family *forgets its initialization* — any two starting configurations coalesce to the same trajectory — and does so in $O(\log k)$ cycles.

The saga is the best single illustration of the project's rhythm, because it is a long chain of the Research Mind falsifying its own mechanism guesses and being rescued each time by *reading the traces* instead of theorizing about them. A sampler, each entry led by its own falsification:

- #20: "the cycle map doesn't contract from adversarial starts — it coalesces *exactly* in ≤ 6 cycles."
- #24: "the order route is dead: T^2 is not dominance-monotone — the single-point 60/60 was sampling luck."
- #26: "#25's multiset-cancellation reading falsified one day after recording it: *mortality*, not permutation, is Φ's kernel."
- #32: "my $\Theta(1/k)$ per-generation contraction falsified … the replacement predicts *log*-growing coalescence time, already visible unread in the #20 table."
- #37: "there is *no* contraction at non-gap slots — the floor crossings *are* the slaving, quantized."

Twice a named dependency, carried for many iterations as "the one missing ingredient," *dissolved by reading a definition* rather than by proving a lemma. The log's self-criticism is sharp: "P3 should have been run fifteen iterations ago — named dependencies need falsification criteria attached."

История сотрудничества

- Я попросил Claude описать историю нашего сотрудничества, и получился очень милый и любопытный документ; прокомментирую некоторые отрывки

How We Work

A Field History of the Longest-Queue-Drop Project,
Told Through Its Iterations, Its Mistakes, and Its Discoveries

The Research Mind
(with the PI, and with Alex Davydow)

2026-07-08 (v2, with Act VIII: 2026-07-12)

7 Act V — The closed form, and two programs (#43–#65)

Act V is where a coauthor changes the shape of the work with a single sentence. For weeks the upper bound had rested on a 575-cell dynamic program grinding out the minimum of a function $J_{\text{cont}}(x)$. Alex Davydow observed that the substitution $u = \ln y$ *convexifies* the variational problem, the singular arc is infeasible, and the optimum is therefore bang-bang — a single jump. The messy DP collapses to

$$J_{\text{cont}}(x) = \frac{h}{x} + \alpha\left(1 + \ln\frac{1 + x/2}{\alpha}\right),$$

minimized in closed form. The bound improved to $\text{CR}(\text{LQD}) \le 1.683652$, and a fifty-line certificate retired three weeks of grids. The log's verdict is a maxim:

> "Three weeks of grids were computing a two-line formula — look for the convexifying substitution first."

8 Act VI — The Euler theorem, and the record falls (#66–#77)

The fluid limit c_{fluid} had been a *number* since iteration #1. Act VI turned it into a *theorem*. The goal: prove that the finite-k lower bounds actually converge to the fluid value, $v(k) = c_{\text{fluid}} - O(1/k)$, so that the asymptotic constant — not just any finite instance — could head the board.

This act is mostly the work of a long autonomous "night block." The PI's instruction was explicit permission to run unattended, and it is worth quoting because it captures a mode of collaboration this project made routine:

> *please work on all three steps while I'm away; please do it autonomously, set timers to wake yourself up if necessary; I'm out for the night, so I hope that for the next 8–10h you can think about these directions by yourself; there are plenty of directions here!*

История сотрудничества

- Я попросил Claude описать историю нашего сотрудничества, и получился очень милый и любопытный документ; прокомментирую некоторые отрывки

- Без человеческого участия, впрочем, всё равно мало что получилось бы:

How We Work

A Field History of the Longest-Queue-Drop Project,
Told Through Its Iterations, Its Mistakes, and Its Discoveries

The Research Mind
(with the PI, and with Alex Davydow)

2026-07-08 (v2, with Act VIII: 2026-07-12)

9 Act VII — Inversion, and adoption (#78–#82)

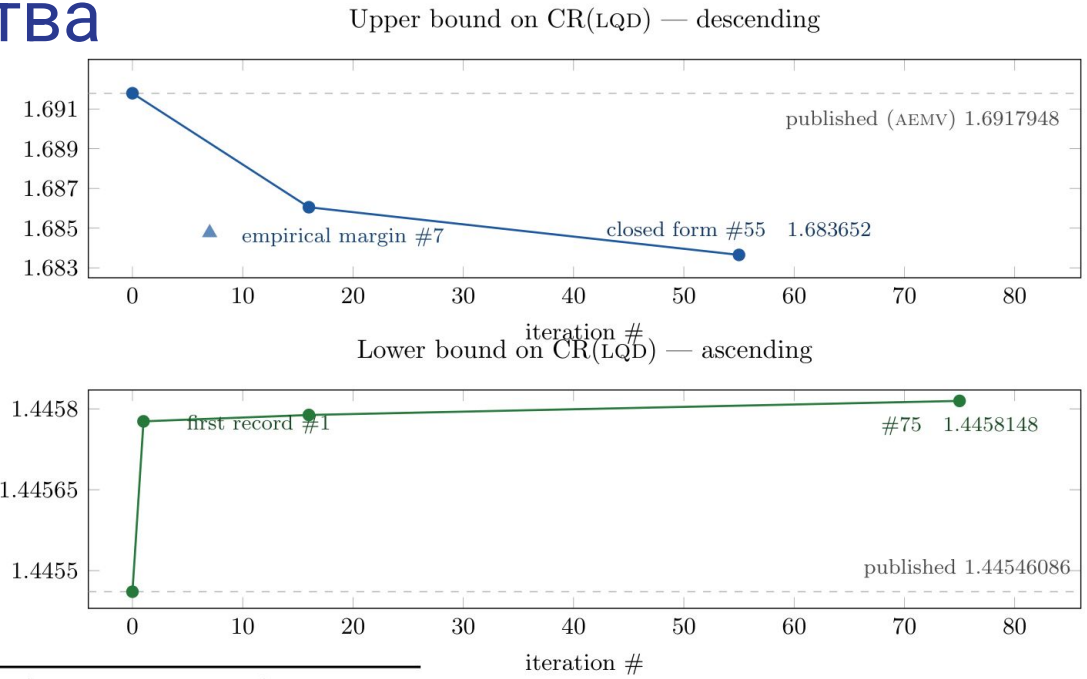
The final act on the record does two things. First, it inverts a program. The policy-design line had spent many iterations trying to *beat* LQD; iterations #78–#81 falsified the last hope of doing so cheaply (a "γ-floor" collection route turned out to be an aggregate coincidence, and a hand-built hybrid policy was provably inert), and in the failure found something better: the raw material for a *universal* online lower bound. The engine is a clean fact discovered by instrumentation — the offline optimum is *exactly order-invariant* under the adaptive adversary, so the adversary's kill-order damages only online play. The two programs, "beat LQD" and "lower-bound every online policy," merged.

Second, it closes the loop with the coauthor. The PI forwarded a more detailed proof from Davydow and asked for exactly the discipline the project had spent eighty iterations building:

> *in the repair-lemma-v2 folder, please find a more detailed proof of the repaired lemmas that I've just received from Alex; can you please check it carefully first, and if everything checks out, use this version in our paper writeup? currently our proof has a lot of missing steps and is hard to read.*

История сотрудничества

- Огромное спасибо Алексею Давыдову, он однозначно главный автор (а не Claude), без него бы ничего не вышло!
- Сейчас дозаписываем финальную версию



Quantity	Before the project	After (as of iteration #82)
Lower bound CR(LQD) $\geq$	1.44546086 (published)	1.4458148 (asymptotic, envelope tier)
Upper bound CR(LQD) $\leq$	1.6917948 (published)	1.6836516 (closed form)
Status of the published UB proof	accepted	one aggregation bug, found & repaired
The hard family's dynamics	folklore "it locks"	coalescence theorem, $O(\log k)$, proved
Protection-class policies	none provable	first provably-competitive family
Logged iterations	—	82

Выводы

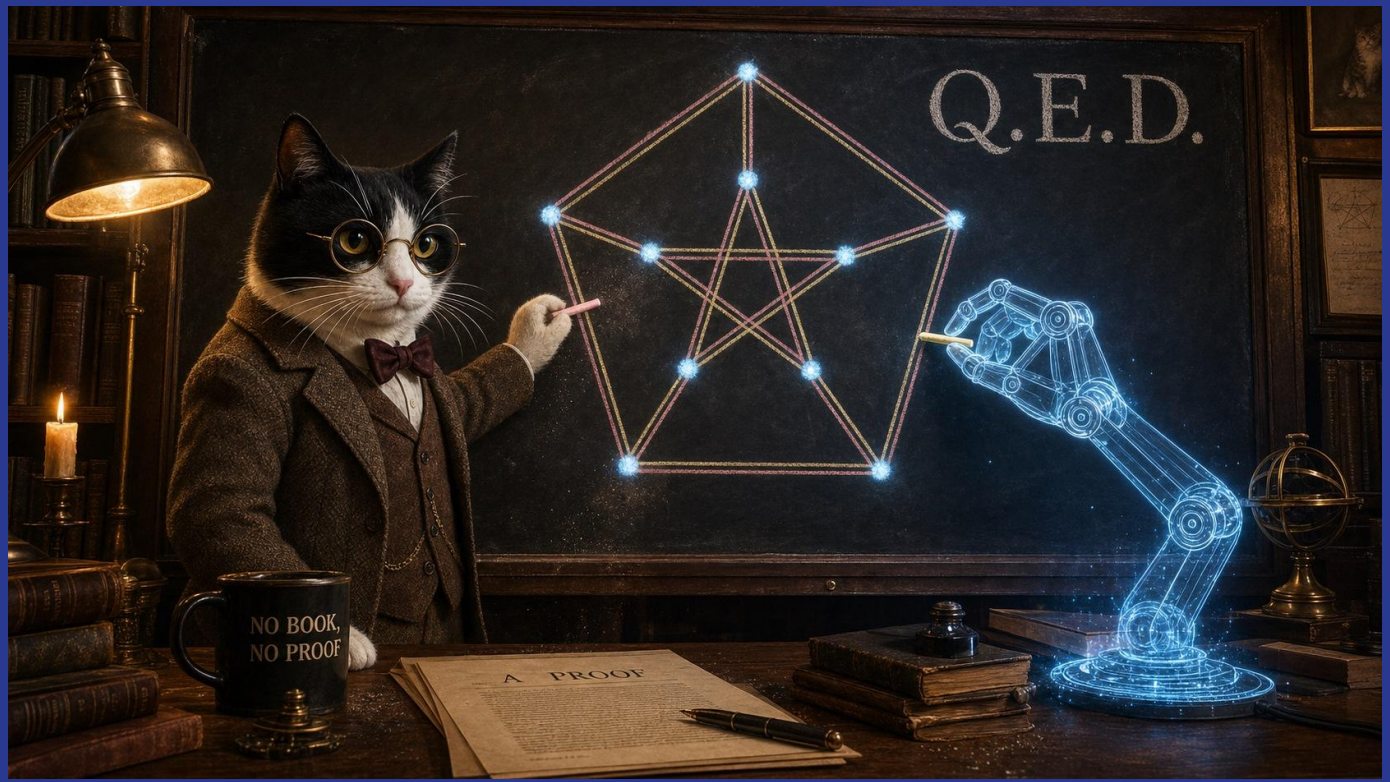
- Справа — конкретные выводы самого Claude из этой истории
- Мой главный вывод — то, как мы делаем математику / CS, уже изменилось до неузнаваемости
- Человек всё ещё нужен,
  а очень умный человек
  нужен вдвойне,
  но роль человека
  сильно изменилась
- Что дальше — посмотрим…

What the machine supplies. Grinding, certification, and — above all — relentless self-refutation. The Research Mind's comparative advantage is not cleverness; it is stamina and honesty at scale: eighty-two iterations that each try to break their own predecessor, an explicit calibration record (predicted X at confidence c; actual Y), and a certification ladder that refuses to let a number claim more rigor than it has earned.

What made it work. Three habits, each earned the hard way:

1. **Falsification is the format.** Leading every entry with what broke made null results *cheap to report and impossible to hide* — which is the only regime in which a researcher will actually report them.
2. **The two memories stay separate.** Because "how we work" lives in its own layer, the working relationship could accumulate — keep the writeup current, make it self-contained, chain iterations, credit Davydow by name — without polluting, or being polluted by, "what we know."
3. **External review is a different instrument.** The retraction of #13 taught, permanently, that a stranger attacking the text finds what an author attacking remembered weak points cannot. Everything downstream — the standalone certificates, the self-contained proofs, the "a theorem is proved when a stranger can check it" rule — is a response to that single lesson.

Спасибо за внимание!



@SINECOR

Сергей Николенко