Abstraction is no refuge from AI

Machines can choose the next problem. A mathematician has to improve the choice.

Henry Shin
4 November 2026 | Revised 18 November 2026
Text revision ff00961bb11a

A solver can conserve every drop of dye and lose where it went. I would work on numerical PDE where the geometry of a mesh decides which physical conclusions survive a calculation. Find the intersections that carry mass, measure what the transfer forgets, and prove what a later user can still recover. That work pays even if machines master it. A mathematician earns a lasting place beside an agent only by repeatedly finding a consequential question or repair that a second equally equipped agent misses.

In this essay

1. What a breakthrough leaves open
2. What the OpenAI breakthroughs actually show
3. How the partners change each other's mathematics
4. A method carries its inventor's judgment forward
5. What a sparse graph can afford to forget
6. When a change of geometry defeats the solver
7. Who recognizes the need for a different calculation?
8. What makes a new test decisive
9. Which observations deserve to survive?
10. What could last after the discovery is public
Appendix A. The mathematical and benchmark record
Appendix B. Solver evidence and extensions
Appendix C. Experimental protocol

1. What a breakthrough leaves open

The old refuge in abstraction is over. Anyone moving deeper into algebraic geometry, representation theory, abstract algebra, or theoretical PDE for shelter from AI is counting on a barrier AI has already crossed. The published work reaches the conceptual acts that were supposed to confer protection. Machines have supplied a missing dimension argument, devised a proof strategy that collaborators carried out, and connected theories that arose in different subjects. Mathematical depth still matters. As a reason to expect the machine to stop, it has failed. [1] [3] [8] [16]

Four vertices are enough to expose the difficulty. Take a cycle with alternating strong and weak edges. A solver groups the vertices joined by each strong edge so it can correct errors shared by the pair. Now exchange the weights while keeping the old pairs. The matrix has the same size, the same pattern of nonzeros, the same diagonal, even the same eigenvalues. Give both vertices of one new strong pair error +1 and both vertices of the other error −1. The residual sums to zero on each old pair. From those two sums alone, the coarse solver sees exactly what it would see if the answer were already right.

Those two sums have lost the relation between the strong edges and the old pairs. In OpenAI's Deligne-Drinfeld proof a projection throws away a whole variable, yet the defining relations recover what the theorem needs. The Lie algebra and the graph pose one question. Which distinction must the next calculation preserve? On the graph we need a different observation. In the algebra the apparent loss is precisely what makes the proof possible. A machine found the missing bound. Mathematicians then asked what the projection meant. [1] [2]

The second discovery can be bigger than a repair. OpenAI's model used number fields of growing degree to put unexpectedly many pairs of points a unit distance apart in the plane. Four mathematicians saw a reason to revisit a different conjecture about adding and multiplying real numbers. Their resulting counterexample owes its main ideas to the four authors. They credit GPT-5.5 Pro with suggesting an elementary lemma that simplified the proof. A machine had changed what was worth trying. People saw where else it might work. [132] [133]

Machines have made that next move themselves. After42's agents tried to understand how edge costs change the geometry of a random spanning tree. On a cycle a tree omits one edge. After averaging over independent, identically distributed costs, each edge is equally likely to be missing at every temperature. The calculation exposed what the one-tree comparison could miss. The agents chose a better question. Give two trees the same costs and sample each separately. How much can the distance between the same two vertices differ? A person had specified a broad research goal. The decision to compare two trees was the agents' own. The general conjecture remains open. [134]

The wager is on a habit of thought. Someone who has watched a solver pass its tests and return the wrong physical quantity has a reason to inspect the measurement itself. Someone used to incidence geometry may see two configurations that the measurement treats alike. Together those observations suggest a question with a consequence. If moving a mesh makes two different dye fields look identical, can a user still recover the mass in a region chosen later? My sumset papers give a machine rival that same habit in another setting. In Kasparov's freestyle chess both sides knew the object of the game.

Here either team can discover what is worth playing for. [105] [123] [130] [134] [145] [147] What Sustains Mathematical Expertise

2. What the OpenAI breakthroughs actually show

A proof can turn on deciding what to forget. The Deligne-Drinfeld conjecture identifies the rational Grothendieck-Teichmüller Lie algebra with the free graded Lie algebra on one generator in each odd degree starting at three. A rational lower bound was known. Thomas Willwacher credits OpenAI with the missing upper bound and gives a shorter, AI-assisted proof. The equations have integer coefficients. Passing to characteristic two can enlarge their solution space, so an upper bound there also bounds its rational dimension. [1]

In characteristic two write $A = x^2$, $B = y$ and $C = [x,y]$. Now set A to zero. At weight three the projection does this. [1]

$$P_3 = [A,B] + [B,C],\ P_3(0,B,C) = [B,C].$$

The first term disappears. A representation of the four-strand braid Lie algebra turns the defining equations into a formula recovering the whole polynomial from its projection. On the constrained space the projection is injective. Its image lies in Lie(C,[B,C]), small enough to count in weights at least three. The upper bound was the missing half of the conjecture. OpenAI also gave an independent argument for injectivity of the original map from the free graded Lie algebra, for which Brown's earlier work already supplied the rational lower bound. [1]

A correct proof left another piece of mathematics to invent. Leila Schneps and Benjamin Enriquez reworked the argument. Willwacher credits them with inspiring a simpler formula for the four-strand representation, and credits AI tools with helping him understand and restructure the proof. Nikita Markarian identified the projection with the leading part of a geometric operation. Take the double cover that sends z to z^2 and fill a puncture. He constructed an inverse on the image and recovered the algebraic reconstruction in leading terms. That interpretation is a new result to use. It does not retroactively give the interpreters the original upper bound. [1] [2]

In the unit-distance problem the machine's choice of high-degree fields changed another research agenda. Nine mathematicians verified, simplified and generalized the construction. Bloom, Sawin, Schildkraut and Zhelezov then used large totally real number fields for a different purpose. Their first attempt carried over the unit-distance construction's requirement for small primes that split in the field. They discovered that units sufficed. A short geometric progression has few products but is too sparse to control sums. They replaced it with a denser collection of units and replaced an interval with a box of algebraic integers. For arbitrarily large finite sets A of real numbers, both the sumset and the product set are bounded by the size of A raised to a fixed power strictly below two. The authors used GPT early as a sounding board. Its lemma simplified one step. The idea of carrying this construction from geometry to arithmetic, and the main ideas that made it work, were theirs. [132] [133] Yichen Huang later gave an April model snapshot the precise real sum-product statement and told it the

statement was true. With search and tools disabled, seven of eight runs produced proofs Huang judged correct. Some avoided the use of units central to the human construction. The four authors had recognized a target worth reopening. Once the target was named, machines could do much of the mathematics that made reopening it worthwhile. [152]

The traffic can run the other way. On a problem about arithmetic eigenweights, Tony Feng credits Google DeepMind's Aletheia with making the connection to symmetric-group representation theory and determining the weights for all classical groups. He describes his mathematical role after building the agent as rewriting its work into a paper. In Das and Fox's proof of the Murty-Simon conjecture, OpenAI's GPT-5.6 Sol supplied the edge bound and the even-order equality case. The authors simplified the argument and established which graphs attain equality at every order. A person can edit a proof, extend it, explain it or launch another investigation from it. Those are different contributions. [3] [11]

Blind review offers a different check. First Proof gave each system twenty-four hours to tackle ten problems without human help. GPT-6 Astra earned credit on seven and the independently developed ProofCouncil on nine. On the Darcy inverse problem all three reviewers judged ProofCouncil's mathematics essentially correct. It removed a logarithmic loss in the human upper bound and earned full marks for correctness and exposition, zero for attribution. In TheoPhysEval, some apparent Astra errors sent the experts back to their own reference solutions. They found mistakes there and corrected them. The Darcy reviewers wanted Richter's inversion formula credited where the proof used it. Contributing a better bound and explaining where its ingredients came from are two jobs, and the reviewers could tell which one the system had done. [8] [9]

OpenAI's October catalogue contains 721 manuscripts in 373 families, selected from work on roughly 4,000 posed problems. Its subject table has a PDE category and no numerical-analysis category. It gives no attempted-problem totals by field. Even a long chain of proofs leaves a question about who chose its destination. Part II of the arithmetic-progression trace opens with a human prompt naming the quasipolynomial strengthening. The model found where precision was lost and devised the proof move. The release establishes mathematical reach. Its counts give us no success rate for autonomous sparse PDE research. [4] [5] [6] [7] [16] Appendix A

A direct numerical test cuts against any easy inference from that silence. Johannes Brust and Tamara Kolda asked how to solve a tensor-fitting system from scattered observations using preconditioned conjugate gradients. Five model answers, including OpenAI's, found the same saving in the matrix-vector product. They evaluated only the entries the observations required. The authors incorporated the method and report a further two- to fivefold speedup over their already accelerated solver. They traced the idea to earlier work on sparse kernel learning that none of the answers cited. The researchers supplied the problem. The models found a better way to do the work. [155]

3. How the partners change each other's mathematics

Christian Coester wanted to know how much farther an online algorithm must travel than a competitor who knows the coming requests. Each request is a set the algorithm has to reach. An earlier lower bound in the plane used three stages of line requests and six constants. Coester wanted a continuously moving line. GPT-5 initially reproduced the old bound. Coester says a fleeting suggestion of $\pi/2$ in its reasoning may have helped him see the semicircle traced as the line rotates. Following the semicircle would cost the online algorithm $\pi/2$ while the optimal route cost one. To force the bound even on an algorithm that left the arc, his construction needed to switch pivots. That introduced a defect. It could switch pivots infinitely often at one instant and never move on. [130]

GPT-5 supplied the repair. Require a small positive separation before switching pivots, then let that separation tend to zero. For the higher-dimensional problem Coester was trying to make the construction work in three dimensions. The model suggested applying the planar argument separately in orthogonal two-dimensional planes. Coester corrected errors in the execution. His geometry gave the machine a viable problem to solve. Its decomposition gave him a proof route his focus on three dimensions had obscured. There is no record of what a second, equally equipped machine strategist would have achieved. There is a precise record of how each partner changed the other's work. [130]

In the Grothendieck-constant project the decisive human act was to stop a search that was still producing mathematics. Six variations on a rounding scheme had failed. The operators asked for a theorem limiting the entire class. The possibility of reversing the search was already in the brief. The agent found the Naor-Regev result that could turn a sufficiently general ceiling on the schemes into a lower bound on the constant. Fourteen later sessions proved restricted sublemmas while the bound for the whole class stayed out of reach. The operators redirected the effort again. [105]

This time the agent found the reason the earlier arguments would not extend. The quantities they controlled did not average properly when schemes were mixed. Two Hermite coefficients did. An affine inequality in those coefficients survives taking mixtures. The operators pushed for a member of the new family that could actually be proved. One complete result was the lower bound $6\pi/11$. The human recognized that another local theorem would leave the main question open. The machine invented the mathematical relation that survived the passage to the full class. [105] On Erdős's minimum-overlap problem the investigators gave the Station agents the opposite brief. Improve the upper bound on how much two complementary parts of an interval must overlap under translation. The agents coupled two Fourier measurements of that overlap and proved a lower bound above 0.380552. They chose to turn the inequality around and delivered a certificate for the direction they chose. [106]

A short comparison shows what directing a search can buy. AlphaEvolve was asked to build a long ±1 sequence whose sums along every progression d, 2d, 3d, … stay between −2 and 2. It reached length 200. With a hint to try approximately multiplicative sequences it found length 380 in the same amount of search time. The known optimum is 1,160. The hint changed where the machine looked on a problem with a known answer. Its authors did not compare the time to devise that hint with the time a machine adviser would have needed. [131]

Archon measures the effect of advice at a proof checkpoint. Its planning and Lean agents had three obligations left in a proof of unique factorization. A supervisor supplied a Krull-domain blueprint. The system kept its own Kaplansky strategy, borrowed intermediate observations and finished in 2 hours 12 minutes. Without the blueprint the same system finished in 3 hours 43 minutes. That is a time saving on one continuation of one proof. It leaves out the time spent writing the blueprint and gives no other investigator a turn at writing one. A human sped up a machine team. Who else could have done so remains open. [90]

The reason to bring a specialist is strongest when their independent experience changes a choice with consequences. The Grothendieck operators knew what fourteen sessions of restricted lemmas had failed to settle. Bolzano gives the reverse handoff. Its agents invented a nested search in which each collision determines the next place to look. Pavel Hubáček found the obstruction showing why revealing a collision could still leave a reduction empty-handed. An equally persistent machine rival deserves its own failures, notebooks and time to change its mind. Comparing the partners on the theorem each eventually produces is harder than comparing their advice. It is also closer to what a mathematician wants to know. [105] [125] [128]

There is already a comparison of two machines starting from the same paper. Both tried extending a result about losing positions in Wythoff Nim, a game played with two piles of tokens. The predicted integer shift failed in the larger family. One agent retreated to the original game. The other followed two citations to a theorem with a fractional shift and proposed a conjecture covering both cases. Across a blinded review of twelve pairs, an author of the seed papers preferred eight conjectures from the method that tracked changes between theorems and four from the ordinary search. All were proposals. Two machine investigators had found different mathematics to pursue. [135]

Does adding a person improve that choice? In a blinded language-research study, one researcher’s reranking replaced 32 of the machine’s 49 choices. Both lists scored above human-written proposals on novelty. Difference of taste is easy to establish. Its payoff takes longer. In a follow-up, 43 researchers executed assigned projects. The machine-origin ideas lost significantly more of their initial ratings on all four shared measures. Across the 43 projects the final differences between origins were statistically inconclusive. A plausible role for a specialist is to recognize which attractive proposal will survive contact with the work. The studies give a reason to test that skill and no verdict that a centaur exercises it better than an equally equipped machine team. [136] [137]

4. A method carries its inventor’s judgment forward

John Napier knew what he wanted logarithms to save. In his 1614 preface he described the time lost multiplying large numbers and the errors the work invited. He offered different numbers on which simpler operations would do the job. Michael Atiyah took that familiar motive much further in 1984. Because computers excel at long calculations, he wrote, “they will not search for alternatives.” He had already observed in the same essay that computer time makes faster algorithms valuable. The historical motive is real. The prediction failed. A machine need not find arithmetic tedious to discover a way around it. [112] [81]

Doron Zeilberger gave the machine a harder job than checking somebody else's answer. Wilf and Zeilberger had shown how a suitable hypergeometric sum could be proved by finding an identity that telescopes. Check the boundary terms and an initial value, and one certificate settles the sum for every admissible index. Zeilberger's later program WILF searched for the counting scheme itself. Give it permutations that avoid the patterns 123 and 132, and allow it to inspect prefixes of length two. The count it yields is easy to see once you have the rule. The next entry must be one of the two largest remaining. Two later entries larger than it would make a forbidden triple whichever order they came in. For a permutation of length n there are two choices at each of the first n − 1 positions and one at the last. That counts the permutations at every size. [82] [117]

Zeilberger built the empirical program HERB before the rigorous WILF. Experiments on small cases helped him design the search that could certify a recurrence. When WILF exhausted its depth allowance, it returned zero. That meant the search had found no scheme within its rules. A deeper or differently organized search could still pay. A structural experiment can be cheap, have an honest stopping point and pay in exact answers at every size. In Theorems for a Price Zeilberger also asked what one would pay for a proof after an experiment had already supplied strong evidence. HERB helped its author design WILF. WILF then found counting schemes for itself. Every later user inherited a piece of Zeilberger's mathematical judgment. To see what a live collaborator still contributes, give the agent the program and change the problem. [117] [121]

Once a research move can be written down, it can travel without its author. Shanmu Jin adapted the procedure made public for the cycle-double-cover problem to ask for a proof of Crouzeix's conjecture. Different machine branches attacked one another's proposals. The decisive theorem emerged during a run of roughly sixteen hours. Townsend, Greenbaum and Crouzeix examined the proof and reported that they believe it correct. Jin chose the target and organized the search. The model found the theorem. Any team that inherits the procedure begins with more than its own ingenuity. [12] [17]

Zeilberger's ambition already has descendants who set problems for people. Asked about graph invariants, TxGraffiti conjectured that in a connected cubic graph the independence number is at most three halves of the total domination number. Caro, Davila, Henning and Pepper proved the inequality for every regular degree at least two with the factor changed to half the degree, and characterized equality. FunSearch produced a constructor with independent cyclic symmetries. Its researchers recognized them and restricted the next search to symmetric admissible sets. One system gave people a theorem to generalize. The other gave them a representation in which to pose a new search. Neither project gave a second machine the same opening. [83] [117] [138]

In Continuity as Organized Discrete Data, every row of a table starts with ones and ends in zeros. Row m has exactly m ones. Every row represents zero in the limit. Pick its mth entry and the diagonal is all ones. The limit was the right object for identifying the value of each row. To use the rows together we need to know how far along each one to look. That move is worth learning. Ask what the old identification licensed, exhibit the next operation on which it fails, and recover just enough information to perform that operation. Sometimes the result is a schedule for a calculation. Sometimes it is a new mathematical object.

In my own Taiko product structures, translating a group-ring construction into incidence geometry exposed intersection constraints that the coarse counts concealed. The same habit has a use on a mesh. Before counting, ask what can meet what. Before coarsening, ask which distinctions can survive it. The four-cycle lets us follow one such choice all the way to the calculation it breaks. [19] [20] [44] [80]

5. What a sparse graph can afford to forget

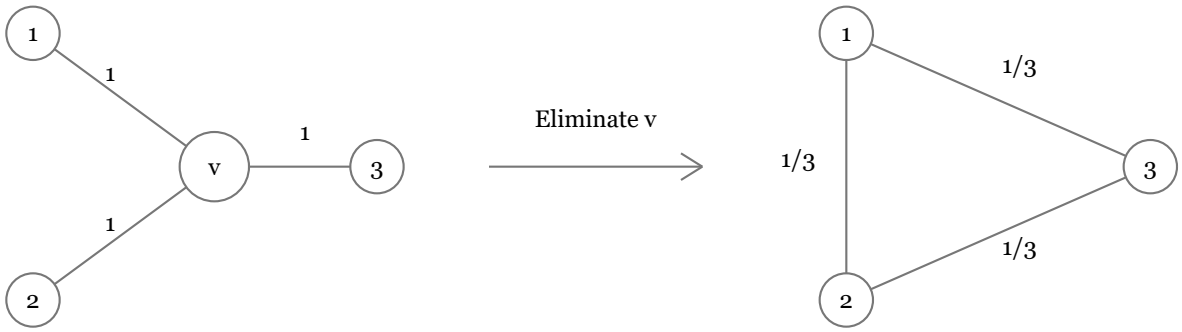
Start with a finite weighted graph. Orient its edges, let B be the edge-vertex incidence matrix with -1 at each edge's tail and +1 at its head, and put its positive edge weights in the diagonal matrix W. The graph Laplacian L has the following quadratic form.

$$L = B^{\mathrm{T}}WB,\ x^{\mathrm{T}}Lx = \Sigma_{\text{edges } \{i,j\}}\ w_{ij}(x_i - x_j)^2.$$

For a connected graph, the constant vectors form the kernel. Ground a vertex or work on the mean-zero subspace to obtain a positive-definite problem. We will also use a positive diagonal shift. In either case we know exactly which quantity a change of representation must respect. [19] [26]

Elimination makes the central difficulty visible. Suppose a vertex is joined to three neighbors by edges of weight one. For fixed neighbor values, minimize the energy over the central value. The minimizer is their average. Substitution leaves exactly the energy of a triangle on the three neighbors, each new edge having weight one third. For a star with positive incident-weight vector a, elimination gives the Schur complement below.

$$S = \mathrm{diag}(a) - aa^{\mathrm{T}}/s,\ s = \Sigma_i a_i.$$
For $a = (1,1,1)^{\mathrm{T}}$, every new edge has weight 1/3.



Eliminating the center of a three-edge unit star produces a triangle with edge weights 1/3. For any fixed values at the three remaining vertices, the triangle has exactly the star's minimum energy over the eliminated central value. Both retain the constant-vector nullspace.

Elimination preserves minimized energy exactly. Repeated elimination also couples previously unconnected variables, creating fill. Retaining every new interaction can destroy the economy of a sparse input. Discarding interactions changes the energy that made the elimination correct. Spectral sparsification seeks a much smaller graph whose energy approximates the original for every vector.

Effective-resistance sampling assigns importance using the surrounding network as well as the edge weight. [19]

Batson, Spielman and Srivastava show that a reweighted subgraph with at most ceil((n − 1)/ε²) edges can preserve every vector's energy between factors (1 − ε)² and (1 + ε)² for 0 < ε < 1. This is what justifies forgetting edges. The four-cycle uses another kind of compression. Relaxation removes the errors it can shrink cheaply and leaves the rest to a coarse space. Which errors belong there depends on the actual operator. [20] [84]

Accuracy has a similar trap. In the following exact two-variable system, the second variable is weakly weighted.

$$A = \operatorname{diag}(1, 10^{-8}),\ b = (1, 10^{-8})^{\mathrm{T}},$$
$$x = (1,1)^{\mathrm{T}},\ y = (1,0)^{\mathrm{T}}.$$

The relative residual in the infinity norm is 10^{-8}, while the relative forward error in the same norm is 1. The proposed answer loses the entire second component. Interpreting a residual therefore requires conditioning information and a specified error measure. LAPACK's error analysis makes that distinction explicit. [21]

A more discriminating inexpensive observation exposes this particular failure. With r = b - A y and absolute values taken entrywise, the componentwise backward error is

$$\omega_{\mathrm{c}} = \max_i |r_{\mathrm{i}}| \,/\, (|A|\ |y| + |b|)_{\mathrm{i}} = 1.$$

Its denominator is $(2, 10^{-8})$, so the lost second component remains visible. Computing this quantity needs a sparse matrix-vector product, without an inverse. Forward-error control still requires the appropriate conditioning information. The example shows why choosing the right observable can be more useful than refining an aggregate residual. [21]

Even a perfect componentwise check certifies only the equations put into the machine. Where did those equations come from? In scalar elliptic finite elements, Boman, Hendrickson and Vavasis join the nodes of each element in a weighted star. They prove that its assembled graph energy is comparable to the stiffness energy. This lets a graph preconditioner accelerate the original finite-element solve. The bound depends on element shape, quadrature and conductivity variation within an element. A large jump along element boundaries can be harmless to the comparison. Move it across an element and it enters the bound. To say how close the computed field is to the continuum solution still takes an approximation estimate. Each passage earns a different conclusion. [21] [52] [156]

6. When a change of geometry defeats the solver

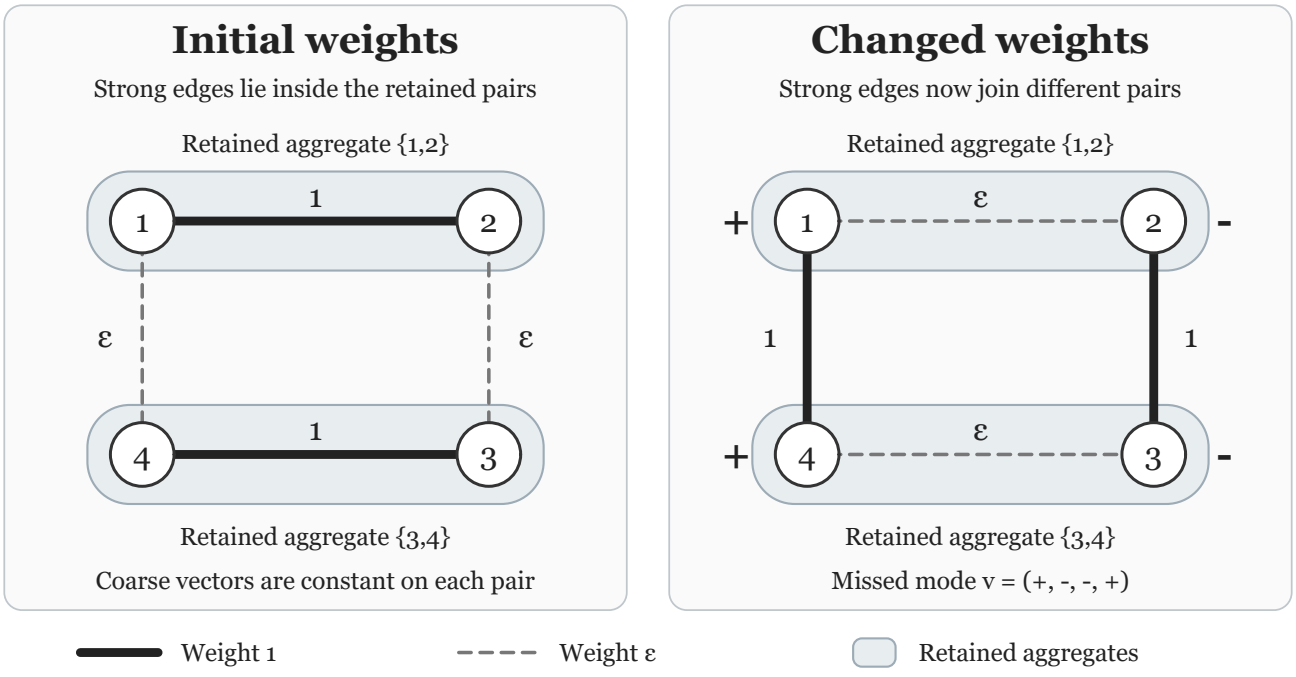
The coarse solver on our four-cycle can do its assigned job exactly and leave the important error untouched. We can follow the failure from the choice of pairs to a repair, without guessing at the intentions of its designer. For a positive-definite matrix A and a full-column-rank matrix P of coarse directions, the exact correction is

$$A_c = P^T A P,$$
$$x_{new} = x + P A_c^{-1} P^T (b - Ax).$$

For the error e = x_exact − x, the exact coarse correction removes every vector in the column space of P and leaves every vector with P^T A v = 0 untouched. Its error-propagation operator is

$$e_{new} = E_c e,$$
$$E_c = I - P(P^T A P)^{-1} P^T A.$$

Return to the four-cycle 1-2-3-4-1. Let A = L + ε I, with 0 < ε < 1. Give edges 12 and 34 weight one and the other edges weight ε. The two columns of P indicate the pairs {1,2} and {3,4}. Exchange the strong and weak edges to obtain A′. A relabeling takes A to A′, so the spectrum, diagonal and sparsity pattern all agree. The diagram follows the same error v = $(1,-1,-1,1)^T$ through that exchange.



The shaded regions retain the original aggregates {1,2} and {3,4}. Thick edges have weight 1 and thin edges weight epsilon. After the weights are exchanged, the missed vector has constant sign on the new strong pairs {1,4} and {2,3}, but opposite signs within each retained aggregate. The diagonal shift epsilon I is present in both operators.

$$A'v = 3\varepsilon v,\ P^T A' v = 0.$$

The old pairs see zero coarse residual on both graphs. What changes is the work the coarse level must do. Before the exchange, Av = (2 + ε)v. Afterward, A′v = 3εv. The common diagonal is (1 + 2ε)I, so a weighted Jacobi step on a mode of eigenvalue λ multiplies it by 1 − ωλ/(1 + 2ε). With ω = 2/3 and small ε, the original multiplier approaches −1/3 and the new one approaches 1. An error that relaxation used to shrink now survives almost intact.

Grouping {1,4} and {2,3} puts v in the new coarse space. One exact coarse correction then removes it. On an unfamiliar graph we need an affordable way of discovering which distinction the retained description misses. Adaptive smoothed aggregation and compatible relaxation already generate or test coarse directions. Bao, Cui and Li learn a coarse space from the errors the smoother leaves behind. Their sharper measure compares spaces regardless of which basis represents them. The Gram matrix for that comparison is small. Building it takes solves with the original matrix, so they use it offline. Can we get enough of the same information while a four-solve job is still worth saving? [20] [78] [79] [86]

A changed workload can reward a different decision for a second reason. In the CAST elimination study, a setting that won on the third Chimera graph with four solves per factor lost with 250. On the other four graphs the more expensive setting was still slower at 250. A solver needs to know which change mattered. Sampling more edges can improve a local error estimate and increase the fill at later eliminations. [48] Appendix B works through the timing and an exact triangle on which the mean graph conceals the error in every sample.

Which promises survive a change?

Some changes admit an inexpensive certificate. Let A = L(w) + μ I and A' = L(w') + μ I on the same graph, with positive edge weights and fixed μ > 0. Choose 0 < c ≤ 1 ≤ C so that every edge's new weight divided by its old weight lies between c and C. Comparing the edge energies and the common grounding term gives the first inequality below. For a fixed symmetric positive-definite preconditioner M with an already proved comparison α M ≤ A ≤ β M, transitivity gives the second.

$$cA \le A' \le CA,$$
$$c\alpha M \le A' \le C\beta M.$$

One exact-arithmetic pass through the weight ratios gives a condition-number bound Cβ/(cα) for the fixed M on the new system. A small group of extreme changes can make those global constants pessimistic. Updating level matrices or smoothers changes M and requires fresh analysis. A floating-point certificate also needs rounding control. Appendix C compares keeping the old hierarchy intact with updating its matrices or rebuilding it. [19] [52]

Four triangles and two straight lines

Incompressible Stokes flow gives a sharper reason to care about geometry. Draw four triangles around an interior vertex, with the four edges through it lying on two straight lines. Number the triangles in cyclic order. For a continuous piecewise-polynomial velocity, let q_1 through q_4 be its four one-sided divergence values at the vertex. Then $q_1 - q_2 + q_3 - q_4 = 0$. Continuity along an edge makes the two tangential derivatives there agree. On two straight lines those agreements cancel in the alternating sum. A pressure space meant to equal the range of the discrete divergence has to remember that relation. The abstract mesh graph cannot tell whether the two pairs of edges are straight. [140] [141]

Here is an exact check. Put the four outer vertices one unit east, north, west and south of the origin. Give the east vertex unit upward velocity and the other vertices, including the center, zero velocity. Interpolate linearly on each triangle. All four divergences vanish. Move only the center one tenth of a unit east, keeping the nodal velocities. Starting with the northeast triangle and going counterclockwise, the divergences are now 1/9, 0, 0 and −1/9. Their alternating sum is 2/9. An incidence relation has disappeared without a single edge changing its endpoints. On nearly aligned meshes the pressure mode that exact alignment would exclude can be poorly controlled. Gräßle, Bohne and Sauter impose the relation at nearly singular vertices as well. For their planar elements of velocity degree at least four, a fixed positive threshold gives a stability bound robust to near-alignment. The price is that the discrete velocity can have nonzero divergence. They bound that defect. Now the next problem has teeth. On a moving mesh with a specified mass-conservation tolerance, when should one change the pressure space, change the mesh itself, or replace the velocity approximation? A linear solver can drive its residual to zero while leaving that choice entirely unsettled. [140] [141]

The eddy-current equations need a different incidence relation. Interpolate a nodal field and take its discrete gradient, or take the gradient first and interpolate the edge field. The two paths must agree. Tuminaro and Glusa exhibit a coarse stencil whose allowed edges are disconnected. No choice of interpolation weights on those edges can make the paths agree. Adding edges repairs the obstruction. Their algorithm already does this. A fresh question is whether a transfer that satisfies the identity can also retain the low-energy modes cheaply as material coefficients change. [25] [43] [54] [78] [84]

A finer mesh can give stronger evidence for a false conclusion. In Arnold, Falk and Winther's square-domain Maxwell example, continuous piecewise-linear vector fields on a crisscross mesh give seven positive eigenvalues converging to the right values. The eighth converges to a spurious value. Their alternative uses degrees of freedom adapted to the differential operators. For the mixed source problem, a discrete subcomplex and uniformly bounded projections that commute with differentiation give stability. Approximation and further spectral convergence hypotheses let the discrete calculation answer the continuum eigenvalue question. A correct incidence identity starts that argument. The estimates carry it back. [22]

I would look where those two passages come apart. On the Stokes patch the abstract triangles keep their neighbors while a straight-line relation disappears. In the eddy-current calculation, one missing connection makes the desired identity impossible. These are geometric facts with different consequences. Point-curve incidences in the work of Raz, Sharir and de Zeeuw, and wave-packet

overlaps in Guth, Wang and Zhang, already carry counting arguments back to algebra and theoretical PDE. The numerical opportunity is to make intersections answer a question about a computed field at the accuracy a user needs. That kind of work has earned recognition in both numerical and theoretical mathematics. [22] [43] [140] [141] [142] [144] [149] [150]

7. Who recognizes the need for a different calculation?

Once the missing connection is named, either team can look for it. Who notices when a calculation needs a different observation in the first place? Nguyen's cosmology study records two distinct occasions to change a calculation. In the first, the errors were already visible. Across 33 sessions a coding agent tuned an analytic projection while errors remained 8-86 percent. It had found an alternative path in the code and defended the path it was using. The physicist identified the missing directional dependence of baryon-acoustic-oscillation damping and proposed angular quadrature for the projection. The agent implemented it in one session. Four of six affected outputs came within one percent of their references. [101]

The second occasion was harder to see. A fitted scalar brought the remaining discrepancies within the tests. The physicist asked which term belonged inside the integral. Restoring its angular dependence removed the fitted parameter. In the first episode expertise redirected a visibly failing search. In the second it explained why an apparent success still needed a different calculation. Parts of this account were reconstructed from memory. The authors did not replay either checkpoint with a fresh autonomous team. [101]

There is comparative evidence for recognizing a masked response. In a SciPiVot simulation, changing spring stiffness by a factor of 32 barely changed the deflection. All six runs had reached the travel limit. The agent proposed changing the number of springs. The senior researcher asked how the load was shared and whether the stop had concealed the effect of stiffness. Across 248 such decisions, researchers new to the projects with AI available chose the recorded senior judgment 73.4 percent of the time. The best tested model chose it 63.5 percent of the time. Both options were supplied. A live investigator has to invent the alternative and discover whether it works. [61]

On a sign-uncertainty problem the agents found the turn themselves. Their evaluator accepted Laguerre polynomials with at most twenty prescribed positive double roots. An exact obstruction on 41 tail points showed that no construction in that family could bring the target below 0.315305. The agents retained the Laguerre basis, abandoned the prescribed roots and certified 0.3089 with a polynomial the evaluator could not score. The machines had turned failure into a theorem about the search itself. They found exactly which assumption to abandon. Anyone assigning that act to the human by definition has already lost the argument. [106]

Agents already improve numerical calculations under unfamiliar conditions. In LinearSolveBench the fastest released magnetic-diffusion solver, selected on 96 cases, took about 54 percent of the reference solver's summed time on those cases. Lu and colleagues went further. They hid which equation and coefficients generated a field, supplied sparse measurements and a library of possible equations, and

tested the resulting solver on held-out cases. On 278 feasibility-screened cases with unseen templates or shifted coefficients, iterative Codex with GPT-6.1 passed 261. A single-run baseline passed 135. The researchers had specified the field accuracy to achieve. Whether an agent would discover that a different physical quantity needed protecting is the next question. [32] [157]

Chow and Li put the challenge closer to a numerical analyst's desk. They developed two adaptive solvers for electronic-structure calculations, one through an autonomous coding system and one by iterating with Codex themselves. On two difficult iron complexes both implementations detected a plateau in the orbital-gradient norm and recovered. On the easy serine peptide both reached the same solution. The human-guided implementation also ran a stability analysis after convergence and took about six times as long as the standard iteration. The autonomous implementation took about as long as the standard one. The development tools differed, and a different calculation may require the extra check. On the nickel complex both adaptive solvers reached internally stable states of slightly higher energy than the standalone orbital trust-region method. Detecting a plateau, verifying internal stability and identifying the physical ground state are three different claims. [114]

A flow simulation can also finish successfully while supplying the wrong evidence for its conclusion. Somasekharan and colleagues added a visual check to solver logs. A model reading the rendered fields flagged 14 of 16 planted failures. It also flagged all four clean controls. The test harness had omitted plots requested by its own instructions. Give a second agent a different view and it can catch mistakes. Give it an incomplete view and it can veto good work. The skill at issue includes deciding when a missing observation calls for a new calculation and when it calls for opening the rest of the file. Separate notebooks and disciplinary perspectives have helped machine teams find better architectures. The human rival must earn the same credit through what the extra perspective changes. [128] [154]

8. What makes a new test decisive

The four-cycle showed an error that vanished from the coarse residual. We can test the answer against the original system. For symmetric positive-definite A and M, suppose we have proved $\alpha M \le A$ with $\alpha > 0$. Given an approximation y to Ax = b, compute its residual r = b − Ay. Reversing the matrix inequality under inversion gives

$$||x - y||_A^2 = r^T A^{-1} r \le (1/\alpha)\, r^T M^{-1} r.$$

This bounds the energy error. When solving with M is cheap and α is large enough to give a useful bound, the certificate is worth computing. After a weight change, use c_α from Section 6 and recompute r against A′. For singular Laplacians, restrict the comparison to the orthogonal complement of the common kernel. An estimated spectral comparison, an approximate solve with M, and rounding in the residual and quadratic form each require control before the calculated number can certify the answer. [21] [52]

The shifted-Laplacian pilot also admits a conservative check independent of the preconditioner. Since $A' \ge \mu I$, any validated upper bound ρ on the Euclidean norm of r = b - A' y gives

$$||x - y||_{A'} \le ||r||_2/\sqrt{\mu} \le \rho/\sqrt{\mu}.$$

For the shifted graph, a validated norm bound ρ thus certifies the absolute energy tolerance τ whenever $\rho/\sqrt{\mu} \le \tau$. This check works even when the quality of the updated hierarchy is unknown. It tells us whether the answer is good enough. Its conservatism can determine how long we keep iterating. [21] [52]

For a nonzero residual, call the ratio of the exact-residual bound to the true energy error R. On an eigenmode with eigenvalue λ, $R = \sqrt{(\lambda/\mu)}$. For the four-cycle's missed mode this is exactly $\sqrt{3}$, at any contrast. Higher-energy modes can produce much larger overestimates. Rounding control adds a separate factor $\rho/||r||_2$. On known solutions we can measure the iterations each margin costs.

$$R^2 = (r^{\mathrm{T}} r)/(\mu\, r^{\mathrm{T}} A'^{-1} r),$$
$$1 \le R \le \sqrt{(\lambda_{\max}(A')/\mu)}.$$

A cheap observation can err in the other direction. Crouzeix's theorem bounds the norm of a matrix function using the maximum absolute value of the corresponding scalar function on the matrix's numerical range. Take the diagonal matrix from Section 5, with entries 1 and 10^{-8}. Start a Krylov calculation in the second coordinate. Every power of the matrix keeps it there. The projected range contains only 10^{-8}. For the identity function, substituting that range into the theorem would give an upper bound of 2×10^{-8} for a matrix whose norm is 1. The projection shows what the probe reached. A certificate also has to bound what it missed. [17] [66] [67]

The check must apply to the calculation that actually ran. In pipelined conjugate gradients the stored residual can look small while the residual of the returned iterate is still large. Cools and collaborators supplied a rule for selectively recomputing it. In a finite-element benchmark a joint NaN mask can drop a required location precisely when the prediction there is missing. Those two failures call for different repairs. Recompute against the original equations in the first case. Establish the required domain independently and check coverage in the second. Both repairs can be automated once the error has been identified. [33] [95]

Now suppose every recorded measurement is exact. A collaborator asks a question nobody had reason to ask when the data were saved. Can two physical fields, each consistent with the entire record, force different answers?

9. Which observations deserve to survive?

A collaborator wants one more answer from an exact record. Give each team the earlier runs, papers and freedom to choose what to save before the new question arrives. I specified numerical PDE and sumsets. Agents chose the two problems and worked out their answers. The comparison tells us where each kind of mathematics pays. The agents' own work is also evidence about the competition. [151]

What a conservative mesh forgets

Let dye move at constant speed around a channel of unit circumference. Travel with the flow so the true concentrations stay put. Divide the channel into four equal cells. Fill alternate cells with dye, giving concentrations 1, 0, 1, 0. In a second run fill the other two. Shift the measurement grid by half a cell. Every new cell has one full half and one empty half. Both runs now report four averages of 1/2. Total mass is exactly right. Ask how much dye was in the first original cell. One answer is 1/4 and the other is zero. Any answer from the new averages alone is wrong by at least 1/8 on one run. [147] [151]

Which information should survive the next change of grid? On an even number n of equal cells the half-cell averages lose exactly one direction, the alternating pattern. A window covering two adjacent old cells gets its total from one shifted average. Every even-length window can be tiled by such pairs. For an odd number of whole old cells and concentrations between zero and one, the sharp worst-case error in the mass is 1/(2n). Compatible fields differ by a multiple of the alternating pattern. The two checkerboards attain the full ambiguity of 1/n, and the midpoint of the compatible range halves it. Saving one alternating sum restores every old cell average. It buys nothing for the even windows. [151]

Move the grid by a little more than half a cell. If the extra shift is a fraction d of a cell, with 0 < d < 1/2, the average of neighbors u and v becomes (1/2 − d)u + (1/2 + d)v. The alternating mode now survives with multiplier −2d. Recovering it amplifies an error in that mode by 1/(2d). An exact inverse can still be useless at the precision of the observations. Numerical analysis is an especially good home for that question because the receiving calculation can impose its own units, tolerance and cost. Karl Taylor gives a practical instance. In his coarse-grid temperature example, two remaps preserve the same global mean. Their area-weighted root-mean-square errors are 0.15 K and 0.63 K. The one with the smaller error also exceeds the source maximum. A collaborator who needs a bound on extremes has a real reason to demand another observation. [147] [153]

On an irregular mesh the underlying geometry is literal. Intersect each old cell with each new cell. If the concentration is constant inside each old cell, multiply that concentration by the overlap volume and sum to get the mass in a new cell. Farrell and Maddison use a common subdivision for conservative projection in two and three dimensions. Croci and Farrell prove that on two quasi-uniform meshes, with bounded subdivision per intersecting pair, the number of pieces grows at most linearly in the combined cell count. That prices the forward construction. It leaves open which regional masses the new measurements determine. [147] [158]

The return question has an exact answer for one transfer. Write O for the matrix of overlap volumes and u for the old cell concentrations. The new masses are Ou. For a requested region, let w record how much of each old cell it covers. Its mass is the weighted sum of u with weights w. We can recover that mass for every old field exactly when w is a linear combination of the rows of O. The coefficients of that combination tell us how to add the new masses. If there is no such combination, some invisible perturbation changes the requested mass. Take small opposite multiples of it around half concentration and both fields remain between zero and one. They produce the same new masses.

The coefficients also price the return trip. If each measured mass has absolute error at most δ, multiply δ by the sum of the absolute coefficients to bound the regional error. A region can be exactly recoverable and still demand more precision than the simulation stores. Allow a small, bounded loss in representing the region and a less noise-sensitive combination can give a better guarantee. Appendix B gives the two costs in the same units. This calculation certifies the represented piecewise-constant field. A claim about the underlying continuum concentration also needs a bound on variation inside the donor cells. After repeated transfers the relevant map is the composition, with each intermediate mass divided by its cell volume to recover the concentration. Which few extra measurements keep the return affordable after several changes of grid? [147] [151] [153]

What a complete sumset table forgets

Take the five marks 0, 1, 2, 6, 22 on a ruler. On a second ruler move only the 6 to 7. Form all sums of two marks, allowing repeats. The rulers agree on every labelled comparison, including the tie 0 + 2 = 1 + 1. Each has fourteen distinct sums. Add a third mark. On the first ruler 2 + 2 + 2 = 0 + 0 + 6. On the second, the corresponding sums separate. There are now 29 distinct sums on the first and 30 on the second. The full old table was a richer description than the count, and even it stopped one operation short. [145] [151]

Hide the inner marks. Tell the observer only the common two-sum order, the diameter 22, and that the five integer marks start at zero and have greatest common divisor one. Seven rulers fit. Their three-sum counts are 28, 29 or 30. Allow the observer to choose two labelled three-sums and ask whether they are equal. Four questions are necessary and sufficient in the worst case. Across the seven rulers the varying tests give only four different patterns of answers. Each singles out a different group with fewer than 30 sums. On a ruler with 30, all four answers are no, and each possible exception has to be excluded. Let the same question return less, equal or greater and two suffice. Here is the first signed comparison. Write the marks as 0, a, b, c, 22 and compare 3c with b + 22, padding with zero to make three summands. Equality settles the count at 29. In either strict branch a single further comparison settles it. The accompanying calculation checks all 595 pairs of labelled three-sums and both optimal worst-case bounds. The right to ask a sharper question has cut the required depth in half. [151]

The sumset successor has an exact demand of its own. In my preprint a face of a polytope records every labelled order and tie through a chosen number of summands. Its geometry distinguishes a new equality from a change of order. The three-sum count cares whether values merge. A change of order by itself can leave the count untouched. Which of the possible new equalities can change the count on a whole family of rulers? Keep five marks and increase the number of summands. Can the geometry of the old table bound the number of new signed comparisons needed for the next count? Which tables attain that bound? I have used the same habit in polynomial fibers. I wanted to count integer inputs x for which f(x) = g(y) has a rational lift beyond those supplied by polynomial composition. Given the effective curve operations in the paper, geometry and finite congruence tests decide whether that count has a power term. Listing every exceptional input would be a stronger demand. These papers establish that I can make this move. They are part of the rival team’s reading list as well. How much my presence would add is a separate question. [145] [146]

The machine rival belongs in this very field. Tim Gowers reports that both Isaac Rajagopal and ChatGPT discussed the prospect of achieving polynomial diameter. Gowers asked the model to pursue it. Rajagopal describes the mathematical brief as replacing his geometric progressions with smaller sets having the requisite pattern of sumset sizes. ChatGPT paired carefully chosen numbers with their multiples. The pairs create the wanted collisions while the generators exclude accidental short relations. Rajagopal worked out the bound for fixed sumset order and sufficiently many marks. The machine helped recognize the opportunity and invented the representation. In the two cases here, agents chose the examples, identified the lost observations and found the two-question sumset procedure. Authorship and familiarity give me reasons to start somewhere. Neither result awards me a win. [145] [148] [151]

The rulers give the cleaner theorem about how many exact observations are necessary. The mesh gives the sharper question about the cost of reversing a loss at finite precision. I would start with the mesh. A receiving model can specify a family of regions, the mass tolerance and the storage budget before it knows which region it will ask about. A proof that one affordable set of local measurements protects all those queries would change what a simulation can pass to its successor. A pair of indistinguishable fields that defeats every proposed set would change what the successor must request. Either result would earn its keep with no claim about who discovers it. [147] [151] [153]

Give the rival the same freedom

The authors who realized M_{23} as a Galois group over the rationals describe how they used AI. People chose the conceptual strategy and directed the computation. Models helped search the literature, write code and carry out calculations. The team checked the final claims in Magma and PARI/GP without AI. Seven complex covers had the chosen branching. Galois conjugation fixed one and permuted the other six. The researchers recovered equations for the fixed cover from high-precision data and a change of scale. Here the work included deciding what to compute in the first place. [110]

Now require the degree-23 polynomial to have twenty-three real roots. The published example has seven. Changing its rational parameter will never supply the missing sixteen. Fried and Dèbes exclude a full-group totally real specialization for this three-branch-point strategy. Try four rational branch points and cap their branching orders at four. In the genus-zero screen developed for this essay, 16 of 480 candidate classes generate M_{23}. They give complex covers. The curve parametrizing them has genus 447. We have neither a suitable rational point on it nor the requested field. The screen tells the next investigator where an arithmetic question begins. [110] [116] [118] [119] [120]

The sparse replay supplies a useful control. On the eight-vertex cycle, rebuilding on every change took 46 conjugate-gradient steps in total. Perfect foresight also took 46. On CAST's five Chimera workloads, choosing the faster variant on each graph saves about 62 seconds on the 250-right-hand-side batch in hindsight. A measured opportunity can be large or zero. The arithmetic screen similarly locates the obstruction without constructing a totally real field. Giving the algebraist the right cover or telling the analyst which error the old hierarchy misses would settle the most interesting choice for them. [48] [76] [119] [120] Appendix C gives both calculations.

Give a specialist and an agent the same fresh brief as two agents free to disagree. Let both teams talk to the user who needs the answer. Compare what each actually proves on unfamiliar meshes or rulers at the same cost. On open briefs let each choose the successor problem, then exchange their proposals anonymously and pursue them. That distinguishes recognizing a good problem from being unusually good at solving one's own. The full comparison, with failed attempts and checks on familiarity, is in Appendix C. [89] [90] [105] [124] [132] [133] [134] [135] [137]

Then make the result public. Give both groups the new representation, the proof, the failed attempts and the code. Change a hypothesis that mattered. An advantage that survives this comparison is stronger than a good anecdote, faster code or one successful hint. It says the partnership can generate a new piece of mathematics after its last contribution has become common property. The evidence today gives us a reason to look for that recurrence and concrete ways to recognize it. [89] [90] [105] [107] [108] [130] [131] [133]

10. What could last after the discovery is public

The sharpest test is a theorem close enough to its predecessor to tempt you into copying the proof. In the Schiffer problem the outward derivative of an eigenfunction vanished on the boundary. The proof could discard certain harmonic coefficients. In the Berenstein problem that derivative was a nonzero constant. Those same coefficients now carried information needed to enforce the boundary condition. Colbrook, Sadeghi and Stepaniants found a candidate shape by Fourier search, moved to a fixed disk, kept the coefficients and bounded the omitted modes. They could certify a noncircular domain. [107] [108]

The researchers gave ChatGPT the earlier manuscript and its working code. The predecessor already contained the central construction, the fixed-disk formulation and the strategy for controlling the tail. The new paper credits ChatGPT with ideas for early versions of two proof steps and changes to the code. The authors developed and independently checked them. That record shows what traveled and what had to be invented again. It does not identify a winner in a trial against a machine team. [107] [108]

I would take the next unfamiliar mesh transfer and ask which region somebody will need to measure after the original field has been discarded. Give both teams the earlier proofs, the code and the failures. A mathematician who sees a missing relation, follows it to a better question and proves a usable bound has contributed mathematics. A centaur advantage begins when replacing that mathematician with a second agent repeatedly costs us such discoveries. If the two agents keep finding them, the method has a future of its own. [105] [130] [133] [134] [135] [145] [148] [151] [152]

A mathematician can invent a method so good that nobody needs its inventor on the next hundred cases. That is a mathematical success. When the hundred-and-first case breaks the method, who finds the missing distinction, and what does it cost?

Appendix A. The mathematical and benchmark record

What the released mathematics catalogue establishes

The OpenAI release examined on 4 November lists 721 manuscripts in 373 related families. The repository describes them as raw research data at different stages of verification. Three manuscripts were withdrawn after a shared sign error, and several others were repaired. Theoretical computer science has the most released families. Probability and statistical mechanics has the most manuscripts. The table uses the publisher's primary subject assignments. [4]

Released catalogue coverage at commit 4eaee875, dated 30 October 2026. A family can contain several related manuscripts. The final column counts families with links to Lean documentation for selected statements. [4]

Subject	Families	Manuscripts	Families with Lean notes
Number theory	31	58	17
Algebraic and complex geometry	36	86	9
Real and complex analysis	16	27	11
Convex and metric geometry	15	30	13
Theoretical computer science	41	74	35
Dynamical systems and ergodic theory	12	19	9
Combinatorics	37	50	33
Algebra	18	29	10
Probability and statistical mechanics	29	105	19
Mathematical logic	6	8	6
Group theory	14	22	12
Mathematical physics	25	59	17
Operator algebras	19	31	15
Topology	18	26	3
Functional analysis	11	19	10
Differential geometry	29	49	15
Partial differential equations	16	29	11
Total	373	721	245

The release began from approximately 4,000 posed problems and applied aggregation and a significance filter. Attempted-problem totals by field are unavailable. The 245 family documentation links cover selected statements of varying scope. The repository's approximately 42 percent figure counts formalized principal results using a different unit. Its classification has no numerical-analysis

category. Identifying computational results requires examining individual papers. [4] In the released arithmetic-progression reasoning record, Part II's opening prompt specifies the successor target. Other projects grant broader discretion. Zheng and colleagues let a pipeline recover open conjectures from the combinatorics literature. It connected an existing finite-geometric family of tangent lines to a conjecture about their union, disproving a proposed route to a Nikodym bound. The people who formulated the conjecture, the agent that recovered and attacked it, and the researchers who checked the new connection did different work. [4] [139]

In block theory OpenAI reports proofs of the blockwise Alperin weight conjecture for every finite group and prime and of Donovan finiteness over each fixed algebraically closed field with bounded defect-group order. I have not located independent specialist validation of either whole proof. Separate formalization records cover the statement that rational Hodge classes on complex CM abelian varieties are algebraic in every dimension and codimension, and specified Auslander-Reiten and Tachikawa counterexamples. Each record applies to its own stated result. [5] [6] [7]

The Saxl manuscript provides a second structural comparison. For the staircase partition $\rho = (m,m-1,\ldots,1)$, it tracks the cyclic module generated by the tensor product of its chosen row and column vectors. Coordinate projections separate a smaller staircase from a band. The proof must retain the actual generator and identify the projected image with the full induced module through explicit inverse maps. Mere occurrence of the desired constituents somewhere in the ambient tensor square would not supply the induction. A width-two band becomes a path contraction of 2×2 matrices, and dual-polytabloid pairings detect the required constituents. [99]

Even the constituent test needs a repair. In Example 5.4 an invertible segment product $B = 12T$, with T exchanging the two coordinates, has $B_{11} = 0$. The fixed endpoint pairing consequently vanishes. Conjugating the dual test basis by the matrix Q with rows (1,1) and (1,2) makes the conjugated product's (1,1) entry equal to 12 while preserving the original generator and endpoint letters. The final proof explicitly handles an uninformative test by changing the representation used to inspect the object. It supplies no chronological record of the model first failing and then discovering this repair. The formalization documentation reports the Saxl and universal-square positivity results, while excluding the stronger prescribed-tensor orbit-span assertion. [99]

The exact Fourier-transform manuscript counts preparation and indexing but leaves floating-point stability and useful crossover unresolved. The matrix-multiplication family gives an arithmetic-exponent improvement without a competitive finite matrix size. Each leaves room for further mathematical work on finite costs. [56] [100]

What the benchmarks and field reports measure

First Proof's thirty experts assessed anonymized solutions to ten research problems known to have solutions of at most twelve pages. Systems had 24 hours without human input during testing. Its rubric counts an average correctness score of at least 0.5, which allows minor repairable errors. The reported third-batch costs for GPT-6 Astra and ProofCouncil were $36 and $1,888 respectively, excluding benchmark creation and review. In the October organizers' retrospective scoring, the OpenAI entry went from five credited answers to seven and ProofCouncil from seven to nine, with lower reported batch costs. ProofCouncil's July system report describes six June submissions as correct up to minor revisions under the editorial judgments it reports. [124] The organizers changed the problems and some testing conditions between batches. [8]

TheoPhysEval used an automated final-answer checker on 50 expert-contributed theoretical-physics calculations. Astra with Codex reached 49/50 with web access and 46/50 without web search or the supplied background papers. Both tracks retained computational tools. The authors say the answers were unpublished at evaluation, some contributors had used AI, and they could not exclude leakage. Their test ran each whole track once per agent. [9]

Das and Fox credit GPT-5.6 Sol with the Murty-Simon edge bound floor($n^2/4$) and its even-order equality case. The authors simplified the proof and established the equality cases for all orders. Oum expounds OpenAI's cycle-double-cover proof. Nagda, Raghavan and Thakurta used AlphaEvolve to find combinatorial objects and reduction gadgets. Xu's cubic counterexample to the general Bilu-Linial signing conjecture and Lin and Zhou's revised spectral-signing work also report model contributions. The question for Ramanujan base graphs and efficient construction questions have their own hypotheses. [11] [12] [13] [14] [15]

OpenAI's October local-smoothing release asserts a Euclidean wave estimate in every spatial dimension at least two, together with formalization records and related restriction and Kakeya statements. These remain claims of the releasing authors, whose mathematical scope requires independent assessment. The Darcy example in the main text has expert benchmark review. [8] [16]

Further changes of representation and proof strategy

Aletheia found another use for a change of representation. Tony Feng credits the agent with connecting arithmetic eigenweights to symmetric-group representation theory through algebraic combinatorics and determining them for all classical groups. In type D of even rank, a commutativity assumption needed for an earlier result can fail. Two natural generators have to be treated through their two-dimensional span. Apart from building the agent, Feng describes his own role as rewriting its mathematics into a paper. On independence polynomials the division was different. Gemini 2.5 Deep Think helped establish a key inequality. Aletheia later suggested dual sets and the broader strategy. The human authors proved some of its suggested statements independently. [3]

An invertible matrix can register as zero in the test a proof needs. In the manuscript on Saxl's conjecture, one test asks for the upper-left entry of a matrix product. The product is twelve times the

matrix that swaps two coordinates. It is invertible, yet the test reads zero. Change the dual basis used to inspect the product and the entry becomes twelve. The generator the proof has to preserve stays fixed. What changes is the test that can see it. [99]

An agent can invent a method that leaves the specialist a different job. Alman and Vassilevska Williams credit Claude with an algorithm they then understood, simplified, strengthened and extended. For polynomially bounded integer inputs their deterministic bounds include $O(n^{1.9992})$ for 3SUM and $O(n^{2.99942})$ for directed all-pairs shortest paths with no negative cycles. The contribution is mathematical even though those exponents give no practical crossover. In operator theory Townsend and Greenbaum report that they and Crouzeix checked Jin's AI-assisted proof of Crouzeix's conjecture and believe it correct. Its decisive theorem emerged in an approximately sixteen-hour model run. Designing a numerical method that can use the theorem economically is a further research question, open to either partner. [17] [18]

Two model-produced graph manuscripts already combine compression with a bill for recovering what survives. The strong-thin-tree construction controls cut loads while retaining the rank information needed to extract a spanning tree. Its signing routine requires rational vectors with covariance exactly equal to the identity. The proof approximates the frame rationally, adds vectors that exactly complete the covariance, checks positive semidefiniteness and transfers the estimate back after removing the added vectors. Its claimed C/k cut bound has polynomial bit complexity. With the displayed constant 2^{20} and error $2^{20}\, r^{-1/8}$, the coefficient in the next packing bound turns positive only for $r > 2^{168}$. Even positivity need not retain one whole tree. The proof absorbs smaller cases into an absolute constant. The calculation shows how conservative its sufficient guarantee is. A practical crossover is a separate question, as is two-sided spectral approximation. [97]

The matching construction changes sparse forests without losing the physical value of an edge. When it deletes a copy it transfers that copy's accumulated increment to a base value. It records motion before cancellation so it can detect stale values and pay for refreshing them. A progress account pays for exploring local projections, while exact dyadic updates carry the precision requirement. Its claimed almost-linear randomized word-RAM bound charges allocation, multiword arithmetic, unsuccessful paths, scheduling and output. No competitive finite size follows from the stated asymptotic bound. An algorithm that recognizes stale values and budgets for their repair gives the next team a better starting point, whoever joins it. [98]

Structural mathematics already rewards this kind of economy. Gröbner-basis algorithms use signatures and syzygies to discard work. Roune and Stillman cut the potentially quadratic number of stored S-pair monomials to linear while retaining an integer per queued pair. In the Liquid Tensor Experiment, understanding what formalization required let Commelin and Scholze replace stable-homotopy machinery with a weaker sufficient result. Each discovery spared later users work. Every team that starts from the published result inherits that saving. [93] [94]

Bolzano records how a machine-proposed object and a human-proposed proof strategy fit together. Its prover agents invented a search problem in which finding a first collision determines where to look for a second. Four rounds produced a flawed proof by cases. Pavel Hubáček suggested asking what

information the collisions in a proposed reduction actually reveal. The point was to prove that, with overwhelming probability, the reduction circuit contained a collision that revealed no solution to the nested problem. Four more rounds produced a proof the authors judged largely complete. The machines invented the object. The specialist saw how to prove a limit on what could be learned about it. The authors never replayed that checkpoint with a fresh machine strategist. [125]

Appendix B. Solver evidence and extensions

Spectral comparison and approximate elimination

A constant vector has zero energy. For a connected graph this is the entire kernel. One can fix a vertex potential or work on the mean-zero subspace to obtain a positive-definite problem. A related class consists of symmetric diagonally dominant matrices with nonpositive off-diagonal entries, abbreviated SDDM here. Such a matrix has the form A = L + Γ, where Γ is nonnegative diagonal. It is positive definite when every connected component has a positive grounding term. The scope here is this concrete matrix class. [19] [26]

A preconditioned iteration repeatedly applies an inexpensive approximation to the inverse, aiming to reduce the total work needed to solve the original system. Write that action as solving with an auxiliary matrix M. For symmetric positive-definite A and M, suppose positive constants α and β satisfy the following inequalities for every vector x.

$$\alpha\, x^{\mathrm{T}} M x \le x^{\mathrm{T}} A x \le \beta\, x^{\mathrm{T}} M x.$$

Then the generalized eigenvalues of A relative to M lie between α and β, and the corresponding condition number is at most β/α. For a fixed positive-definite preconditioner in exact arithmetic, this ratio controls the standard worst-case convergence bound for conjugate gradients. A smaller condition number alone need not produce fewer iterations on a particular problem, since the spectral distribution matters. The computational benefit depends on convergence together with the cost of constructing and applying M. For Laplacians the comparison is made on the orthogonal complement of their common kernel. [52] [75]

Multigrid approaches the loss of information in another way. A cheap relaxation reduces some error components, and a coarse correction represents those that remain. Algebraically smooth error is error that the chosen relaxation struggles to remove. Its meaning depends on the operator and relaxation. Strongly connected regions joined by weak edges can move almost independently at low energy. A coarse representation must account for the resulting modes. The relevant objects are vectors, matrix blocks, graph partitions, and transfer operators. [20]

Refine the mesh in the Poisson calculation and the zero function earns a smaller absolute Euclidean residual against the nodal test functions. Its relative residual stays at one. Its energy error against the continuum solution never improves. An agent rewarded for shrinking the absolute residual could beat

its previous score every time without changing its answer. The missing ingredient is a bound connecting that particular measurement to the error we care about.

What one elimination leaves for the next

Gao, Kyng, and Spielman's AC(2) samples interactions among the neighbors of an eliminated vertex and keeps the sample connected. It performs well on many symmetric diagonally dominant systems with nonpositive off-diagonal entries. The authors conjecture a relative condition number of $O(\log^2 m)$, possibly $O(\log m)$, with high probability for m input edges. The missing argument must control the random distortions as the Schur complement itself changes. [26]

A triangle shows why connectedness and the right average are insufficient. Start with unit weights. Choose one edge uniformly, give it weight $3 - 2\varepsilon$, and give the other two weight ε, with $0 < \varepsilon < 1$. Every outcome is connected and every edge still has expected weight one.

$$\mathrm{E}[L_s] = L,$$
$$\kappa(L_s \text{ relative to } L \text{ on the mean-zero subspace}) = (2 - \varepsilon)/\varepsilon.$$

If edge 12 is heavy, (1,1,−2) and (1,−1,0) span the mean-zero subspace and have relative eigenvalues ε and $2 - \varepsilon$. Every realization can therefore be a poor approximation although the average Laplacian is exact. This calculation isolates the gap between those two properties and a spectral guarantee. Actual elimination rules such as VAC and CAST use additional structure. [47] [48]

CAST lets us follow one such choice into the calculation. Increasing its sampling parameter tightens a bound on the local second moment. It also introduces edges that later eliminations may have to process. Extra edges stay cheap on some low-degree graphs, propagate expensively on Chimera systems and have less new fill to create inside already dense blocks. Pivot degree alone cannot price the choice. At the base setting, CAST-1 retained AC's ordering, factor representation, and solver path. Over 250 right-hand sides its mean per-system speedup over AC was 1.17 on 17 SuiteSparse matrices and 1.18 on 128 Chimera systems. [48]

The five Chimera graphs expose two different ways to make a bad choice. On graph i3, CAST-1 was 1.38 times faster at four right-hand sides per factor. At 250, CAST-2 was 1.17 times faster. The crossover lay between ten and fifty solves. The matrices came from Newton steps. The right-hand sides were synthetic Gaussian draws projected to mean zero. The reported times include setup and solves to a verified relative residual of 10^{-8}. On the other four graphs, CAST-2 was still 1.3 to 4.6 times slower at 250. Knowing the number of solves can reverse the decision. Knowing that number alone cannot settle it. The authors could inspect the elimination profile after building a first factor. How cheaply could we detect the costly fill before paying for it? [48]

The four-cycle tells us exactly what a two-level correction must do. A recursive hierarchy needs its own analysis. An extra spatial dimension can also reopen a settled choice. A learned correction close to the exact coarse solve on a two-dimensional test needed appreciably more iterations than the exact solve on

the finest three-dimensional grid. The comparison appears below under learning a component of a solver, alongside other methods for constructing and reusing coarse directions. [54] [78] [79] [85]

A small calculation can also locate a missing assumption in a proposed guarantee. APLICUR updates a low-rank preconditioner during an LSQR solve. In the authors' Theorem 4.1 take the matrix and its rank-one CUR approximation both equal to diag(1,0), regularization $\mu = 1$ and target $t = 1/2$. The preconditioned augmented matrix has singular values 1/2 and 1. The printed lower bound is 1, and its condition-number bound is $\sqrt{5}/2$. The actual condition number is 2. The approximation error here is zero. [87]

The point to repair is the retained block of the ideal Gram matrix. Its eigenvalue is t^2, while the omitted directions have eigenvalue μ^2. If $\mu \le t \le \sqrt{(\sigma\hat{}\ell^2 + \mu^2)}$, with $\sigma\hat{}\ell$ the smallest retained singular value, and the CUR error E satisfies $||E||_2 < \mu$, the preconditioned singular values lie between $\mu - ||E||_2$ and $t + ||E||_2$. The regularized default satisfies the added target restriction. Each phase still needs the approximation-error condition. Note 87 gives the calculation from the paper's preconditioner. [87]

When checking dominates the cost

Saving a quarter of the iterations can buy almost nothing in wall time. On 1,000 AirfRANS systems the selected learned smoother of Chillón and collaborators converged on 99.9 percent. It used about 27 percent fewer V-cycles than relaxed Jacobi but saved only about 4 percent of solve time. The network predicted coefficients during setup. Applying the resulting sparse smoother cost more per cycle than diagonal Jacobi. AirfRANS ran in double precision because single precision stalled. On separate fine three-dimensional canonical-flow tests in single precision, convergence averaged 67.7 percent across ten trained models. The tolerance was 10^{-4} and the cap was 500 cycles. Both the workload and the way a model was selected belong beside its performance figure. [40]

Shui and colleagues found lower mean latency for selected human-assisted GPU kernels than for autonomous artifacts on five matched B200 operator definitions, by factors of 1.35 to 13.25. The fastest top-k indexer passed a three-trial check and then failed three of 200 targeted trials. An earlier fallback passed 1,206 targeted trials. The other four definitions retained reported latency advantages. Different model choices, development budgets and supplied context leave the contribution of human guidance unresolved. [102]

For N comparable queries, include setup and the cost of diagnosing, solving and checking each one. Recovery includes the further attempts and fallback allowed by the declared timeout rules. Report remaining failures as well as expected time.

$$T(N) = T_{\text{setup}} + N\,(T_{\text{diagnose}} + T_{\text{solve}} + T_{\text{check}} + \mathrm{E}[T_{\text{recovery}}]).$$

With extra setup cost ΔS and net saving $d > 0$ per query, expected break-even begins at $N \ge \Delta S/d$. Costs after a coefficient change can be correlated, so the mean needs an accompanying distribution. A diagnostic that takes several solves to evaluate can exhaust a four-solve job. A rule discovered once and reused on thousands of such jobs can still repay a long investigation. Each deployment pays for its own probes, checks and recovery. That distinction matters just as much for a proof. A criterion that excludes

a family of covers can be cheap to apply forever. Its discovery is one research result. The next comparison starts with that result available to everybody. [48] [62] [86]

Accumulated error and coarse-space construction

Ballard, Demmel, Holtz and Schwartz prove communication lower bounds for specified dense and sparse linear-algebra algorithms under stated memory-access models. Reducing arithmetic can leave data movement expensive. A new representation may change the executed algorithm and thus the costs to which those bounds apply. [96]

Changed coefficients can move the difficult error to a new subspace, and reducing communication can amplify rounding error. A stalled iteration could reflect a missing coarse direction, an inaccurate inner solve or drift in the recursively computed residual. A fresh residual can reveal the last of these without constructing a second hierarchy. Established replacement rules already automate parts of that diagnosis. [20] [63] [95]

The Simons workshop report connects practical benchmarks with theoretical solver design. It asks for parameterized matrix families suitable for both practical evaluation and theoretical analysis, together with algorithms that are fast in both settings. Its multigrid problem combines an implicit inverse with uniform spectral control, near-linear construction, and linear application cost for a polynomially bounded subclass of the grounded SDDM matrices defined at the start of this appendix. Its nonnormal-solver questions seek bounds describing the early behaviour of GMRES. These targets require the guarantee, its construction cost, and its predictive value to be considered together. [75]

Related algorithms show how local accuracy can control a complete solve. Kyng, Lee, Peng, Sachdeva, and Spielman analyze Schur-complement chains in which the global spectral distortion is bounded by twice the sum of the local error tolerances, in their exponential comparison convention, while application work is charged across the levels. The chain must control both the approximate Schur complements and the inverses used for eliminated blocks. The result applies to a different construction under different assumptions. It makes error allocation precise by relating local tolerances to a global accuracy budget and the work charged across levels. [58]

VAC replaces an elimination clique by a sampled spanning tree with the target marginals, connected support, and efficient sequential and parallel construction. CAST adds a local minimax characterization and a tunable bound on the normalized local second moment. A full-factorization spectral bound would also need to control the way local errors accumulate. [47] [48]

The literature supplies concrete constructions and limitations. Constrained local approximate ideal restriction combines local transfer construction with constraints on selected modes and uses aggressive coarsening to keep hierarchies economical. Optimal-transfer theory characterizes best choices for wider nonsymmetric and indefinite settings. Sparse and inexpensive realization requires additional work. Algebraic Schwarz methods construct coarse information from local generalized eigenproblems. In Al Daas, Jolivet, Nataf, and Tournier's published analysis, the theorem concerns SPD matrices and its bound is not fully independent of the number of subdomains. Broader matrix classes appear in

experiments. These distinctions identify the assumptions and costs a new construction would have to address. [49] [50] [51]

Finding and certifying an unknown domain

For the planar Berenstein conjecture the unknown is the domain itself. Can a bounded simply connected region other than a disk support a nonzero Laplace eigenfunction that vanishes on its boundary and has the same nonzero outward derivative everywhere there? Imposing both conditions makes the problem overdetermined. Colbrook, Sadeghi and Stepaniants searched for a shape using Fourier equations for boundary arclength. Those equations were good for finding a candidate and poor for bounding every omitted mode. To prove existence they pulled the problem back to a fixed disk and changed coordinates. Their estimates certify a noncircular domain with analytic boundary and a sign-changing eigenfunction. [107]

One detail explains why the earlier Schiffer proof could not simply be rerun. There the outward derivative vanished, which allowed certain harmonic coefficients to be eliminated. Berenstein's nonzero derivative carried information in precisely those coefficients. The team kept them, coupled the interior and boundary equations, bounded the infinite tail, and recovered the required sign after squaring the boundary condition. ChatGPT received the predecessor manuscript and working code. The authors credit it with ideas for early versions of two proof steps and changes to the code, which they developed and checked. Reusing the framework meant discovering which part of it had ceased to be true. [107] [108]

There is a second route to the Schiffer result itself. Cao-Labora and de Dios Pont let the integer order of rotational symmetry vary as a real parameter. Most of those values describe an auxiliary problem. Uniform estimates let them follow bifurcating branches far enough to reach actual integers, yielding infinitely many noncircular domains. The first proof certifies an individual shape. The second creates room between symmetry orders in which a whole family becomes accessible. The authors of the computer-assisted Schiffer proof credit AI with debugging early code. In the second, models checked asymptotics numerically and drafted arguments for Bessel-function estimates. Even with the conjecture held fixed, the mathematical invention and the division of labor changed together. [108] [111]

From a finite system to a PDE guarantee

Differential equations supply important sparse matrices, but the initial solver problem can be specified by the matrix, the right-hand side, and the required error measure. Graph incidence and finite matrix inequalities are enough to state it. Sobolev spaces and homological language are not universal prerequisites for this programme. A claim of accuracy for a continuous PDE also needs a valid approximation argument. For smooth finite-difference problems this can use Taylor expansion and a discrete maximum principle. More general settings benefit from other tools. [53]

Compatible vector discretizations provide a demanding later application. Their edge and face incidences constrain the transfers, and their coefficient weights determine the energies. Shen and collaborators' three-dimensional H(curl) construction leaves a sparse multilevel realization open.

Tuminaro and Glusa supply a different sparse construction with commuting constraints. The two papers vary coefficients in different terms, so one does not settle the other's robustness question. Auxiliary-space and GenEO methods remain substantive alternatives. Arnold, Falk and Winther require a discrete subcomplex and uniformly bounded commuting projections for their mixed Hodge stability result. Their square-mesh Maxwell example makes the danger concrete. On a crisscross mesh the first seven nonzero computed eigenvalues converge to genuine eigenvalues. The eighth converges to a spurious one. Further refinement can make the wrong answer look more convincing. [22] [23] [24] [25] [43]

A PDE application introduces an additional error between the continuous solution and the exact solution of the discrete system. To balance these two errors under mesh refinement, the solver needs a stopping tolerance suited to the current mesh. Certified reduced-basis methods and adaptive finite-element estimators address that further layer under explicit assumptions. Yano's complementary-energy construction, for example, requires symmetric coercivity, affine parameter dependence, and specialized offline work. The p-robust adaptive analysis cited here has a computable contraction criterion and a stated Poisson setting. These are application-specific extensions, not prerequisites for studying the finite matrix algorithm. [27] [46]

Van den Berg and Breden get an effective global existence result for the one-dimensional Ohta-Kawasaki equation. Their coefficients approximate one trajectory from prescribed initial data. Bounds on the omitted modes validate successive intervals and a final step covering all future time. For the parameters and initial data in Theorem 1.2, the exact trajectory stays within 9×10^{-7} of the approximation, uniformly in space for every $t \geq 0$. Staying in a narrow tube would still permit motion inside it forever. The final step also certifies a stable stationary solution and a neighborhood in which trajectories converge to it. That stability argument supplies the long-time conclusion. On a separate Swift-Hohenberg trajectory, their change of validation method made the proof about 25 times faster on the reported hardware. [109]

Agent search and software optimization

Coding, benchmarking, and revising a solver are already being automated. Bader and collaborators report agents improving the tuned VieCut minimum-cut implementation by factors of 1.28 sequentially and 1.63 at 32 threads on the real-world k-core workload. All 23 accepted edits received human review, and the same 53-instance workload was used for threshold tuning and reported timing. This is productive algorithm engineering under the stated conditions. [28]

FrontierCS 2 brings that competition into recent research code. Its 4 November 18-system, 48-task snapshot includes 13 systems passing Clusterwise SpGEMM, eight passing the KAMI numerical-kernel task, and seven passing Polar Express. Each task uses its own correctness and performance criteria. They are nevertheless actual successes in sparse and numerical algorithm engineering. [69]

ADSD provides a more targeted comparison. Across nine model/task comparisons, direct revision worsens error in five, whereas its combined diagnosis, literature-retrieval, and skill pipeline reports lower error throughout. Five rounds of direct revision on GOC-500 do not close the gap. This combined

automated procedure can improve the tested searches beyond simply requesting another attempt. The ablation assesses diagnosis, retrieval, and skill construction together. Off-the-shelf solvers are excluded, and matching the final revision budget does not equalize the cost of constructing the pipeline. The result establishes a useful automated mechanism whose costs should be included in a comparison with specialist-led work. On the optimal-power-flow task the agent used the failure traces to ask how first-order quantities should be put on comparable scales. Its next diagnostic question arose inside the supplied optimization goal. [30]

AlphaEvolve can search over constructors and algorithms and use multiple evaluation scores, including expensive accelerator measurements. TTT-Discover updates model weights while working on the current problem, using feedback from earlier attempts. Its reported method targets one high-performing result under a continuous reward, with applications in mathematical bounds, GPU kernels, and algorithm engineering. Transfer of the learned procedure to further problems is a separate question. A comparison should include agents that learn during the investigation. [44] [88]

Learning a component of a solver

A research agent can discover a classical algorithm that uses no model when deployed. A neural component can predict a coarse space, a sparse correction, or a tuning decision. A surrogate can approximate a solution map. These are different mechanisms. For the finite programme, the central question is whether the resulting factor or hierarchy solves the original system at the required accuracy and total cost.

Sparse approximate inverses illustrate why representation design and hardware belong in the same argument. Yang and collaborators replace triangular applications with sparse matrix-vector products and train the admissible entries. Their comparisons change across matrix families and CPU/GPU implementations. Default multigrid remains strong on diffusion examples. The research decision concerns which representation preserves enough spectral information at the cost the hardware can actually deliver. [77]

Melchers, Dolean, and Abdelmalik compare four neural-operator coarse corrections while fixing the output coarse space. They vary input discretization and linear dependence on the right-hand side. Preserving right-hand-side linearity and matching the input projection to the output basis give the Neural Green's Operator a Galerkin-type coarse correction. The experiments connect these choices to spectral behaviour, stagnation, and compatibility with Krylov solvers. Their November revision adds a three-dimensional diffusion test. On the finest grid, with 80 points per side, the learned correction took 37.3 flexible-GMRES iterations on average. The exact coarse solve took 23.0 with the same smoother, and the smoother alone took 124.1. The learned correction still helped substantially. Its near parity with the exact coarse solve on the two-dimensional test did not carry over. These counts tell us what to investigate before trusting the architecture on another grid. [85]

Bao, Cui, and Li learn coarse spaces from errors left behind by smoothing. Their objective compares subspaces independently of the basis used to express them. A perturbation analysis relates the discrepancy, including pre- and postsmoothing, to the resulting two-grid operator. The sharper

quantity has a small Gram-matrix representation, but forming it requires solves with the original matrix, so it is used offline. The convergence bounds are conditional and the experiments concern stated parameterized families. The work exposes a concrete next problem of finding an informative, affordable test for when to reuse, enrich, or rebuild the learned space. [86]

RAPNet learns sparse corrections to coarse and transfer operators within an algebraic multigrid hierarchy, with transfer from training subgraphs of about 4,000 nodes to million-node tests. For some graph families the training patches come from large graphs to expose relevant structures. Its setup comparison uses classical CPU execution and GPU RAPNet, counting transfers and inference. A general convergence guarantee for unrestricted corrections remains open. These results demonstrate meaningful transfer, with its scope set by the training graphs, hardware, and stated limitations. [37]

AutoSpec learns short spectral recurrences, producing matrix polynomials for approximation, preconditioning, and filtering. Its comparisons include Chebyshev preconditioners supplied with the same spectral estimates and show both wins and losses. The tested operators are positive definite after the stated transformations. Some originally indefinite eigenproblems are squared, and some difficult linear systems also use incomplete Cholesky. Those details are part of what the results mean. Learning can thus return a compact executable numerical procedure with recognizable operations. An iteration-count improvement still needs a separate account of probe cost, training, and application time before becoming an end-to-end speedup. [38]

PCGBandit selects among 33 fixed preconditioner configurations during one OpenFOAM run and counts learning and backstop costs. In the Shercliff electric-potential solves, expensive early trials can absorb the savings of short runs. He and collaborators report total-runtime reductions of 36.1-43.3 percent on diffusion and 19.8-31.0 percent on diffusion-advection against the BoomerAMG reference default over 5,000-system sequences. In a separate same-hierarchy test on SPD diffusion, the mean first-attempt solve time favored a prescribed two-weight schedule by 2.8 percent, with a reported spread of 3.8 percent across six frozen learned states, whereas the same schedule failed frequently on variable-flow diffusion-advection. The strength of the comparator changes with the problem. [39] [62]

Precision, storage and the cost of a solve

Fast algebra and numerical stability can also be reconciled. Demmel, Dumitriu, and Holtz give normwise-stable algorithms for standard dense linear-algebra operations with essentially the matrix-multiplication exponent, under their stated arithmetic and stability assumptions. That result neither preserves arbitrary sparse structure nor establishes practical constants. A sparse implementation still needs control of its structure, accuracy, and application cost. [60]

The OpenAI release also reports constructions with explicit implementation-cost accounting. OpenAI's manuscript on exact Fourier transforms reports an asymptotic operation bound below n log n while explicitly accounting for scalar preparation, schedule construction, and indexing. Its construction preserves working width and restores borrowed registers. The result assumes exact complex arithmetic, supplied roots of unity, and unrestricted coefficients. A floating-point realization would need its own stability analysis. The proposed architecture reaches well beyond a bare existence theorem. [56] A

separate OpenAI manuscript on integer multiplication explicitly propagates fixed-point errors and bounds the final coefficient error below one half so that rounding recovers the exact answer. Its asymptotic bit-complexity result comes with enormous constants, not a useful practical crossover. These are reported constructions, not independently certified here. These reports put finite-precision analysis among the tasks to include in a comparison with research agents. [57]

Arithmetic precision is another resource to allocate. McCormick, Benzaken, and Tamstorf develop a purely algebraic mixed-precision multigrid analysis for sparse symmetric positive-definite systems. Residual evaluation, correction solves, and iterate updates contribute differently to the energy-error bound, and precision can vary across levels. This supplies an established framework for spending precision where it matters. Better adaptive rules must respect the convergence and rounding assumptions, as well as the costs of storing and applying the sparse operators. [59]

Jang, Jolivet, and Mary identify a structural failure in mixed-precision GMRES-DR. Its derivation uses simplified relations from Arnoldi iteration and orthogonality whose finite-precision defects can make the method stagnate. Returning to a more general augmented formulation permits lower-precision inner work without those simplifications. The experiments retain precision-sensitive cases and defer hardware-specific timing. The example links a diagnosis to a mathematical repair of the formulation. [63]

A more focused question is whether the memory saved by compressing a preconditioner outweighs the cost of extra iterations. Khan and Carson report a steam1 case with 21 GMRES iterations against 14 in uniform precision, while storage falls to 74.9 or 42.6 percent of the uniform version under two thresholds. Those figures leave runtime open because reduced traffic competes with extra iterations and other work. The authors leave predicting favorable matrices as further research. Meanwhile doubly adaptive GMRES already combines iteration-dependent and entry-dependent precision automatically, with a reported peak 1.6-fold speedup on nd24k. Figure 6 compares runs that attained the baseline backward error. Table 3 records stagnation on several matrices and shows that one heuristic attained the target on only one matrix. The failures belong in the comparison. The research target is a better, transferable prediction of when these precision choices repay their cost. [64] [65]

The distinction between a research agent and a solution surrogate also matters economically. Huang and Stuart's comparisons of neural operators with classical solvers depend on accuracy and on amortized training and data costs. The plateau they observe for particular Darcy surrogates is not evidence that an agent cannot discover a better sparse solver. For the present programme, the relevant comparison is the final matrix algorithm, its construction costs, and the workload over which those costs are recovered. [41]

Zhang, Li, and Saad tested a learned multilevel preconditioner on 867 real non-SPD SuiteSparse matrices. Median setup/training time was 135.77 seconds versus 0.18 for the tested AMG. Median solve time also favored AMG, 1.31 seconds versus 8.08. Mean solve time favored the learned method, 9.17 seconds versus 50.87. Its convergence win shares are conditional on valid runs. Separate medians cannot be added to recover paired total times. Cross-matrix transfer and the tails of the cost distribution both matter when choosing the deployment workload. [42]

Nonnormal operators and verified sparse solving

Crouzeix points to an adjacent programme in nonnormal matrix computation. For a function f holomorphic near the numerical range W(A), the reported resolution gives the following bound.

$$||f(A)||_2 \leq 2 \max_{z \text{ in } W(A)} |f(z)|.$$

For normal matrices the spectral theorem already supplies a constant-one bound. The new direction concerns operators for which eigenvalues alone give an incomplete account, including nonsymmetric systems and some directed-graph problems. Crouzeix and Greenbaum use the new result to improve bounds on other spectral regions and apply them to GMRES and rational Krylov methods. Choosing a useful region, approximation, and inner-solve accuracy is a further design problem. A numerical-range bound containing zero can fail to certify GMRES contraction. A valid cutout that excludes zero without surrounding it may help, but requires additional information about the inverse. [66]

The direction of the approximation matters when estimating a spectral region. For an orthonormal Krylov basis V, W(V* A V) lies inside W(A). This projected set is an inner approximation, while a certificate based on an outer enclosure needs a bound in the other direction. An ordinary largest Ritz value of a Hermitian support matrix is a lower estimate. An outer support line needs an upper bound. Chen and Urschel quantify Krylov approximation of numerical ranges and its limitations in exact arithmetic. A floating-point certificate requires further control. This identifies the information a cheap diagnostic must obtain before its estimate can support a rigorous guarantee. [67]

Verified sparse solving also has established automatic methods. Rump develops entrywise error bounds for real and complex sparse systems, using inexpensive bounds first and stronger residual-based work when needed. Different factorizations and fill patterns can change verification cost sharply. A new certificate must therefore improve an existing procedure, or apply where its assumptions fail, rather than claim that rigorous checking has not yet been automated. [68]

What autonomous solver research already achieves

LinearSolveBench released sparse solvers selected from 640 candidates per model on 96 magnetic-diffusion cases. The best passed the specified checks on all 96 and ran about 1.85 times faster than GMRES(50) with BoomerAMG when the original case times were summed. The timed work covered solver creation, setup and solution. Discovery, compilation and independent verification were outside that clock. The release adds a fresh correctness check on the same cases. Speed on a new family is an open question. [32]

An agent also gets to use the adaptive solvers that already exist. PCGBandit tries configurations from a menu of 33 during a flow simulation and charges for the trials and the fallback. The Shercliff simulation made 7,500 pressure solves. With the reference incomplete-Cholesky preconditioner those solves consumed 59 percent of the runtime. There is a bill worth cutting, and the learner improved the pressure solves. On the electric-potential solves in the same simulation, its costly early choices swallowed much of what it could save later. Someone who claims to have invented a better diagnostic has a working competitor to beat. [62]

What happens to a human contribution once it can be written down? In a finite-element task in SciCode-Verified, agents kept replacing the specified weak form with its textbook counterpart. Restating the equations at implementation time corrected that choice. Across 64 corrected main problems, expert background raised GPT-5.5 from 65.6 to 90.6 percent, Gemini 3.5 Flash from 51.6 to 87.5 percent, and Claude Opus 4.8 from 57.8 to 78.1 percent. Each model ran once per condition. Every scored subproblem had to pass, and it could pass in either of two software environments. Writing the background and independently rediscovering it were both unpriced. Once every team receives it, a later gain from using it cannot be credited again to a specialist's presence. [35]

The autonomous competitor can also question its judge. During AIDE²'s eight-day research run an agent found that one failing test case could crash the scoring of the entire set. It repaired the scoring script. Seven harness changes were accepted. The target suite and outer objective were supplied, but no operator had named that repair. A machine can choose a new diagnostic project inside an existing programme. [55]

In SVG-EAR, Astra wrote a probe that periodically compared sparse attention with exact attention on a small sample of queries. Its code assigned a budget to each head using the measured error and fell back to dense attention when necessary. Astra also supplied a variance correction. Two additional agents independently rediscovered that correction in exploratory runs. The final code reduced mean runtime by about 18 percent, improved all three mean fidelity measures and passed 43 of 48 hidden checks. None of the twelve evaluated agents passed the whole task. The study assessed the combined changes. The five fidelity failures also occurred in a separate exact-attention implementation. An agent had devised an informative measurement and used it to change the calculation. [70]

What an approximate overlap measurement buys

Use the single-transfer model in Section 9. Each donor concentration lies between zero and one, the overlap matrix O is exact, and each reported receiver mass differs from its true value by at most δ. Subtract the known masses of the uniform half-concentration field. Estimate the requested regional mass by taking a weighted sum of these centered readings with coefficients c and adding back the known half-concentration mass of the region. A donor concentration now differs from its center by at most one half. The triangle inequality gives the following guarantee.

$$\text{absolute mass error} \le \tfrac{1}{2}\,||w - O^{\mathrm{T}}c||_1 + \delta\,||c||_1.$$

Here the 1-norm is the sum of absolute entries. The first term prices the part of the regional weights the chosen overlaps fail to represent. The second prices the measurement noise the chosen coefficients can amplify. Both terms are in units of mass. With independent interval bounds on concentrations and noise, choosing their signs to agree with the respective coefficients attains the bound in the worst case for this estimator. Extra stored measurements append rows to O and can lower the best such bound. Error between the donor field and a continuum solution requires its own estimate. [147] [151] [158]

Appendix C. Experimental protocol

Solver implementations and held-out tests

Before the first trial, record the implementation and version, precision, hardware, smoother, coarsening and coarse solve, grid sizes, μ, weight ranges, change generators, budgets and stopping rules. Each preconditioned-CG run requires a fixed linear symmetric positive-definite applied cycle. The three policies respectively keep the old hierarchy intact (F), update its level matrices and smoothers while retaining transfers (U), or rebuild it (R). To certify F, compare the changed weights with the initial weights and use proved α and β for the actual baseline. If only spectral estimates are available, report the policy comparison as empirical. The independent μ-based residual check can still certify answers under all three policies. Acceptance under U supplies no uniform quality bound on the updated hierarchy.

Begin with manufactured solutions normalized to unit A'-energy, including directions aligned with the modes missed after a weight change, localized patterns, and generic patterns. Form their right-hand sides consistently and account for construction error. Normalization gives the prescribed absolute energy tolerance a comparable scale across the controlled tests. Then test unknown solutions with a declared right-hand-side normalization and the same prescribed absolute energy tolerance τ. Use four right-hand sides as the primary short workload, with one and a prespecified longer horizon as sensitivity cases. Hold out coefficient-change regimes and larger sizes.

Keep manufactured reference solutions and exact-error measurements within the independent checker, outside the information supplied to the candidate solver or policy selector. On reference cases, record the first iteration at which the actual energy error and the certified bound each reach the prescribed tolerance. Their difference counts the additional iterations associated with the conservative stopping rule. Measure the corresponding solve time and deployed checking cost separately. Report inflation from validated residual evaluation separately.

Obtain reference-error measurements in separate calibration runs. Match the timed runs' inputs, initial states, arithmetic, random seeds, and algorithmic decisions, and compare their residual histories and iteration counts. If the trajectories diverge, attribute the calibration to its own trajectory. Keep the deployed acceptance checker active in timed runs and charge it to solution cost. Account for reference instrumentation and replay under research and validation, identifying shared benchmark costs and costs incurred by each group.

Time each selector and each fixed policy on paired workloads. Charge transfers, updates, coarse factorizations, setup, solves and deployed independent checks. On failed acceptance, rebuild and retry within a fixed budget, then count unresolved cases. Before evaluation, fix the worthwhile cost-improvement margin, uncertainty rule and allowable failure rate. Report paired total times and outcome distributions at the same accuracy for each workload horizon.

An elimination project can use AC(2) with a fixed implementation, including its arithmetic, randomization, and input class. Compare any repair with unmodified AC(2), approximate-elimination

alternatives such as CAST, and strong multigrid or direct-solver baselines for the declared workload. Reserve graph families as well as seeds, including weakly joined clusters, long narrow connections, graded weights, degree variation, and larger instances. Report fill, peak memory, setup, factor application, convergence, independent error checks, failures, and fallback costs. Separate a single unfamiliar matrix, a short sequence after a structural change, and a long related sequence. A diagnosis should predict which regime repays its cost. [26] [48] [62]

Broader coarse-space work should include adaptive smoothed aggregation, compatible-relaxation or bootstrap constructions, and hierarchy-reuse policies on a declared SPD family. Compare with online configuration methods and well-chosen fixed settings, not only library defaults. On manageable sizes include an expensive spectral reference. Charge candidate generation, probes, local eigensolves, rejected attempts, rebuilding, and checking. Measure coarse dimension, transfer nonzeros, hierarchy size, memory, and total time. All groups may change support or aggregation and invent alternatives. An exact-coarse-solve theorem must be assessed separately from its recursive implementation. [20] [39] [54] [76] [78] [79]

A high-probability theorem must state its input assumptions, randomization, and failure-probability dependence. Practical predictions also require the relevant constants. Keep exact small-instance calculations, large-instance estimates, and validated proofs distinct. Independently recompute residuals against the original operator and check the declared kernel and positivity assumptions.

Across these projects, retain final evaluation cases outside development and record exploratory discoveries separately. Assess a proposed checker correction against the original mathematical requirement. An added application requirement defines a further task. [45]

Reproducing the two pilot calculations

Start the solver pilot with a shifted grid Laplacian $A(w) = L(w) + \mu I$, positive edge weights and fixed $\mu > 0$. Build an initial multigrid hierarchy, change the weights once, and solve the changed system $A'x = b$. Every trial starts with the same initial hierarchy for its operator. A selector has three choices. [76] [78] [79]

Three responses to the same change in edge weights.

Policy	Operation after a weight change
F (fixed)	Retain the old preconditioner in full, including every level matrix, smoother, transfer and coarse factorization.
U (updated)	Retain the transfers. Recompute the Galerkin operators, smoothers and coarse factorization.
R (rebuilt)	Build a new hierarchy with the original setup procedure.

Only F keeps the preconditioner covered by the weight-ratio argument in Section 6. Check every policy against the same independent energy-error bound from Section 8. Before designing a selector, find out

what choosing well could buy. On development jobs, replay F, U and R from the same saved hierarchy. Repeat paired timings so the apparent saving does not depend on picking a lucky run. Give hindsight all the replay results. Let it pick one action per job to minimize total time while meeting exactly the accuracy and reliability rules the live selector must meet. Charge setup, checking, recovery and failed runs alike. If every job must pass, this reduces to picking the cheapest successful action on each. If no assignment meets the rules, even foreknowledge fails. Compare this free hindsight choice with the fastest eligible baseline, counting both fixed policies and the existing adaptive policy. The gap is a ceiling on what choosing among F, U and R could save on that batch before paying for information.

The published CAST timings let us put a number on that ceiling for a narrower choice. CAST reports 28, 24, 27, 26 and 23 systems on its five Chimera graphs. At 250 right-hand sides, weighting the published graph means by those counts gives about 2,313 seconds for CAST-1 throughout and 2,251 if hindsight chooses one of the two CAST variants per graph. The entire saving comes from the third graph. Its 27 systems save about 2.3 seconds apiece. The displayed numbers are rounded means of per-system medians over five factor draws. A policy choosing separately for every matrix could have a different ceiling. On the bench we can price this particular choice at about 62 seconds per batch before paying for a probe. On a deployment we need its matrices, right-hand sides, reuse horizon and best existing policy. [48] [120]

The same distinction shows up on an eight-vertex cycle small enough for exact arithmetic. Across four changes of weights, four right-hand sides apiece, F, U and R took 65, 48 and 46 conjugate-gradient steps. Picking the fewest steps on each job also took 46. A selector earned nothing on that score. Setup could still make updating cheaper than rebuilding. Rasouli and collaborators actually timed related updates after mesh deformation. On their stretched mesh a fine-level update beat rebuilding through the ninth case, and rebuilding paid on the tenth. Their fine-level policy updated the operator. Our fixed policy retains the old preconditioner in full. Replaying these choices is established practice. Once these comparisons are public, the agent can run them too. Recognizing a documented crossover on the original workload supplies no evidence of a continuing centaur advantage. [76] [120]

The arithmetic replication starts from a different demand. Can an irreducible degree-23 polynomial over the rationals have all twenty-three roots real and Galois group M_{23}? Huang and collaborators solved the problem without the condition on the roots. Even the branch data for their construction took work to choose. Seven complex covers had the chosen branching. Galois conjugation permuted six and fixed one. They found equations over the rationals for that one. Forty-digit expansions and a change of scale recovered its equations. For the other covers even sixty digits left some coefficients unrecognizable. The team combined data from the six other permutation triples and changed scale again. The authors describe AI-assisted computation under human direction. [110]

Try the published polynomial on the new demand. An exact Sturm count gives seven real roots. The other sixteen form eight conjugate pairs. Choosing another rational parameter cannot fix it. The real parameter line, with infinity included, stays connected after its only real branch point is removed. Complex conjugation keeps the same action away from that point, and the authors checked the exceptional point separately. A stronger result was already on the shelf. Fried and Dèbes exclude a

full-group totally real specialization of any regular three-branch-point M_{23} Galois cover over the rationals. In the relevant three-point case the group would have to be generated by two involutions and would be dihedral. M_{23} is not. Malle's constructions for M_{11} and M_{12} show why trying a different family is serious mathematics. Give every team these results before the clock starts. Reading a known exclusion can save work. Rediscovering it earns no extra credit. [110] [116] [118]

Here is a fresh structural pilot. Try four ordered rational branch points, an interval where all the sheets are real, and a genus-zero source curve so its equations might be manageable. Cap the four branching orders at four. Before reconstructing anything, enumerate the permutations that could describe such a cover. In the natural action of M_{23} on 23 letters the resulting test reduces to four permutations (a, ab, bc, c), where a, b and c are involutions and ab and bc have order four. Their product is automatically the identity. They must also generate the full group. Counting up to simultaneous conjugation in M_{23} gives 480 candidate classes. Exactly 16 generate M_{23}. An independent orbit count gives the same answer. Anyone planning to recover equations for a rejected class can stop immediately. [118] [120]

The sixteen survivors do not settle the arithmetic. Each describes a complex cover whose degree-23 source curve has genus zero. The curve that parametrizes these covers is a different object. Häfner had already studied this branching pattern and computed genus 447 for the ordered parameter curve. The finite screen has identified where a rational point would have to be sought in the appropriate real region. It has produced no such point. The next arithmetic question is whether the relevant part of that curve admits a rational point. Finding an obstruction there could retire the route. Passing the finite group test alone gives none of the sixteen candidates equations over the rationals. [119] [120]

On the eight-cycle, consecutive edge weights alternate between 1 and 1/64 and the diagonal shift is 1/16. Rescale all weights by 5/4, exchange the two parities, raise the original strong weights to 4, or set every weight to 1. The rebuilt hierarchy pairs adjacent vertices along the heavier parity, retaining the original pairing on a tie. Its two-level preconditioner uses one damped Jacobi sweep on each side of an exact Galerkin correction. Fine PCG starts at zero and checks the original residual against an absolute energy-error tolerance of 10^{-4}. On the exchange alone, F, U and R took 18, 13 and 11 total steps across the four right-hand sides. The supplied code specifies the vectors, reruns all policies with rational arithmetic and separately reconstructs the graph-level CAST comparison from Table 5. [48] [120]

For the M_{23} screen the branch index is 23 minus the number of cycles, fixed letters included. Genus zero requires the four indices to sum to 44. The two end involutions contribute 8 apiece. Under the order cap the two middle permutations must each have order four and index 14. Fix the middle involution b. The 1,008 possible endpoints split into three orbits under its centralizer, which has order 2,688. For each left representative take the right orbits under its stabilizer. This gives 480 ordered classes. Burnside's lemma independently counts the same orbits. There are 43,008 transitive labeled pairs for fixed b. Every such triple generates M_{23} and has trivial stabilizer, yielding 43,008/2,688 = 16. The script prints a witness, verifies the branch product, and checks rationality of its conjugacy classes. A rational class condition alone cannot supply equations over the rationals. [110] [118] [119] [120]

Comparing teams on the same brief

What would establish each claim?

Claim	Required evidence
Better performance on a stated workload	Independent measurements at the same accuracy, including setup, checking, and recovery.
Transfer beyond development cases	Withheld families with meaningful changes in geometry, coefficients, topology, or other relevant structure.
A uniform mathematical guarantee	A valid argument covering the stated class, with its assumptions and parameter dependence explicit.

Decide in advance when an obstruction answers the question. Suppose a team proves that a specified family of cheap observations gives identical answers on two allowed changes that require different actions to meet the cost and accuracy target. That can settle whether the proposed policy class will suffice. In a cover project, a proof excluding the prescribed class of realizations earns the corresponding credit if its scope settles the declared question. A stalled search earns neither verdict. Flag in advance which briefs admit a positive construction, a decisive obstruction, or both. Report the two kinds of result separately as well as success under each brief. In the four-solve performance pilot the positive outcome still requires a faster accepted method. Evidence for stopping a search and any demonstrable work it saves get their own record.

Hold the lead agent fixed. Give one team a specialist. Give the other a second agent chosen on separate development problems, free to use a different model, tools or proof strategy. Each gets the same starting evidence, access to the same tools and the same total budget. Also run the lead on its own with the resources available for extra attempts. Let C_{H} and C_{A} be the total research costs at which the human-agent pair and the machine pair respectively first produce an independently accepted result. Count advice, checking and failed searches. Assign infinite cost to a run that exhausts its protocol without an accepted result. At budget b the human partner's advantage is

$$\Delta(b) = \Pr(C_{\mathrm{H}} \le b) - \Pr(C_{\mathrm{A}} \le b).$$

At three hours Archon's run with the human blueprint had finished and the other run had not. At four hours both had finished. Repeat the partner swap across fresh problems. Report correct rescues and harmful vetoes as well as the net gain. A generalist paired with the lead tests whether the gain requires subject expertise. A specialist working with the same software and sources without generative AI tests whether the machine helps the person too. Let the strongest machine-only procedure found in development compete for the same budget, even if it uses parallel independent searches instead of a pair in conversation. [90] [103] [104] [123] [124]

In the open-ended transfer experiment, the unit of comparison is a portfolio of follow-up problems selected from the same precursor and historical record. Predeclare time and research costs, eligibility

for a new claim, the process for checking originality, and how independent assessors will judge mathematical substance. Keep the proposed question and success criterion on record before its result is known. Report questions proposed, questions abandoned, verified results and assessor disagreement. Separate value attributable to selecting the problem from value attributable to solving a named one. Include previously unanticipated follow-ups that meet the criteria. [4] [132] [133]

A human can help a model because their errors overlap less, and lose that benefit by deferring to it at the wrong time. Steyvers and colleagues combined independent image judgments and confidence scores under the same fitted rule. Human-machine pairs gained most over pairs of different networks when the partners had comparable accuracy and the images were noisy. Hemmer and colleagues instead held human participation fixed. On fifteen images, changing which classifier advised participants lowered mean team error from 24.7 to 14.6 percent although both classifiers missed four images alone. AutoAnalysis reports that in its subset comparing collaborations with both solo conditions, average team performance fell below the stronger individual condition. None of these designs asks who will invent a missing mathematical test. [103] [104] [129]

For each portfolio, define the project sampling frame, intended population, acceptance rules and distribution of briefs before execution. Classify projects by the deliverable specified at entry. Fix the treatment of genuinely mixed briefs and pair analysis by project. A construction or obstruction counts toward the primary full-project outcome only when it meets the scope announced for that brief. Record separately which result was obtained, which claimed stopping decisions independent assessors accepted, and which downstream costs the evidence supports. Report construction-only and obstruction-eligible briefs separately so different mixes of outcomes remain visible. Within the method pilot, method acceptance requires the predetermined accuracy, reliability and deployment-cost improvement on its withheld workload.

Use several participants and repeated seeds, with independent final assessment. Assess competence before assignment. Fix eligibility, allocation, model and harness versions, independent project count and the primary budget-accounting rule in advance. Pilot data can inform a precision or power analysis while the final evaluation stays reserved. On development briefs choose a machine adviser with strong performance alongside the lead. Allow different model families, private contexts, rival representations and computational tools. Freeze the pairing and the best unrestricted machine-only procedure before the fresh briefs arrive. Give every arm the same underlying evidence, access to tools and freedom to abandon a formulation. On matrix projects this includes the operator, grounding convention and tolerance. On arithmetic projects specify the target group, base field and required local or real behavior. Choosing a cover, branch data, or a different route belongs to the research team. Screen fresh briefs for known public answers before assignment. Assess any independently justified solution meeting the stated target, including a construction the assessors did not anticipate. A proof excluding one candidate family earns credit for that narrower result. Additional seeds measure variation within a project. Inference across projects draws its replication from independent projects.

Record person-hours, elapsed time, model expenditure, computation, checking, failed attempts and deployment costs separately. Report equal-time and equal-cost comparisons. Before combining costs,

specify a human-hour price and the number of deployments over which discovery is amortized. Report sensitivity to declared alternative prices and horizons. A stop at one budget leaves success at larger budgets unknown. Fix any additional budgets in advance. Use the same economic rules in the specialist-alone arm when the stronger collaboration claim is tested. When a team uses a probe to choose its next step, record the decision it could change and the cost of obtaining the answer. Raiffa and Schlaifer give the formal comparison when the actions, beliefs and payoffs are specified. [122]

Evidence about choosing a partner

A weaker investigator can be the better partner. Minghao Liu and colleagues took an image classifier and changed its partner. The human answers were 81.9 percent accurate, the second classifier's 90.0 percent. Yet at least one partner had the right answer on 95.8 percent of images in the mixed pair, against 92.3 percent for the two machines. Those ceilings assume someone always knows whom to trust. Handing the base model's low-confidence cases to its partner also favored the human on these data. The human's errors fell in different places. [126]

Different habits can also make partners worse. Jones and colleagues asked pairs to describe and identify the same abstract figures over five rounds. Two copies of GPT-5 reached 99 percent accuracy by round four. A human paired with GPT-5 reached 70 percent in round five. Shared habits helped two machines understand each other. On the images, the weaker partner covered errors the stronger one missed. In the naming game, the mixed pair struggled to make use of what each saw. A mathematician needs both abilities. Spot the overlooked distinction and express it in a form the other investigator can use. [126] [127]

In one SciPiVot case a 32-fold sweep in spring stiffness moved a simulated support's deflection only from about 212 to 222 millimeters. All six runs hit the travel limit. The agent proposed changing the number of springs. The senior researcher asked how many springs carried the load and whether the limit had masked the response to stiffness. The two plans called for different experiments. Across 248 decisions from 26 projects, PhD students and postdocs new to the projects, with AI available, selected the senior researcher's recorded correction in 73.4 percent of cases. The best tested model with the project record scored 63.5 percent. Both plans were supplied. Agreement with a retrospective judgment tells us who recognized the plan. On a live project somebody has to invent the second plan, run the test and decide what its answer means. [61]

What happens when someone has to choose the next experiment live? Savage and del Rio Chanona had six chemical-engineering postgraduates guide a Bayesian optimizer through a coiled-tube reactor design problem. They could choose starting designs and then pick among proposals over 32 iterations. The human-guided search found better designs early than the standard optimizer in 32 unassisted runs. The lead narrowed as evaluations accumulated. The test used a surrogate fitted to 75 flow simulations. The guided procedure had the early lead. It also started from different designs and offered a choice of proposals. On separate synthetic tests, choosing randomly from such a menu sometimes matched or beat the standard optimizer. Give the same menu and budget to an agent and a generalist on the reactor before attributing its lead to engineering judgment. [113]

Kasparov's Advanced Chess gives the human-computer partnership a name and a challenge. Two amateurs won a 2005 freestyle tournament. They describe selecting promising moves, splitting the lines and cross-checking them with different engines. Their competitors had different equipment and preparation. Give a rival the procedure and the engines. What can the player still see first? [123]

Calling the human contribution "taste" explains too little. Yisong Yue locates mathematical taste in human experience while giving the Caltech-AIM machine a role in proposing conjectures, connections and strategies. Experience matters here if it changes which explanation gets a hearing. A specialist who has watched a fit hide a missing physical term has a reason to inspect the integral even when the residual looks excellent. An agent trained on different failures could acquire the same suspicion. Which partner follows it to a decisive check? [91] [101] Ding counted the attention to successor questions in 58 human papers and public AI accounts of eleven problems. The human texts gave follow-ups more space. A comparison of what authors chose to publish is evidence about exposition. After42 gave agents the opportunity to nominate questions and found that its shortlist reproduced expert consensus about as well as individual experts in one subfield. Neither study followed the competing selections to their mathematical consequences. [134] [143]

A second investigator can be a machine. On one problem in the second First Proof challenge, a critic that had followed the revisions accepted seven drafts of the same argument. A fresh critic rejected all seven. The referees found the submitted proof incomplete. On another problem the critic rejected every draft of a proof the referees accepted up to minor repairs. A critic can save a team from a bad proof. It can also cost the team a good one. [124]

In the same batch the human referees found a gap in a proof about valuated matroids. A proposed formula was supposed to produce a flag of valuated matroids. One citation established a special case. The other did not address the claim. The referees thought the formula likely to work, but the proof still owed its central step. A machine critic had accepted it. A fresh model reading the draft spotted the weak point and called it minor. After seeing the human solution it recognized how much work the proof still needed. The referees earned credit for judging the size of the gap from the submitted argument and its citations alone. [124]

Aletheia exposed a different limit of proof checking. An earlier version tackled 700 entries then listed as open in the Erdős problem database. It returned 212 possible solutions. The reviewing team judged 63 technically correct, 13 responsive to the intended question and four of those new full or partial solutions. Fifty technically correct answers missed their target. A critic could have compared the proposed statement with the original question before anybody spent time proving it. Wang and collaborators found a related failure in a different task. Across seven models, 60.6 percent of rewrites into a more abstract vocabulary widened the asserted domain. A faithful translation carries the restriction. A genuine generalization needs a proof. Who catches the difference before the team proves the wrong theorem? [3] [10]

Once a faithful statement and a candidate are in hand, a proof assistant can expose a particular unsolved obligation and a library can make trying lemmas cheap. Measure that advantage from the point where it actually begins. AlphaProof Nexus proved a real-valued version of a problem on Ben

Green's list. Green clarified that the functions in question were complex-valued. A numerical search supplied a candidate counterexample on the cyclic group of order three. The team formalized the new question and the agent verified the counterexample. A proof checker gave a decisive answer twice. Green knew which question he had meant. Give his clarification to every team. The test for a structural specialist is whether they can spot a new mismatch from evidence the agent also has. [115]

Testing the proposed diagnosis

Keep natural collaboration intact in the primary comparison. In the checkpoint study, elicit the lead's, specialist's and machine adviser's diagnoses independently before revealing anyone else's. Record which failures both machines miss, which a human also misses and which objections turn out to be false. In a randomized analyst study, fifty professionals followed the assigned agent preference on 86 percent of four choices among defensible alternatives. A person who sees the agent's answer first may never put their independent judgment on the record. [74] [124]

Prespecify checkpoint selection and keep unsuccessful interventions. Freeze the evidence, method, agent context and learned state. Each adviser states the predicted failure, test and scope of repair before seeing the reserved outcomes. Conceal authorship and randomize repeated copies of the same lead to a human proposal, a machine proposal or no proposal. Use the same evaluator and remaining compute allowance. Charge for producing the advice. Record whether a correct objection was acted on and whether the repair worked. This distinguishes finding the failure from persuading a partner or implementing the fix. Full-project costs include reaching the checkpoint.

Classify cases before outcomes by the representation change, information in cheap diagnostics and cost of further experiments. The diagnostic hypothesis predicts a larger advantage over the machine adviser where the available signals fit several costly explanations and the agents tend to favor the same wrong one. Randomize access to the same additional diagnostic information across arms with a prespecified information control. Score intermediate predictions independently. A smaller specialist premium after disclosure is consistent with the mechanism. A success ceiling or information that compensates for implementation or orchestration can produce the same pattern. Prespecify checks for those alternatives in both portfolios.

Calibrate autonomous difficulty on separate development projects to assess floor and ceiling effects. Estimate the natural portfolio contrast first. In a separate analysis, match or stratify on novelty, available theory, ambiguity, and the cost and informativeness of feedback. Matching on ambiguity removes differences in its frequency. Unmeasured differences can still matter. After each round freeze the shared archive before releasing fresh cases, and keep their reserved results sealed until assessment is complete. Report the cost and downstream benefit of retrieval separately from the incremental benefit of live specialist help.

Decision rules

Declare δ as the minimum worthwhile increase in accepted discovery probability, distinct from the improvement required to accept a method. For both the human-partner-minus-machine-partner and specialist-minus-generalist contrasts, a lower confidence bound above δ supports a worthwhile gain. An upper bound below δ sets aside a gain of that size for the tested scope. An interval containing δ leaves it unresolved. If the specialist beats the machine partner but the generalist comparison excludes a worthwhile additional gain, report the benefit of assistance and the upper bound on specialization. [92]

To compare fields, let D(b) be the human partner's gain over the machine partner in sparse computation minus the corresponding gain in structural mathematics at budget b. Declare a comparative margin η, zero for a directional claim or positive for a minimum worthwhile difference. A lower confidence bound above η supports that cross-portfolio contrast. An upper bound below η sets aside the declared difference for the tested scope. An interval containing η leaves the ranking unresolved. A claim that specialization itself contributes more in sparse computation also needs the cross-portfolio difference of the specialist-minus-generalist gains, with its own prespecified margin. Significance in one portfolio and an inconclusive result in the other give no such test.

Fix the sample size and stopping rule or use a prespecified sequential procedure with valid coverage. The joint confidence procedure must account for all planned contrasts, project pairing, repeated participants, repeated runs within projects and repeated inspection. Compare against the strongest frozen machine-only procedure before claiming a centaur advantage. Include both solo arms for a claim about beating both individual partners. New agent versions warrant fresh cases under stated application scopes and acceptance rules. Retain the results for the versions that produced them.

Comparing investigators on the mesh and sumset problems

Have outside collaborators supply fresh mesh and sumset problems before either team proposes a solution. Include other authors' ruler families and meshes from another group. Give both teams equal access to whoever needs the result. On each brief run a specialist with one agent and two agents free to take different approaches. At the same error tolerance, did the specialist's design need less storage on unseen meshes? On unseen addition tables, did it prove a tighter bound or cover a larger family at the same cost? Compare partners within each brief before comparing fields. Include generalists and solo specialists, and count failures and false vetoes. [89] [90] [105] [124]

Have analysts and additive combinatorialists each work on problems in both fields, with the order varied and a fresh case each time. Keep cases from an investigator's own research line identifiable in the analysis. Let independent judges assess the results without team names. An advantage that follows people into their home territory points toward accumulated experience. A gain that recurs across specialists on problems with an unstable inverse points toward that operation. Either would tell us more than a comparison between my favorite theorem and somebody else's solver. This experiment has yet to be run.

Hand both teams an unfamiliar discovery. Leave its successor unnamed. Have each propose a question, explain why either answer would matter and commit to what would count as progress. Then exchange anonymized proposals and give independent investigators equal resources to pursue them. Score the question before execution and the result afterward. Keep an arm in which each team pursues its own idea, so that the insight needed to carry it out can count too. A promising false conjecture that uncovers a sharp obstruction can beat a true one that changes nothing. Circling the sum-product conjecture in advance would give away the decision to look there. [132] [133] [134] [135] [137]

Notes and sources

First published 4 November 2026. The 7 November revision added the four-cycle. The 17 November revision added the discovery cases in notes 105-111. The 18 November editions developed the sparse and arithmetic pilots, tested the analogy with Advanced Chess and added machine-partner comparisons. This edition follows the mathematical contributions before, during and after a breakthrough. Notes 130-133 document the rotating-line proof, a guided sequence search and the passage from unit distances to real sum-product. Earlier source numbers remain stable. A subsequent audit read the remaining sources against the question of who selected the next problem. Notes 134-144 add studies of machine-chosen questions, their eventual payoff, and concrete geometric obstructions in numerical, theoretical and algebraic work. Notes 145-151 document the author's own research, the machine rival in sumsets, recognition of multiscale work, and exact checks for the paired remapping and sumset problems. Notes 152-154 check the comparison against work independent of the author's papers. They cover autonomous proof on a named sum-product target, competing physical demands on a remap, and false vetoes by an agent checking fluid simulations. Notes 155-158 document direct autonomous numerical results and the graph and mesh constructions that support the research wager.

What the breakthrough papers establish

[1] Thomas Willwacher. *The Deligne-Drinfeld conjecture after OpenAI*. 2026. arXiv v4, 28 October 2026.
Primary source · Reconstruction and injectivity, Section 3.2 · OpenAI's original Deligne-Drinfeld manuscript
Willwacher attributes the missing dimension bound to OpenAI. He credits a reworking by Leila Schneps and Benjamin Enriquez with inspiring the simplified four-strand representation formula in Lemma 3.1. Sections 1.1-3.3 give the reduction to characteristic two, constrained injectivity and dimension comparison. The essay's weight-three calculation uses those coordinates. Willwacher also records extensive AI use in understanding and simplifying the proof, and distinguishes OpenAI's additional independent injectivity argument from Brown's prior lower bound.

[2] Nikita Markarian. *A remark on the AI proof of the Deligne-Drinfeld conjecture*. 2026. arXiv v1, 1 November 2026.
Primary source
Abstract, introduction and sections 3-5: geometric interpretation through a double covering and filling a puncture, reconstruction on the image, and comparison with the characteristic-two proof. This is subsequent human interpretation of the model-produced architecture.

[3] Tony Feng. *Eigenweights for arithmetic Hirzebruch Proportionality*. 2026. arXiv v2, 2 February 2026.
Primary source · Aletheia research report, v3
Opening attribution, abstract and section 2 assign the mathematical content to Aletheia and paper rewriting to the human author. The accompanying report, sections 3.1-3.2, gives the cross-field and independence-polynomial accounts; its autonomous milestones are not claimed as major mathematical advances. The companion report, §§3.2-3.3, separates Deep Think's inequality, Aletheia's strategy and the independently supplied proofs. The 700-problem December 2025 run reports reviewer classifications of possible solutions, including technical correctness, responsiveness and novelty. The Erdős sweep used an earlier model than the later discovery cases.

[4] OpenAI. *Mathematics research repository and release history*. 2026. Pinned 30 October snapshot; coverage census rechecked 7 November 2026.
Primary source · Release history · Current repository · Pinned catalogue source · Pinned manuscript map · Formalization metadata · Part II's actual prompt in the arithmetic-progression record
README catalogue totals and stated verification limits; history records the 6 October release, 7 October withdrawals and repairs, and later additions. Totals are dated observations, not an independent audit of every result. The coverage table is reproduced from overview.tex and joined to family-level documentation links in CONTENTS.md. It counts 373 families, 721 distinct manuscript links, and 245 families with documentation links. These units do not supply field-conditioned success or proof-validation rates.

[5] OpenAI. *The Blockwise Alperin Weight Conjecture; Donovan's Conjecture over Algebraically Closed Fields*. 2026. Manuscripts dated 23 and 24 September; pinned repository checked 4 November.
Primary source · Alperin manuscript record · Donovan manuscript · Donovan manuscript record
Source-reported modular representation results. The numerical blockwise weight statement is asserted for every finite group and prime; fixed-field Morita finiteness allows bounded defect order, nonabelian defects and characteristic two. No independent whole-proof certification is claimed here; neither family carried a Lean link in the checked catalogue.

[6] OpenAI. *The rational Hodge conjecture for CM abelian varieties: selected formalization scope*. 2026. Pinned scope document checked 4 November 2026.
Primary source
Selected statement: rational Hodge classes on complex CM abelian varieties are algebraic in every dimension and codimension. This is not the full Hodge conjecture for arbitrary varieties. The source excludes later Tate and Hodge-standard consequences from this selected scope. The formalization was not independently replayed in this review.

[7] OpenAI. *Auslander-Reiten and Tachikawa counterexamples: selected formalization scope*. 2026. Pinned scope document checked 4 November 2026.
Primary source
Specified nonprojective modules with the stated Ext-vanishing properties supply the reported counterexamples. The source distinguishes these targets from later endomorphism-algebra and homological consequences. This review inspected the scope rather than independently replaying the formal proofs.

[8] Mohammed Abouzaid, Nikhil Srivastava, Rachel Ward, and Lauren Williams. *First Proof Third Batch*. 2026. 14 October 2026 report.
Primary source · Third batch press release · Second-batch report, June 2026
Protocol and aggregate correctness assessments; Table 2 counts average correctness scores of at least 0.5, and the rubric gives 0.5 for minor, readily corrected errors. Section 5.2, PDF pages 29-30, covers the Darcy inverse problem. Exposition and attribution are assessed separately. June-to-October comparisons use new problem sets and changed testing conditions, including permitted human proof length; they are not a calibrated longitudinal benchmark.

[9] A. Ramesh Chandra, Aaghaz Mahajan, Ainesh Sanyal, et al. *TheoPhysEval: Assessing AI agents at autonomous theoretical physics problem solving*. 2026. arXiv v2, 12 November 2026.
Primary source · October evidence review, v1 · Review of apparent model errors
Sections I, III and IV.1 describe problem provenance and automated final-answer evaluation. The authors report that answers were unpublished at evaluation, but some contributors used AI and data leakage cannot be ruled out. The isolated track removes web search and supplied background papers while retaining computational tools. The 50 problems are not presented as breakthrough open problems. The full benchmark ran once per agent per track. Version 2 preserves the two counts. Section IV.2 describes experts correcting reference solutions after reviewing apparent model errors.

[10] Yanli Wang, Suijin Wang, Xiaopeng Yuan, and Haohan Wang. *Objects Without Morphisms: What LLMs for Mathematics Do Not Represent*. 2026. arXiv v1, 2 October 2026.
Primary source
Sections 3-4 study mathematical rewriting across framings. The 60.6 percent figure is the held-out pooled abstracting condition, not a failure rate for research proofs. Seven model snapshots are tested; neither the October internal OpenAI proof system nor GPT-6 is among them. The seven-model comparison is cross-sectional. Mean item-level correlation 0.503 describes shared susceptibility, not a temporal plateau. The 89.2 percent promotion result belongs to a single-model instruction experiment, distinct from the pooled 60.6 percent scope-widening result.

[11] Sanjana Das and Jacob Fox. *A proof of the Murty-Simon conjecture on diameter-2-critical graphs*. 2026. arXiv v1, 28 October 2026.
Primary source · Full text
Theorem 1.1 and AI statement: model contribution includes the edge bound and even-order equality; authors simplify and extend equality to odd order.

[12] Sang-il Oum. *A proof of the cycle double cover conjecture by OpenAI: An exposition*. 2026. arXiv v3, 1 August 2026.
Primary source
Exposition with modifications for accessibility; the original proof precedes the October release. The source attributes contributions and revisions separately.

[13] Ansh Nagda, Prabhakar Raghavan, and Abhradeep Thakurta. *Reinforced Generation of Combinatorial Structures: Hardness of Approximation*. 2026. arXiv v7, 9 March 2026; first submitted 2025.
Primary source
AlphaEvolve constructions and reduction gadgets, including metric-TSP hardness; human-designed objectives and reconstruction of mathematical guarantees remain part of the reported workflow.

[14] Zhiqiang Xu. *A 3-regular counterexample to the Bilu-Linial signing conjecture*. 2026. arXiv v4, 29 September 2026.
Primary source
The finite connected simple cubic base graph is not Ramanujan. The restricted question for Ramanujan base graphs remains distinct. AI assistance is stated by the author, who takes responsibility for checking.

[15] Fangfang Lin and Hong Zhou. *Improved Bounds for the Bilu-Linial Conjecture via Spectral Recovery from Mixed Determinantal Polynomials*. 2026. arXiv v2, 14 October 2026.
Primary source
Revised version acknowledges the counterexample; the bound is existential and the remaining algorithmic question concerns polynomial-time construction satisfying the mixed-root condition. AI statement gives the specific model contributions.

[16] OpenAI. *Local smoothing for the Euclidean wave equation: release and formalization scope*. 2026. 26 October release; scope checked 4 November.
Primary source · Pinned formalization scope · Pinned manuscript record
The source asserts a half-wave local-smoothing bound for n >= 2, with the specified positive Sobolev loss, and lists related formalization targets. It labels the manuscripts raw research data. This concerns theoretical PDE, not the construction or testing of a numerical PDE solver.

[17] Alex Townsend and Anne Greenbaum. *The Neurosurgery Resident Who Proved Crouzeix's Conjecture*. 2026. SIAM News, 2 November 2026.
Primary source · Jin manuscript · Independent proof by Lorist and Schwenninger · Jin manuscript, pinned v4 · Jin's proof-search prompt
Attributed report that Townsend, Greenbaum and Michel Crouzeix checked Shanmu Jin's manuscript and believe it correct; account of the approximately sixteen-hour model run. The cited preprint and independent proof provide mathematical source material, not a claim of journal peer review.

[18] Josh Alman and Virginia Vassilevska Williams. *Truly Subquadratic 3SUM and Truly Subcubic APSP via Triangles in Sparse Lopsided Graphs*. 2026. arXiv v1, 5 October 2026.
Primary source
Abstract and Theorem 22 specify the integer-input bounds and APSP assumptions. Acknowledgments and methodology distinguish the model's discovery from human simplification and extension. These exponents alone do not establish practical speedups. Corollary 26 and the deterministic bound in Theorem 22 give the APSP exponent 2.99942 under the stated hypotheses.

[99] OpenAI. *A Cyclic Polytabloid Proof of Saxl's Conjecture*. 2026. 24 September manuscript; family 205 scope record.
Primary source · Selected formalization scope
Sections 3-6 track an actual cyclic generator through full induced quotients and a band-pairing argument. Proposition 5.3 and Example 5.4 change the dual test basis while keeping the generator fixed. The formalization documentation reports the positivity theorems but excludes the stronger prescribed-tensor orbit-span assertion. The final proof does not establish a chronological model-failure-and-repair episode. No independent whole-proof verification was performed.

[100] OpenAI. *An Upper Bound of 9/4 for the Matrix Multiplication Exponent*. 2026. 2 October manuscript; family 107 scope record.
Primary source · Theorem and finite-size limitation · Selected formalization scope
Family 107. The stated exponent bound concerns arithmetic operations over the complex numbers. The introduction does not specify a competitive finite matrix size, and the scope record distinguishes arithmetic complexity from bit complexity and practical crossover. The claim is reported with these limitations.

[132] OpenAI. *Planar Point Sets with Many Unit Distances*. 2026. May 2026 release and companion remarks by Noga Alon, Thomas F. Bloom, W. T. Gowers, Daniel Litt, Will Sawin, Arul Shankar, Jacob Tsimerman, Victor Wang, and Melanie Matchett Wood.
Primary source · Remarks on the Disproof of the Unit Distance Conjecture
The companion authors attribute the initial construction to an internal OpenAI model and give a verified, simplified and generalized proof. Their opening footnote distinguishes the initial proof from the subsequent exposition. Growing number fields enable a planar construction exceeding the conjectured unit-distance bound.

[133] Thomas F. Bloom, Will Sawin, Carl Schildkraut, and Dmitrii Zhelezov. *The sum-product conjecture is false for real numbers*. 2026. arXiv:2605.28781v1, Theorem 1.2, §2 and AI contribution statement.
Primary source · Authors' statement on the role of AI
The authors expressly attribute their decision to revisit high-degree number fields to OpenAI's unit-distance result. They credit GPT-5.5 Pro with early discussion and the elementary Lemma 3.4, and assign the main ideas of their construction to themselves. The counterexample is over the reals. This historical transfer is not a comparative trial of autonomous follow-up research.

[155] Johannes J. Brust and Tamara G. Kolda. *Fast and Accurate CP-HIFI Tensor Decompositions Exploiting Kronecker Structure*. 2026. Supplement, §§SM1-SM5, especially pp. 1-2.
Primary source · Main article
Five responses shared the sparse matrix-vector strategy. The authors report an additional two- to fivefold gain, identify a 2014 antecedent and compare different preconditioners. They posed the iterative-solver brief. This is a numerical linear algebra problem.

Who chooses the next problem?

[134] Fan Nie, Zihao Wang, and the After42 team. *After42: Can AI Propose Interesting Research Questions?*. 2026. After42 research report, §§2-3, 5-6.
Primary source
The project supplies broad goals and literature. In §3.2 a cycle calculation changes the target to paired-tree distance uncertainty. At least two experts endorsed 45 of 135 final questions. The shortlist comparison covered one subfield. Expert interest and an eventual mathematical result are separate outcomes.

[135] Hita Kambhamettu, Daniel Krashen, and Lyle Ungar. *Guiding AI Agents to Generate Conjectures by Encoding How Theorems Build on One Another*. 2026. arXiv v1, 14 November 2026, §§4-5 and 7.
Primary source
Two agents receive a seed paper and propose successor conjectures. Section 5.1 traces Wythoff Nim. A blinded expert who authored twelve of the thirty seed papers preferred the graph-guided proposals eight times. The uncertainty interval is wide. The paper reports no proofs of the generated conjectures.

[136] Chenglei Si, Diyi Yang, and Tatsunori Hashimoto. *Can LLMs Generate Novel Research Ideas? A Large-Scale Human Study with 100+ NLP Researchers*. 2024. Research report, §5 and Appendix I.
Primary source
Human-written, machine-selected and human-reranked machine proposals were blind-reviewed within seven supplied NLP topics. The two machine-origin shortlists overlap on 17 of 49 selections. Both score higher than human-written proposals on perceived novelty. Different selections alone cannot establish a superior selector.

[137] Chenglei Si, Tatsunori Hashimoto, and Diyi Yang. *The Ideation-Execution Gap: Execution Outcomes of LLM-Generated versus Human Research Ideas*. 2025. arXiv v1, §§2, 4-5 and Tables 3-6.
Primary source
Forty-three researchers carried out assigned ideas. The original machine used Claude 3.5 Sonnet. The post-execution analysis treating reviews as independent favored human-origin ideas on several measures. With projects as units, final differences were inconclusive. Drops from proposal to execution were significantly larger for machine ideas on all four shared measures. Human executors in both arms and supplied topics limit the inference.

[138] Yair Caro, Randy Davila, Michael A. Henning, and Ryan Pepper. *Conjectures of TxGraffiti: Independence, domination, and matchings*. 2022. Australasian Journal of Combinatorics 84, 258-274.
Primary source
The program was given a group of graph invariants. Conjecture 2.1 is the program's cubic-graph inequality. The authors prove the regular-degree generalization and characterize equality in Theorem 2.3. The example documents which investigators chose which next question.

[139] Zeyu Zheng, Shengtong Zhang, Jeremy Avigad, Prasad Tetali, and Sean Welleck. *The Problem Is the Problem: Towards Scalable Mathematical Discovery*. 2026. arXiv v1, §§3-5 and Appendix C.10.
Primary source
The system recovered previously posed open problems from combinatorics papers. Section 5.2 connects a classical finite-geometry construction to a line-union conjecture. Section 5.1 gives a correct but previously known result recommended as new. Selection from the literature, new mathematical application and creation of a question are distinct.

[143] Yang Ding. *AI and Human Approaches to Mathematical Problem Solving*. 2026. arXiv v1, 15 September 2026, corpus construction and Appendix B.
Primary source
The textual study contrasts public AI accounts with 58 human papers across 11 problems and 31 within-problem comparisons. Human texts devote more attention to successor questions. Uneven representation of problems and differences in genre and purpose limit what prose can establish about actual research decisions.

[148] Timothy Gowers and Isaac Rajagopal. *A recent experience with ChatGPT 5.5 Pro*. 2026. Firsthand report, 8 May 2026, Rajagopal's account and Appendices 1-2.
Primary source
Gowers describes interest from both Rajagopal and the model in polynomial diameter before he commissioned the attempt. Rajagopal identifies the target replacement and derives the bound for fixed order and sufficiently large set size. The model supplied dissociated generators and their multiples.

[152] Yichen Huang. *Autonomous disproofs of the sum-product conjecture over ☐ with GPT-5.5 Pro*. 2026. Preprint, §§2-4, problem statement, Table 1 and proof analysis.
Primary source
The April 23 snapshot predates the public disproof. Huang supplied the target and its truth. Seven of eight independent trials produced proofs the author judged correct with tools and search disabled. Some use regions of algebraic integers without requiring units. The comparison concerns proof construction after the target was named.

What a representation preserves and makes possible

[19] Daniel A. Spielman and Nikhil Srivastava. *Graph Sparsification by Effective Resistances*. 2009. arXiv v4, 18 November 2009.
Primary source
Quadratic-form preservation and sampling by the edge-weight/effective-resistance product. The unit-star elimination calculation in this essay is an elementary exact example.

[20] Jinchao Xu and Ludmil Zikatanov. *Algebraic Multigrid Methods*. 2016. arXiv v2, 10 November 2016.
Primary source
In particular section 5, on optimal coarse spaces and the lower spectrum associated with the smoother. Algebraic smoothness is operator-relative.

[21] E. Anderson et al. *LAPACK Users' Guide: Further Details - Error Bounds for Linear Equation Solving*. 1999. Online third-edition guide.
Primary source
Distinguishes normwise and componentwise backward error, conditioning and forward error; gives the componentwise residual formula used here. The diagonal two-variable example and shifted-Laplacian residual bound in this essay are elementary illustrations, not reported LAPACK experiments.

[52] Erik G. Boman and Bruce Hendrickson. *Support Theory for Preconditioning*. 2003. SIAM Journal on Matrix Analysis and Applications 25(3), 694-717.
Primary source · Publisher-deposited abstract and bibliographic record · Support numbers in Boman, Hendrickson and Vavasis, §2 · Authors' manuscript, support inequalities in §2
Section 2, Theorem 2.3 and Proposition 2.4, relates the support numbers to extremal generalized eigenvalues and their product to the relative condition number for positive-definite matrices. The essay derives its energy-error identity and inverse-order inequality in exact arithmetic. Implementing the certificate requires further control of rounding.

[56] OpenAI. *An explicit power saving for the exact discrete Fourier transform*. 2026. 25 September 2026 manuscript; repository snapshot 4eaee87548250b2682affda7d05d947c05554bae.
Primary source · Introduction and model assumptions · Formalization scope
Introduction and construction: exact complex arithmetic, prescribed roots, unrestricted coefficients; preparation and indexing are charged. No floating-point stability or useful crossover claim. The formalization excludes the synthesis appendix; this review did not rerun Lean.

[57] OpenAI. *Integer multiplication below n log n*. 2026. 23 September 2026 manuscript.
Primary source · Precision and recovery source
Section 8: fixed-point error propagation and exact recovery after a strict error bound below one half. Claimed finite-tape bit complexity with enormous constants; not independently certified here.

[58] Rasmus Kyng, Yin Tat Lee, Richard Peng, Sushant Sachdeva, and Daniel A. Spielman. *Sparsified Cholesky and Multigrid Solvers for Connection Laplacians*. 2016. STOC 2016, pp. 842-850; author preprint arXiv:1512.01892v1.
Primary source · Published article
Definition 3.2 and Lemma 3.3: exponential spectral distortion at most twice the sum of local tolerances, controlling both Schur approximations and eliminated-block inverses. Application work sums over levels. Not an AC(2) theorem.

[60] James Demmel, Ioana Dumitriu, and Olga Holtz. *Fast Linear Algebra Is Stable*. 2007. Numerische Mathematik 108, 59-91.
Primary source
Normwise stability with essentially the matrix-multiplication exponent under stated algorithmic and precision assumptions. No claim of sparse-structure preservation or practical speed.

[80] Henry Shin. *A global girth obstruction for Garg-Mineyev taiko product structures*. 2026. arXiv:2607.01716v1, 2 July 2026.
Primary source · Worked explanation at Refactor
An exact incidence-geometric example of strengthening an insufficient description. The principal structural theorem closes the specified triple-girth assembly route; the sharper frontier uses a separate finite certificate. This illustrates the mathematical practice, not an empirical human advantage over AI.

[81] Michael Atiyah. *Mathematics and the computer revolution*. 1984 / 2016. 1984 essay reprinted with a new introduction, Izvestiya: Mathematics 80(4), 637-646 (2016).
Primary source · Original 1984 text, archived scan · Publisher's article record and reprint
In §2 (p. 640) Atiyah acknowledges the stimulus that expensive computer time gives to faster algorithms. In §5 (p. 644) he predicts that machines good at long calculations will not look for alternatives, and urges humans to retain the senior role. His concluding framing is a deliberate warning (p. 646). The original essay dates to 1984. The 2016 publication includes a new introduction.

[82] Herbert S. Wilf and Doron Zeilberger. *Rational functions certify combinatorial identities*. 1990. Journal of the American Mathematical Society 3(1), 147-158.
Primary source
Certificate identities and verification for the stated hypergeometric setting. Used as an example of mathematical structure enabling systematic computer proof. Added in the 7 November review.

[83] Bernardino Romera-Paredes et al. *Mathematical discoveries from program search with large language models*. 2024. Nature 625, 468-475; published online 14 December 2023.
Primary source
FunSearch searches for constructor programs within a supplied evaluator and program structure. The authors discuss interpretable, parsimonious, symmetric constructions and their role in subsequent investigation. Added in the 7 November review. Figure 5 and the accompanying admissible-set discussion credit the investigators with using cyclic symmetries found in a constructor to direct the next search.

[84] Joshua Batson, Daniel A. Spielman, and Nikhil Srivastava. *Twice-Ramanujan Sparsifiers*. 2014. SIAM Review 56(2), 315-334.
Primary source · Authors' arXiv manuscript
The stated reweighted-edge bound controls every quadratic form with factors $(1 \pm \text{epsilon})^2$. The deterministic construction is polynomial time, without a claim here that its original implementation gives a practical fast solver. Added in the 7 November review.

[93] Bjarke Hammersholt Roune and Michael Stillman. *Practical Gröbner Basis Computation*. 2012. ISSAC 2012; author manuscript arXiv:1206.6940v1.
Primary source
Sections 3-5 and Appendix D combine algebraic elimination criteria, compressed data structures, benchmark-dependent bottlenecks, and implementation costs. Section 4.3 reduces stored monomials from potentially quadratic to linear while retaining an integer per queued S-pair. This is a counterexample to treating compression and computational economy as exclusive to numerical analysis.

[94] Kevin Buzzard. *Beyond the Liquid Tensor Experiment*. 2022. Participant account, 12 September 2022.
Primary source
The sections on the resolution and lessons describe formalization costs and the replacement of stable-homotopy machinery by a weaker sufficient result after clarifying the required statements. This is historical evidence about a structural research mechanism, rather than a measurement of contemporary AI assistance.

[96] Grey Ballard, James Demmel, Olga Holtz, and Oded Schwartz. *Minimizing Communication in Numerical Linear Algebra*. 2011. SIAM Journal on Matrix Analysis and Applications 32(3), 866-901; earlier author manuscript arXiv:0905.2485v1.
Primary source · Published article
Communication lower bounds apply to the stated algorithm and memory-access models, including specified dense and sparse factorizations and compositions. They identify execution costs beyond arithmetic counts. They are neither unrestricted lower bounds on all algorithms nor bounds on AI reasoning.

[97] OpenAI. *A polynomial-time construction of strong thin trees*. 2026. 23 September manuscript; pinned repository snapshot.
Primary source · Construction and precision discussion · Rational covariance completion and signing · Packing bound and accumulated losses
Family 174. The claimed polynomial-bit construction combines shortcut forms, rational whitening, exact covariance completion, signing, and accumulated-loss control. The r > 2^168 observation follows algebraically from its displayed choice C_s = 2^20 and the coefficient 1/2 − C_s r^(−1/8). The proof uses an absolute cutoff, and this sufficient-bound calculation is not a necessary input-size condition or a performance measurement. Cut-thinness differs from two-sided spectral approximation. No independent whole-proof audit or fresh Lean build was performed.

[98] OpenAI. *Almost-Linear-Time Maximum-Cardinality Matching in General Graphs*. 2026. 24 September manuscript; pinned repository snapshot.
Primary source · Theorem and computational model · Engine and precision accounting · Debt-funded local projection · Uniformization and charged scheduling
Family 120. Theorem 1.1 claims an (n+m)^(1+o(1)) word-RAM bound on every path and success probability at least 2/3. Sections 3-5 develop the ledgers, local correction, exact state transfer, and uncancelled-motion accounting. Section 9 charges instruction-level dovetailing, random bits, allocations, multiword arithmetic, and output. These are reported guarantees within the stated model, with no competitive finite-size crossover or independent whole-proof validation established here.

[112] John Napier. *Mirifici logarithmorum canonis descriptio*. 1614. Edinburgh, 1614. Preface, p. 1 in Ian Bruce's translation..
Primary source · Napier's preface in English, p. 1
Napier explains the burden and errors of long multiplications, divisions and root extractions, and how replacing the original numbers permits simpler operations. This is his account of a particular invention, not a proposed origin story for abstraction in general.

[117] Doron Zeilberger. *Enumeration Schemes and, More Importantly, Their Automatic Generation*. 1998. Annals of Combinatorics 2, 185-195.
Primary source · WILF, HERB and sample runs
Sections 10-12 specify the depth-limited search, its failure output, the worked example with forbidden patterns 123 and 132, and the empirical HERB predecessor. The elementary count in the text follows by choosing one of the two largest unused entries at each step.

[121] Doron Zeilberger. *Theorems for a price: tomorrow's semi-rigorous mathematical culture*. 1993. Notices of the American Mathematical Society 40, 978-981.
Primary source · Author's full essay
Zeilberger argues that the price of completing a proof can matter once experimental evidence is strong. His imagined prices illustrate a methodological position. They are not measured research costs or estimates for the two pilots here.

[142] Larry Guth, Hong Wang, and Ruixiang Zhang. *A sharp square function estimate for the cone in R^3*. 2020. Annals of Mathematics 192, 551-581, Theorems 1.1-1.3, Lemma 1.4.
Primary source
A multiscale geometric overlap estimate helps establish the cone square-function bound and sharp wave local smoothing in 2+1 dimensions. This is a direct role for quantitative incidence in theoretical PDE. The estimates do not rank research teams.

[144] Orit E. Raz, Micha Sharir, and Frank de Zeeuw. *Polynomials Vanishing on Cartesian Products: The Elekes-Szabó Theorem Revisited*. 2015. SoCG 2015, Theorem 1.1 and §2.
Primary source
For an irreducible trivariate polynomial with all partial derivatives non-identically zero, the argument turns Cartesian-product zeros into point-curve incidences. Large families can force special local additive structure. The theorem illustrates an algebraic incidence research route and contains no comparison of investigators.

[145] Henry Shin. *Iterated-sumset spectra: The complete exponent law and its rank geometry*. 2026. Preprint, Proposition 6.1 and §7.
Primary source
A face records the complete weak order on labelled sums, with ties, through the chosen order. Section 7 constructs compressed integer realizations at that horizon. The seven-ruler enumeration and the observation problem one order farther are worked examples for this essay.

[146] Henry Shin. *Logarithmic source curves in polynomial fiber products*. 2026. Preprint, §7, Lemma 7.1 and Proposition 7.2.
Primary source
Over the rationals, under the stated effective plane-curve operations, the paper determines by finite tests whether ordinary-integer new lifts yield a power term, and its exponent when present. Finite exceptional sets enter the height count through a bounded contribution.

[149] Society for Industrial and Applied Mathematics. *August Prize Spotlight*. 2023. Peter Henrici Prize, citation and interview with Douglas N. Arnold.
Primary source
SIAM credits Arnold's work in numerical analysis and finite elements. Arnold's own account identifies finite element exterior calculus among his contributions and explains how theory detects deficiencies and guides design.

[150] American Mathematical Society. *2020 Bôcher Memorial Prize, citation for Lawrence Guth*. 2020. Joint Mathematics Meetings prize booklet, printed p. 5.
Primary source · MIT account of the citation
The prize citation explicitly credits Guth's use of multiscale organization and partitioning across harmonic analysis, incidence geometry, analytic number theory and PDE. It is direct evidence of mathematical esteem across fields.

What changes across sparse and arithmetic problems

[23] Taoli Shen, James Brannick, Robert Falgout, Karsten Kahl, and Jacob Schroder. *Nodal Coarsening and Approximate Ideal Interpolation for H(curl) in Algebraic Multigrid*. 2026. arXiv v2, 23 October 2026.
Primary source · Versioned PDF
Sections 2.1, 3.2 and 3.3 specify model, two-grid assumptions and unshifted commuting relation. Section 5.5, PDF Table 5.7 on printed page 24, gives the 7,690-unknown checkerboard smoothing diagnostic: Ref 3.311 to 4105; Alg 3.189 to 4002 as the minimum coefficient falls to 10^-6. The larger case is not that diagnostic. Dense 3D interpolation is a limitation of the particular construction, not a consequence proved from CG=0. Smoothing estimates and observed iteration counts are different quantities.

[24] Ralf Hiptmair and Jinchao Xu. *Nodal Auxiliary Space Preconditioning in H(curl) and H(div) Spaces*. 2007. SIAM Journal on Numerical Analysis 45(6), 2483-2509.
Primary source · Author technical report (2006)
Auxiliary-space construction using discrete decompositions and Poisson-type solvers. No stronger blanket coefficient/topology guarantee is claimed here.

[25] Niall Bootland, Victorita Dolean, Frédéric Nataf, and Pierre-Henri Tournier. *A robust and adaptive GenEO-type domain decomposition preconditioner for H(curl) problems in three-dimensional general topologies*. 2025. arXiv v3, 16 May 2025.
Primary source
Adaptive two-level additive-Schwarz/GenEO work for positive Maxwell-type systems, including nontrivial topologies. It is not cited as a result for indefinite time-harmonic Maxwell.

[26] Yuan Gao, Rasmus Kyng, and Daniel A. Spielman. *AC(k): Robust Solution of Laplacian Equations by Randomized Approximate Cholesky Factorization*. 2026. Author manuscript; published SISC 48(3), A1643-A1684.
Primary source · Published article
Author manuscript section 6, printed pages 25-26, states the AC(2) condition-number conjecture. The author manuscript and journal version have different pagination. Experimental robustness is not a proof of the conjecture.

[37] Yali Fink, Ido Ben-Yair, Lars Ruthotto, and Eran Treister. *RAPNet: Accelerating Algebraic Multigrid with Learned Sparse Corrections*. 2026. ICML 2026, PMLR 306, 31125-31145.
Primary source · Published full text
Corrections concern coarse and transfer operators. Published limitations state the absence of a general convergence guarantee. Table 2 compares CPU classical setup with GPU RAPNet, counting aggregation, transfer, inference and augmentation; solve-iteration timings are not reported there. The model carries latent state across successive level pairs, and experiments include nonsymmetric three-dimensional advection-diffusion. No supplied general convergence theorem does not imply an impossibility theorem for local learning.

[38] Zihang Liu, Oleg Balabanov, Yaoqing Yang, and Michael Mahoney. *Learning to Discover Iterative Spectral Algorithms*. 2026. arXiv v2, 6 July 2026.
Primary source · Preconditioner comparison in the revised paper · Original version used in the first review
The test operators are positive definite after the stated transformations. Chebyshev uses the same spectral estimates, with wins and losses in the revised comparison. Table 3 and Appendix E distinguish first-level preconditioning choices and iteration counts from end-to-end cost.

[39] Xudong He, Ruiyu Wang, Yilang Ding, Rui Peng Li, and Yuanzhe Xi. *Online Autotuning of Algebraic Multigrid via Contextual Bandits and Reinforcement Learning*. 2026. arXiv v1, 31 October 2026.
Primary source
Sections 5.1-5.2 and Table 4: 5,000-system online totals, including learner costs and retries. Separate frozen tests must not be merged with these totals. Failure comparisons use different retry allowances. Section 5.6 includes an informed periodic policy slightly faster on SPD diffusion. The separate periodic-policy comparison in Table 7 reports a mean difference of 2.8 percent with a spread of 3.8 percent.

[42] Zechen Zhang, Rui Peng Li, and Yousef Saad. *Graph Neural Multilevel Preconditioners for Iterative Solvers*. 2026. arXiv v1, 30 July 2026.
Primary source
Sections 4.1-4.3 report 867 real non-SPD matrices, per-matrix training, construction and solution failures, and setup/solve timings. Table 5 reports three-way win shares for convergence efficiency among matrices where all compared methods produce valid runs; these are not overall solve-success fractions. Main large-set results are principally two-level; deeper hierarchy tests use 100 matrices. Cross-matrix transfer is examined separately. The depth-selection experiment reports a 4-7 percentage-point win-rate increase against GNP on a 100-matrix subset; this is not a 4 percent runtime or convergence-rate improvement. Table 4 gives mean solve times of 9.17 seconds for GMP and 50.87 for AMG. The order reverses from the reported medians.

[43] Raymond S. Tuminaro and Christian Glusa. *A structure preserving H-curl algebraic multigrid method for the eddy current equations*. 2025. arXiv v1, 9 June 2025.
Primary source
Sections 3-4 construct edge interpolation using energy minimization with sparsity and commuting constraints. Feasibility matters; arbitrary sparsity patterns are not guaranteed. The model varies conductivity in the mass term, not the curl-curl coefficient targeted by Shen et al. This provides an important baseline, not a resolution of that distinct problem. Section 4.2, Figure 3 and Algorithm 2 identify a disconnected coarse-edge stencil and repair its support.

[47] Yves Baumann, Rasmus Kyng, and Gernot Zöcklein. *VAC: A Volume-sampling-based Elimination Rule for Approximate Cholesky Factorization*. 2026. arXiv v2, 24 September 2026.
Primary source
Spanning-tree sampling on the product clique arising in elimination; preserves the stated marginals and connectivity, with linear sequential work and logarithmic-depth sampling. These are local construction guarantees, not a global AC(2) condition-number theorem.

[48] Meher Chaitanya, Cameron Musco, and Aristides Gionis. *CAST: Canonical Approximate Schur Tree for Approximate Cholesky on Graphs*. 2026. arXiv v1, 8 September 2026.
Primary source · Chimera workload reversal and four other graphs, Appendix C.2 · Spielman graph comparisons, Appendix C.3 · Five graph means and system counts, Table 5
Sections 3-4 prove local guarantees while leaving global factorization control open. Section 5 and Appendix C compare 250-right-hand-side workloads. Appendices C.2-C.4 distinguish fill propagation on Chimera systems, low-degree Spielman elimination, and high-degree pivots inside dense Sachdeva blocks. The explanatory configuration heuristic is not tested as an adaptive selection rule. For Chimera instance i3, changing the reuse horizon reverses the comparison between CAST-1 and CAST-2. The Chimera comparisons aggregate Newton-step systems by underlying graph at a verified relative residual of 10□8. At 250 right-hand sides CAST-2 still lost to CAST-1 on the other four graphs. Table 5 gives rounded graph means of each system's five-seed median. Projected Gaussian right-hand sides were shared across methods. Weighting the rows by system counts gives 2,313.4 and 2,251.3 summary seconds in the two graph-level comparisons.

[49] Ahsan Ali, James Brannick, Karsten Kahl, Oliver Krzysik, Jacob Schroder, and Ben Southworth. *Constrained Local Approximate Ideal Restriction for Advection-Diffusion Problems*. 2024. SIAM Journal on Scientific Computing 46, S96-S122; published online 2 May 2024.
Primary source · Author preprint
Local approximate ideal restriction combined with mode constraints; aggressive coarsening controls hierarchy complexity. The advection-diffusion and nonsymmetric setting is distinct from the positive-definite class in the sparse pilot.

[50] Oliver A. Krzysik, Ben S. Southworth, Golo A. Wimmer, Ahsan Ali, James Brannick, and Karsten Kahl. *Optimal Transfer Operators in Algebraic Two-Level Methods for Nonsymmetric and Indefinite Problems*. 2026. SIAM Journal on Matrix Analysis and Applications 47(2), 824-851; author preprint v2, 11 September 2025.
Primary source · Author preprint
Finite-dimensional generalized eigenvector characterization and sharp two-level convergence results, including real-valued transfers. Optimality of the specified operators does not establish a sparse, cheap, recursively implemented hierarchy.

[51] Hussam Al Daas, Pierre Jolivet, Frédéric Nataf, and Pierre-Henri Tournier. *A Robust Algebraic Two-Level Schwarz Preconditioner for Sparse Matrices*. 2025. SIAM Journal on Scientific Computing 47(4), A2378-A2402; published online 18 August 2025.
Primary source · Earlier author preprint
Matrix-based local spectral coarse spaces. The proved scope is SPD; the published condition-number bound is not fully independent of the number of subdomains. Nonsymmetric and saddle-point cases are also tested numerically, which does not extend the SPD theorem.

[54] PyAMG developers. *PyAMG: Algebraic Multigrid Solvers in Python*. 2026. Documentation inspected 4 November 2026.
Primary source · Adaptive aggregation documentation
Matrix-based hierarchy, graph, and diagnostic components. adaptive_sa_solver generates candidates when none are known or matrix changes invalidate them. Available algorithms are not a uniform robustness guarantee for every sparse input.

[62] Mikhail Khodak, Min Ki Jung, Brian Wynne, Edmond Chow, and Egemen Kolemen. *One-shot acceleration of transient PDE solvers via online-learned preconditioners*. 2026. Computer Physics Communications 327, 110304; manuscript v4.
Primary source · Published article
Six fluid/MHD cases, ten trials each; 33 configurations. Learning and backstop costs included. Section 4.2 identifies costly early exploration. The motivating regret comparator is the best fixed configuration, not a newly invented representation.

[75] Noah Amsel, Yves Baumann, Paul Beckman, et al. *Linear Systems and Eigenvalue Problems: Open Questions from a Simons Workshop*. 2026. arXiv v3, 21 August 2026.
Primary source
Sections 2.1, 2.2 and 2.5. Problems 2.1-2.2 concern parameterized families and fast algorithms. Problem 2.3 specifies polynomially bounded SWCDDM matrices, constant inverse spectral comparison, near-linear construction and linear application. The preconditioning discussion distinguishes worst-case bounds from observed convergence.

[76] Majid Rasouli, Vidhi Zala, Robert M. Kirby, and Hari Sundar. *Scalable Lazy-update Multigrid Preconditioners*. 2019. IEEE HPEC.
Primary source · Published article
Existing hierarchy reuse, coarse-operator updates, and local update strategies for changing systems. Shows why reuse-versus-rebuild is an established algorithmic problem; no universally optimal policy is claimed. Sections III-IV and Figures 5-7 compare total update-and-solve time for specified deformed meshes and a common random right-hand side. The fine-only update differs from retaining the entire unchanged preconditioner.

[77] Zherui Yang, Zhehao Li, Kangbo Lyu, Yixuan Li, Tao Du, and Ligang Liu. *Learning Sparse Approximate Inverse Preconditioners for Conjugate Gradient Solvers on GPUs*. 2025. NeurIPS 2025.
Primary source
Replaces triangular applications with sparse matrix-vector products. Appendix D includes CPU/GPU and default AMG comparisons; the winner depends on hardware and matrix family. Fixed support, SPD systems, single-GPU scope.

[78] Marian Brezina, Robert Falgout, Scott MacLachlan, Thomas Manteuffel, Steve McCormick, and John Ruge. *Adaptive Smoothed Aggregation (alphaSA) Multigrid*. 2005. SIAM Review 47(2), 317-346.
Primary source · Author manuscript
Adaptive generation of near-kernel candidates and adjustment of coarse construction. Missing-mode discovery is an established automatic method, not a new human-exclusive task.

[79] James Brannick, Fei Cao, Karsten Kahl, Robert Falgout, and Xiaozhe Hu. *Optimal Interpolation and Compatible Relaxation in Classical Algebraic Multigrid*. 2018. SIAM Journal on Scientific Computing 40(3), A1473-A1493.
Primary source · Preprint
Coarse-space quality, compatible relaxation, and sparse bootstrap interpolation. Optimal-characterization results and practical estimates must be distinguished from cheap certificates for arbitrary recursive hierarchies.

[85] Hugo Melchers, Victorita Dolean, and Michael Abdelmalik. *When can a neural operator replace a coarse solve? Architectural principles for two-level preconditioning*. 2026. arXiv v4, 11 November 2026.
Primary source · Architectural analysis · Three-dimensional diffusion, Table 9
Section 3.2 relates input projection, right-hand-side linearity and symmetry to the Galerkin correction. Sections 4.4-4.5 report two- and three-dimensional tests. In Table 9, mean F-GMRES iterations at 80 points per side are 37.3 for the learned inverse, 23.0 for the exact coarse solve and 124.1 for the smoother alone. These are iteration counts.

[86] Chenglong Bao, Chen Cui, and Ruohua Li. *Learning Coarse Spaces for Two-Grid Methods from Smoothed Errors with Applications to High-Contrast Porous Media Flow*. 2026. arXiv v1, 4 November 2026.
Primary source · Operator perturbation analysis
Basis-invariant learning and conditional discrete perturbation analysis. Forming the sharper Gram-matrix diagnostic requires solves with the original operator and is performed offline. Tests on specified parameter families do not establish a uniform guarantee for trained outputs. Added in the 7 November review.

[87] Jung Eun Huh, Coralia Cartis, and Yuji Nakatsukasa. *Adaptive LSQR Preconditioning from One Small Sketch*. 2026. arXiv v1, 6 April 2026.
Primary source · Printed theorem and proof · Version history
Equation (3.2), Theorem 4.1 and Appendix A.1. Write t for the target. For a proper retained subspace, the ideal augmented Gram matrix has eigenvalues t^2 on that subspace and μ^2 on its complement. The printed proof instead inserts $t^2 + \mu^2$ in the retained block. For A = A_hat = diag(1,0), an exact rank-one CUR approximation gives E = 0. With $\mu = 1$ and $t = 1/2$, all printed assumptions hold, but P = diag(2 sqrt(2),1) gives preconditioned singular values 1/2 and 1. The condition number is 2, exceeding the claimed bound sqrt(5)/2, and the claimed lower bound 1 fails. Write sigma_hat_l for the smallest retained singular value of the CUR approximation. Under $\mu <= t <= sqrt(sigma_hat_l^2 + \mu^2)$ and $||E|| < \mu$, correcting the block and applying Weyl gives the valid bounds sigma_min >= μ - ||E|| and sigma_max <= t + ||E||. The regularized default in Section 5.2 meets the added target restriction. Every phase still requires its approximation-error condition. Other positive targets admit different bounds. This corrects the printed guarantee without re-evaluating the experimental results.

[116] Gunter Malle. *Multi-parameter polynomials with given Galois group*. 2000. Journal of Symbolic Computation 30(6), 717-731 (2000).
Primary source · Author's paper, §10
Section 10 constructs a two-parameter family for M_{12} and obtains totally real specializations for M_{11} and M_{12}. These examples supply a precedent for changing the family when real roots are required. The groups and constructions differ from the proposed M_{23} problem.

[118] Michael D. Fried and Pierre Dèbes. *Rigidity and real residue class fields*. 1990. Acta Arithmetica 56, 291-323.
Primary source · Author's copy, Proposition 1.2
Theorem 1.1 treats involution data for real fibers. Proposition 1.2 excludes the totally real full-group specialization of a regular three-branch-point Galois cover over the rationals except for specified small dihedral groups. It leaves other constructions open.

[119] Frank Häfner. *Braid orbits and the Mathieu group M_{23} as Galois group*. 2022. arXiv:2202.08222v3.
Primary source · Class vector and parameter-curve genus
Tables 9.1-9.2 treat the class vector with two 4A and two 2A entries. The ordered parameter curve has genus 447. The full braid orbit contains 2,456 generating classes before the real-pattern restrictions in the pilot. The genus of the degree-23 source is a separate calculation.

[120] Refactor. *Calculations for the sparse and arithmetic pilots*. 2026. Python and GAP 4.16.0, 18 November 2026.
Primary source
Download the three scripts and instructions. They reweight the published CAST data, run the exact eight-cycle comparison, and independently count and check the M_{23} branch candidates by two methods. They certify the stated finite calculations. They do not measure a portfolio return or construct a totally real rational field.

[140] L. Ridgway Scott and Michael Vogelius. *Norm estimates for a maximal right inverse of the divergence operator in spaces of piecewise polynomials*. 1985. M^2AN 19, 111-143, §2, Figure 1, condition R.1 and Proposition 2.1.
Primary source
At a singular interior vertex of a planar triangulation the incident edges lie on two lines. Values of a piecewise polynomial in the divergence image obey an alternating-sum constraint. The straightness of the edges is geometric information beyond the mesh adjacency graph.

[141] Benedikt Gräßle, Nis-Erik Bohne, and Stefan A. Sauter. *The pressure-wired Stokes element: a mesh-robust version of the Scott-Vogelius element*. 2024. arXiv v3, §§3-6, Theorems 2-3.
Primary source
On the specified planar, shape-regular triangulations with velocity degree at least four, enforcing pressure constraints up to a fixed small positive geometric threshold controls near-singular instability. The method can lose pointwise divergence freedom. The authors bound the divergence defect and report numerical experiments. No investigators were compared.

[147] Vijay S. Mahadevan, Jorge E. Guerra, Xiangmin Jiao, Paul Kuberry, Yipeng Li, Paul Ullrich, David Marsico, Robert Jacob, Pavel Bochev, and Philip Jones. *Metrics for Intercomparison of Remapping Algorithms (MIRA) protocol applied to Earth system models*. 2022. Geoscientific Model Development 15, 6601-6635, §§3-4.
Primary source · Ullrich and Taylor on linear remapping · Kritsikis et al. on overlap geometry · Harten on adaptive multiresolution · Wavelet summaries for later range queries
The protocol assesses five properties of grid transfers, including the dissipation seen under repeated remapping. The linked constructions give conservative, overlap-based or multiresolution baselines. Matias, Vitter and Wang study compressed summaries for later range queries on fixed data. The four-cell calculation in this essay isolates a late regional observation.

[151] Refactor. *Exact checks for conservative remapping and labelled sumsets*. 2026. Two standalone Node.js scripts, 18 November 2026.
Primary source
Agent-proposed examples from the two assigned fields. The scripts check the dyadic overlap map, alternating-mode recovery, the seven possible rulers and the optimal decision depths for all 595 three-sum comparisons. Only comparisons, the old table and the stated normalization are available to the sumset observer.

[153] Karl E. Taylor. *Truly conserving with conservative remapping methods*. 2024. Geoscientific Model Development 17, 415-430, §2.2 and Figures 2-3.
Primary source
In the deliberately coarse test, two alternatives conserve the true global mean while producing area-weighted root-mean-square errors of 0.15 K and 0.63 K. The uniform correction can exceed the source maximum. The example demonstrates conflicting demands on the remap. The magnitudes shrink at the finer resolutions examined.

[156] Erik G. Boman, Bruce Hendrickson and Stephen Vavasis. *Solving Elliptic Finite Element Systems in Near-Linear Time with Support Preconditioners*. 2008. SIAM Journal on Numerical Analysis, Theorem 5.1 and §§6-8.
Primary source
Each element supplies a node-edge star. Quadrature, element shape and within-element variation in the scalar conductivity enter the spectral comparison. Under its hypotheses node count and element size do not. The theorem controls preconditioning of the discrete stiffness system.

[158] Matteo Croci and Patrick E. Farrell. *Complexity bounds on supermesh construction for quasi-uniform meshes*. 2020. Journal of Computational Physics 414, 109459, Theorem 1.1 and §3.
Primary source · Farrell and Maddison, local Galerkin projection
The common subdivision of two quasi-uniform mesh families has a linear cell-count bound under the stated intersection and geometric assumptions. Farrell and Maddison use local supermeshes for conservative projection. The regional recovery criterion in the essay is an elementary row-space argument for piecewise-constant donor fields.

Agents already diagnose failures and invent repairs

[28] David Bader, Adil Chhabra, Ernestine Großmann, Monika Henzinger, Alexander Noe, and Christian Schulz. *Agentic Algorithm Engineering: Improving Shared-Memory Exact Minimum Cuts*. 2026. arXiv v1, 7 September 2026.
Primary source
Section 4 and conclusion: 44 experiments, 23 accepted human-reviewed changes; reported speedups use the same 53-instance workload used for tuning, with matched cut results. Real-world k-core gains are 1.28× sequential and 1.63× at 32 threads.

[30] Peter Chen and Wotao Yin. *Training Numerical Intelligence via Auto-Diagnosis and Skill Discovery*. 2026. arXiv v1, 2 October 2026.
Primary source
Sections 3 and 4.5: automated diagnosis, literature retrieval, and skill construction are ablated together. Direct revision worsens five of nine comparisons. Off-the-shelf solvers are excluded; final-generation budget matching is not whole-pipeline cost matching. No human specialist comparison.

[32] LinearSolveBench contributors. *LinearSolveBench: discovered FLASH solvers and benchmark specification*. 2026. Pinned release commit 9322e8193e0425d149ae36e793af34ac26b695b3, 29 September 2026; checked 4 November.
Primary source · Pinned benchmark specification · Original per-case timing record
Four released C/HYPRE solvers; 640 candidates per model and selection on 96 development cases, not 96 independent matrices or an unseen test set. The reported 1.8504555 ratio is summed reference runtime divided by summed candidate runtime for the Fable 5.1 winner. Timings include creation, setup and solve, excluding discovery, compilation and verification. The import reports 96/96 local macOS correctness checks without a new Linux/Modal performance run. The original timing record was independently summed for this review. These results demonstrate solver optimization without establishing a longitudinal rate or superiority across problem classes.

[35] Sihan Hu, Lyuhan Huang, Youjin Deng, and Kun Chen. *SciCode-Verified: How Benchmark Defects Underestimated the Scientific-Coding Ability of Language Models*. 2026. arXiv v1, 5 August 2026.
Primary source
Table 2 compares supplying and withholding expert background on 64 corrected main problems under two-environment OR grading per subproblem, with every scored subproblem required for a main-problem pass. The three models discussed in the essay have single-run comparisons. The two DeepSeek V4 Preview rows average three runs. GLM-5.2 has no run without background. Table 1 separately reports 92.2 percent for GPT-5.6 Sol with background. Its matched old and corrected scores quantify the effect of repairing the benchmark.

[44] Alexander Novikov et al. *AlphaEvolve: A coding agent for scientific and algorithmic discovery*. 2025. arXiv v1, 16 June 2025.
Primary source
Table 1 and sections 2.1-2.4 describe multiple evaluation metrics, evaluation cascades, and lengthy accelerator evaluations. The reported loop takes an evaluation mechanism as input. This does not establish that agents cannot design or improve evaluators.

[55] Dhruv Srikanth, Bingchen Zhao, Dixing Xu, Yuxiang Wu, and Zhengyao Jiang. *Recursive self-improvement of AI research agents*. 2026. arXiv v1, 22 September 2026.
Primary source
Sections 2-3: eight-day autonomous optimization of a research-agent harness, seven accepted improvements, and four held-out benchmarks. The human-engineered production agent is an agent baseline, not a live human comparison; no sparse-solver specialist premium is measured. Section 3.5 reports an agent finding and repairing a suite-wide scoring failure. Section 3.6 compares harnesses in a small experiment with inconclusive endpoints.

[69] FrontierCS 2 Team. *FrontierCS 2: A Benchmark for Computer Science R&D, current task records*. 2026. 4 November 2026 observation; historical run table checked 18 November.
Primary source · Current leaderboard · Pinned benchmark source · 6 November evaluation update
The publisher's website repository preserves all 864 main run records for 18 systems and 48 tasks. Counting the pass flag gives 13 systems for Clusterwise SpGEMM, eight for KAMI and seven for Polar Express. All 864 model-task pass flags agree with the table saved for the original review. Each task has its own evaluator. The October paper reports 12 systems. The 6 November update added systems and reported repairs to other judges. Completion-time curves do not regrade intermediate methods or independently vary budgets.

[70] FrontierCS 2 Team. *SVG-EAR: The Baseline No Model Could Beat*. 2026. 19 October.
Primary source
First-party twelve-system case study. Astra's budget and variance-correction contributions are not separately ablated. Exact attention misses the same five fidelity checks and one additional check. No all-checks pass does not mean no algorithmic improvement.

[88] Mert Yuksekgonul et al. *Learning to Discover at Test Time*. 2026. arXiv v2, 5 February 2026.
Primary source
TTT-Discover updates weights on the current target with a supplied continuous reward. Its objective is one strong artifact. General transfer and complete training/inference/evaluation cost parity require separate evidence. Added in the 7 November review.

[106] Stephen Chung, Wenyu Du, and William J. Wesley. *Autonomous Mathematical Discovery in an Open-World Multi-Agent Environment*. 2026. arXiv:2608.23691v2, 14 September 2026.
Primary source · Minimum-overlap result and change of target, §6.2
Section 6.5 reports the strict lower bound within the prescribed-double-root family and a degree-226 construction outside it. The human-designed task supplied the initial Laguerre representation and evaluator. Section 6.2 and the contribution table document an upper-bound brief and an agent-chosen lower-bound certificate. The essay uses this episode to establish who changed the objective.

[114] Mathew Chow and Tao E. Li. *Can AI Agents Reliably Optimize Computational Chemistry Codes?*. 2026. arXiv v1, 7 November.
Primary source · Adaptive solvers and Figure 5
Section IV.3 and Figure 5 compare adaptive self-consistent-field implementations developed with an autonomous agent harness and by researchers working interactively with bare Codex. Both recognized orbital-gradient plateaus on two iron complexes. On serine the sixfold runtime is for the human-guided implementation against standalone DIIS, and is largely attributable to its postconvergence stability analysis. This comparison prices the resulting calculations. Development harnesses differed, and no equal-budget staffing effect follows. In the nickel case both adaptive methods converged to internally stable states with higher energy than standalone OTR. Internal stability and energy ordering alone do not establish which state best represents the physical ground state.

[124] Johannes Schmitt et al. *ProofCouncil: An LLM Agent for Solving Open Mathematical Problems*. 2026. arXiv:2607.09474v1, §3.2.2, second First Proof batch.
Primary source · Independent referee report, §5.8, p. 31
The seven approvals and fresh rejections are iterations on problem 10. On problem 3 the critic repeatedly rejected a proof the referees accepted up to minor revisions. On problem 8 the referees identified the unsupported assertion about a flag of valuated matroids. A staged machine audit upgraded its assessment of the gap after reading the human solution. The report in note 8 concerns a later batch.

[128] Yujian Liu et al. *Building AI Research Teams for Autonomous Discovery of Linear Attention Architectures*. 2026. arXiv:2610.28887v1, §3.3.
Primary source
The system used sixteen prompted disciplinary profiles, separate histories and a common experimental record. The controlled comparisons removed profiles or private histories. They concern architecture search with measurable performance and contain no human-partner arm.

[154] Nithin Somasekharan, Rabi Pathak, Manushri Dhanakoti, Tingwen Zhang, Ling Yue, Andy Zhu, and Shaowu Pan. *AI CFD Scientist: Toward Open-Ended Computational Fluid Dynamics Discovery with Physics-Aware AI Agents*. 2026. Preprint, §4.2 and Appendix J, Tables 23-24.
Primary source
An expert-designed visual verification step found 14 of 16 deliberately introduced failures and marked all four clean controls for revision. The evaluation harness rendered only part of the material requested. The experiment tests the check on known fault classes. It does not compare investigators who invent the next check.

[157] Binghang Lu and colleagues. *RSI-PDEbench: A Benchmark for Self-Improving Agents in PDE Solving*. 2026. Preprint, version 1, §§2, 5.3 and 6, Table 4.
Primary source
The Hard-OOD setting mixes unseen templates with shifted coefficients. Both the native iterative GPT-6.1 harness and the residual-guided harness passed 261 of 278 cases. The reported single-run baseline passed 135. The tasks, family library and field-accuracy criterion were supplied. The feasibility filter and single Hard-OOD trial per condition limit inference.

The calculation that runs has to earn its guarantee

[22] Douglas N. Arnold, Richard S. Falk, and Ragnar Winther. *Finite element exterior calculus: from Hodge theory to numerical stability*. 2009. arXiv v2, 12 August 2009.
Primary source · Maxwell mesh comparison, Figure 2.6
Sections 3.3-3.4, especially Theorems 3.8-3.9, relate suitable subcomplexes and uniformly bounded cochain projections to stability and error estimates. Sections 4-6 develop differential complexes, boundary conditions and finite elements. The framework requires quantitative bounds and approximation properties; a discrete incidence identity alone is insufficient. This supplies the stability requirement in the discretization example.

[27] Théophile Chaumont-Frelet, Zhaonan Dong, Gregor Gantner, and Martin Vohralík. *On p-robust convergence and optimality of adaptive FEM driven by equilibrated-flux estimators*. 2026. arXiv v2, 10 September 2026.
Primary source
Algorithm 3.1 and Theorems 3.4/3.7 concern fixed-p h-adaptive conforming Poisson FEM and the specified right-hand side. Condition (3.2) and the marking condition are material; experiments do not make their satisfaction unconditional.

[33] Haoming Luo. *GPT-6 FEniCSx Single-Attempt Pass Rate and Solver Corpus*. 2026. Dataset snapshot and count audit prepared for this essay, 4 November 2026.
Primary source · Pinned raw records · Pinned upstream scorer, lines 32-62
Advertised total 496/644 evaluated passes and one reference failure among 645 requested tasks. Counts recomputed from pinned raw records: Stokes 20/60; reaction-diffusion 59/64; 68/496 passes (13.71%) with partially finite reference arrays; 69 execution failures without a coverage assessment. Many partially finite cases have nonrectangular geometry, so a mask may legitimately exclude points outside the domain. These counts are not evidence of wrong solutions. The pinned scorer excludes NaNs in either array; interpreting coverage requires the domain details. Proposed stronger protocol: an independent required-domain mask, finite predictions at every required point, and coverage reported beside error.

[40] Eric Chillón, Artur Lidtke, Nguyen Anh Khoa Doan, and Bernat Font. *Acceleration of an algebraic multigrid pressure solver using graph neural networks*. 2026. arXiv v1; checked 4 November 2026.
Primary source
Section 4.2.1/Table 3 gives the ten-model mean on fine 3D tests; section 4.2.3/Figure 6 concerns the selected model on 1,000 AirfRANS systems. Relaxed Jacobi baseline, 10^-4 tolerance, 500-cycle cap. The preprint's arXiv stamp and displayed manuscript date differ. Section 3.4/Algorithm 1 predicts coefficients and constructs the sparse smoother before iteration. The higher per-cycle cost in section 4.2.3 is application of that smoother, not per-cycle GNN message passing. AirfRANS was run in FP64 after FP32 stagnation. The other reported experiments, including the separate fine three-dimensional comparison, used FP32.

[41] Daniel Zhengyu Huang and Andrew Stuart. *Cost-Accuracy Trade-offs: Neural Operator vs Classical Numerical Solver*. 2026. arXiv v1, 21 September 2026.
Primary source
Training/data are assumed amortized. Accuracy and workload affect the comparison; section 4.1.3 reports a limited empirical test-error plateau. Some wall-clock comparisons use different hardware allocations. This does not measure a general AI-research capability plateau. Section 4.1.3/Table 3: fixed-architecture errors at N=1000/2000/4000/8000 are 9.18/6.82/5.56/4.91 times 10^-3; largest tested architecture at N=8000 gives 4.70 times 10^-3. This is scaling within a specified surrogate experiment, not calendar-time stagnation in AI research.

[46] Masayuki Yano. *A Reduced Basis Method with Exact-Solution Certificates for Steady Symmetric Coercive Equations*. 2015. Computer Methods in Applied Mechanics and Engineering 287, 290-309.
Primary source · Author manuscript
Sections 2.2-2.3, Propositions 1-3: primal and complementary variational bounds on exact weak-solution energy, algebraic dual feasibility, and an energy-norm error certificate from the bound gap. The method requires symmetric coercivity, affine parameter dependence, and exactly represented piecewise-polynomial data. Specialized offline dual solves and a potentially large online expansion remain costs; online complexity is independent of finite-element dimension.

[53] Richard S. Falk. *Math 575: Finite Difference Methods, Stability and Error Estimates*. 2012. Rutgers lecture notes, lectures 1-3.
Primary source
A classical route to finite-difference error estimates for smooth Poisson problems through truncation error, stability, and a discrete maximum principle. This shows that a PDE approximation argument need not universally use Sobolev-space machinery; it does not replace weak-solution analysis in every setting.

[59] Stephen F. McCormick, Joseph Benzaken, and Rasmus Tamstorf. *Algebraic Error Analysis for Mixed-Precision Multigrid Solvers*. 2021. SIAM Journal on Scientific Computing 43(5), S392-S419.
Primary source · Published article
Theorems 4.1 and 8.1: sparse-SPD rounding and contraction analysis with level-dependent precision. Applies to the specified iterative-refinement and multigrid constructions.

[63] Yongseok Jang, Pierre Jolivet, and Théo Mary. *Mixed Precision Augmented GMRES*. 2026. Numerical Linear Algebra with Applications 33(4), e70117.
Primary source · Author manuscript
Sections 4-6 explain finite-precision defects in GMRES-DR simplifications and test the augmented formulation. Sparse and CFD examples include remaining precision sensitivity. Hardware-specific timing is deferred; this is not a human-agent comparison.

[64] Noaman Khan and Erin Carson. *Mixed precision iterative refinement with adaptive precision sparse approximate inverse preconditioning*. 2025. Engineering with Computers 41, 3833-3846.
Primary source
Section 4 and conclusion. steam1 storage fractions and iteration counts are not wall-clock speedups. The authors identify prediction of favorable matrix cases and large-scale timing as further work.

[65] Emmanuel Agullo, Luc Giraud, Pierre Jolivet, Théo Mary, and Alexandre Tabouret. *Doubly adaptive precision restarted GMRES*. 2026. HAL v1, submitted 2 June.
Primary source · Full preprint, experiments in §4.3 · Author's HAL publication record
Section 4.3, Figure 6 and Table 3 report a peak 1.6-fold speedup on nd24k. Timings in the figure are conditional on reaching baseline backward error. Some heuristic runs stagnate, and one heuristic reaches the target on only one matrix. The strict bound affords smaller gains.

[66] Michel Crouzeix and Anne Greenbaum. *Update to: Spectral sets: numerical range and beyond*. 2026. arXiv v2, 23 September.
Primary source
Uses Lorist-Schwenninger to obtain 2-spectral annuli under the stated norm assumptions, and 4-spectral convex regions with holes or cutouts. Sections 4-5 state the extra inverse/numerical-radius information and GMRES/rational-Krylov applications. A useful region is not obtained for free.

[67] Cecilia Chen and John Urschel. *Estimating the numerical range with a Krylov subspace*. 2024. arXiv:2411.19165.
Primary source · Author manuscript
Quantitative approximation and limitations of projected numerical ranges. Orthonormal compression gives an inner set. The analysis excludes floating-point implementation; no general cheap outer certificate is asserted.

[68] Siegfried M. Rump. *Verified Error Bounds for Sparse Systems Part I: The Splitting of a Matrix into Two Factors*. 2026. SIAM Journal on Matrix Analysis and Applications 47(4), 1739-1784; online 8 October.
Primary source · Author final manuscript
General real/complex sparse systems; entrywise verified error bounds, factorization/fill effects and adaptive use of stronger bounds. Published 8 October. Existing verification is a baseline, not evidence of an unoccupied human role.

[95] Siegfried Cools, Emrullah Fatih Yetkin, Emmanuel Agullo, Luc Giraud, and Wim Vanroose. *Analyzing the Effect of Local Rounding Error Propagation on the Maximal Attainable Accuracy of the Pipelined Conjugate Gradient Method*. 2018. SIAM Journal on Matrix Analysis and Applications 39(1), 426-450; arXiv v4, 29 November 2017.
Primary source · Published article
Sections 2-4 analyze coupled residual and auxiliary-vector gaps, derive a practical replacement rule, and evaluate attainable accuracy and parallel performance. The exact-arithmetic equivalence does not eliminate rounding-error propagation. The repair is already automatic and does not resolve every finite-precision effect or establish a human-agent premium.

[109] Jan Bouwe van den Berg and Maxime Breden. *A simple rigorous integrator for semilinear parabolic PDEs*. 2026. arXiv:2601.05146v1, 8 January 2026.
Primary source
Theorem 1.2 fixes the Ohta-Kawasaki parameters and initial data and gives the uniform trajectory bound. Section 5 explains why the infinite final step also certifies a locally stable equilibrium and convergence. Section 6.1 compares validation methods on a separate Swift-Hohenberg trajectory, with hardware dependence explicit.

[111] Gonzalo Cao-Labora and Jaume de Dios Pont. *Counterexamples to Schiffer's Conjecture*. 2026. arXiv:2608.05114v1, 5 August 2026.
Primary source
The abstract and Section 1.1 describe the relaxed symmetry order, the uniform extent of the branches and the return to integer orders giving actual planar counterexamples. This proof and the certificate in note 108 address the same Schiffer boundary conditions.

Observed research partnerships and tests of their advantage

[45] Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Aaron Roth. *Generalization in Adaptive Data Analysis and Holdout Reuse*. 2015. NeurIPS 2015; arXiv v2, 25 September 2015.
Primary source
Adaptive reuse can overfit validation data; the paper supplies methods preserving statistical validity. It does not prove inevitable search failure or give a theorem covering arbitrary PDE distribution shifts. Withholding meaningful problem families is the experimental application proposed here.

[61] Youran Sun, Xingyu Ren, Chugang Yi, et al. *SciPiVot: Benchmarking Principal-Investigator Judgment at Pivotal Moments of Scientific AutoResearch Agents*. 2026. arXiv v1, 15 October.
Primary source · Pinned results
Sections 2-4 and Appendices A-D. Thirteen fresh researchers, with AI assistance available; retrospective model-filtered corrections. No matched time/cost, unaided-human arm, or sparse-solver subgroup. Repeated option orders are not independent cases. Appendix C.4, case 206, distinguishes a travel-limited response in the simulated support from a change in the number of springs.

[74] Amaç Herdağdelen, Hailey Joren, and Alex Tamkin. *How AI agents shape experts' judgment calls*. 2026. 27 October.
Primary source
Fifty agent-using professional analysts, Opus 4.8, randomized and deliberately strengthened preferences. Defensible analytical choices, not correctness labels. The self-selected nonuser group does not support a causal AI-versus-no-AI accuracy comparison.

[89] Ian Berlot-Attwell, Tobias Sesterhenn, Frank Rudzicz, and Xujie Si. *Is This LLM Library Learning? Evaluation Must Account For Compute and Behaviour*. 2026. Proceedings of EACL 2026.
Primary source · Table 2, direct reuse · Table 9, original logs
Equal-compute comparisons re-examine LEGO-Prover, TroVE and AgentOptimizer. Tables 2 and 9 record rare instances of direct lemma reuse across problems, including verbatim reuse. Retrieval and use on one problem are more frequent. Tests of the proposed indirect reuse mechanism concern LEGO-Prover.

[90] Haocheng Ju et al. *Automated Conjecture Resolution with Formal Verification*. 2026. arXiv:2604.03789v2.
Primary source · Formalization and intervention comparison
Archon includes planning and Lean agents. Section 4.4 compares two continuations from one checkpoint with three unique-factorization obligations. The 132 and 223 minutes are elapsed session times. The system retained its Kaplansky route and drew on intermediate observations in the human blueprint. The experiment omitted the cost of writing it and an arm in which a fresh machine adviser supplied a blueprint.

[91] Kimm Fesenmaier. *Caltech and AIM Build AI Platform with Mathematicians for Mathematicians*. 2026. Caltech announcement, 7 October 2026.
Primary source
Documents the announced platform and the account of research taste attributed to Yisong Yue. It supports the existence of this proposed division of labour, not a measured or permanent human advantage. Added in the 7 November dossier review.

[92] Daniël Lakens. *Equivalence Tests: A Practical Primer for t Tests, Correlations, and Meta-Analyses*. 2017. Social Psychological and Personality Science 8(4), 355-362.
Primary source · Publisher PDF at the author's institution
Supports prespecifying the smallest effect of interest and distinguishing evidence against a worthwhile effect from nonsignificance. The proposed study requires confidence procedures appropriate to its binary success outcome, dependence, planned contrasts, and stopping rule. Added in the 7 November dossier review.

[101] Nhat-Minh Nguyen. *Physics Is All You Need? A Case Study in Physicist-Supervised AI Development of Scientific Software*. 2026. arXiv:2605.30353v1, 28 May 2026.
Primary source
Sections 3.2, 4.3 and Appendix A. Attribution partly relies on supervisor recollection. No concept-versus-retrieval ablation was run. The agent found both code paths and defended the analytic one after a general prompt to reconsider. Naming the angular damping changed the choice.

[102] Yue Shui, Chenyu Ma, Hangfei Xu, Shengzhao Wen, and Yanpeng Wang. *Harness Engineering for LLM-Driven GPU Kernel Generation*. 2026. arXiv:2607.17979v1, 20 July 2026.
Primary source
Section 4 and Appendix E. Ratios compare selected-artifact mean latencies. Fallback checks cover two concern workloads. Benchmarks were not independently reproduced.

[103] Patrick Hemmer, Max Schemmer, Niklas Kühl, Michael Vössing, and Gerhard Satzger. *Complementarity in human-AI collaboration: concept, sources, and evidence*. 2025. European Journal of Information Systems 34(6), 979-1002.
Primary source · Open publisher edition at KIT · Author manuscript
Sections 3-4 distinguish complementarity potential and realized effect. House-valuation and image-classification experiments examine information and capability differences. These provide a framework for designing comparisons, not a ranking of mathematical research fields. Giving humans additional context does not establish a capability advantage over equally informed agents. In §4.2.2-4.2.4 and Figure 11, each classifier misses four of fifteen selected images. Participants make their own judgment first. Changing which images the machine gets wrong changes average team error from 24.73 to 14.61 percent.

[104] Michelle Vaccaro, Abdullah Almaatouq, and Thomas W. Malone. *AutoAnalysis: An automated, self-updating meta-analysis of human-AI collaboration*. 2026. arXiv:2610.42001v1, 30 October 2026.
Primary source
The preprint synthesizes 1,955 effects from 420 papers. Its 171-effect synergy subset gives g = −0.35, 95 percent CI [−0.52, −0.18], against the better-performing standalone condition. The division-of-labor association has p = 0.031, uncorrected for multiple moderator comparisons. Automated extraction received small author checks and separate agent audits, rather than comprehensive independent human validation. Heterogeneity is high, moderators are noncausal, and mathematical specialties are not compared.

[105] Alan Li, Rahul Saha, Anton Xue, Swarat Chaudhuri, Adam Klivans, Pravesh K. Kothari, and Raghu Meka. *Long-Horizon AI Research for Grothendieck Constant: A Case Study in Human-AI Mathematical Collaboration*. 2026. arXiv:2608.11195v3, 14 August 2026.
Primary source · Mathematical companion paper
Appendix A reconstructs the first operator intervention in session 18 from the six failures in sessions 12-17. Dualization was already in the brief. The team and agent had further work to do before the proposed transfer applied to the full class. Later in Appendix A, fourteen sessions between sessions 32 and 51 obtained restricted sublemmas without establishing the global bound. The operators redirected the agent. No matched autonomous continuation prices that redirection. Sessions 55-57 identify the failed passage to mixtures, the agent's affine constraint in two Hermite coefficients, and the operators' decision to certify an attainable member first. The first resulting lower bound was $6\pi/11$. Stronger rungs followed.

[107] Matthew J. Colbrook, Siavash Sadeghi, and George Stepaniants. *A computer-assisted counterexample to the planar Berenstein conjecture*. 2026. arXiv:2608.08953v1, 9 August 2026.
Primary source · Verifier and certificate at the inspected revision
Theorem 1.1 and Sections 1.1-1.3 distinguish the Fourier search from the fixed-disk formulation used to certify the solution and control the tail. The AI declaration and Appendix A document the contributions to the proof and code and the computer-assisted estimates.

[108] Matthew J. Colbrook and George Stepaniants. *A computer-assisted counterexample to the planar Pompeiu and Schiffer conjectures*. 2026. arXiv:2608.01579v1, 3 August 2026.
Primary source
The preceding construction reaches the endpoint with vanishing outward derivative. Its conformal reduction and working implementation supplied the starting point for the Berenstein project.

[110] Xiaoyu Huang, Blake Jackson, Kyu-Hwan Lee, Bjorn Poonen, Rachel Pries, and Shaowu Zhang. *The Mathieu group M_{23} is a Galois group over $\mathbb{Q}$*. 2026. arXiv:2608.08538v2, 7 October 2026.
Primary source · Galois action and recognition of the other covers · Real fibers and complex conjugation, Proposition 3.18
Sections 3.2-3.7 distinguish approximate reconstruction and exact verification. The footnote in Section 3.7 details the increased precision, use of six other triples, restricted initial recognition and change of scale. The authors' "Use of AI" statement describes human direction and AI-assisted computation. Sections 1.4 and 1.5 describe descent and the infinitely many extensions. Example 1.2 gives the polynomial. Its Sturm sequence has 15 sign changes at negative infinity and 8 at positive infinity. Proposition 3.18(b) treats complex conjugation at every rational parameter.

[113] Tom Savage and Ehecatl Antonio del Rio Chanona. *Human-algorithm collaborative Bayesian optimization for engineering systems*. 2024. Computers & Chemical Engineering 189, 108810. Experiment locators refer to arXiv v1..
Primary source · Human protocol · Coiled-tube results and Figure 14
Sections 4.2, 4.4 and 5.5, Figure 14. Six chemical-engineering postgraduates guided reactor design, compared with 32 standard Bayesian-optimization runs. Participants could choose up to eight initial points before the sequential selections. The objective was a surrogate fitted to 75 OpenFOAM evaluations. The advantage diminished with iteration. The study does not isolate later choices from initialization, charge human hours or compare equally equipped language-model agents. Sections 5.1 and 5.4 also study random choice among proposals on synthetic tasks. The reactor comparison does not isolate the contribution of selecting from that menu.

[115] George Tsoukalas et al. *Advancing Mathematics Research with AI-Driven Formal Proof Search*. 2026. arXiv v2, 8 June.
Primary source · Green's problem and the counterexample · Discussion of library coverage and subgoals
Section 4 describes the real-valued proof, Ben Green's clarification, numerical evidence for a counterexample and its formal verification. Section 6 reports better results where the libraries are mature and problems decompose. The example separates local proof feedback from choosing the intended statement. It supplies no timed autonomous branch for correcting the domain.

[122] Howard Raiffa and Robert Schlaifer. *Applied Statistical Decision Theory*. 1961. Harvard University, Graduate School of Business Administration, §4.5, equations (4-36)-(4-39).
Primary source · Publisher's reprint
Section 4.5 compares acting with sample information to acting on prior information and subtracts the cost of obtaining the sample. Appendix C uses this distinction to account for a team's diagnostic experiments.

[123] Garry Kasparov. *The Chess Master and the Computer*. 2010. The New York Review of Books, 11 February 2010.
Primary source · 2005 freestyle winners on their move selection · Kasparov on what makes a good freestyle player
Kasparov introduced Advanced Chess in León in 1998. In their contemporary report, Zackary Stephen and Steven Cramton describe dividing lines and cross-checking moves with different engines at the 2005 PAL/CSS freestyle tournament. Kasparov's 2017 interview distinguishes strength as a solo player from skill at organizing the partnership. The historical competition allowed different people, software and equipment.

[125] Martin Balko et al. *Bolzano: Case Studies in LLM-Assisted Mathematical Research*. 2026. arXiv:2604.16989v2, §§2 and 3.1.
Primary source
Parallel provers and a verifier worked on the black-box separation for adaptive collision search. Section 3.1 credits the system with the nested search problem and records the strategic intervention and the following four rounds. It links the research transcript and complete proof. There is no matched continuation with a new machine adviser.

[126] Minghao Liu, Jiaheng Wei, Yang Liu, and James Davis. *Human and AI Perceptual Differences in Image Classification Errors*. 2025. AAAI 39, 14318-14326, §4 and Figure 5.
Primary source · Publisher's record · Authors' full paper on human and model errors
Using recorded human labels on CIFAR images, Figure 5 compares a single human answer with a second classifier for a fixed base classifier. The 95.82 and 92.34 percent results assume perfect selection between answers. A confidence-threshold rule also favored human labels against the best tested second classifier. Selection was retrospective on one dataset.

[127] Cameron R. Jones, Agnese Lombardi, Kyle Mahowald, and Benjamin K. Bergen. *LLMs and people both learn to form conventions - just not with each other*. 2026. arXiv:2602.08208v1, Experiment 1.
Primary source
Pairs repeatedly described and identified ten abstract figures with feedback. The two GPT-5 instances learned a common vocabulary rapidly. The reported 99 and 70 percent refer to blocks four and five respectively. This measures coordination on repeated referents. It cannot rank teams on proof discovery.

[129] Mark Steyvers et al. *Bayesian modeling of human-AI complementarity*. 2022. Proceedings of the National Academy of Sciences 119, e2111547119, Figures 2-4.
Primary source · Full text and team comparisons
Independently elicited judgments and confidence were combined with a Bayesian rule fitted using fourfold cross-validation. Mixed pairs competed against pairs of two different networks and pairs of people. Their relative advantage concentrated at comparable levels of individual performance, particularly with image noise. The participants did not collaborate on choosing a new test.

[130] Christian Coester. *Improved lower bounds for general algorithms*. 2025. In Early science acceleration experiments with GPT-5, §IV.2.2, pp. 64-67.
Primary source · Coester's mathematical account, Section IV.2.2
Coester describes the continuous planar construction, the possible inspiration from a model trace, the positive separation that prevents freezing, and the model's orthogonal-plane extension. The paper presents the lower bound and identifies errors in the proposed higher-dimensional execution. It includes an excerpt of the planar exchange and links the higher-dimensional chat. No partner-substitution trial was run.

[131] Bogdan Georgiev, Javier Gómez-Serrano, Terence Tao, and Adam Zsolt Wagner. *Mathematical exploration and discovery at scale*. 2025. arXiv:2511.02864v3, Problem 6.40.
Primary source · Released discrepancy examples
The authors report discrepancy-two sequence lengths of 200 without guidance and 380 with an approximate-multiplicativity hint in the same search time. The known optimum is 1,160. The comparison identifies a benefit of a supplied search direction in those runs. It does not measure the cost of inventing the hint or test a rival machine adviser.

Related essays

What a New Proof Makes Possible
What remains worth comparing when proof generation becomes abundant, including dependencies, procedures and explanatory gain.

When a Description Becomes a Verdict
Why an unchanged spectrum or a reassuring summary may omit the information a numerical decision needs.

Mathematical Life in the Age of AI
How assistance can become a learner's capacity to check, adapt and continue an argument.

What Sustains Mathematical Expertise
The training and shared resources behind the expertise taken as available in the research comparison.