AI research and the remaining human advantage

Why abstraction is no refuge, and where sparse numerical computation offers a stronger research bet

Henry Shin
4 November 2026
Text revision 047145c9215b

The old refuge in abstraction is over. My judgment is that algebraic geometry, representation theory, and related structural subjects no longer offer a credible shelter for a distinctive human research role: AI already contributes at the conceptual level that was supposed to protect them. I would redirect research effort to the design and repair of sparse numerical methods, where spectral graph algorithms, matrix analysis, and numerical analysis meet. The central task is to diagnose what a cheap representation fails to preserve, then construct an economical correction. This is a definite strategic choice, not a measured ranking of complete research workflows. Evidence checked through 4 November 2026.

In this essay

1. A rapidly changing research advantage
2. Beyond verbal imitation
3. Where does human expertise still matter?
4. Finite structure under computation
5. Diagnosing and repairing sparse representations
6. Automation already reaches numerical research
7. From a promising method to a reliable solver
8. Testing the proposed advantage
9. Where to concentrate research effort

1. A rapidly changing research advantage

The old refuge in abstraction is over. I would no longer advise a mathematician to seek a durable human research advantage by moving further into algebraic geometry, representation theory, abstract algebra, or another technically forbidding structural subject. AI systems now contribute substantial proofs, useful organizing constructions, and connections between theories. Moving to more remote objects or more intricate arguments does not restore a division of labor in which humans supply the conceptual architecture and machines merely carry out calculations.

My assessment is that the human advantage supplied by abstract proof construction is eroding rapidly. A subject can produce more mathematics, attract more users, and even employ more people while requiring less specialist input for a comparable result. Its intellectual growth does not restore the scarcity of a contribution that machines can supply. The strategic claim concerns that loss of protection; it does not require every research stage to have been automated or every mathematician to have become replaceable. The studies below establish important capabilities, not a field-wide replacement rate.

The former protection rested on the belief that sufficiently sophisticated abstraction required conceptual command that machines could imitate verbally but could not exercise at the research frontier. The surviving evidence now contradicts that belief. Machine-generated work is being reconstructed by specialists, developed into further mathematics, and checked through independent routes. The failures matter, but they do not erase the contributions that survive examination. A presumption of human conceptual exclusivity is no longer a credible basis for investing years of mathematical training.

I would concentrate that effort on the design and repair of sparse numerical methods at the intersection of spectral graph algorithms, matrix analysis, and numerical analysis. A change in graph weights, connectivity, or arithmetic precision can invalidate the representation that made a computation cheap. The research task is to identify what has been lost, construct a correction, and establish that the corrected method repays its cost. Approximate elimination and algebraic coarse spaces make the choices concrete: which interactions to retain, which modes to represent, and how accurately to compute with them.

By a human research advantage I mean the additional progress that specialist participation makes possible when strong AI is already available, with time, computation, checking, and failures included in the cost. The comparison is with the same agent working autonomously and with a mathematically capable generalist using it. The proposed contribution is the discovery of better diagnostic and repair rules under unfamiliar conditions, rather than the routine execution of a rule once it is known. My judgment is that this numerical work offers a stronger prospect of such a contribution than reliance on abstract proof production. I would act on that judgment now and test the comparison directly.

Much of the mathematics collected at Refacte is algebraic and structural. I continue to develop and explain it because its constructions are worth understanding. Their value does not depend on a human monopoly on discovery. The proposed direction carries the concern with explicit constructions into

sparse computation, where a useful approximation must explain both what it preserves and what it makes possible at a given cost.

2. Beyond verbal imitation

The Deligne-Drinfeld case reaches the heart of the structural objection. The conjecture identifies the rational Grothendieck-Teichmüller Lie algebra with a free graded Lie algebra on one generator in each odd degree 3, 5, 7, and so on. Willwacher attributes the missing dimension bound to OpenAI and reconstructs its proof. The crucial move is to work in characteristic two, rewrite the problem in new generators, and project away one generator. The defining equations permit reconstruction, so this apparent loss of information yields a bound on the entire solution space in every weight. This is a choice of mathematical representation with a proof that it retains what matters. It occurs inside the machinery of associators and Lie algebras, at the level where conceptual sophistication was supposed to provide insulation. [1]

Markarian's November note makes the case stronger. It interprets the projection geometrically through the double covering that sends z to z^2 and the filling of a puncture, and makes the reconstruction explicit. This geometric interpretation is Markarian's contribution. It develops the proof architecture attributed to OpenAI into further structural mathematics. Human explanation can now follow machine discovery in this way too. 'Humans supply structure, machines perform calculations' is no longer a dependable division of labor. [2]

Aletheia provides another kind of evidence. In Eigenweights for arithmetic Hirzebruch Proportionality, Tony Feng attributes the mathematical content to Aletheia and describes his role, apart from building the agent, as rewriting its output into a paper. The agent connected arithmetic eigenweights to symmetric-group representation theory through algebraic combinatorics and determined them for all classical groups. The accompanying report says the imported techniques were unfamiliar to the authors of the originating work; it also describes an independence-polynomial collaboration in which the agent supplied a broader roadmap and the human collaborators supplied the proofs. These are not claimed as landmark advances. They are direct evidence that useful cross-field organizing ideas are no longer exclusively human contributions. [3]

OpenAI's larger catalogue is evidence of breadth, with a different evidential status. The 4 November snapshot lists 721 manuscripts in 373 related families and approximately 42 percent of principal stated results formalized, from a broader evaluation posing approximately 4,000 problems. These are not 721 independent breakthroughs, and 42 percent is not a success rate. The repository calls the material raw research data at different verification stages. Its history records three withdrawals after a shared sign error and repairs to other manuscripts. An unexamined entry cannot carry the weight of an independently reconstructed proof. The catalogue should widen the investigation, not replace it. [4]

Modular representation theory supplies further examples of this breadth. OpenAI reports proofs of the blockwise Alperin weight conjecture for every finite group and prime, and of Donovan finiteness over fixed algebraically closed fields with bounded defect-group order. These results are reported in

preprints; I have not located an independent specialist validation of either whole proof. The distinction matters, but so does the choice of targets: central block-theoretic problems are now being attacked directly. The accompanying formalization records also 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 in homological algebra. These records do not establish that every surrounding conjecture is settled. They make a strategy of retreating into supposedly protected abstraction increasingly implausible. [5] [6] [7]

First Proof supplies a more bounded test. Thirty expert reviewers 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. The third batch reports 7/10 correct for GPT-6 Astra and 9/10 for ProofCouncil, with reported ten-problem batch costs of $36 and $1,888 respectively. Here, correct means an average correctness score of at least 0.5 under the report's rubric, which includes minor, repairable errors. Those figures exclude the cost of creating and reviewing the benchmark and do not price the whole research process. They make it untenable to assume that obtaining a substantial research proof must require a human researcher to construct it. Exposition and attribution remained weaker than correctness. [8]

The June and October reports also record improved results and lower reported batch costs: ProofCouncil rises from 7/10 to 9/10 correct, and the direct OpenAI entry from 5/10 to 7/10. These are fresh problem sets with changed testing conditions, not repeated measurements on a fixed exam. They document successive capabilities without supplying a calibrated rate of progress. [8]

TheoPhysEval adds a distinct kind of evidence: under an automated final-answer checker, GPT-6 Astra with Codex answered 49 of 50 problems correctly with web access and 46 without web search or the supplied background papers; computational tools remained available. These were unpublished expert calculations, not a test of inventing physical theories or resolving breakthrough open problems. The result is narrower than full proof review, but broader than memorized textbook fluency. Together with the proof results, this demonstrates machine capability on technically demanding, unfamiliar problems. [9]

3. Where does human expertise still matter?

A common response locates the human advantage one level above the last demonstration: first in proofs rather than calculations, then in concepts rather than proofs, then in taste and problem selection. These can be substantive activities. Naming another activity, however, does not establish that it preserves a substantial specialist advantage once strong AI is available. A research programme may still benefit from human selection and interpretation while requiring much less human proof construction. The relevant question is how much the specialist changes the attainable result once the machine contribution is available.

Preserving hypotheses remains a significant difficulty. Wang and collaborators test translations between mathematical framings and find that seven general-purpose models often fail to carry implicit

hypotheses into a more general setting. On their held-out benchmark, 60.6 percent of abstracting rewrites widen the asserted domain. This matters directly to structural mathematics: getting the objects right is not enough if a passage from one theory to another silently changes the theorem. Preservation of hypotheses is therefore worth testing separately. The study evaluates specified rewriting protocols and older model snapshots, not the October internal proof system; it neither sets a ceiling on improvement nor cancels the reconstructed research contributions. [10]

The distinction is between a valuable contribution and a protected domain. Human reconstruction of a machine proof remains mathematics. A human who detects a lost hypothesis may be indispensable to a project. Neither fact restores an exclusive human supply of the proof idea itself. The same standard must apply on the numerical side: pointing to a hard implementation problem is not yet evidence that a specialist outperforms a well-equipped agent on it.

Nor can graph theory inherit the protection that abstraction is losing. Das and Fox credit GPT-5.6 Sol with the Murty-Simon edge bound $\text{floor}(n^2/4)$ and the even-order equality case; the authors simplified the argument and supplied equality for odd order too. Oum's earlier exposition develops OpenAI's cycle-double-cover proof. Nagda, Raghavan, and Thakurta use AlphaEvolve to find combinatorial objects and reduction gadgets. Xu's cubic counterexample to the general Bilu-Linial signing conjecture and Lin-Zhou's revised spectral-signing work also report model contributions. The restricted Ramanujan-base question and efficient construction questions must be distinguished from the statements already addressed. [11] [12] [13] [14] [15]

The same change reaches theoretical PDE. In First Proof's Darcy inverse problem, ProofCouncil removed the logarithmic loss in the human upper bound, with all three reviewers finding the mathematics essentially correct. OpenAI's October local-smoothing release asserts the Euclidean wave estimate in every spatial dimension at least two, with accompanying formalization records and related restriction and Kakeya statements. The former is an expert-reviewed benchmark result; the latter remains an attributed research release, not an independent certification by this essay. Neither supports the assumption that sophisticated PDE analysis is inherently protected. [8] [16]

Crouzeix's conjecture brings the warning into operator theory and numerical analysis. Townsend and Greenbaum report that they and Crouzeix checked Jin's AI-assisted proof and believe it correct; the decisive theorem emerged from an approximately sixteen-hour model run. Turning such a theorem into an economical numerical method is a further task, but it cannot be reserved for humans merely by calling it implementation. [17]

Subtle algorithm design is exposed too. Alman and Vassilevska Williams credit Claude with an underlying algorithm that they understood, simplified, strengthened, and extended. For polynomially bounded integer inputs, their stated bounds include $O(n^{1.9992})$ for 3SUM and $O(n^{2.9995})$ for directed all-pairs shortest paths with polynomially bounded integer weights and no negative cycles. The exponents do not establish practical speedups. The contribution nevertheless rules out an easy fallback to 'machines cannot invent the subtle algorithm.' [18]

The next question is how a promising construction becomes a dependable numerical method. A sparse approximation can preserve the wrong quantity; a coarse representation can miss the error directions that matter; a cheap arithmetic step can break an identity used elsewhere in the algorithm. These are different failures, requiring different repairs. The proposed specialist contribution is to discover which explanation fits the evidence and turn it into a better construction. A list of interacting requirements does not establish that contribution; an informative diagnosis and a successful repair can.

SciPiVot supplies direct but conditional evidence about choosing a better next step in research. It reconstructs 248 decisions from 26 scientific projects. Fresh PhD students and postdocs, with AI assistance available, selected the supported next step in 73.4 percent of cases; the best model tested with access to the project record reached 63.5 percent. Both candidate plans were supplied, so these scores measure recognition of a better-supported correction, not generation of an original diagnosis. The cases include questioning an unjustifiably restricted physical model and an uninformative stiffness sweep. These are selected historical corrections, not a random sample of research decisions, and the study supplies no matched-cost comparison or sparse-solver subgroup. It supports investigating researcher-led diagnosis without establishing a durable premium in this particular field. [61]

Abstract researchers also choose problems, invent definitions, connect theories, and reconcile demands across a larger construction. A fair comparison of complete projects must include that work. The existence of interacting constraints on both sides does not establish equal rates of substitution. The cases above directly undermine abstract sophistication as a source of protection, even though they do not measure the relative productivity of complete abstract and numerical projects.

There is a constructive connection between the two directions. A structural proof can supply a smaller representation and establish what it preserves. In the Deligne-Drinfeld argument, the projection is lossless on the constrained solution space because reconstruction is possible. Sparse elimination begins with an exact energy identity, then permits controlled approximation to save work. The question becomes how much information can be discarded while the retained structure still controls the required answer. The common move is to choose a representation around the information the result actually needs; sparse computation adds the demand that retaining it be affordable. [1] [19] [58]

My work on Taiko product structures gives an exact example of this practice. A proposed group-ring construction becomes a finite incidence problem. Coarse size counts leave apparently feasible profiles; pressure and intersection constraints expose the obstruction. The numerical programme carries this concern with retained information into problems where approximation and computational cost must also be controlled. [80]

I would pursue the advantage in discovering cheap, reusable diagnostic rules. For an unfamiliar matrix or a short sequence after a change, little subsequent work may remain to recover the cost of a broad search or an unnecessary rebuild; PCGBandit's costly early trials give a concrete example. [62] Short workloads constrain the deployed diagnostic, not the research effort that discovers it. A costly investigation can repay its cost across many later workloads, whether a specialist or an agent produces the rule. Which probe distinguishes a missing coarse mode from ineffective relaxation? Which graph change predicts that sampled elimination will become unreliable? Which discrepancy calls for a

different arithmetic formulation? These opportunities recur as operators, retained representations, arithmetic, and workloads change: a previously adequate approximation can fail for a different reason, while a previously worthwhile repair can cease to repay its cost. My bet is that expertise linking graph structure, matrix estimates, and numerical error helps identify which cheap test will distinguish the competing explanations. The proposed contribution is to discover such tests under unfamiliar conditions and make them reusable, so that later solves need less search. Its economic comparison must charge specialist-led and autonomous discovery over the same family of workloads, including the cost of producing and transferring each rule.

The Simons workshop on linear systems and eigenvalue problems places this programme at an established intersection of theoretical computer science and numerical analysis. Its agenda joins parameterized sparse systems, graph-Laplacian techniques, spectral approximation, and iterative convergence. The open questions include descriptive bounds for nonnormal solvers and the gap between an effective method and a method with useful guarantees. They identify a substantial mathematical agenda on which to test the proposed specialist contribution. Their difficulty alone does not establish that contribution. [75]

4. Finite structure under computation

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 and its energy are:

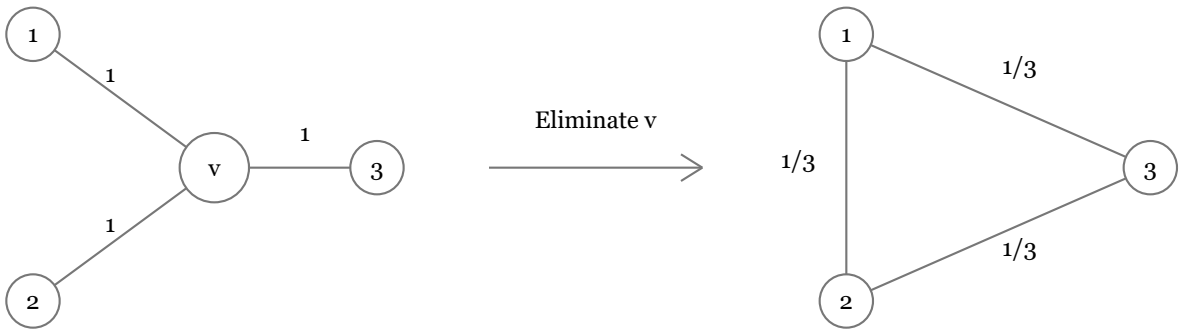
$$L = B^{\mathrm{T}}WB,\ x^{\mathrm{T}}Lx = \Sigma_{\text{edges } \{i,j\}}\ w_{ij}(x_i - x_j)^2.$$

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. These are concrete matrix classes, not a claim about every system produced by a PDE. [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. More generally, the Schur complement of a star with incident-weight vector a is:

$$S = \operatorname{diag}(a) - aa^{\mathrm{T}}/s,\ s = \Sigma_i a_i.$$

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, but repeated elimination creates fill: previously unconnected variables become coupled. 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 is a foundational construction: the importance of an edge depends on the network around it, not only on its own weight. [19]

For two symmetric positive-definite matrices A and M, the useful comparison is quantitative. 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. If M is inexpensive to solve with, it can be a useful preconditioner for A. A good bound with an expensive M may be useless computationally; an inexpensive M with a poor bound may leave a slow iteration. For Laplacians the same comparison is made on the orthogonal complement of their common kernel. [52]

Multigrid approaches the loss of information in another way. A cheap relaxation reduces some error components; a coarse correction represents those that remain. Algebraically smooth error means error that the chosen relaxation struggles to remove, not necessarily a slowly varying function of physical coordinates. 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]

Accuracy has a similar trap. Consider this exact two-variable system, where the second variable is weakly weighted:

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

The relative residual in the infinity norm is 10^{-8}, but the relative forward error in the same norm is 1: the proposed answer loses the entire second component. This elementary calculation does not discredit residual tests. It shows why they must be interpreted through conditioning and a specified error measure. LAPACK's error analysis makes that distinction explicit. [21]

These are finite mathematical problems. Their treatment requires spectral inequalities, approximation, probability, and careful arithmetic, but does not require beginning with a theory of continuous function spaces. The research question is what information a sparse construction must retain to produce the required answer reliably and economically.

5. Diagnosing and repairing sparse representations

The programme I would choose is the diagnosis and repair of sparse representations when the conditions that made them effective have changed. Begin with Laplacians and positive-definite SDDM matrices. A change in edge weights or connectivity can make an elimination rule expensive or leave a coarse space unable to represent an important mode. Acquire enough information to choose a correction without spending more on diagnosis and reconstruction than the remaining workload can repay. Approximate elimination and algebraic multigrid provide two concrete routes: improve the retained interactions or the retained directions, while controlling spectral distortion, fill, memory, and total solution cost.

The choices are often explicitly combinatorial: select an elimination order, sample a spanning tree, choose coarse vertices, group vertices into aggregates, or prescribe a sparse pattern of connections. Matrix analysis supplies the criterion those choices must meet. A connected sample or an economical partition is useful only if the resulting operator retains the required quantitative control. The graph construction and the numerical guarantee are parts of the same problem. [20] [47]

Approximate elimination: local rules and accumulated error

Gao, Kyng, and Spielman's AC(k) work attaches a precise mathematical question to an implemented solver. Its approximate-Cholesky construction sparsifies the interactions created during elimination while maintaining connected sampled support among a removed variable's neighbors. The parameter controls the sampling effort. The experiments show favorable robustness and performance on broad SDDM workloads, alongside cases where established competitors are faster. These are existing effective solvers, not evidence that the matrix class has remained unsolved. [26]

The economical AC(2) construction still lacks the global preconditioning guarantee the authors seek. Section 6 conjectures a relative condition number $O(\log^2 m)$, possibly $O(\log m)$, with high probability, where m is the input edge count. The construction has proved local and structural properties, but those properties do not supply the conjectured spectral bound. The question is how random distortions accumulate through a sequence of changing Schur complements. [26]

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. This does not prove the AC(2) conjecture: it establishes a different construction under different assumptions. It does make error allocation a precise research problem: where can a looser approximation save work without exhausting the global accuracy budget? [58]

A useful project could prove a restricted guarantee, find an obstruction to a proposed argument, or identify which graph structures require more sampling. A diagnosis should distinguish a deficient sampling rule from a poor ordering or a tolerance that is unnecessarily strict. Its repair might increase work locally while reducing fill or iterations elsewhere. The proof problem and the computational problem inform one another. A standalone proof of the AC(2) conjecture has no special claim to human protection; agents should be allowed to attack it directly.

Why the obvious invariants are insufficient

An elementary example isolates the spectral obligation. Let L be the Laplacian of a triangle whose three edges have weight one. Choose an edge uniformly at random, give it weight 3 - 2 ε, and give the other two edges weight ε, with 0 < ε < 1. Every sampled graph remains connected and has the same constant-vector kernel. Each edge's expected weight is one, so the sampled Laplacian L_s is unbiased:

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

To check the second claim, suppose the heavy edge joins vertices 1 and 2. The vectors (1,1,-2) and (1,-1,0) are orthogonal eigenvectors of both L and L_s. Their generalized eigenvalues for L_s relative to L are ε and 2 - ε, respectively, and they span the mean-zero subspace. The relative condition number is therefore (2 - ε)/ε and becomes arbitrarily large. Connectivity and unbiasedness survive while spectral quality deteriorates. This is an original illustrative calculation, not a counterexample to AC(2), VAC, or CAST.

Recent constructions already go well beyond those two elementary properties. Baumann, Kyng, and Zöcklein's VAC uses spanning-tree sampling to replace the clique arising at an elimination step, preserving the desired marginals and connectivity with efficient sequential and parallel sampling. Chaitanya, Musco, and Gionis's CAST develops a related tree construction, a local optimality characterization, and a tunable bound on the normalized local second moment. CAST explicitly separates these local guarantees from a worst-case spectral theorem for the complete factorization. Unbiasedness and connected support alone are therefore not new contributions. A new sampler must improve the construction, its cost, or a quantitative guarantee; understanding and controlling accumulated distortion is a particularly direct target. [47] [48]

CAST supplies a concrete example. Compared with AC, CAST-1 changes the local sampling rule while retaining the ordering, factor representation, and PCG path. Its mean per-system speedup ratios are 1.17 on 17 SuiteSparse matrices and 1.18 on 128 Chimera-IPM systems, with each factor reused for 250

right-hand sides. The relative performance of CAST-2 and AC(2) reverses across graph families. The authors connect that reversal to whether additional edges propagate through later elimination steps. Their CAST-1 elimination profile helps explain the regimes on the tested corpus, but obtaining it requires a first factorization, and they do not test it as an adaptive selection rule. The opportunity is to obtain the decisive information more cheaply: predict when additional sampling will improve the preconditioner enough to repay its construction and fill costs. A proposed rule must beat fixed settings and existing automatic selectors on unseen families after its own observation and setup costs are charged to the declared workload. [48]

A second route: retain the difficult modes on a coarse level

For a symmetric positive-definite matrix A, let the columns of a full-column-rank matrix P describe coarse corrections. The standard coarse matrix and exact coarse correction are:

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

This is finite-dimensional linear algebra. The hard design question is which columns and entries P needs. Global spectral information can identify useful coarse directions, but computing it may already be expensive; a dense P can also make P^T A P costly to form and solve. Local matrix information and graph aggregation offer cheaper approximations. Their adequacy depends on the relaxation they accompany and the modes they retain. A two-level result with an exact coarse solve is not automatically a theorem about a recursively applied multilevel implementation. [20]

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; it does not make those choices sparse or inexpensive. 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]

A bounded project can fix one matrix family and one baseline, then ask whether a smaller support, cheaper local eigensolve, or different partition preserves the relevant convergence behavior at lower total cost. If it cannot, a small counterexample may identify the missing mode and suggest a repair. These mechanisms already have automatic implementations: adaptive smoothed aggregation generates missing near-nullspace candidates, compatible relaxation helps assess coarse choices, and hierarchy-reuse methods update changing systems without repeating every setup operation. A new result must improve on these methods, not simply rename what they already do. [54] [76] [78] [79]

A changed graph can defeat an unchanged coarse space

Consider the four-cycle 1-2-3-4-1 with A = L + ε I, where 0 < ε < 1. Initially give edges 12 and 34 weight one, and edges 23 and 41 weight ε. Let P contain the indicator vectors of the two strong pairs, {1,2} and {3,4}. Now interchange the strong and weak weights. The new matrix has the same sparsity pattern, diagonal, and eigenvalues as before: a relabeling relates the two graphs. But the old coarse space has become misaligned with the strong connections. For v = $(1,-1,-1,1)^T$, the changed matrix satisfies:

$$Av = 3\varepsilon v,\ P^T A v = 0.$$

The exact coarse correction therefore leaves v untouched. The diagonal of A is (1 + 2 ε) I, so a weighted Jacobi step multiplies this error by 1 - 3 ω ε/(1 + 2 ε). For a fixed admissible relaxation weight, that factor approaches one as ε decreases. Changing the aggregates to {1,4} and {2,3} puts v in the coarse space, and the exact coarse correction removes it. This is an elementary illustrative calculation, not a new multigrid result or a test of AI capability.

The example identifies the missing information. Matrix size, sparsity, diagonal, and even the spectrum alone do not describe the relation between the current operator and the retained coarse directions. On a large graph, the question is how cheaply that mismatch can be detected and repaired. A useful rule could choose among reuse, different relaxation, local enrichment, and rebuilding. Automatic competitors must be allowed the same choices. The case matters especially when only one or a few solves remain, leaving little opportunity to recover the cost of unsuccessful experiments.

From matrix solvers to PDE applications

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. If a result additionally claims accuracy for a continuous PDE, it 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. Finite element exterior calculus explains a broad route from continuous structure to stable discretization, but that route need not organize the finite solver research itself. [22] [23] [24] [25] [43]

What specialist judgment has to accomplish

• Separate failure mechanisms. Determine whether a slow solve comes from scaling, a poorly represented mode, insufficient sampling, an ineffective ordering, or expensive hierarchy construction. Design a test that distinguishes those explanations.

• Connect structure to a quantitative claim. Turn a graph pattern or a matrix decomposition into an energy bound, a convergence estimate, or a prediction that can be checked on new instances.

• Construct revealing examples. Vary bottlenecks, degree distributions, weight contrasts, and size to expose the mechanism. A small example that defeats a proposed explanation can be more informative than another large timing table.

• Account for the complete calculation. Establish whether a repair helps after setup, factor or hierarchy size, memory, repeated solves, checking, and failures have been included.

Turn a successful diagnosis into code, a proved bound, or a reproducible decision procedure that other users and agents can apply. Test its predictions on fresh cases and its cost against the strongest available alternatives. The specialist contribution is the discovery of a better rule under new conditions. Once that rule is expressed, routine use of it should become automatic. A useful new inequality can be a substantial contribution; its assumptions must match the construction it controls.

6. Automation already reaches numerical research

A serious case for this research direction has to start from the strongest competition. 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. It rules out 'doing the implementation' as a sufficient human strategy. [28]

LinearSolveBench brings the challenge closer to the proposed specialty. Its released C/HYPRE solvers were selected from 640 candidates per model using 96 magnetic-diffusion development cases. The best reported result is about a 1.85-fold speedup over GMRES(50) with BoomerAMG in the original measurements, with all 96 cases passing the specified checks. The timing includes solver creation, setup, and solve, but excludes discovery, compilation, and independent verification. These are development-set selections, not measurements on an unseen workload; the released verification record reports a fresh correctness check without a fresh Linux performance comparison. Even with those limits, they demonstrate actual sparse-solver optimization. A human advantage cannot be inferred merely from working in this area. [32]

FrontierCS 2 brings that competition into recent research code. Its current 18-system, 48-task record 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. The choice of subject offers no protection from this competition. [69]

The SVG-EAR case reaches the proposed diagnostic role directly. Astra introduced a probe that periodically compares sparse attention with exact attention on a small sample of queries, then chooses a per-head budget from the measured error, with a dense fallback when no tested budget passes. It also supplied a variance correction absent from the reference method. The final implementation reduced mean runtime by about 18 percent, improved all three mean fidelity measures, and passed 43 of 48 hidden checks. The diagnostic and correction were not separately ablated, and a passed probe did not ensure every final fidelity check passed. Its five fidelity failures also occurred in a separate exact-attention implementation, so they cannot simply be assigned to sparse approximation. The agent nevertheless performed the sequence of measurement, diagnosis, and adaptive repair. A specialist advantage must improve on that capability, not assume it is missing. [70]

Numerical decisions are being automated at several scales. NextEval reports effective agent-directed selection among numerical tools on 92 optimization problems of dimension at most five; its advantage concerns objective-evaluation efficiency, excluding language-model time and cost. DecompGym requires prescribed Benders or column-generation structure as well as numerically acceptable answers. Astra and GPT-6.1 Sol pass all 81 instances on their first attempts. Within these settings, agents already make adaptive choices and coordinate numerical subproblems. [71] [72]

Mathematical control of implementation cost is also exposed. 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; it asserts no floating-point stability. 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. They prevent treating finite-precision analysis itself as a human refuge. [57]

Nor is there a general opposition between fast algebra and stability. 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. The target here is a useful sparse method at the required accuracy and cost, rather than stability in isolation. [60]

Other numerical research environments show both capability and limits. GAV supplies iterative experimentation; AutoPDEBench reports improvements over selected neural-solver architectures; FEniCSx and PDEAgent-Bench examine the translation of continuous problems into numerical code. Their tasks, models, and evaluation rules differ. Formulation failures cannot establish a specialist advantage in a matrix-only contest. The relevant question is how much a specialist improves a method when its input problem and available numerical tools are already clear. [29] [31] [33] [34]

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. Targeted numerical knowledge can therefore improve a search beyond simply requesting another attempt, and the diagnosis supplying it is already automated. 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; it does not measure a human specialist premium. [30]

SciCode-Verified supplies a related comparison after correcting defective specifications and tests. On its 64 main problems, expert-written background raises 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. These are single-run results under the paper's two-environment OR grading rule. They measure the value of supplied background, not a live specialist's advantage or the agent's inability to produce the same information. The background sometimes supplies a derivation or prevents substitution of a familiar formulation for the stated one. Knowledge can matter greatly while becoming reusable once expressed. [35]

Removing finite-element assembly and library conventions lowers the entry cost for both people and agents. The matrix interface makes the comparison cleaner; it does not make the work harder to automate. AlphaEvolve supports multiple evaluation scores and evaluations lasting hours on accelerators. An agent can exploit expensive feedback too. The question is whether a better diagnostic saves enough of that expense to improve the complete investigation. [44]

Proof automation reaches the finite guarantees, and agents can improve their own search procedures. AIDE2 reports seven successive improvements during an eight-day autonomous run, with gains transferring to four held-out benchmarks and its final agent matching or exceeding the human-engineered production-agent baseline. That comparison concerns agents, not live sparse-solver specialists. It rules out reserving research-process design for humans. The specialist must improve what happens next even when the agent can revise its own methods. [55]

Long investigations can still lose the connection between an appealing result and sufficient evidence. In Into the Unknown, none of 135 agent submissions on fifteen open scientific follow-ups meets every strict requirement. Under the study's automated grading, Astra nevertheless passes 97.77 percent of individual integrity checks, showing how an all-or-nothing headline can hide substantial progress. The experiment supplies no timed human comparison. It motivates diagnosis of unsupported claims and incomplete tests; it does not establish a frontier plateau. [73]

Adaptive reuse by either people or agents can overfit a validation set. Dwork and collaborators analyze this problem and give methods for preserving validity under reuse. Separate development feedback from final evaluation and distinguish performance on a tested workload, transfer to unseen families, and a uniform mathematical bound. More successful search on the same cases does not establish the latter two. [45]

The matrix interface removes an easy source of ambiguity, but not every evaluation error. A residual can be measured against the wrong operator; nullspace handling can alter the problem; failures can

disappear from an average; an eigenvalue estimate can be presented as a certified bound. The specialist must demonstrate a useful contribution after the task and these evaluation rules are shared. A proof kernel likewise checks the encoded proposition, so a purported theorem still has to match the actual algorithm and its assumptions. [10] [21]

7. From a promising method to a reliable solver

Once a candidate method exists, the question becomes what the implementation actually preserves. A learned representation may discard an important mode; rounding may invalidate an identity; a small residual may fail to control the required error. The task is to connect representation, arithmetic, and checking so that the resulting solver meets its accuracy requirement at an acceptable total cost.

Learned components and the operator they produce

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]

Liu, Au Yeung, Chung, and See learn approximations to local spectral coarse spaces for two-level Schwarz preconditioning of heterogeneous Darcy systems. Their training objective measures a coefficient-weighted discrepancy between subspaces, and their large two- and three-dimensional tests report reduced setup and total solution costs. They also propose a bound connecting this discrepancy to the preconditioned condition number. That guarantee needs clarification: the printed local preconditioner can be nonsymmetric despite the stated symmetry claim, and the supporting argument leaves a projection metric unclear. These observations concern the written formulation; they do not establish a failure of the implementation or the reported experiments. The research question remains useful: which approximation criterion can be checked economically and shown to preserve the required convergence? [36]

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;

unrestricted corrections lack a general convergence guarantee. 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]

Adaptation within a changing calculation is already a practical method. PCGBandit chooses among 33 preconditioner configurations during one OpenFOAM run without an offline training corpus. It counts learning and backstop costs. Its detailed analysis identifies costly early trials in Shercliff electric-potential solves, before enough evidence accumulates. The 33-configuration action set is fixed; the method does not construct a new coarse representation. This isolates a useful problem: better early diagnosis when the workload is too short to amortize unsuccessful exploration. [62] He and collaborators also learn BoomerAMG hierarchy choices and within-solve relaxation online. Their 5,000-system sequences report total-runtime reductions of 36.1-43.3 percent for diffusion and 19.8-31.0 percent for diffusion-advection against the reference default. Repeated use can pay for information that is too expensive to acquire for one solve. [39]

These methods suggest a promising approach: retain operations whose role is understood, and learn decisions that are expensive to make exactly. Retaining classical components does not, by itself, certify the learned ones. If a network predicts a coarse space, the question is whether the approximation is good enough and how to tell when it is not. If it chooses sparsification or relaxation parameters, the question is which structural properties survive the choice. A fallback to a trusted calculation helps only if the detection rule and fallback together provide the required protection at an acceptable cost. A confidence score is a prediction to evaluate, not a certificate merely because it is called one.

Precision can change the mathematical construction

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; their 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 still contain precision-sensitive cases and

defer hardware-specific timing. The example captures the proposed research activity: discover which relation has failed, then repair the formulation rather than merely run more iterations. It does not establish that agents cannot discover the same repair. [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 do not decide runtime: 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 an author-reported maximum 1.6-fold speedup. The research target is a better, transferable prediction of when these precision choices repay their cost. [64] [65]

A finite error bound that can be used in a computation

Suppose A and M are symmetric positive definite and a proved lower comparison gives $\alpha M \le A$ for $\alpha > 0$. For an approximate solution y to $A x = b$, form the residual $r = b - A y$. In exact arithmetic the squared energy error has the following bound:

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

This follows by reversing the positive-definite matrix inequality under inversion. If M is inexpensive to solve with and α is useful, the bound turns a residual into an energy-error certificate. For Laplacians it applies on the subspace where the common kernel has been removed. Support theory is one established algebraic framework for bounding such comparisons. [52]

To use this bound, the lower comparison α must hold for the actual construction; a few spectral probes do not certify it. If M is solved only approximately, that error needs its own control. A rigorous floating-point certificate must also account for rounding in the residual and quadratic form. The cost of certification must be justified by the required assurance and the declared workload. A sharper bound or a cheaper valid check can itself be a research result. [21]

A PDE application introduces an additional error between the continuous solution and the exact solution of the discrete system. 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]

Nonnormal matrices: a theorem still needs a useful diagnostic

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 bound:

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

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): the projected numerical range is not a certified outer enclosure. An ordinary largest Ritz value of a Hermitian support matrix is a lower estimate, whereas an outer support line needs an upper bound. Chen and Urschel quantify Krylov approximation of numerical ranges and its limitations; their analysis does not supply a universal floating-point certificate. Diagnosing when a cheap spectral picture is inadequate is a precise task, not a license to relabel an estimate as a 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]

Count the full cost and the failed cases

Performance claims need a stated workload and an account of failures. Chillón and collaborators' learned smoother makes the point sharply. On fine unstructured three-dimensional canonical-flow tests, the fraction of successful solves averages 67.7 percent across ten trained models. A separate evaluation selects the best complete-data model and tests it on 1,000 AirfRANS systems: it converges on 99.9 percent of those systems, uses about 27 percent fewer V-cycles, and gives only about a 4 percent solve-time speedup. The baseline is relaxed Jacobi, with relative tolerance 10^{-4} and a 500-cycle cap. [40]

The network predicts coefficients during setup; applying the resulting sparse smoother costs more per cycle than diagonal Jacobi, absorbing much of the iteration saving. The two evaluations use different tests and aggregations. Together they show why iteration savings, model selection, and the cost of applying the learned operator must all be reported. They do not establish a general limit on learned smoothers. [40]

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 test a learned multilevel preconditioner on 867 real non-SPD SuiteSparse matrices. Median per-matrix setup/training time is 135.77 seconds against 0.18 for the tested AMG; median solve time is 8.08 against 1.31 seconds. These separate medians cannot be added to obtain a paired comparison of total time. The paper also reports useful cross-matrix transfer. Its convergence win shares are conditional on compared methods producing valid runs. Construction reliability, solution accuracy, and convergence speed are different outcomes, and a method intended for one solve faces different economics from a reusable trained construction. [42]

For a fixed workload of N comparable queries, include setup, the initial solve attempt, checking, and expected recovery in the time model. Recovery includes any additional failed attempts and the specified fallback. Fix timeout and fallback rules before estimating these costs, and report the fraction of queries that remain unresolved; an average over successful runs alone omits part of the workload:

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

On this expected-time model, if the new method has extra setup cost ΔS and saves $d > 0$ per query after recurring costs are counted, expected break-even occurs at the smallest integer at least $\Delta S/d$. If setup is already cheaper and $d > 0$, it pays off from the first query. Coefficient changes can force rebuilding, shared failures can change recovery costs, and correlated failures can increase the risk of a large overrun even when expectations remain additive. Report the success rate and the spread of costs alongside the mean. Hardware and the number of required outputs also affect the comparison. Discovery and deployment have different economics: a costly research campaign can produce an inexpensive method used millions of times. Counting only model tokens, iteration counts, or the fastest successful solve answers a narrower question.

The checker must evaluate the supplied problem independently of the submitted answer. For a matrix task, fix the operator, right-hand side, gauge, and required components before running the candidate; reject nonfinite outputs on those components. The FEniCSx scorer illustrates why this matters: its joint NaN-exclusion mask can omit points where a prediction is NaN, although missing reference values can legitimately indicate locations outside a physical domain. The lesson is to define the checked domain independently and report coverage, not to treat every missing reference value as a defect. [33]

A final performance table is convincing only when its rows answer the same question. Setup, checking, rebuilding after a coefficient change, and recovery from failure can determine which method is preferable. Accounting for them is part of the mathematical claim, not a secondary implementation detail.

8. Testing the proposed advantage

The central test is what a sparse-computation specialist adds to the strongest available agent. Give an autonomous agent, a specialist using that agent, and a mathematically capable generalist using it comparable total budgets. Include human time, model use, computation, checking, and failed attempts. Use several participants and repeated seeds; retain specialist-alone work as an additional baseline. Select evaluation families and participant criteria before inspecting comparative runs, and keep final assessment independent of the competing teams. Report deliberately adverse stress tests separately from performance on the declared application workload. Specify what improvement would justify the additional specialist cost, and report uncertainty. An alternative algorithm counts as success if it meets the same specification.

Every group receives the same matrix interface: the input matrices and right-hand sides, nullspace or grounding convention, accuracy criterion, baseline source, development cases, papers, and diagnostic playbook. All may change representations, prove lemmas, construct adversarial graphs, or replace the expected algorithm. Record model and harness versions, compute, human interventions, and failed attempts. Report equal-time and equal-cost comparisons. This measures the productivity of the specified resources; it does not equalize years of human or model training.

Human participation alone does not establish independent judgment. In a randomized study, fifty professional analysts using agents followed deliberately assigned agent preferences on 86 percent of four analytical choices. The alternatives were defensible, so this is influence evidence rather than an accuracy ranking. Record a participant's initial diagnosis before revealing the agent's recommendation, retain the proposed test, and measure whether the intervention improves the final evidence. An approving signature is not the contribution being claimed. [74]

Allow every group to improve its search procedure, diagnostic tools, and evaluation harness within the same budget. Specify in advance how decision points will be selected for intervention tests, and retain unsuccessful interventions. At a selected checkpoint, freeze the solver, evidence, and agent context before introducing the proposed diagnosis or repair. Compare repeated continuations with and without that intervention under equal remaining compute budgets, charging its production cost to the group that supplied it and allowing the controls to generate their own alternatives. Record the diagnosis and its prediction before opening the reserved cases. These continuations estimate what the intervention changes at those particular points. The full-project comparison, including the cost of reaching those states and producing the intervention, tests the broader specialist advantage.

After each completed round, add useful procedures, counterexamples, and diagnostic rules to a shared archive before testing fresh cases. A gain that disappears once those artifacts are supplied establishes the value of producing them, but not a continuing need for specialist intervention on later tasks. A persistent specialist premium must survive this transfer of knowledge. Distinguish the cost of discovering a useful intervention from the gains after it becomes available to everyone; keep each round's final cases unseen until that round ends.

Repeated success against a fixed agent does not establish a lasting human advantage. Repeat the comparison as autonomous systems improve, giving every group access to the updated systems. Keep the declared application scope and success criteria fixed, and select fresh cases using a rule established before comparative results are known. The proposed specialist premium must survive these improvements in the competition as well as the sharing of earlier discoveries. If it disappears, the numerical programme remains valuable, but its claim to be a human moat has failed on that workload.

Assess two kinds of work separately: improving a method under a mature shared evaluation procedure, and identifying a deficiency in that procedure and constructing a valid, affordable extension. Fix the mathematical problem class, accuracy requirements, budgets, and acceptance rules before development. All groups may revise their development tests and propose new representations. Independent final assessment must determine whether a correction meets the original mathematical requirement. Repairing a checker that accepted an inaccurate answer enforces the existing task; adding a new application requirement creates a further task and must be reported separately. Charge one-time construction and recurring checking costs over the same declared workload.

Different claims require different evidence; success in one row does not establish the others.

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.

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: 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]

A claimed high-probability theorem must state its input assumptions, randomization, and failure-probability dependence. Using the bound to predict practical performance also requires the constants needed for that prediction. One unlucky seed does not disprove such a theorem, and many successful runs do not prove it. A counterexample to an intermediate argument is useful without being a counterexample to the algorithm. Small-instance eigenvalue calculations can test a proposed

explanation; large-instance spectral estimates must be labeled as estimates. Distinguish these from a validated proof of the full bound.

A first pilot can use shifted grid Laplacians $A(w) = L(w) + \mu I$, with positive edge weights and fixed $\mu > 0$. Build an initial hierarchy using one fixed symmetric multigrid implementation. After a weight change, choose between retaining its transfer operators while recomputing Galerkin operators and smoothers, and rebuilding with the same setup procedure. Solve four right-hand sides for that changed matrix. Reserve coefficient-change regimes and larger sizes for within-family testing. Compare the diagnostic with always-reuse, always-rebuild, and a specified existing adaptive or reuse policy. Measure post-change total time, including probes, updated coarse factorizations, setup, solves, checking, failures, and fallback. Declare the accuracy, reliability, and worthwhile improvement margin before evaluation. A successful diagnostic establishes a useful rule on this workload; the participant comparisons and later shared-archive rounds test who can discover such rules and whether a specialist premium persists.

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]

Independent checks should target different failure modes. Two implementations can share a reference bug; two models can repeat the same proof error. Recompute the residual against the original operator, check the declared kernel and positivity assumptions, examine small cases independently, and review the proof against the implemented algorithm. Record what supports a theorem, a program, a numerical output, or a PDE application. Those are different objects of verification.

Keep final evaluation data separate from development. An agent allowed to optimize repeatedly on the final family is reporting tuning performance even if its experiments are sophisticated. At the same time, preregistration should not forbid following an unexpected valid discovery. Fix budgets, primary claims, holdouts, and acceptance rules; report exploratory results as exploratory. Release all runs and enough artifacts to inspect the interventions. Logs can establish who proposed a test, supplied a correction, or changed code. They need not pretend to reveal every private source of an idea or assign a simplistic percentage of credit. [45]

Repeat the comparison on complete abstract projects, allowing definition and conjecture formation, imported theory, counterexamples, proof construction, and independent assessment. Otherwise the experiment establishes a result about a solver task, not a measured ranking of fields. The strategic judgment is broader than any one trial and must be revised as the evidence changes.

9. Where to concentrate research effort

For someone choosing where to invest years of mathematical training, I regard the strategy of seeking protection in deeper abstraction as spent. Algebraic geometry, representation theory, abstract algebra, and theoretical PDE offer no exemption merely because their objects are remote or their arguments sophisticated. The demonstrated machine contributions already reach the conceptual level invoked to justify that exemption. The continued existence of difficult open problems does not restore it.

I would redirect effort to the design and repair of sparse numerical representations, where spectral graph algorithms, matrix analysis, and numerical analysis meet. Begin with a strong elimination or multilevel method and a declared matrix family. Identify what a change in weights, connectivity, or precision has made unreliable; construct a diagnostic that predicts fresh cases; and turn it into an economical repair. Count the probes, failed attempts, rebuilding, and checks as part of the result. This puts the research effort into discovering better transferable rules, with existing automatic procedures and agent-discovered alternatives as serious competitors.

This is the human advantage I would pursue: the ability to discover what a cheap computational representation fails to preserve, and to repair it without sacrificing the saving that justified approximation. The evidence supports this technical programme and gives reason to test researcher-led diagnosis; a lasting sparse-specialist advantage remains to be demonstrated. A human earns that advantage by improving the result beyond what the strongest agent and a capable generalist can obtain with the same methods and resources. Share each successful diagnosis, then test what the specialist contributes next. I would concentrate research effort there now. The old refuge in abstraction is over; the next research direction has to demonstrate its value.

Notes and sources

Sources and versions checked on 4 November 2026. These notes distinguish reported experiments, specialist assessments, and the record calculations performed for this essay.

[1] Thomas Willwacher. *The Deligne-Drinfeld conjecture after OpenAI*. 2026. arXiv v4, 28 October 2026.
Primary source
Specialist reconstruction and simplification of the proof; cited as evidence of substantive mathematical engagement, not as certification of the entire OpenAI catalogue.

[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.

[4] OpenAI. *Mathematics research repository and release history*. 2026. Pinned snapshot, checked 4 November 2026.
Primary source · Release history · Current repository
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.

[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 v1, 9 October 2026.
Primary source · Abstract and full author list
Sections III and IV.1 describe the automated final-answer evaluation. 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.

[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
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.

[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. The diagonal two-variable example in this essay is an original elementary illustration, not a reported LAPACK experiment.

[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
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 is mathematical support for the proposed specialism, not an experiment establishing a human advantage.

[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.

[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.

[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.

[29] Muheng Li and Wenlong Mou. *GAV: A scientific experimentation environment for PDE agents*. 2026. arXiv v1, 21 October 2026.
Primary source
Sections 3-4 and Appendix C.1 distinguish model/task configurations and repeated exploration/refinement. Navier-Stokes uses Sol; kinetic experiments include both Sol and Astra. Table 2 compares final Sol/Astra scores on the generated kinetic tests; Appendix C.1 gives six three-hour development rounds per agent. Appendix D.1 reports independent-test transfer failures, including a worse final than first retained Astra program. Model settings and effort differ; this is not an equal-token-cost comparison.

[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.

[31] Ruoyan Li, Wei Wang, and Yizhou Sun. *AutoPDEBench: Benchmarking LLM Auto-Research for Neural PDE Solver Design*. 2026. arXiv v1, 26 September 2026.
Primary source · Versioned full PDF
Section 4.2, PDF pages 8-9/Figure 4 compares agent designs with NeuralCFL (HighSpeedFlow), CAPE (MixedParam), and UPT (LagInletOutlet). These are three selected architecture comparisons, not a timed human-research contest or all 25 datasets. The main agentic baseline uses GPT-5.5.

[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.

[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.

[34] Zhen Hang et al. *PDEAgent-Bench: A Multi-Metric, Multi-Library Benchmark for PDE Solver Generation*. 2026. arXiv v1, 10 May 2026.
Primary source
Section 4.4 and Appendix M give the executable-template ablation, preserving numerical decisions as model responsibilities. Gemini 3.1 Pro gains 2.2, 2.9 and 3.5 percentage points on the three libraries. This is an older-model controlled ablation, not a comparison with timed human experts or current frontier systems.

[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
Section 3, Table 2: expert-background ablations on 64 corrected main problems, single runs with two-environment OR grading. They measure supplied-context effects, not human-versus-agent productivity. Table 1 separately gives GPT-5.6 Sol 92.2 percent with background; benchmark repair is not longitudinal model improvement.

[36] Yucheng Liu, Tak Shing Au Yeung, Eric Chung, and Simon See. *Graph-Based Meshfree Multi-scale Coarse Space Approximation for Two-Level Schwarz Methods*. 2026. arXiv v1, 27 March 2026.
Primary source · Formulation check
Section 4.2 tests Cartesian TPFA Darcy systems, including 2048^2 and 256^3 grids. The proposed condition-number bound is Theorem 3.4, equation (9). Equation (2) applies partition weights on one side of each local inverse and need not be symmetric; adding the symmetric coarse correction does not remove this issue. Lemma 3.3 also needs its projection metric specified: the Euclidean interpretation admits a counterexample, while a weighted interpretation can repair that objection. The linked formulation check gives exact examples. These findings concern the printed formulation, not a demonstrated failure of the code or experiments. The guarantee requires clarification. The arXiv history still listed only v1 on 4 November 2026.

[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 v1, 10 February 2026.
Primary source
Section 5 and Appendix E use positive-definite test operators after stated transformations. Comparisons with Chebyshev use the same spectral probes and include wins and losses. Exact degree matching is not asserted here.

[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.

[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.

[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.

[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.

[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.

[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.

[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.

[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.

[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
Sections 3-4 give local guarantees, not a global factorization bound. Section 5 and Table 1 report paired 250-right-hand-side workloads. The proposed explanation of configuration preference is not tested as an adaptive selection rule.

[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 proposed initial SPD matrix class.

[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.

[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
Algebraic techniques for spectral comparisons and preconditioning. The energy-error identity and inverse-order inequality displayed in this essay are elementary exact-arithmetic consequences, not claims of an implemented floating-point certificate.

[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.

[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.

[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.

[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.

[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.

[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.

[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.

[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.

[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
Official HAL abstract describes temporal error relaxation and spatial precision assignment, plus an author-reported maximum speedup. Full PDF was inaccessible in this review; the numerical tables were not independently inspected.

[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.

[69] FrontierCS 2 Team. *FrontierCS 2: A Benchmark for Computer Science R&D, current task records*. 2026. Live asset checked 4 November 2026.
Primary source · Pinned benchmark source · Data asset
Live asset checked 4 November: 18 systems, 48 tasks, one main run each. SpGEMM, KAMI and Polar Express have different evaluators. The October paper has 12 systems; 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.

[71] Sikai Cheng, Tinghan Ye, Arnaud Deza, Xiaonan Qi, Miao Li, Litong Liu, and Changkun Guan. *DecompGym: Benchmarking LLM Agents on Designing Decomposition Algorithms*. 2026. arXiv v1, 28 October.
Primary source
Eighty-one instances in nine families. Generated-solver runtime at most 60 seconds, feasibility, GPT-6 Astra method audit, and objective within one percent of stored optimum. Prescribed decomposition tasks, not unrestricted research. Accepted submissions were not independently human-audited; the false-acceptance rate is unmeasured.

[72] Shuo Chen, Li Cui, Cunxin Huang, Ya-Feng Liu, Minghang Xie, Meng Xu, and Zaikun Zhang. *NextEval: An Agent-Based Solver and Benchmark for Derivative-Free Optimization*. 2026. arXiv v1, 27 October.
Primary source
Ninety-two problems, n <= 5, budget 50n objective calls; ten repetitions in the main rotated comparison. Normalized profile areas are neither success percentages nor wall-clock speedups. Planned mechanism ablations are not completed results.

[73] Seungone Kim, Jiayi Geng, Young-Jun Lee, et al.. *Into the Unknown: Benchmarking AI agents on open scientific research problems*. 2026. arXiv v1, 13 October.
Primary source
Fifteen shared research directions, nine systems, no fixed wall-clock limit, up to five continuations. Scientist-specified rubrics and repeated automatic judgments. Strict submission gates differ from node-level rates. Some prose retains earlier counts; the 135-submission result uses Table 2. No timed human competitor.

[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.

[75] Noah Amsel, Yves Baumann, Paul Beckman, et al.. *Linear Systems and Eigenvalue Problems: Open Questions from a Simons Workshop*. 2026. arXiv v1, 5 February.
Primary source
Sections 2.1 and 2.5 connect graph-Laplacian techniques, numerical linear algebra, parameterized test families, and nonnormal convergence questions. Open problems establish a research agenda, not human comparative superiority.

[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.

[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.

[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 Refacte
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.