Forecasting Computer Science Breakthroughs in 2027-2031: An Evidence-Grounded Prospective Research Portfolio

Ada Mao | **GPT6-luna**

Prospective Research Forecast / Research Portfolio / Methodological Study
Primary field: Computer Science
Secondary fields: Artificial Intelligence; Algorithms and Complexity; Scientific Forecasting
Date and Version: 2026-10-17 | Version 1.0
Evidence cutoff: 2026-10-17
Research status: qualitative, uncalibrated; not peer reviewed

---

Abstract

Which computer-science research questions are most likely to see substantial, independently verifiable technical progress in 2027-2031? This paper presents an evidence-grounded prospective portfolio rather than a probability model. We screened the project's ten prospectively frozen qualitative research commitments, re-reviewed twelve earlier radar candidates, and assessed seven additional primary-source leads. We grouped dependent questions into research families, matched source claims to their stated models and endpoints, recorded counterevidence, and ranked opportunities separately for 2027-2028 and 2029-2031. The near-term portfolio prioritizes repository-scale code-and-proof synthesis, constrained large-language-model inference scheduling, intermediate-regime adaptive privacy filters, approximate fully dynamic maximum matching, and artifact-level multi-agent verification. The medium-term portfolio emphasizes the dynamic-matching/ordered Ruzsa-Szemerédi extremal chain, planar Three-Sets Cut-Uncut, planar discrete Fréchet distance under translation, two-stage integer-program optimization, and research agents that produce reproducible evidence. Several reported systems results make these areas concrete, but they are source-author results, not independent replications. The ten frozen commitments are distinct from the ten ranked portfolio entries, and the two rankings are not independent samples. Paper 3 has no eligible, independently adjudicated computer-science Target A or Target B evaluation; Paper 4 has no demonstrated action-selection advantage. We preserve an unresolved ORS deadline defect and an observation-window ambiguity in the three original events. The result is a transparent, falsifiable editorial portfolio, not evidence that AI can predict breakthroughs.

> ***Forecast status.*** *This paper reports a QUALITATIVE_UNCALIBRATED_FORECAST. It does not claim calibrated probabilities, validated Paper 3 performance on computer-science outcomes, Paper 4 planner superiority, independent human peer review, or that any forecast breakthrough has already occurred.*

1. Introduction

Research attention is easier to measure than scientific progress. Publication volume, citation growth, benchmark popularity, and topical similarity can rise without a theorem being proved, an algorithm improving against a matched baseline, or an empirical result surviving independent replication. Conversely, a technically important result may appear in a new formulation or in a venue that simple retrieval misses. Forecasting breakthroughs therefore requires an event that is specified before its outcome, a point-in-time evidence boundary, and a rule for distinguishing resolution from partial progress and missing evidence.

This paper addresses a narrower question than "Can AI predict science?": given the evidence available on 2026-10-17, which specific computer-science problem families merit the strongest qualitative expectation of independently verifiable progress by 2031, and what observations would change that ordering? We use "progress" in a technical sense and retain uncertainty. The portfolio is purposive and researcher-authored; it is not a representative sample of computer science.

The project's Paper 3 method studies question-level outcomes using frozen, point-in-time features and temporal validation. Its astronomy results cannot be transferred to computer science without eligible CS outcomes. Paper 4 organizes competing hypotheses, evidence gaps, and candidate research actions; its utility estimates are not observed scientific value. We use these methods as design constraints and vocabulary, not as validated forecasting or planning engines in this study.

2. Background and related work

Scientific forecasting and question outcomes

Science-of-science research has forecast citation impact and bibliometric "breakthrough" proxies, including work using early citation-network structure. PreScience provides a large, time-snapshotted scientific forecasting benchmark for paper- and topic-level outcomes. Neither establishes the exact task here: forecasting whether a frozen, theorem-level computer-science question will be resolved or substantially advanced under predeclared assumptions. The distinction does not make this portfolio the first scientific forecasting benchmark.

The project's question-discovery and astronomy historical-backtesting work is closer in unit and workflow: freeze a question and evidence cutoff, search later literature, and adjudicate outcomes. Its prospective astronomy instance is same-project prior work, not independent corroboration. It also documents a serious challenge: outcome categories may be difficult for careful reviewers to apply consistently. ResearchBench evaluates inspiration retrieval, hypothesis composition, and ranking; PaperBench evaluates replication of published AI research. These are relevant but distinct from prospective theorem-resolution forecasting.

Paper 3 and Paper 4

Paper 3's baseline uses question-by-cutoff observations, features limited to information available at each historical cutoff, and temporal rather than random validation. Future attention, direct resolution, and premise revision are distinct outcome dimensions. Its quantitative astronomy attention results do not imply CS predictive skill. Here we preserve two targets separately: **Target A**, direct proof or refutation of the frozen proposition, and **Target B**, substantial independently verifiable progress even if the proposition remains open. Target B may include a verified theorem or bound, matched-baseline algorithmic improvement, independent replication, independently checked counterexample, justified premise revision, or substantive narrowing. Attention is not a success label for either target.

Paper 4 represents a question through competing hypotheses, a provenance-linked evidence graph, missing or contradictory support, candidate actions, and declared cost and information-gain assumptions. That structure is useful for documenting what evidence would change a ranking. It does not show that the highest-ranked action is optimal or that the planner outperforms simple baselines. No information-gain probabilities or efficacy results are inferred here.

3. Research design

Evidence boundary and selection

The evidence cutoff is **2026-10-17**. We reviewed the final portfolio dossier, the ten-record prospective cohort and its manifest, earlier CS radar and source-screening material, Paper 3 and Paper 4 methodology records, and novelty/readiness audits. We screened primary-source leads in formal verification, inference systems, adaptive privacy, dynamic algorithms, approximation, geometry, integer programming, and scientific-agent evaluation. Candidate claims were checked against the source's actual graph/problem class, parameterization, baseline, metric, and assumptions. Where a reported benchmark result is not independently reproduced in this project, it is explicitly attributed to its authors.

We grouped related questions before ranking. In particular, approximate dynamic maximum matching and the ordered Ruzsa-Szemerédi (ORS) extremal parameter form one dependent chain, not two independent confirmations. The same principle applies to overlapping serving-systems and repository-verification questions. We did not treat absence of a retrieved paper as evidence that a question is neglected or unresolved.

Ranking method

The portfolio contains five entries for 2027-2028 and five for 2029-2031. Each entry had to have a source-backed baseline, a technically interpretable progress endpoint, an identifiable obstacle, counterevidence, and a plausible independent verification route. We then made a qualitative ordering using the clarity of the evidence and endpoint, a plausible technical path, horizon tractability, potential significance, and strength of counterevidence. We used no numerical weights, probabilities, attention counts, or citation counts. The order is a decision-priority judgment by the research authors, not an estimate of success probability.

The final portfolio is separate from the ten frozen commitments described in Section 7. Seven commitments were added to the original three in a prior registration step. The final 5+5 rankings neither create ten new events nor replace the frozen texts. No new event passed the threshold and independence gate in the final portfolio round.

4. What counts as a breakthrough or progress?

We use "breakthrough" in the title as shorthand for substantial technical progress on a specified research question; it is not a formal label assigned to the listed opportunities. A qualifying progress observation must be independently verifiable and may be one of:

1. a theorem or bound that advances the exact stated model;
2. an algorithmic improvement against a matched, explicitly stated baseline;
3. an empirical improvement independently reproduced under a comparable protocol;
4. an independently checked counterexample to a stated claim;
5. a justified refutation or revision of a key premise; or
6. a material narrowing of the open question that rules out a meaningful class of approaches.

These categories can coexist and should not be collapsed without a separate protocol. A direct proof or refutation of the frozen proposition is **Target A**. Substantial progress short of full resolution is **Target B**. Attention, citation or publication activity is tracked separately; practical computational impact is also separate. Missing or inaccessible evidence is unadjudicated, not a negative outcome. An algorithmic proof establishes correctness relative to its formal specification, not that the specification captures intended behavior.

5. Near-term portfolio: 2027-2028

5.1 Repository-scale code-and-proof synthesis

Question. Can automated systems produce buildable implementations and machine-checked proofs across multi-module repositories with cross-module invariants, rather than solve isolated proof obligations?

Baseline and evidence. Vero reports 43 repository instances; its authors report 27 complete code-and-proof solves for the strongest evaluated configuration and ten instances resistant to all tested configurations. VeriSoftBench reports difficulty with deep repository dependencies. CoCo-Prover reports improved success-cost results, including up to 26/43 Vero solves, in its own evaluation. These are author-reported benchmark results; this project has not reproduced them.

Expected progress and obstacle. The qualitative expectation is further benchmark-specific improvement in solve coverage and cost-quality trade-offs, not reliable autonomous software engineering in general. Cross-module invariants, reusable proof structure, complete builds, benchmark curation, shifting models and compute budgets, and correspondence between formal specification and intended behavior are major obstacles.

Counterargument and significance. Gains may be benchmark- or model-specific, and a proof kernel certifies only the encoded specification. Progress would matter for software assurance if it transfers to new repositories and independently audited specifications.

Independent verification. Lock repository versions, tasks, model/tool versions and budgets; test hidden repositories; check proofs with the trusted kernel; require clean builds; separately audit specifications; and report cost and failure modes. A further model evaluation would require its own authorization and is not claimed here.

5.2 LLM inference scheduling under KV-cache pressure and speculative execution

Question. Can a serving policy jointly schedule requests, speculative work, and KV-cache residency under changing arrivals to improve goodput or tail latency at fixed output quality, memory, and service-level constraints?

Baseline and evidence. Online scheduling results depend on the arrival model; one fully adversarial model rules out a constant competitive ratio. Spexis introduces speculative lookahead scheduling and reports up to a 34% request-rate increase on evaluated GPU/workload configurations. Heterogeneous KV placement is an adjacent line, not evidence that all components compose. The Spexis result is not independently replicated here.

Expected progress and obstacle. We expect configuration-specific integration and scheduling gains rather than a universal online guarantee. Apparent improvements can disappear when memory, model quality, workload, preemption, energy, and p95/p99 latency are matched. Hardware and model generations also change quickly.

Independent verification. Replay fixed arrival traces on matched hardware, model, memory, and quality settings; disclose code and configurations; report goodput, tail latency, energy, eviction/recomputation, and failures. Separate an algorithmic online guarantee from measured systems performance.

5.3 Fully adaptive privacy filters at intermediate sampling rates

Question. Can valid, quantitatively useful privacy filters be proved for fully adaptive compositions of Poisson-subsampled Gaussian mechanisms in intermediate sampling-rate regimes not covered by current approximate constructions?

Baseline and evidence. Tran et al. show that a natural fully adaptive (f)-DP filter can be invalid, state validity conditions, and construct approximate GDP filters in low-rate ($q<0.2$) and high-rate ($q>0.8$) regimes under their assumptions. This makes the middle-regime question precise but does not establish that a uniformly useful extension exists.

Expected progress and obstacle. A restricted theorem, sharper characterization, or lower bound is plausible. Full adaptivity, approximation error, and numerical soundness are difficult; existing RDP or privacy-loss-distribution accountants may be stronger in some settings. The frozen event's comparator and improvement region remain flagged as underspecified.

Independent verification. Check assumptions and proof line by line, establish explicit error bounds, compare against named accountants under identical composition models, and use certified arithmetic only for numerical claims. A numerical experiment cannot substitute for the theorem.

5.4 Sublinear updates for approximate fully dynamic maximum matching

Question. Can a general-graph fully dynamic data structure maintain an explicit ((1-epsilon))-approximate maximum matching in (O(n^1-)) amortized update time for fixed (arepsilon,>0), under the frozen adversary and output assumptions?

Baseline and evidence. FOCS 2024 and SODA 2025 work connects the approximate maximum-matching update time to ordered-RS density. Bernstein et al. later show sublinear update time for **maximal** matching. This is adjacent and important, but maximal matching is not approximate maximum matching. The ORS extremal question remains part of the same research family.

Expected progress and obstacle. Further structural or parameterized bounds are plausible; exact resolution by the horizon is uncertain. A better extremal bound only helps if the reduction preserves the desired exponent and all approximation, adversary, amortization, and explicit-output guarantees.

Independent verification. Match the graph model, update sequence, adversary, probability guarantee, preprocessing/space, output representation, and exponent theorem by theorem. Keep ORS and dynamic matching linked in analyses and outcomes.

5.5 Artifact-level verification of multi-agent work

Question. Can a verifier predict independently checked final-artifact correctness and improve end-to-end task success after accounting for cost, false positives, false negatives, and context?

Baseline and evidence. MAS-ProVe reports that process verification does not consistently improve multi-agent performance and may exhibit high variance. EIG and ResearchBench assess ideation or benchmark-native quality; PaperBench tests replication. These benchmarks do not jointly measure independently verified scientific discovery. The astronomy backtesting work also reports low inter-annotator agreement under its specific outcome taxonomy; this is a warning about measurement, not an estimate for CS.

Expected progress and obstacle. We expect task-specific verifier designs and clearer negative boundaries, not a universal process verifier. Shared models, benchmark contamination, insufficient traces, and costly external ground truth may make predictive verifier scores unreliable.

Independent verification. Freeze verifier outputs before outcome adjudication; use independent tests or reviewers for the final artifacts; report error rates, agreement, compute, and paired task success. A judge score is not itself a correctness label.

6. Medium-term portfolio: 2029-2031

6.1 Dynamic matching and the ORS extremal bottleneck

The specific extremal question is whether the ordered Ruzsa-Szemerédi parameter at fixed linear matching density has polynomial growth strong enough to imply a fixed-exponent sublinear update bound for the approximate maximum-matching target. This is one connected research family, not two observations. The 2024-2025 algorithmic reductions make the chain explicit, but an ORS improvement must be transferred through the exact reduction and quantifiers. The frozen ORS event has an unresolved date conflict (Section 7); we preserve it rather than repair it retrospectively.

6.2 Planar Three-Sets Cut-Uncut with terminal sizes ((1,1,2))

Can unweighted planar Three-Sets Cut-Uncut with two singleton terminal sets and one two-vertex terminal set that must remain connected be solved exactly in polynomial time? Bentert et al. give a randomized fixed-parameter algorithm for the planar two-set version and explicitly leave this three-set case as an open question in their 2024 paper. Ordinary planar Multiway Cut does not enforce the required connectivity of the two-vertex set. The endpoint is theorem-level and verifiable, but the added connectivity constraint may be the central difficulty. Verification must compare the exact terminal pattern, graph class, costs, connectivity condition, exactness, and infeasibility handling.

6.3 Planar discrete Fréchet distance under translation

The frozen question concerns exact discrete Fréchet distance under translation between two planar polygonal curves, with a target (O(n^4polylog n)) bound in a comparable computational model. The audited baseline reports an approximately (~ O(n^14/3)) algorithm and an SETH-based (n^4-o(1)) lower bound for the specified problem. Thus the remaining exponent room is narrow and conditional. A matching theorem must resolve model, arithmetic, exact-value versus decision, and bit-complexity details.

6.4 Two-stage integer programming with unrestricted global coefficients

Can the FPT feasibility algorithm for two-stage block-structured integer programs extend to exact linear optimization when local coefficient parameters remain bounded but global-block coefficients are unrestricted? Cslovjecsek et al. establish a neighboring feasibility result and identify optimization as a distinct next question. A later FPT theorem parameterized by all coefficient sizes does not settle the unbounded-global-coefficient regime. The important counterargument is that large global coefficients may encode the hard part. The endpoint must distinguish exact optimization, infeasibility, and unboundedness under the same encoding and parameterization.

6.5 Scientific agents that produce reproducible evidence

Can research agents freeze a falsifiable question, execute a discriminating study, preserve auditable raw artifacts, and obtain independent reproduction more reliably than simple baselines? Current tools separately measure ideation, replication, or process verification; they do not establish a common outcome for original, independently confirmed scientific progress. This is the least mature portfolio entry. A useful advance requires an independently adjudicated outcome taxonomy, agreement and missingness reporting, raw-artifact reproduction, and comparison against simple baselines. No new forecast event is frozen for this family.

7. Existing prospective forecast registry

The project registry contains **ten frozen qualitative commitments**: three original records and seven later registrations. They are commitments within a research program, not ten newly generated portfolio rankings and not ten independent observations. The records retain source URLs, event definitions, evidence standards, source cutoffs, issue timestamps, and content/snapshot hashes. Their full immutable text and metadata are in the cohort registry at the research checkpoint; the linked activation record preserves the three earlier entries.

Frozen event	Issue time (UTC)	Structured deadline	Compact event endpoint	Primary-source provenance
`PLANAR-THREE-SETS-CUT-UNCUT-112`	2026-10-17 02:47:37	2030-12-31	Exact polynomial-time algorithm for the planar unweighted ((1,1,2)) Three-Sets Cut-Uncut case.	Bentert et al., ICALP 2024
`DISCRETE-FRECHET-TRANSLATION-PLANAR-N4`	2026-10-17 02:47:37	2031-10-16	Exact planar discrete Fréchet under translation in (O(n^4polylogn)) time under comparable assumptions.	Bringmann, Künnemann & Nusser

Frozen event	Issue time (UTC)	Structured deadline	Compact event endpoint	Primary-source provenance
K-HOP-DOMINATION-LOW-EXPANSION-NO-GIRTH	2026-10-17 02:47:37	2031-10-16	Remove the high-girth condition while preserving the stated constant-round bound for bounded (k)-hop expansion.	Lenzen & Wenning
MAXBISECTION-GW-APPROX-001	2026-10-17 03:15:09	2031-12-31	Reach the Goemans-Williamson MAX-CUT ratio for general unweighted MAX BISECTION.	Brakensiek et al., SODA 2026
DYNAMIC-MATCHING-POLY-EXPONENT-001	2026-10-17 03:15:09	2031-12-31	Explicit ((1-epsilon))-approximate maximum matching with fixed-exponent sublinear fully dynamic update time.	Behnezhad & Ghafari, FOCS 2024
ORS-POLYNOMIAL-GROWTH-001	2026-10-17 03:15:09	**2030-12-31 in structured deadline; 2031-12-31 in frozen event text**	Polynomial growth of ordered-RS with fixed linear matching density.	Assadi, Khanna & Kiss, SODA 2025
SUBSET-SUM-EXPONENT-IMPROVEMENT-001	2026-10-17 03:15:09	2031-12-31	Fixed exponential-base improvement over (2^n/2 for worst-case exact Subset Sum.	Chen et al., ESA 2025
TWO-STAGE-IP-OPT-UNBOUNDED-GLOBAL-COEFF-001	2026-10-17 03:15:09	2030-12-31	Extend FPT feasibility to exact optimization without parameterizing unrestricted global coefficients.	Cslovjecsek et al., SODA 2024
SHALLOW-QC-IDENTITY-PTAS-001	2026-10-17 03:15:09	2031-12-31	PTAS for diamond-norm distance of constant-depth 1D local quantum circuits from identity.	Bravyi, Parham & Tran
ADAPTIVE-FDP-GDP-FILTER-MID-Q-001	2026-10-17 03:15:09	2030-12-31	Valid approximate GDP filters for the frozen intermediate sampling-rate regime under full adaptivity.	Tran et al.

The original three records have the same recorded issue timestamp, 2026-10-17 02:47:37 UTC, but their individual `observation_start` fields are null. Their issue date precedes the global 2027-01-01 forecast-horizon start; the interval between issuance and the stated horizon is therefore ambiguous. This paper preserves that defect and does not retroactively choose whether results in that interval count.

The later seven records were issued at 2026-10-17 03:15:09 UTC and have source-only evidence cutoffs recorded at issuance. They remain qualitative. Two independent human reviewers are required for final event adjudication; an AI review is not an independent human review. The original three records and later seven records retain separate provenance. Full-text source bytes were not archived or hashed; recorded source hashes cover citation/metadata records and registry snapshots, not the exact paper content.

The ORS record has a specific unresolved integrity defect: the structured deadline is 2030-12-31, while its frozen event text states 2031-12-31. We neither select a preferred deadline nor rewrite the record here. If the conflict affects adjudication, the outcome must remain unresolved until a provenance-based correction process is authorized.

8. Ranking robustness

We compared three qualitative priority scenarios while holding candidate families fixed:

- **Proof-tractability priority** raises explicit theorem targets such as adaptive privacy, planar cut-uncut, and two-stage IP.
- **Computational-impact priority** raises inference scheduling and repository-scale code/proof work.
- **Source maturity and independent-verifiability priority** raises exact theorem endpoints and lowers multi-agent process verification and autonomous science, whose outcome labels are least settled.

Leading family membership is more stable than the exact order; the medium-term ordering is particularly sensitive. The sensitivity table in the companion audit records the variant orders. These are ordinal scenario comparisons, not statistical sensitivity estimates, calibration intervals, or probabilities.

9. Scientific validation framework

Future evidence monitoring should preserve the exact question, source cutoff, issue timestamp, family identifier, and predeclared outcome criteria. A theorem-level match requires independent reviewers to compare the result's object, assumptions, graph/problem class, computational model, parameters, quantifiers, and endpoint. For empirical claims, independent verification requires versioned code/data, matched baselines, comparable resource budgets, uncertainty reporting, and reproducible artifacts. Reviewers should record disagreement rather than resolve it by silent consensus; search failure remains missing evidence.

For **Target A**, evaluate direct proof/refutation of the frozen proposition. For **Target B**, independently adjudicate verified progress categories without collapsing progress into full resolution. Attention and impact remain separate. Any future Paper 3 model comparison needs sufficient eligible outcomes, point-in-time features, family-grouped chronological splits, an untouched final time period, and simple baselines such as publication volume, field momentum, and source-only text features. No CS score is valid before those gates pass.

Paper 4 should freeze competing hypotheses, evidence provenance, action choices, costs and information-gain assumptions before action. After execution, feasibility must be separated from realized scientific evidence yield. Compare prospectively selected actions to simple baselines only when outcomes are comparable. The present portfolio does not measure planner efficacy.

10. Limitations

This is one purposive qualitative synthesis, not a representative cohort or a calibrated model. The ten frozen commitments are small, heterogeneous, and mostly mature only in 2030-2031. No outcome labels have matured; no independent human event adjudication is complete; no independent panel reviewed the final ranking. Candidate families are dependent, and publication, selection, and reporting bias are plausible. Recent Vero, CoCo-Prover, and Spexis measurements are author-reported and not independently reproduced in this project. Some empirical topics need more precise cross-source endpoints. Source hashes cover metadata records, not full-text bytes. The original three have the observation-window ambiguity; the ORS record has its deadline mismatch. No calibrated forecasting accuracy, Paper 3 CS predictive performance, or Paper 4 action-selection superiority has been demonstrated. This is not peer-reviewed work.

11. Conclusion

The evidence supports a ranked set of questions that merit attention as prospective research opportunities, not claims of predictive success. Repository-scale code-and-proof synthesis and constrained inference scheduling have the strongest qualitative near-term case because they have visible technical progress and objective artifact-level tests, yet their reported gains remain configuration-sensitive. Adaptive privacy filters offer a precise theory problem in an intermediate regime. Approximate maximum matching is linked to the unresolved ORS extremal bottleneck; maximal matching progress is not a substitute. The medium-term list includes exact planar, geometric, and integer-programming questions with theorem-level endpoints, while reproducible scientific agents remain particularly uncertain because reliable outcome adjudication is itself unresolved.

The ten frozen commitments remain distinct from the final 5+5 portfolio. Paper 3 has not shown CS predictive power, and Paper 4 has not shown planner superiority. The contribution is a transparent, falsifiable research portfolio with explicit counterarguments and verification criteria. Whether this portfolio has independent methodological novelty or publication value remains open to external review.

Author contributions and AI disclosure

Ada Mao directed and supervised the research program. GPT6-luna, an AI research agent, performed literature synthesis, candidate screening, evidence organization, and prospective qualitative forecast development under human supervision. GPT6-luna is an AI research collaborator, not a human researcher. The manuscript and portfolio remain uncalibrated and have not received independent human peer review.

Data and reproducibility

The portfolio index, canonical detailed dossier, hash manifest, and frozen registry are publicly inspectable in the research repository at commit `db6aaa4`. The manifest verifies exact local artifact bytes and the unchanged frozen inputs; it does not hash cited full-text papers. Repository validation checks structure and integrity, not the truth of future forecasts. No experiments, model training/inference, new labels, or new forecast commitments were generated in the final portfolio round.

References

1. Min, C., Bu, Y., & Sun, J. (2021). Predicting scientific breakthroughs based on knowledge structure variations. *Technological Forecasting and Social Change*, 164, 120502. https://doi.org/10.1016/j.techfore.2020.120502
2. Ajith, A., et al. (2026). PreScience: A dataset and benchmark for scientific forecasting. arXiv:2602.20459.
3. Mao, H. (2026). Evidence-Based Scientific Question Discovery: A Framework with Historical Backtesting. arXiv:2608.09968.
4. Mao, H. (2026). Historical Backtesting for Scientific Question Discovery: A Protocol and Astronomy Pilot. arXiv:2608.16795.
5. Liu, Y., et al. (2026). ResearchBench: Benchmarking LLMs in Scientific Discovery via Inspiration-Based Task Decomposition. *Findings of ACL 2026*. https://aclanthology.org/2026.findings-acl.644/.
6. Starace, G., et al. (2025). PaperBench: Evaluating AI's Ability to Replicate AI Research. arXiv:2504.01848.
7. Ye, Z., et al. (2026). Vero: Can AI Agents Build Formally Verified Software Repositories? arXiv:2608.13522.
8. Xin, Y., et al. (2026). VeriSoftBench: Repository-Scale Formal Verification Benchmarks for Lean. arXiv:2602.18307.
9. Yao, S., et al. (2026). Cost-Efficient Theorem Proving via Agent Orchestration in Program Verification. arXiv:2610.09681.
10. Jaillet, P., et al. (2025). Online Scheduling for LLM Inference with KV Cache Constraints. arXiv:2502.07115.
11. Jung, H., et al. (2026). Spexis: Speculative Lookahead Scheduling for LLM Inference. arXiv:2609.34370.
12. Tran, H., et al. (2026). f-Differential Privacy Filters: Validity and Approximate Solutions. arXiv:2602.06756.
13. Behnezhad, S., & Ghafari, M. (2024). Fully Dynamic Matching and Ordered Ruzsa-Szemerédi Graphs. *FOCS 2024*. https://doi.org/10.1109/FOCS61266.2024.00027.
14. Assadi, S., Khanna, S., & Kiss, P. (2025). Improved Bounds for Fully Dynamic Matching via Ordered Ruzsa-Szemerédi Graphs. *SODA 2025*. https://doi.org/10.1137/1.9781611978322.96.
15. Bernstein, A., et al. (2025). Deterministic Dynamic Maximal Matching in Sublinear Update Time. arXiv:2504.20780.
16. Venkataramani, V., et al. (2026). MAS-ProVe: Understanding the Process Verification of Multi-Agent Systems. arXiv:2602.03053.

17. Bentert, M., et al. (2024). Two-Sets Cut-Uncut on Planar Graphs. *ICALP 2024*. https://doi.org/10.4230/LIPIcs.ICALP.2024.22.

18. Bringmann, K., Künnemann, M., & Nusser, A. (2019). Fréchet Distance Under Translation: Conditional Hardness and an Algorithm via Offline Dynamic Grid Reachability. arXiv:1810.10982.

19. Lenzen, C., & Wenning, S. (2024). Tight Bounds for Constant-Round Domination on Graphs of High Girth and Low Expansion. arXiv:2408.12998.

20. Brakensiek, J., Huang, M., Potechin, A., & Zwick, U. (2026). MAX BISECTION Might Be Harder to Approximate than MAX CUT. *SODA 2026*. https://doi.org/10.1137/1.9781611978971.197.

21. Chen, L., Jin, C., Randolph, W., & Servedio, R. (2025). New Algorithms for Pigeonhole Equal Subset Sum. *ESA 2025*. https://doi.org/10.4230/LIPIcs.ESA.2025.86.

22. Cslovjecsek, J., Koutecký, M., Lassota, A., Pilipczuk, M., & Polak, A. (2025). Parameterized Algorithms for Block-Structured Integer Programs with Large Entries. *TheoretiCS*, 4. https://doi.org/10.46298/theoretics.25.15.

23. Bravyi, S., Parham, T., & Tran, M. C. (2024). Identity Check Problem for Shallow Quantum Circuits. arXiv:2401.16525.