Can AI Do Research On Its Own?

Towards end-to-end automation of AI research
– Lu et al,. Nature (2026)

The Idea in Brief

01 Before This Paper

Earlier AI solved narrow problems: protein structures, math proofs, etc. LLMs now help with single steps, but no method exists to automate the entire research process.

02 Goal Of Paper

The AI Scientist runs the end-to-end research process: ideation, lit review, experiments, analysis, manuscript writing, and review.

03 Scope

ML experiments happen entirely on a computer. The system can write the code, run it, measure results, and debug with no human or physical lab in the loop.

The Pipeline: Idea to Reviewed Paper

01

IDEATION

Propose research ideas. Check each against Semantic Scholar, then score and archive.

02

Experimentation

Write, run, debug, and improve experiment code. Template free does it in four stages.

03

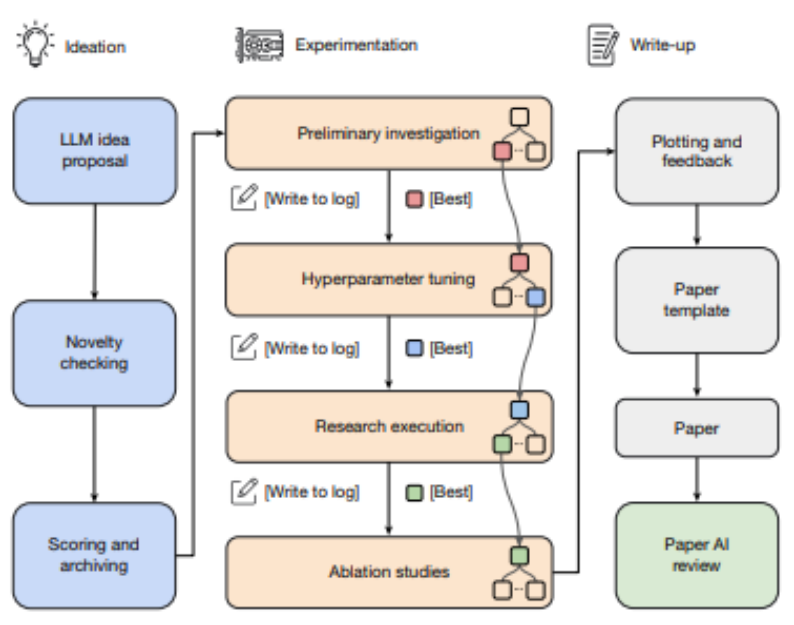
Write-Up

Plot results, fill a LaTeX conference template, and add citations.

04

Review

An LLM reviewer grades the paper with NeurIPS review guidelines.

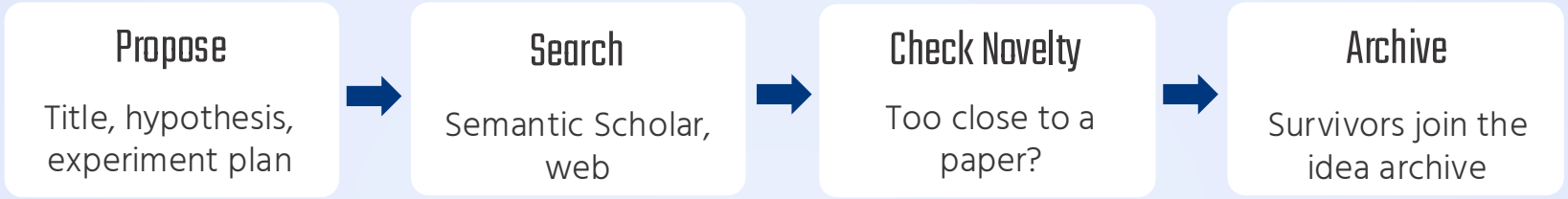


Two Versions Of The System

	Template-based	Template-free
Starting point	Human code that reproduces a known training run	No code given. It writes its own first script.
Research scope	A specific topic set by the template	Open-ended ideas within a broad theme
Writing code	Aider, an open-source coding assistant, edits the template	Claude Sonnet 4 writes a new script at every tree node
Running experiments	Up to 5, run in a fixed order	Parallel tree search across 4 stages
Datasets	The benchmarks that come with the template	Can query the Hugging Face Hub for datasets

Template-based is a focused mode. Template-free trades that structure for wider exploration and spends extra compute on search.

Step 1: Ideas and Novelty Checks



Template-Based

- An LLM acts as a mutation operator: it proposes variations and extensions of ideas already in the archive.
- Each idea scores itself from 1 to 10 on interestingness, novelty and feasibility.
- The novelty check runs up to 10 rounds of literature search, refining the query each round.

Template-free

- Ideas are not tied to any code. Each reads like a paper abstract: a problem, a new method and the expected outcome.
- A literature-review module looks for knowledge gaps. Reflection rounds then revise the proposal using the search results.

Step 2: Writing and Running the Code

Template-Based

- Plans up to 5 experiments and runs them in order. Aider edits the code for each one.
- On a crash, Aider gets the failing code and error log and writes a patch. Up to 4 retries.
- Each run times out after 7,200 s (2 hours).
- Every result is logged in the experiment journal.

Template-free

- Writes its first script from scratch.
- Finds data by querying the Hugging Face Hub. Ten example datasets are listed in its prompt.
- Explores many versions at once in a tree (next two slides).

Model	Job
OpenAI o3	Generates ideas. Critiques the code after each run.
Claude Sonnet 4	Writes each experiment's plan and code for every node.
GPT-4o	Critiques plots and captions. Ranks nodes to expand next.
OpenAI o1	Writes the paper in LaTeX and revises it.
O4-mini	Runs the automated review.

Step 3: Four Stages of Experiments

01

Preliminary Investigation

Build a baseline that shows the idea can be tested.

02

Hyperparameter Tuning

Tune settings to improve existing model.

03

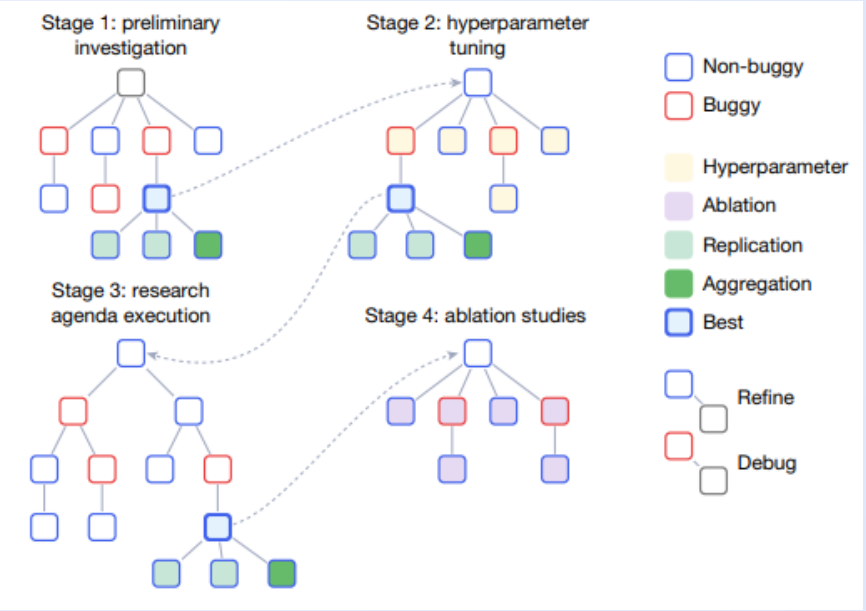
Research Agenda

Runs the main experiments to test hypothesis.

04

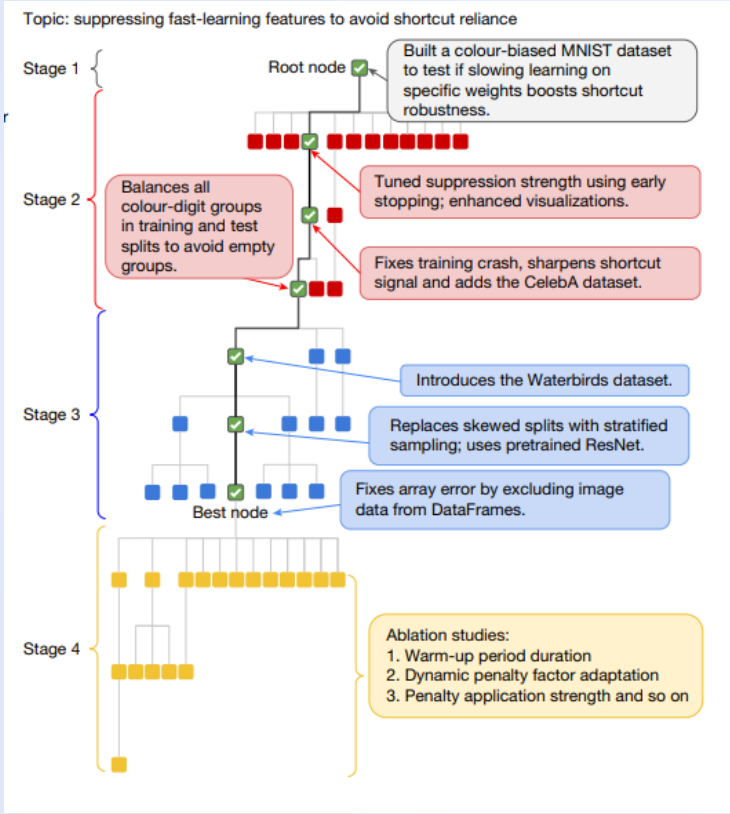
Ablation Studies

Remove or vary components to see which ones matter



After each stage, an LLM judge picks the best leaf node, which seeds the next stage.

Step 4: Agentic Tree Search



Each Node Is One Experiment

It stores the plan, script, error trace, metrics, plots, o3's critique of the code, and 4o's feedback

When a Node Is Buggy

Its code crashed, or 4o flagged its plots for problems like unclear labels or misleading visuals

How The Tree Grows

Each round expands several nodes at once. With a preset prob., it picks a buggy node to debug. Else, it ranks working nodes and refines the best one.

Special Nodes

Hyperparameter and ablation nodes track what was already tried. Replication nodes rerun a result with new seeds. Aggregation nodes plot the mean and std. dev..

Step 5: Turning the Journal Into a Paper

01

Journal

Notes, metrics, and plots frome very experiment.

02

Figures

4o verifies results and merges results from several stages.

03

Draft

System fills a blank template using notes and plots

04

Citations

Up to 20 rounds of Semantic Scholar search. Each citation comes with a justification.

05

PDF

Editing passes, then compile to PDF. Up to 5 iterations for errors.

Automated evaluation and human peer review

The automated reviewer uses o4-mini and NeurIPS review guidelines.

Reviewer output	How the authors use it
Accept or reject	The authors assess agreement with ICLR decisions.
Manuscript score	The authors compare scores across system configurations.

The reviewer assesses methodological soundness, presentation and contribution.

The authors submitted three selected manuscripts to ICLR 2025 ICBINB, a workshop on negative results in deep learning.

Evaluating review decisions

Balanced accuracy assigns equal weight to accepted and rejected papers.

Balanced accuracy =
(accuracy on accepted papers + accuracy on rejected papers) ÷ 2

Review method	Overall accuracy	Balanced accuracy
Baseline that rejects every paper	65%	50%
Baseline reviewer	65%	**72%**

The rejection baseline achieves 65% overall accuracy but only 50% balanced accuracy.

Lu et al., Nature 651 (2026). Supplementary Table A4. Reported means.

Agreement with conference decisions

The 2025 evaluation uses papers from after the model's stated knowledge cutoff.

Reviewer and paper set	Balanced accuracy
Automated Reviewer, ICLR 2017–2024	**69% ± 4 pp**
Automated Reviewer, ICLR 2025	**66% ± 3 pp**
Human review consistency, NeurIPS 2021	**66%**

The human comparison uses a separate dataset from NeurIPS 2021.

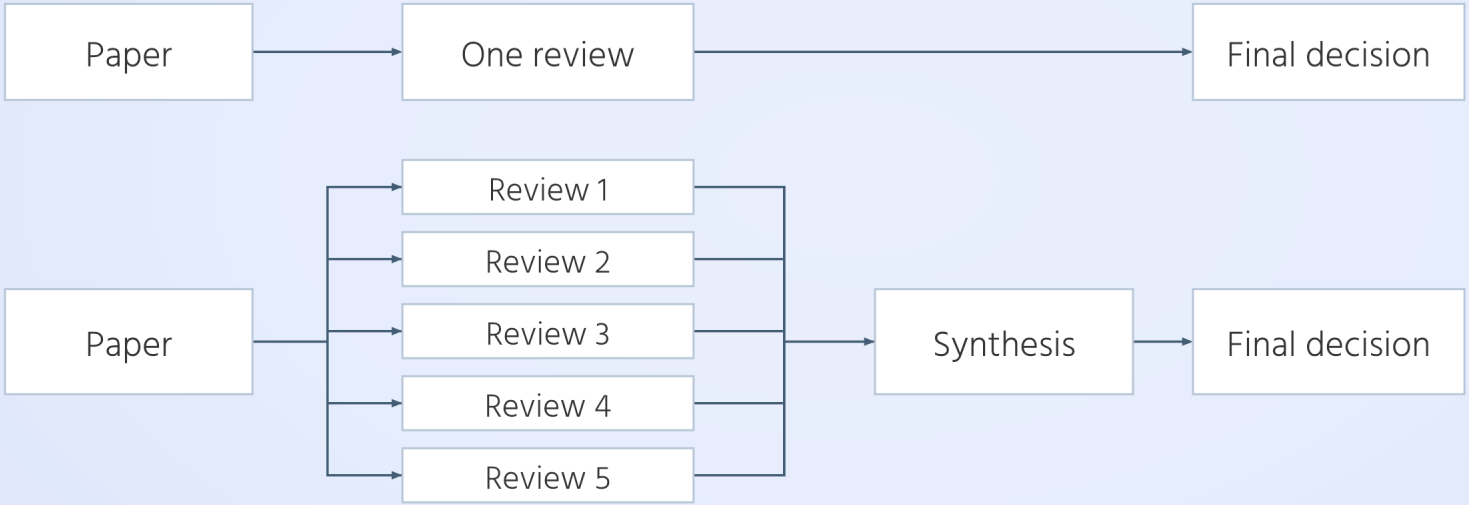
The ± values are percentage points and show 95% bootstrap confidence intervals.

Lu et al., Nature 651 (2026), Table 1.

Single review and synthesis of five reviews

Both settings use o4-mini and the same base prompt.

Balanced accuracy



Synthesizing five reviews increased the reported mean balanced accuracy.

Lu et al., Nature 651 (2026). Supplementary Table A4. Reported means.

Figure access and self-review

Each ablation adds one component to the baseline reviewer.

Review process	Mean balanced accuracy
The reviewer evaluates the text.	**72%**
The reviewer also evaluates the figures.	69%
The reviewer reassesses its initial review.	72%

Figure access and self-review showed no improvement in mean balanced accuracy.

The reported means do not establish the statistical significance of individual differences.

Lu et al., Nature 651 (2026). Supplementary Table A4.

A stricter review prompt

The additional instruction favors rejection when the reviewer is uncertain.

Error relative to conference decisions	Base prompt	Stricter prompt
False acceptance The model accepts a paper the conference rejected.	45%	**18%**
False rejection The model rejects a paper the conference accepted.	15%	**41%**

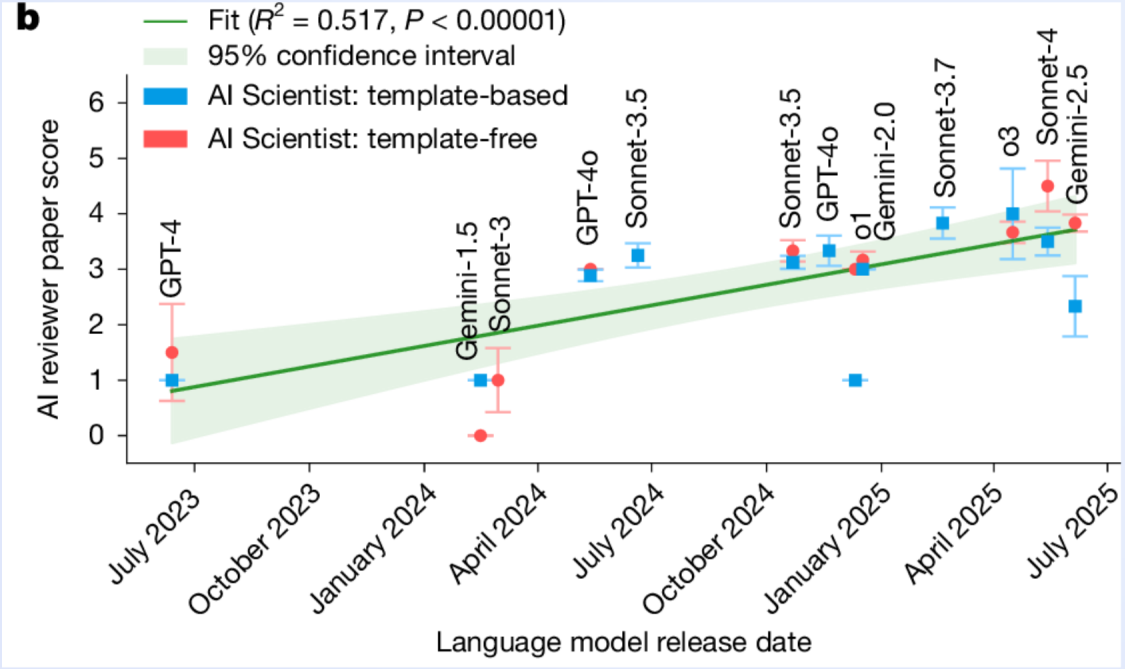
The stricter prompt reduced the false acceptance rate at the cost of a higher false rejection rate.

Row 1 uses rejected papers as its denominator. Row 2 uses accepted papers.

Lu et al., Nature 651 (2026). Supplementary Table A4. Reported means.

Model release dates and manuscript scores

Later model release dates were associated with higher mean manuscript scores.



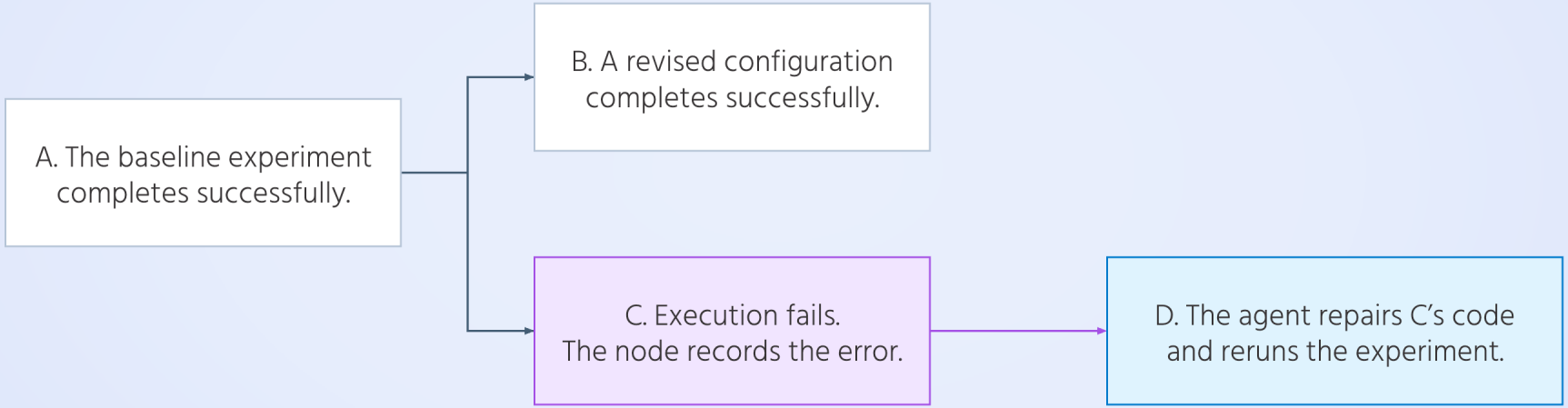
Each point averages 3 manuscripts in the template-based setting or 6 in the template-free setting.

Error bars denote standard errors. The green band denotes a 95% confidence interval.

Lu et al., Nature 651 (2026), Fig. 1b and SI A.2.9. Figure: CC BY 4.0.

An experiment node

Each search node records the code and outcome of an experimental attempt.



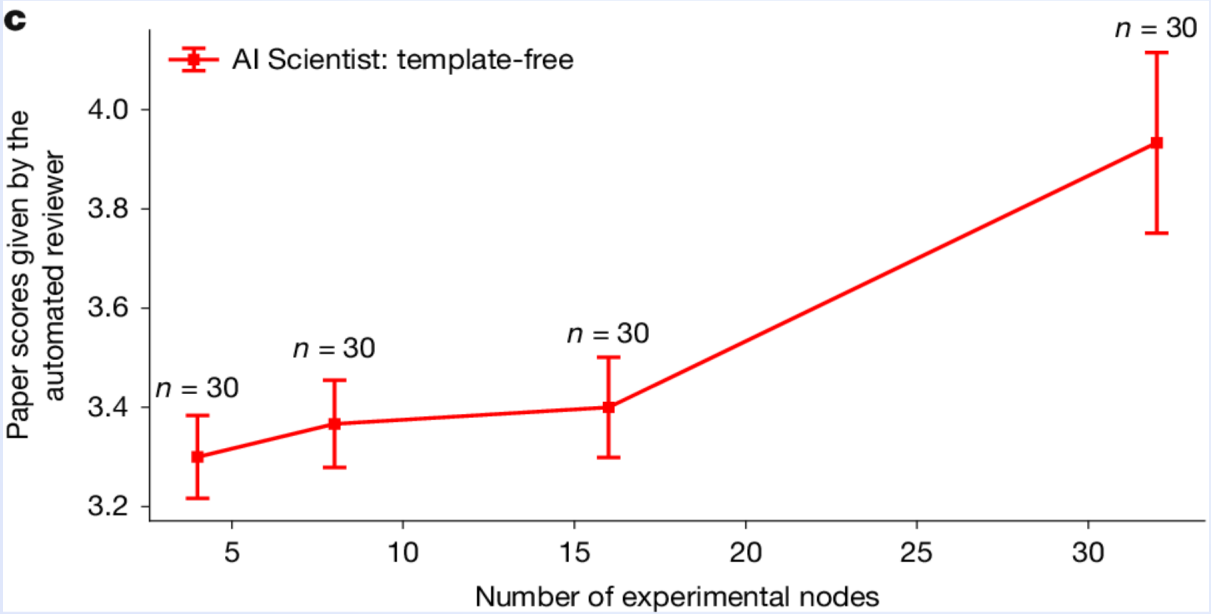
Recorded execution errors provide feedback for subsequent debugging.

This is an illustrative workflow. Some nodes aggregate existing records without a new experiment.

Lu et al., Nature 651 (2026). Methods: parallelized agentic tree search.

Experiment budget and manuscript scores

The comparison fixes Claude Sonnet 4 and increases both research and ablation budgets.



Stage 3 + Stage 4

Total	Research	Ablation
4	2	2
8	4	4
16	8	8
32	16	16

The first two stages are excluded from this budget.

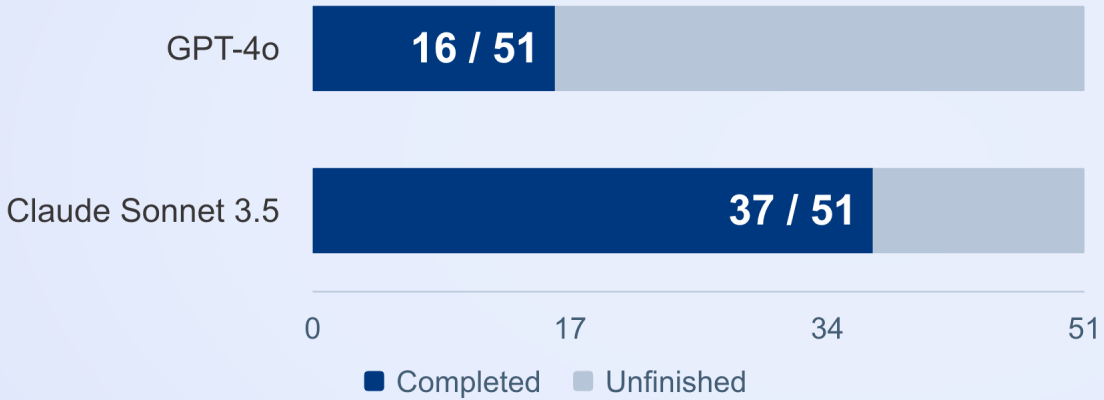
Mean score = (sum of 30 final review scores) ÷ 30

3 ideas × 2 write-ups × 5 complete reviews = 30 scores. Error bars show SE.

Lu et al., Nature 651 (2026), Fig. 3c and SI A.2.9. Figure reproduced under CC BY 4.0.

Completion and manuscript scores

Each model generated 51 ideas using the same diffusion template.



Scores of completed papers

Model	Mean score
Claude 3.5	2.97
GPT-4o	2.81

Mean score = total score of completed papers ÷ number of completed papers

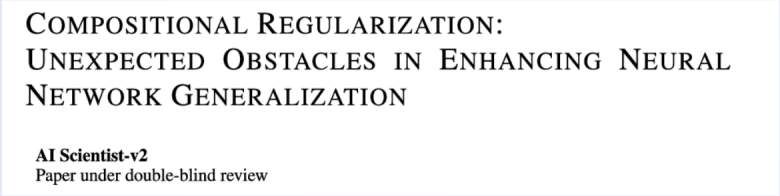
Mean scores are conditional on manuscript completion and exclude unsuccessful runs.

Lu et al., Nature 651 (2026), SI Table B6. Completion requires a compilable manuscript. Idea sets differ.

Human review at ICLR 2025 ICBINB

The authors selected three manuscripts for a workshop on negative results in deep learning.

Generated manuscript


COMPOSITIONAL REGULARIZATION:
UNEXPECTED OBSTACLES IN ENHANCING NEURAL NETWORK GENERALIZATION

AI Scientist-v2
Paper under double-blind review

Experiments on arithmetic expressions showed no consistent generalization gains from the proposed regularization.

Manuscript	First-round assessment
Compositional regularization	6, 7, 6 Above threshold
Label noise and calibration	Below threshold
Crop pest and disease classification	Below threshold

$(6 + 7 + 6) \div 3 = 6.33$

Selected manuscripts underwent no human editing. The authors withdrew all three submissions after review.

Limitations of the evaluation

1. The human evaluation is limited to manually selected submissions at a workshop focused on negative results.

2. The paper includes no comparison of tree search and linear execution using the same model and budget.

3. The stability of model rankings and component effects across newer systems remains untested.

The authors' internal review placed all submissions below the ICLR main-conference standard.

Lu et al., Nature 651 (2026). SI A.2.9 and A.4. Later harness example: Auto-RecSys (2026).