

LECTURE 02

Patterns of discovery: analogy, recombination, anomaly, mechanism

Analogy as a mapping of relations with explicit failure conditions.

Reasoning agents, feedback, and scientific evidence

A delay with several possible causes

Imagine a membrane responds slowly after its surroundings change. A delay alone leaves several explanations open: transport, reaction or a slow instrument.

Making the membrane thicker could change the delay. That observation suggests investigating how a spatial process creates a timescale.

The question becomes which familiar systems combine local transfer, conservation and a length scale. This creates a reason to look for a structural analogy.

Finding a useful source analogy

Heat spreading through a slab is a candidate because local transfer and conservation can produce a delay that depends on thickness.

Map temperature to concentration, thermal diffusivity to molecular diffusivity, and slab thickness to membrane thickness. Check the boundary conditions before transferring a law.

Under diffusion assumptions, doubling thickness at fixed diffusivity multiplies the characteristic time by four. A different scaling invites another mechanism or a revised assumption.

Choosing a mechanism to investigate

Candidate	Reason to choose it	Cost or limitation
Instrument response	Delay persists in a reference measurement	Changing the sample may be irrelevant
Reaction kinetics	Transformation limits the response	Thickness alone may not control the time
Diffusive transport	Spatial transfer suggests a length ² scale	Flow, reactions or variable diffusivity alter it

Pursue diffusion when the spatial mapping is credible. Compare thickness and reference measurements before treating the analogy as an explanation.

Pedagogical reconstruction

[Gentner \(1983\)](#)

Recombination needs compatible interfaces

A hypothesis can combine a mechanism from one domain with an intervention or measurement from another. Compatibility depends on units, timescales, entities, and background assumptions.

Typed slots make these checks explicit. They are an implementation aid rather than a universal definition of scientific thought.

Preserve the sources of each component and identify the new relation produced by their combination. Renaming familiar components alone provides little evidence of conceptual novelty.

An anomaly is model-relative

$$\text{Surprise}(y) = -\log p(y \mid \text{model, experiment})$$

A surprising result is evaluated against a predictive distribution, not merely against a familiar narrative. Large residuals may indicate measurement error, a missing variable, or a wrong mechanism.

First check the observation pipeline; then compare model revisions. A flexible model can absorb an anomaly without learning an explanatory structure.

Record which prediction failed and what additional test would distinguish an instrument problem from a scientific limitation.

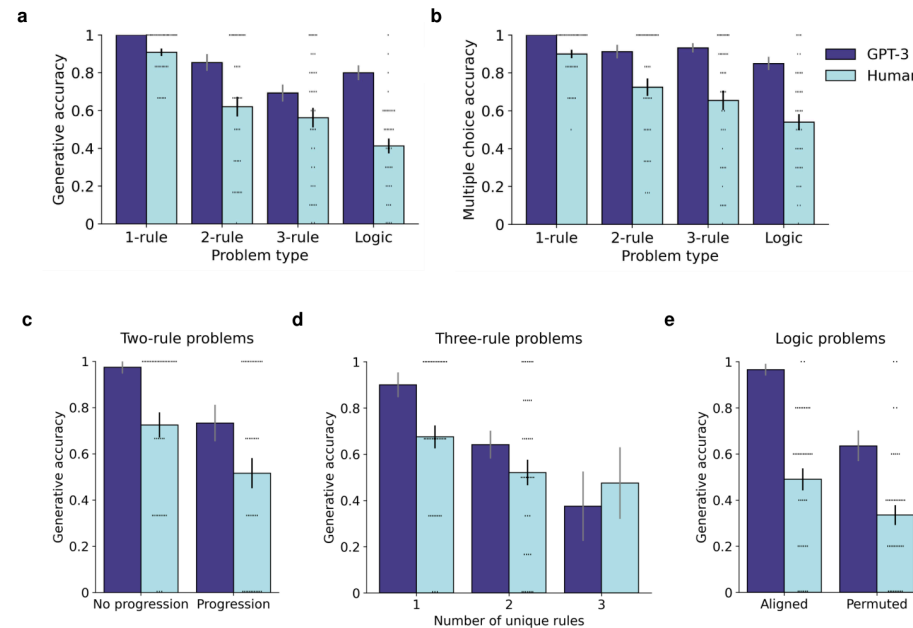
Analogical reasoning: study design

The study evaluates language-model performance on controlled analogical reasoning tasks and compares it with human performance under specified conditions.

Task variants alter relational complexity, response format, and surface representation. These manipulations help distinguish successful pattern use from broad claims about scientific analogy.

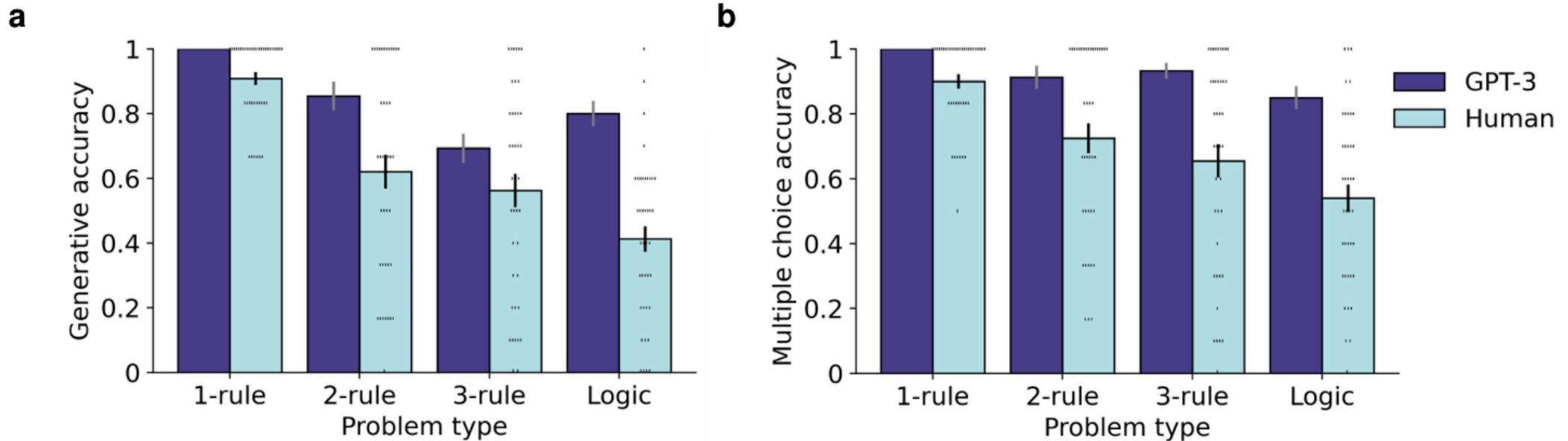
A benchmark result establishes behaviour on the tested problems; transferring the method to mechanism discovery requires an additional operational evaluation.

Analogical reasoning: task-level results



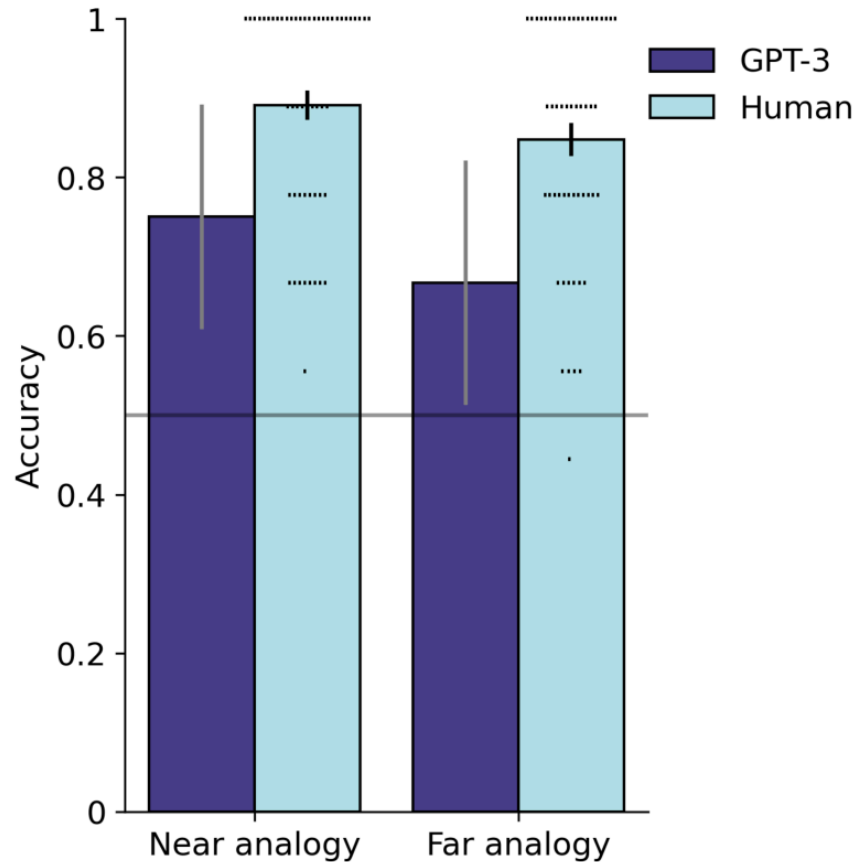
Generative accuracy and multiple-choice accuracy measure different tasks. Rule count and problem family also vary across panels, so the bars do not measure one interchangeable analogical capability.

Analogical reasoning: rule-complexity comparison



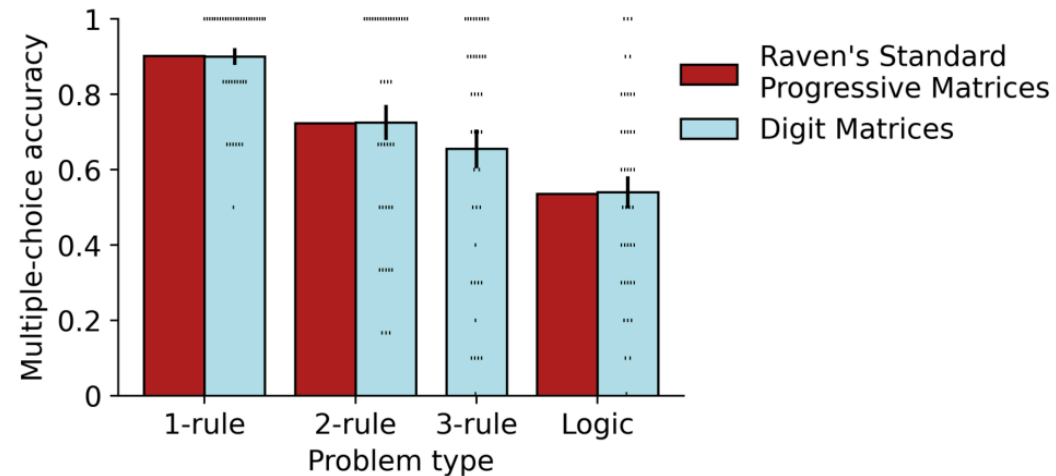
The rule-complexity comparison varies the conditions identified by the bars. An observed performance difference is evidence about those conditions; an explanation of the difference requires additional support.

Evidence: transfer changes the task



The story-analogy task has chance accuracy 0.5. GPT-3 reaches approximately 0.75 on near analogies and 0.67 on far analogies. Human error bars show standard errors and model error bars show binomial confidence intervals. The task concerns analogy identification.

Emergent analogical reasoning: interpretation and limits



Human and model conditions, response formats, training familiarity, and the tested task family limit broad interpretations. Scientific analogy additionally requires testable predictions and evidence about a target system.

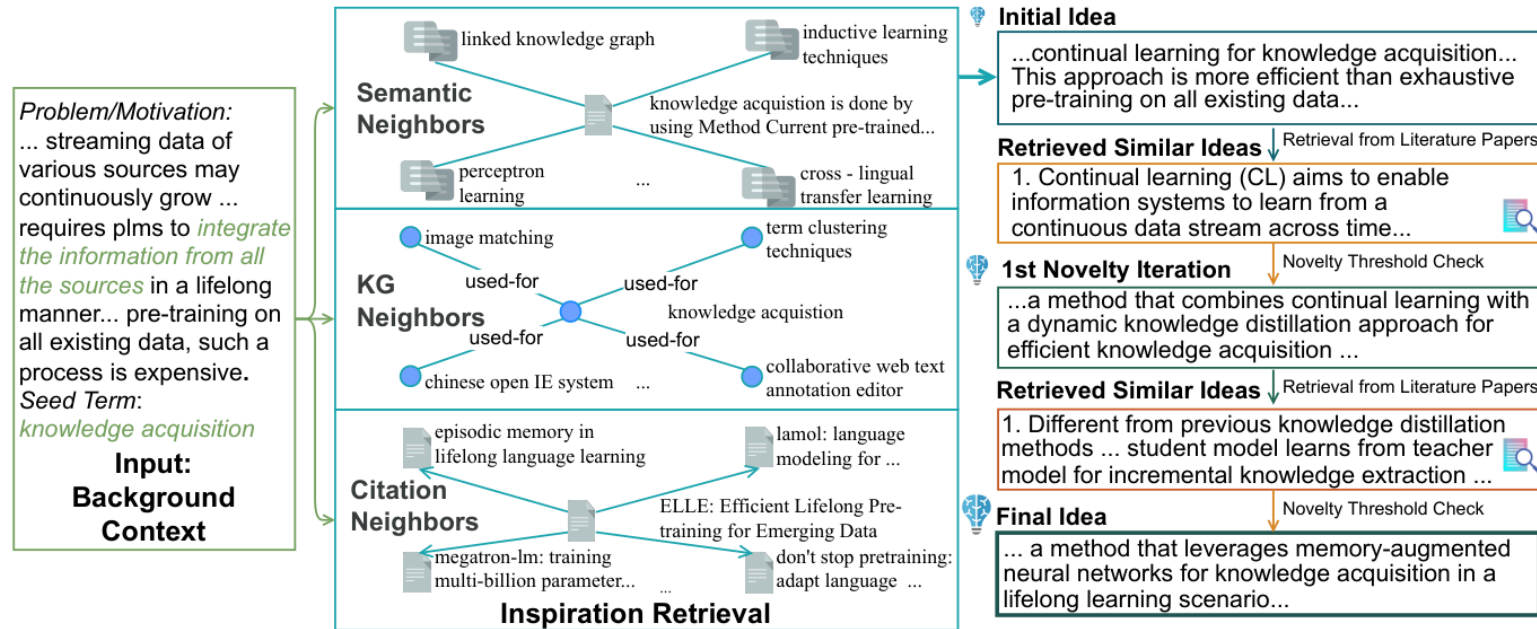
SciMON: the method

SciMON retrieves scientific inspirations from semantic neighbours, knowledge-graph relations, and citation links, then generates a candidate idea from a background problem.

Iterative revision uses a novelty-related comparison against retrieved literature. The design turns retrieval and novelty control into explicit parts of the generator.

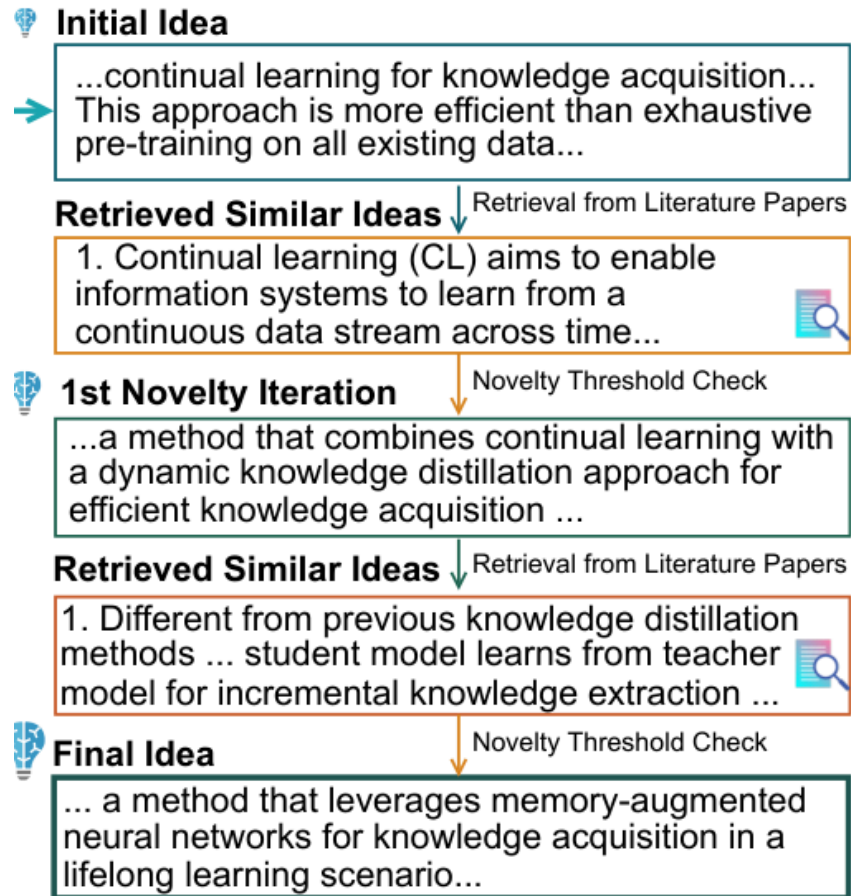
Retrieval quality and the chosen similarity criterion shape which ideas are treated as new, so the novelty objective remains a proxy.

SciMON: retrieval and revision workflow



Background context and three inspiration sources feed iterative idea revision. Retrieval of similar prior ideas can change a proposal. Novel wording and a new scientific relation remain distinct outcomes.

SciMON: mechanism detail



The initial idea is compared with a retrieved predecessor before revision. A changed scientific assumption or relation can support conceptual novelty; different wording alone establishes only textual revision.

Evaluation scope

The study evaluates generated ideas and novelty-oriented revisions under its dataset and human-evaluation protocol.

The table or example compares particular system variants along the stated evaluation dimensions.

Better novelty-related scores indicate improvement under that protocol; they do not independently establish feasibility, conceptual originality against the entire literature, or successful experiments.

SciMON: interpretation and limits

Type	GPT4FS	+SN	+CT	+KG
1st Novelty Δ (%)	+54.4	+55.6	+47.8	+46.7
2nd Novelty Δ (%)	-	+57.8	-	-
1st new terms Δ	+23.1	+22.8	+22.1	+21.9
2nd new terms Δ	-	+21.5	-	-

An incomplete literature index can miss predecessors, while embedding distance can reward superficial changes. Type compatibility, feasibility, and empirical usefulness require additional checks beyond the retrieval objective.

Novelty has several reference frames

Textual novelty concerns wording; bibliographic novelty concerns overlap with known publications; conceptual novelty concerns a new explanatory relation; empirical novelty concerns a supported new observation.

These criteria can disagree. A familiar method applied to an unexplored system may produce valuable empirical knowledge. A highly unusual paragraph may be scientifically empty.

Evaluate novelty alongside plausibility, feasibility, and expected learning rather than treating distance from existing text as a sufficient score.

Diversity is a property of a set

Ten separately plausible ideas can still form a redundant portfolio. Measure whether candidates cover different mechanisms, predictions, or experimental outcomes.

Pairwise embedding distance can screen for repeated wording. A scientific diversity measure should reflect distinctions relevant to the next decision.

A useful ablation compares unconstrained sampling with an operator that explicitly changes one scientific assumption while holding the rest fixed.

A proposal operator must earn its complexity

Choose a fixed task set, literature snapshot, proposal budget, and evaluator.
Compare retrieval alone, unconstrained generation, relation mapping, and typed recombination.

Evaluation distinguishes testable candidates and successful validation from infeasible proposals, duplicated work and the cost of novelty assessment.

If an operator improves judge scores but produces no better experiments, the evidence supports ideation quality rather than discovery performance.

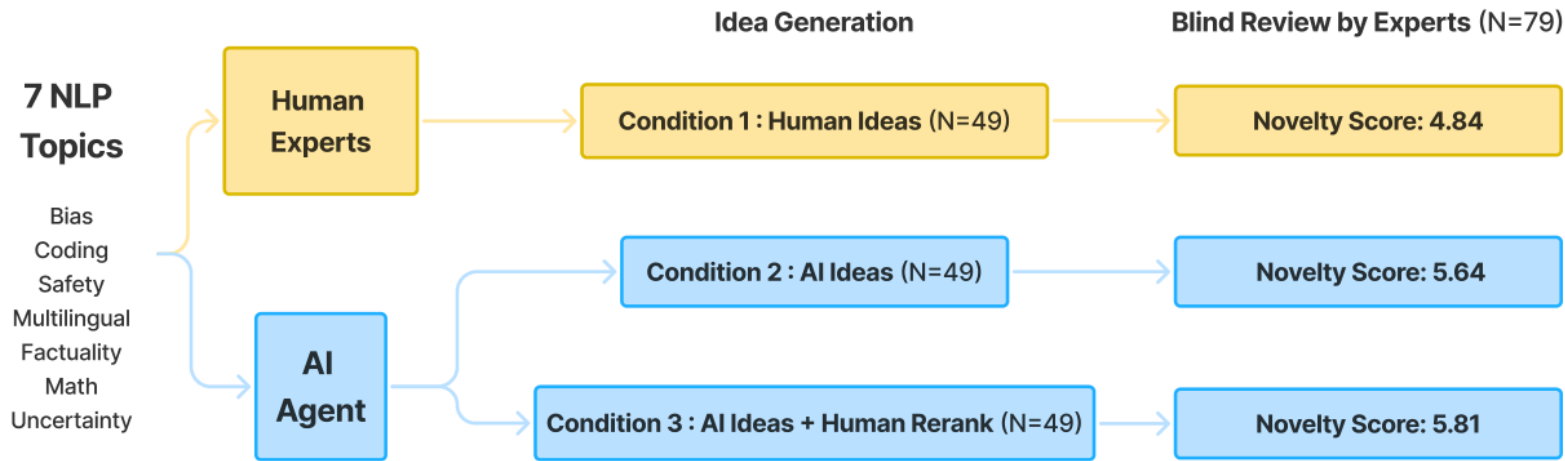
Research ideas: expert evaluation protocol

The study compares human-written and AI-generated NLP research ideas using a structured expert evaluation. A further condition includes human reranking of generated candidates.

Presentation is standardised and reviewers are blinded to the origin of the ideas. Multiple dimensions include novelty, feasibility, and expected effectiveness.

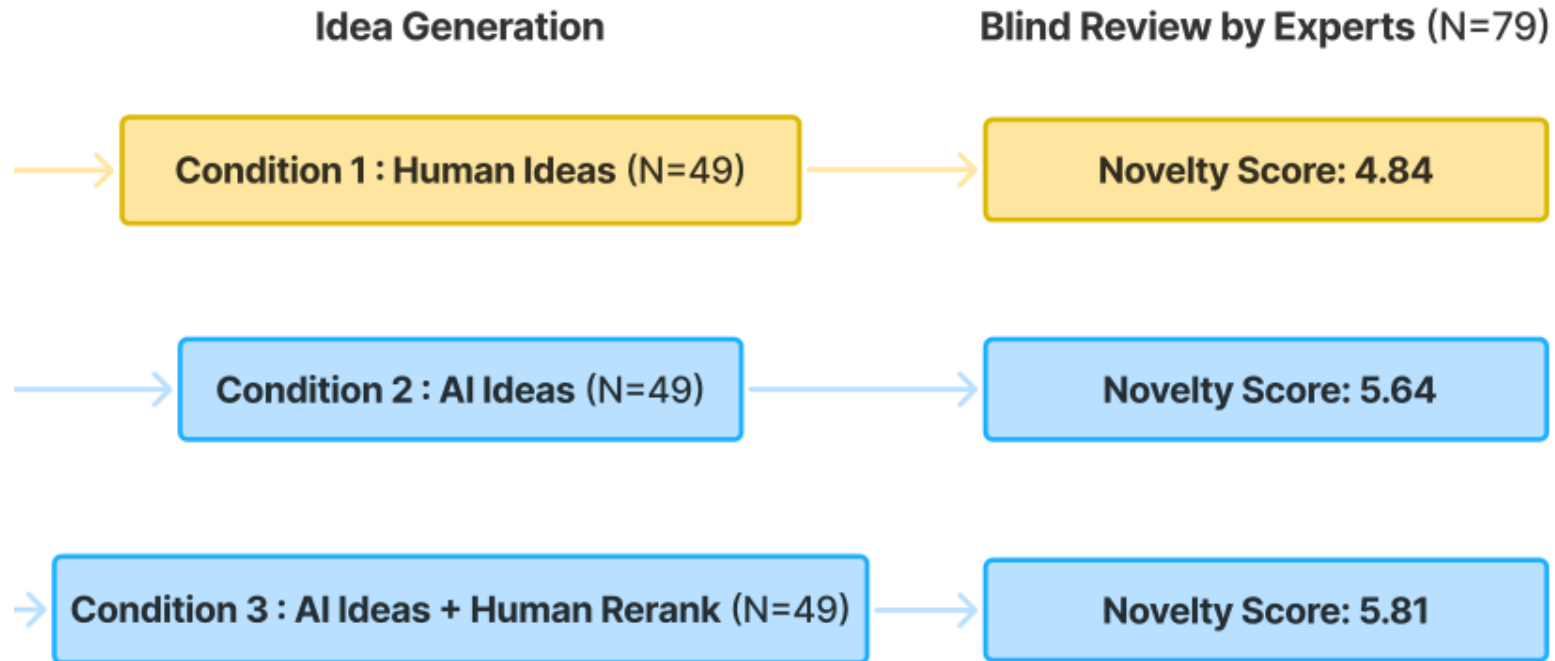
The protocol treats ideation quality as something to measure rather than assuming that a plausible proposal is scientifically valuable.

Research ideas: blinded study design



Three experimental conditions feed into blinded review. Idea authors and idea reviewers are distinct groups. Counts of ideas and counts of reviewers therefore describe different units of analysis.

Research ideas: conditions and novelty ratings



Each of the three idea-generation conditions contributes 49 ideas to the depicted review design. The reported novelty scores describe that comparison; the aggregate scores alone do not display uncertainty, feasibility or execution outcomes.

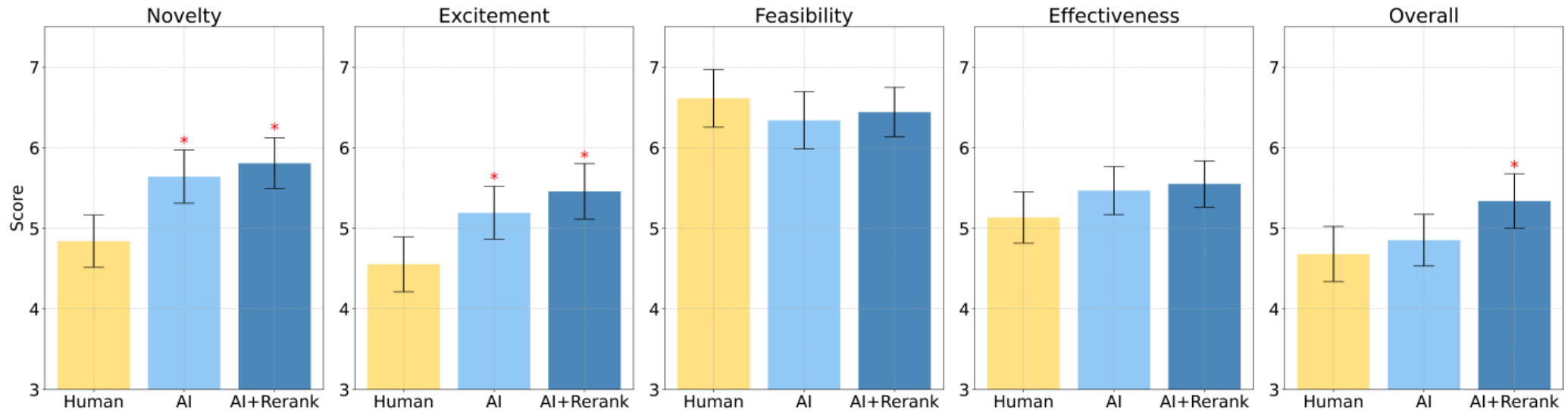
Evaluation scope

The displayed results separate novelty, excitement, feasibility, effectiveness, and overall assessment. AI-generated ideas can score differently on these dimensions.

The evaluation scale and uncertainty intervals constrain the interpretation of differences between conditions.

Expert judgement of an unexecuted idea is informative about perceived quality, but it is not a measurement of the result that its proposed experiment would produce.

Expert evaluation of research ideas: interpretation and limits



The study's topics, participant pool, generation budget, and idea-selection procedure bound the conclusion. Blinding reduces some biases but does not turn subjective ratings into direct empirical validation.

Variable choice precedes hypothesis ranking

A search procedure can compare only hypotheses expressible in its representation. An omitted variable can make several plausible mechanisms observationally indistinguishable.

Problem formulation specifies the observed quantities, the quantities that may vary, and the relation being explained.

Three kinds of revision

Revision	Changed object	Required comparison
New parameter	Value within one model	Fit and uncertainty
New variable	Available state information	Added measurement value
New relation	Functional or causal structure	Discriminating prediction

A more flexible representation may fit past observations better while making less useful future predictions.

Units constrain a proposed relation

Suppose time T depends on length L and a transport coefficient D with units length squared per time. The ratio TD/L^2 is dimensionless.

A proposed explanation should state when this ratio is constant and which boundary conditions justify that claim. Units constrain candidate forms but do not identify the correct mechanism.

An omitted condition can mimic a new law

Condition	Length L	D	$T = L^2 / D$
A	1	1	1
B	2	2	2
C	4	4	4

Synthetic values omit a common geometry factor. T appears proportional to L because D also varies with L. The underlying conditional relation remains quadratic in L.

A prediction difference becomes a likelihood ratio

Operation	Calculation or state	Consequence
Model A	Predicted time 4	Quadratic scaling example
Model B	Predicted time 2	Linear scaling example
Observation	Time 3.8; noise SD 0.2	Residuals -0.2 and 1.8
Log evidence ratio	$(1.8^2 - 0.2^2) / (2 \times 0.2^2) = 40$	Favors A within this model pair

Synthetic Gaussian observation model with fixed predictions and equal noise variance. Evidence compares two specified predictions under matched conditions.

Worked analogy: diffusion across two domains

Constructed analogy: imagine a membrane's delayed response to a concentration change and compare it with heat diffusion through a slab. The example illustrates relational mapping, not a reported experiment.

Propose a mapping: concentration corresponds to temperature, molecular diffusivity to thermal diffusivity, and slab thickness to membrane thickness.

The analogy predicts a characteristic relaxation time proportional to squared thickness. It assumes a diffusion-dominated regime, approximately constant coefficients, and comparable boundary conditions.

A prediction from dimensional analysis

$$\tau \propto L^2/D, \quad \tau(2L)/\tau(L) = 4$$

For diffusivity D with units length squared per time and thickness L with units length, L^2/D has units of time.

This argument suggests a scaling law; it does not determine the numerical prefactor or prove the diffusion equation applies.

Doubling thickness at fixed diffusivity predicts a fourfold characteristic time. A reaction-limited process can violate this relationship and thereby distinguish a competing mechanism.

Translate the analogy into a discriminating test

The illustrative comparison uses otherwise matched membranes with thickness L and $2L$, holding composition, temperature and boundary concentrations comparable.

The comparison depends on the definition of relaxation time, propagation of measurement uncertainty and possible thickness-dependent changes in material structure.

A ratio near four supports this scaling regime; it does not uniquely establish diffusion because other mechanisms may produce the same dependence over a limited range.

An operator-level evaluation protocol

Matched observations and a shared source analogy separate differences in the proposed mapping, its assumptions and its discriminating predictions.

Blinding reduces system-identity effects in evaluation. Relational consistency, feasibility and novelty remain separate judgments, while execution provides additional evidence within the specified public-data or simulation setting.

Retrieval and dimensional-analysis baselines distinguish a proposed reasoning contribution from access to an appropriate source.

Timescale implied by the boundary problem

$$\begin{aligned}\partial_t u &= D \partial_{xx} u, & u(0, t) &= u(L, t) = 0 \\ u(x, t) &= A \sin(\pi x / L) e^{-t/\tau}, & \tau &= L^2 / (\pi^2 D)\end{aligned}$$

Let u be departure from equilibrium, D constant diffusivity, L slab thickness, and A the initial mode amplitude. The endpoint departures are fixed at zero.

Substitute the sine mode: its second spatial derivative is $-\pi^2 u / L^2$ and its time derivative is $-u / \tau$. Equating them fixes the prefactor.

At long times, this slowest mode dominates if present initially. Different boundary conditions or transport mechanisms can change the result.

A scaling exponent does not identify a mechanism

$$\alpha = \frac{d \log \tau}{d \log L} = 2 - \frac{d \log D}{d \log L} = 2 - \beta$$

Let $D(L)=D_0(L/L_0)^\beta$ at reference length L_0 . With fixed boundary conditions, logarithmic differentiation gives the time exponent $\alpha=2-\beta$.

Both constant-speed transport and diffusion with $\beta=1$ predict $\alpha=1$. A measured twofold time ratio cannot distinguish these explanations.

Independent diffusivity information and observations at additional lengths distinguish the proposed scaling law from changes in transport properties. The failed fixed-D prediction remains part of the evidence.

An analogy needs a relation-preserving map

Source: heat transport	Target: molecular transport	Candidate correspondence
Temperature field $T(x,t)$	Concentration field $c(x,t)$	A scalar varies over space and time
Heat flux proportional to $-\nabla T$	Molecular flux proportional to $-\nabla c$	Flux follows a local gradient
Conservation + constitutive law	Conservation + constitutive law	A diffusion equation can result
Insulated boundary	Impermeable boundary	Zero normal flux in both models

A lecturer-created mapping: shared equations suggest a transferable method only after the physical assumptions have been checked.

Where the analogy stops working

Changed target condition	Why the old mapping breaks	Consequence
Bulk flow is important	Transport includes advection	Add the flow term before transferring results
Molecules react or are consumed	Local conservation has a source or sink	Estimate reaction and diffusion jointly
Transport is actively driven	Gradient-only flux can be wrong	A matching diffusion fit may mislead
Only a narrow time range is observed	Different mechanisms can fit equally well	Choose a discriminating measurement

The useful transfer is conditional: a mathematical resemblance proposes a model; a new observation tests its applicability.

Recent methods in context

ResearchBench (ACL Findings 2026). Separates inspiration retrieval, hypothesis composition and ranking in literature-derived scientific tasks. Retrospective reconstruction and a publication-date filter do not establish prospective novelty or eliminate contamination.

Model Discovery Agent (preprint; revised 2026). Combines Bayesian model/parameter inference, LLM proposals of model structures and value-of-information experiment choice. Version 4 reports fixing prompt leakage and rerunning benchmarks; comparisons must use the corrected version.

AutoDiscovery (NeurIPS 2025). AutoDiscovery combines hypothesis exploration with Bayesian surprise. It links an observed belief change to further search, while surprise and the prospective value of a future experiment remain different quantities.

References (1/2)

[Dedre Gentner \(1983\). Structure-Mapping: A Theoretical Framework for Analogy](#)

[Taylor Webb, Keith J. Holyoak, Hongjing Lu \(2022\). Emergent Analogical Reasoning in Large Language Models](#)

[Qingyun Wang, Doug Downey, Heng Ji et al. \(2023\). SciMON: Scientific Inspiration Machines Optimized for Novelty](#)

[Chenglei Si, Diyi Yang, Tatsunori Hashimoto \(2024\). Can LLMs Generate Novel Research Ideas? A Large-Scale Human Study with 100+ NLP Researchers](#)

[Tarun Gupta, Danish Pruthi \(2025\). All That Glitters is Not Novel: Plagiarism in AI Generated Research](#)

References (2/2)

[Chenglei Si, Zitong Yang, Yejin Choi et al. \(2026\). Towards Execution-Grounded Automated AI Research](#)

[Yujie Liu, Zonglin Yang, Tong Xie et al. \(2025\). ResearchBench: Benchmarking LLMs in Scientific Discovery via Inspiration-Based Task Decomposition](#)

[Kevin Murphy \(2026\). Model Discovery Agent: LLM-assisted Bayesian experiment design for data-efficient discovery of mechanistic world models](#)

[Dhruv Agarwal, Bodhisattwa Prasad Majumder, Reece Adamson et al. \(2025\). AutoDiscovery: Open-ended Scientific Discovery via Bayesian Surprise](#)