THE SCIENTIFIC RESEARCH TASTE BENCHMARK (SCITASTE-BENCH): A MULTI-DISCIPLINARY BENCHMARK FOR ASSESSING LARGE LANGUAGE MODELS' JUDGMENT OF RESEARCH TASTE IN SCIENTIFIC PAPERS

Anonymous authors
Paper under double-blind review

ABSTRACT

Whether large language models can assess the innovativeness of academic publications is important in literature reviews, research funding screening, and academic recommendations. We introduce the **Scientific Research Taste Benchmark (SciTaste-Bench)**, a gold-standard benchmark covering **six disciplines**: physics, mathematics, chemistry, biology, medicine, and finance. SciTaste-Bench measures research taste, that is, the ability to judge the scientific value and potential of a paper, of which innovativeness is only one dimension among the seven assessed. SciTaste-Bench integrates metadata from NASA ADS, OpenAlex, and arXiv, aligns papers to field-specific classification systems, and evaluates each paper using the seven-dimensional *Academic Taste Index*. The weights for each dimension are derived from a bi-objective machine-learning derivation (five regressors, including random forests) and are *shared across all six disciplines*. This benchmark includes **300 selected papers (50 per discipline)** and a backup pool of **5,700 papers** (950 papers per field) covering 1996–2023. To allow external knowledge to influence model decisions, SciTaste-Bench supports four evaluation modes: `pure_qa`, `native_search`, our self-built structured tool-calling agent DIMAGENT-V3 (9 bibliometric tools in ADS/OpenAlex), and `dimagent_v3_native`. Model outputs are evaluated with two complementary schemes: one combines correlation with dimension-level accuracy, and the other is based on the WAPE relative error. In the evaluation of ten state-of-the-art large language models, the tool-based structured evaluation generally outperformed the toolless and raw web search modes, but the improvement was model-dependent. Physics is the easiest subject, while chemistry, biology, and medicine are still more challenging. By measuring how well large language models judge the value potential of academic work across fields, SciTaste-Bench supplies the value-judgment substrate that auto-research systems need to screen, rank, and prioritize candidate ideas. We release the dataset, the evaluation harness, and the scoring code to support reproducible research on scientific-innovation assessment; all data and code will be made publicly available on Hugging Face.

Keywords: large language models; research-paper innovativeness; scientific benchmark; scientometrics; multi-disciplinary evaluation

1 INTRODUCTION

The evaluation of the innovativeness of research papers sits at the intersection of scientometrics and natural language processing. Citation-based metrics have a long history (Funk & Owen-Smith, 2017; Park et al., 2023), and large language models are already used to produce scientific feedback (Liang et al., 2024), generate and rank research ideas (Si et al., 2025), and act as reviewers or

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

judges (Zheng et al., 2023; Nguyen & Ahmadi, 2026). Most existing evaluations, however, focus on a single discipline—often physics or NLP—tend to use small or loosely structured samples, and rarely control whether the model has access to external knowledge. Three questions therefore remain open: whether a model's ability to judge innovativeness varies significantly across disciplines; how its performance changes when it is equipped with retrieval tools, native web search, or structured bibliometric database queries; and whether models show systematic bias in data-sparse or sensitive domains such as biomedicine and the social sciences.

SciTaste-Bench is designed to address these questions. The main contributions are:

1. **A unified and reproducible construction pipeline** that integrates multi-source metadata, classifies papers by discipline-specific taxonomies, and scores them with a seven-dimensional Academic Taste Index whose machine-learning-derived weights are uniform across all six disciplines (Sec. 3).
2. **Cross-discipline comparability under one fixed weight vector** covering physics, mathematics, chemistry, biology, medicine, and finance, with 300 selected papers and 5,700 backup papers (950 per discipline) under unified metadata (Sec. 4).
3. **DIMAGENT-V3, a structured tool-calling agent we designed and built**, exposing nine bibliometric tools over NASA ADS and OpenAlex and so isolating the effect of structured external knowledge (Sec. 5).
4. **A dual scoring protocol** combining a correlation/dimension composite with WAPE relative-error grades, plus an empirical study of ten LLMs across four modes (Sec. 6, 7).
5. **A foundation for auto-research:** because SciTaste-Bench provides fixed, gold-standard judgments of scientific value across six disciplines, it can serve as the value-prioritization substrate for autonomous research systems.

2 RELATED WORK

Scientometric measures of impact and disruption. Citation dynamics is heavy-tailed, often described as a discretised lognormal or a hooked power law depending on field and age (Thelwall, 2016). Beyond raw counts, the CD index and its variants measure whether a paper consolidates or disrupts existing knowledge (Funk & Owen-Smith, 2017), and large-scale analysis shows papers and patents becoming less disruptive over time (Park et al., 2023).

LLMs for scientific evaluation. Large language models can generate plausible peer-review feedback (Liang et al., 2024) and novel research ideas (Si et al., 2025), LLM-as-a-judge paradigms are now common in open-ended evaluation (Zheng et al., 2023), and a growing literature surveys the reliability of LLM-based peer review (Nguyen & Ahmadi, 2026). We complement this line with a structured, multi-discipline gold-standard set carrying fixed numerical ground truth rather than free-text review judgments.

Tool-augmented LLMs. Tool calling over large numbers of real-world APIs has been studied at scale (Qin et al., 2024). DIMAGENT-V3 applies the idea to bibliometrics: instead of open web search it uses nine deterministic tools to retrieve raw metadata from curated academic databases, and the model must reason from that evidence to produce a score.

3 CONSTRUCTION OF THE SIX-DISCIPLINE EVALUATION SET

3.1 OVERALL ARCHITECTURE

Figure 1 illustrates the pipeline: the indexing layer collects records from three sources, the fusion layer resolves conflicts, the classifier assigns discipline and sub-discipline labels, the index engine computes the seven-dimensional index, and a final quality checker separates papers into the selected evaluation set (300 papers) and the backup set (5,700 papers).

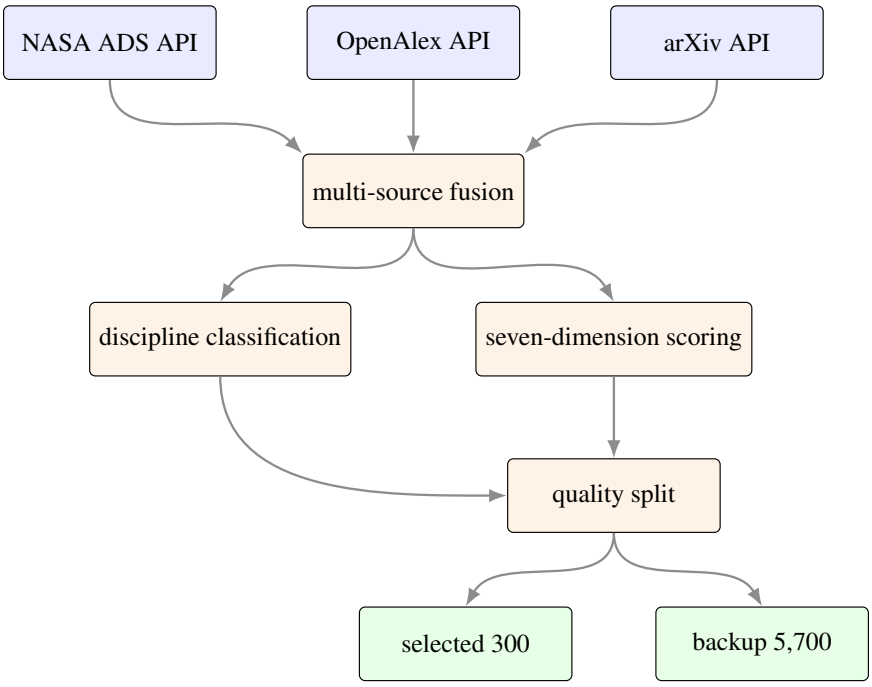

Figure 1: The data collection and routing architecture for the evaluation set fuses three sources, classifies and scores them on seven dimensions, and then splits the data into the selected evaluation set and the backup evaluation set.

3.2 Multi-source data collection

Three sources feed the fusion layer. NASA ADS is the authoritative backbone for physics (Accomazzi, 2024) and supplies `bibcode`, `title`, `abstract`, `doi`, `bibstem`, `year`, `citation_count`, `authors`, `references`, and `orcid_pub`. OpenAlex is the backbone for the five non-physics disciplines (Sec. 4) and contributes what ADS lacks (Priem et al., 2022): the full citing and referenced-work lists that the CD index needs for disruptiveness, the Concepts/Topics taxonomy for conceptual innovation, and the authorship records with institution and country for team collaboration. The arXiv XML API adds preprint metadata, whose subject-class field is the highest-priority signal for discipline classification. When the sources disagree we apply fixed precedence rules: title, publication year, and citation count follow ADS; abstracts follow ADS $>$ OpenAlex $>$ arXiv; discipline labels follow arXiv $>$ journal inference $>$ keyword matching $>$ title matching.

3.3 Discipline classification

Each academic discipline is organized according to a second-level sub- discipline taxonomy (Table 4 in Appendix A). This taxonomy is aligned with the graduate-education discipline catalogue while being granular enough to provide each sub-discipline with a roughly balanced pool of papers. Because physics and the other five disciplines rely on different backbone databases, they use two different labeling mechanisms.

Physics: a four-stage cascade over ADS/arXiv metadata. Eight categories are defined in physics: classical, condensed matter, quantum, applied, high energy, computational, and interdisciplinary physics, plus astrophysics. A paper's label is set by a four-stage cascade (Table 3 in Appendix A): arXiv subject classification first, then matching of the publication venue against discipline-characteristic journals, then controlled keywords in the title and abstract, and finally title matching; the first stage that fires decides. Papers that fail all four stages remain unclassified and are routed to the backup evaluation set.

Table 1: The seven dimensions of the Academic Taste Index and their final normalized weights.

ID	Dimension	What it measures	Weight
A	Citation Dynamics	post-publication attention and impact	29.5%
B	Knowledge Breadth	breadth/diversity of integrated references	11.8%
C	Semantic Novelty	semantic distinctiveness relative to prior work	11.6%
D	Disruptiveness	degree of breaking with the existing literature	8.9%
E	Conceptual Innovation	novelty of concept/topic combinations	12.9%
F	Team Collaboration	internationalization and scale of the team	10.1%
G	Journal Impact	venue prestige (IF / JCR / CAS partition)	15.3%

The other five disciplines: taxonomy-driven collection via OpenAlex. Mathematics, chemistry, biology, medicine, and finance have no equivalent of arXiv's subject classes, so their labels are assigned *by construction* rather than post hoc. Each sub-discipline in Table 4 is anchored to an OpenAlex subfield identifier and a set of controlled keywords describing the area: article-type records whose primary topic belongs to that subfield and whose text matches the keywords are collected in descending citation order, and the subfield filter alone back-fills the pool when the keyword search returns too few records. After cross-sub-discipline deduplication, in which each paper belongs to only one sub-discipline, roughly 125 papers per sub-discipline are retained by stratified sampling across composite-score quartiles and labeled by the collection cell they were drawn from. This avoids between-sub-discipline misclassification while the subfield filter and keyword query preserve topical purity within each cell. The result is eight sub-disciplines for mathematics, chemistry, biology, and finance, and seven for medicine.

3.4 Seven-dimensional Academic Taste Index

The core scoring engine assigns each paper a 0–100 innovation score as a weighted sum over seven interpretable dimensions.

$$\text{innovation_score} = \sum_{i=1}^{7} W_i \cdot \text{DimensionScore}_i, \qquad \sum_i W_i = 1.0, \tag{1}$$

where each $\text{DimensionScore}_i \in [0, 100]$. The same seven dimensions and the same weight vector are applied identically to every discipline, and this is what makes cross-discipline comparison possible. Table 1 lists the dimensions, what they measure, and their final weights.

The weight vector is itself data-driven. The training corpus is the 998-paper quantum-physics V5 evaluation dataset from an earlier single-discipline study. Two regression analyses were run: one predicting citation impact from the other six dimensions, and one predicting journal impact from the other six. Each analysis trained five regressors—linear, ridge (α=1.0), lasso (α=0.1), random forest (100 trees), and gradient boosting (100 trees)—on standardized dimension scores with an 80/20 train–test split (seed 42) and five-fold cross-validation. Per-model feature contributions (absolute standardized coefficients for linear models, feature importances for tree ensembles) were averaged across the five models; dimensions shared by the two analyses received the mean of their two contributions, and the raw weights were normalized to sum to one, yielding the final vector in Table 1. The random forest attained a test R^2 of 0.736 for citation impact but only 0.170 for journal impact, indicating that citation behavior is largely predictable from the remaining dimensions whereas journal placement is comparatively independent.

From feature contributions to weights (dual targets). Which dimensions deserve the largest weights? In everyday scholarly practice two community-sanctioned outcomes shape perceptions of a paper's value: how often it is cited and where it is published. Citation impact (A) and journal impact (G) therefore each act as the prediction target of one of the two analyses, and a target cannot be weighted from its own analysis: A's weight is taken from its contribution to predicting journal impact (37.99%) and G's from predicting citation impact (19.73%), while dimensions B–F use the arithmetic mean of their contributions across the two analyses. The raw weights sum to 128.86% and are renormalized to 100% (Table 5 in Appendix B).

3.5 DIMENSION-TO-SCORE MAPPING

Each dimension's raw sub-parameters are mapped onto a 0–100 score; the complete formulas and parameter lists are given in Appendix C. Citation dynamics blends a log-normalized citation count with an empirical percentile (0.6/0.4). Knowledge breadth, semantic novelty, disruptiveness, conceptual innovation, and team collaboration are each built from several secondary parameters that are imputed, standardized, projected onto the first principal component, and mapped to 0–100 by mean-rank normalization. Journal impact is a fixed-weight blend of the impact factor (70%) and the JCR and CAS partition schemes (15% each).

4 SIX-DISCIPLINE COVERAGE AND UNIFICATION

SciTaste-Bench covers physics, mathematics, chemistry, biology, medicine, and finance. Physics is collected through NASA ADS and organized into eight sub-disciplines (Sec. 3); the five non-physics disciplines are aligned to the same framework in five stages: (1) define a second-level taxonomy aligned with the graduate-education discipline catalogue; (2) collect candidate papers through SCI replacement and OpenAlex retrieval, align them by title/DOI/bibcode, and assign an internal work identifier linked to the corresponding OpenAlex record when no ADS bibcode exists; (3) recompute the seven-dimensional scores with the unified scoring pipeline; (4) apply stratified sampling and manual verification; and (5) produce unified output for cross-discipline merging.

Two principles keep the merged set comparable. **Weight consistency:** all six disciplines share the same weight vector (Table 1), making their scores comparable. **Field mapping:** Chinese dimension names are mapped to stable English keys (`citation_impact`, `journal_impact`, etc.), raw dimension scores are scaled to 0–100, and the composite score is recomputed. Within each discipline the 50 selected papers cover the major second-level sub-disciplines, balance journal tiers Q1–Q4 under a combined CAS-2025 and JCR-2025 tier map, emphasize 2000–2020 papers with relatively stable citation data, and span the score range so that the set retains discriminative power; each sample is then manually checked for title, abstract, discipline, and journal consistency.

The released data set consists of a *selected* split (300 papers, 50 per discipline) and a *backup* split (5,700 papers, 950 per discipline) that supports refresh and re-sampling; Table 6 in Appendix D summarizes both. Each selected record is additionally assigned `discipline`, `subcategory_id`/`name`, `journal_tier`, `journal_name`, and `openalex_id` to ensure reproducibility, and a complete example record—with the ground-truth total and the seven per-dimension scores—is given in Appendix D.

5 TOOL-CALLING AGENT DIMAGENT-V3

5.1 DESIGN

In agent mode, the model does not evaluate papers based solely on parametric memory. Instead, it issues *tool calls* to nine special tools. DIMAGENT-V3 is not an off-the-shelf component: the agent framework was designed and implemented in-house by the authors—the nine-tool interface, the two-round call-then-score loop, and the reviewer system prompt are all our own—rather than adapted from an existing agent codebase. Each tool queries NASA ADS or OpenAlex for non-physics domains in real-time and returns *raw* evidence rather than precomputed results. The model must integrate this evidence and decide the score itself based on it. Since the dataset provides `bibcode`, a separate search step is not required.

5.2 THE NINE TOOLS

Table 7 in Appendix E lists each tool, the raw data it returns, and the dimension it supports; each tool receives one `bibcode` argument.

5.3 AGENT LOOP

The loop is deliberately compact (Figure 3 in Appendix E). The system prompt casts the model as a senior academic reviewer and prescribes three steps: run all nine tools in parallel, analyze the returned raw data, then score each dimension 0–100 and calculate a weighted composite score. Given a `bibcode`, nine simultaneous calls are sent in Round 1; after the raw data payload returns, Round 2 outputs an exact JSON object containing the composite score, seven dimension scores, a brief reasoning trace, and self-reported confidence. This isolates structured bibliometric data as the only external input.

5.4 BACKEND ROUTING

Physics papers are requested directly from NASA ADS. Non-physics papers are routed through OpenAlex: `openalex_bibcode_map` converts the internal identifier to an OpenAlex work ID before calling the OpenAlex API. To comply with OpenAlex's rate limits, we enforce a minimum request interval and include a `mailto` field in the User-Agent to join the polite pool.

6 EVALUATION PROTOCOL AND SCORING

6.1 FOUR EVALUATION MODES

SciTaste-Bench separates the data set from the evaluation harness and supports four complementary modes (Table 8 in Appendix E). The two `dimagent` modes are driven by DIMAGENT-V3, the agent we designed and built for this benchmark (Sec. 5), so the tool interface and the scoring prompt are held fixed across models; the verbatim prompt templates and a fully worked example are in Appendix J.

6.2 SCORING SCHEME 1: CORRELATION + DIMENSION ACCURACY

This scheme combines an overall and a detailed assessment. Part A (50%) measures the Spearman coefficient ρ (15%), the Pearson coefficient r (10%), mean absolute and mean square error (15%), and |Bias| (10%). Part B (50%) estimates the seven dimensions through the weighted mean Spearman ρ (30%) and the weighted mean absolute error (20%), using the weights of Table 1. The final score is Part A $\times$ 50% + Part B $\times$ 50%; grades are ≥ 80 excellent, ≥ 60 moderate, ≥ 40 poor, and < 40 unusable.

6.3 SCORING SCHEME 2: RELATIVE ERROR (WAPE)

The simple mean absolute percentage error $\frac{1}{n}\sum_i |p_i - g_i|/g_i$ is dominated by papers with low true values, so we adopt the weighted absolute percentage error used in forecasting evaluation (Makridakis et al., 2022):

$$\text{WAPE} = \frac{\sum_i |p_i - g_i|}{\sum_i g_i}. \tag{2}$$

WAPE weights the total error by the true values, preventing poorly scored papers from dominating the result. It maps to a 0–100 score by grade interval: A ($< 5\%$), B (5–10%), C (10–20%), D (20–30%), E (30–50%), and F ($\geq 50\%$) (e.g., $< 5\% \rightarrow$ 90–100).

7 EXPERIMENTS AND RESULTS

7.1 SETUP

We evaluated ten frontier large language models (qwen-3.8-max, qwen3.7-plus, qwen3.7-max, claude-fable-5, claude-opus-4-8, gpt-5.6-sol, gemini-3.5-flash, kimi-k3, Doubao-Seed-2.1-Pro, DeepSeek-V4-Pro) through a unified API gateway, under all four modes on the 300-paper six-discipline set, with reasoning or thinking enabled where supported: high for the Qwen and GPT

models, adaptive for the Claude models, standard for Gemini, Kimi, Doubao, and DeepSeek. For each model–mode combination we recorded the percentage of successfully scored papers.

7.2 Main results

Table 2 reports the Scheme-1 and Scheme-2 scores for all ten models; gpt-5.6-sol was fully re-run after a parameter-conflict fix (Sec. 7.5) and gemini-3.5-flash is newly added in this round. Structured tool grounding (`dimagent_v3`) remains the strongest single intervention: for the three Qwen models and claude-fable-5, Scheme-2 scores rise from roughly 20–33 in the tool-free modes to 47–52 (grade E $\rightarrow$ D), whereas claude-opus-4-8 improves only to 34.5 (E) in the pure tool mode, its combined mode being its best configuration (40.7). There are two exceptions: gpt-5.6-sol's Scheme-2 score collapses to 10.6 (F) under `dimagent_v3` but recovers to 28.7 (E) when native search is added, and gemini-3.5-flash is the weakest overall, scoring F in all four modes with only a marginal gain from structured tools. Among the newly added models, kimi-k3 stays at grade E throughout (structured tools lift Scheme 2 from 25.0 to 33.4); Doubao-Seed-2.1-Pro is the strongest newcomer, reaching D (44.0) in its best configuration, the combined mode; and DeepSeek-V4-Pro is the least sensitive to external grounding—its tool-free `pure_qa` score (30.2) is already its best and structured tools add little (34.5), so its parametric prior is comparatively well matched to the task while bibliometric grounding fails to compound.

7.3 Cross-discipline analysis

Table 9 in Appendix G breaks down the `dimagent_v3` mode by discipline for qwen-3.8-max. Physics is by far the easiest (ρ=0.78, WAPE 13.1%, Scheme-2 70.4, grade C), reflecting richer parametric coverage and cleaner ADS grounding. Finance is the strongest non-physics discipline. Medicine has the weakest correlation (ρ=0.31), while chemistry and biology have the largest relative error (WAPE $\approx$ 29%). The cross-model breakdown (Table 10 in Appendix G) shows that the physics advantage is shared by most models: the three Qwen models, claude-fable-5, and kimi-k3 all reach Scheme-2 between 46.9 and 70.7 on physics. Doubao-Seed-2.1-Pro inverts this pattern, scoring its best in finance (58.6) and medicine (49.8) and its worst in physics (15.8), a reminder that the discipline gap is partly a property of each model's training mix rather than of the benchmark. The claude-opus-4-8 physics cell (6.3) is an instability of that model's tool run (ρ=$-$0.16, WAPE 74.4%), not a reflection of discipline difficulty.

7.4 Findings

Structured tools beat raw web search. Across models, `dimagent_v3` is usually the strongest or near-strongest mode. Native search alone barely improves `pure_qa` and is sometimes worse: unstructured web results mix signal with noise, which often makes the model alternate between the paper's own abstract and distracting external hits, whereas structured tools return more precise bibliometric evidence better matched to this task. The advantage is not universal—gpt-5.6-sol collapses in the pure tool setting and gemini-3.5-flash improves only marginally (Sec. 7.5).

Adding native search to tools rarely helps. For the three Qwen models, `dimagent_v3_native` tracks `dimagent_v3` closely, within about one point on both schemes. The pattern diverges for the Claude models: claude-fable-5 drops from 47.0 to 39.3 on Scheme 2 while claude-opus-4-8 rises from 34.5 to 40.7, making the combined mode its best configuration. The most extreme case is gpt-5.6-sol, where native search pulls the model back from a collapsed pure-tool run (Scheme-2 10.6) to 28.7, its only grade-E configuration; for gemini-3.5-flash the two tool modes are essentially tied. Among the newly added models, Doubao-Seed-2.1-Pro is the clearest beneficiary of stacking search on tools (39.2 $\rightarrow$ 44.0, E $\rightarrow$ D), whereas kimi-k3 (33.4 $\rightarrow$ 30.1) and DeepSeek-V4-Pro (34.5 $\rightarrow$ 33.7) gain nothing from the combination.

Discipline gap persists even with tools. Physics holds a large advantage in most models and modes while biology and medicine remain difficult, pointing to uneven training coverage and to the limits of ADS-style grounding outside physics. The collapsed physics sub-run of gpt-5.6-sol's combined mode is a notable exception (Sec. 7.5).

Table 2: Main results on the six-discipline 300-paper set. The Scheme-2 grade is given in brackets; superscripts mark run-level issues described below.

Model	Mode	Scheme 1	Scheme 2	Grade	Success
qwen-3.8-max	pure_qa	21.10	28.10	E	300/300
	native_search	23.20	31.60	E	300/300
	dimagent_v3	28.00	51.80	D	300/300
	dimagent_v3_native	27.10	50.30	D	300/300
qwen3.7-plus	pure_qa	19.40	23.20	E	300/300
	native_search	19.80	20.80	E	300/300
	dimagent_v3	27.80	47.60	D	300/300
	dimagent_v3_native	27.70	48.10	D	300/300
qwen3.7-max	pure_qa	19.30	20.10	E	300/300
	native_search	20.20	21.10	E	300/300
	dimagent_v3	27.60	47.70	D	300/300
	dimagent_v3_native	27.40	47.20	D	300/300
claude-fable-5	pure_qa	22.30	27.60	E	269/300[†]
	native_search	24.60	32.70	E	280/300[†]
	dimagent_v3	26.80	47.00	D	289/300[†]
	dimagent_v3_native	25.20	39.30	E	291/300[†]
claude-opus-4-8	pure_qa	20.20	25.50	E	300/300
	native_search	21.90	27.00	E	300/300
	dimagent_v3	23.70	34.50	E	300/300
	dimagent_v3_native	25.20	40.70	D	300/300
gpt-5.6-sol	pure_qa	19.50	19.20	F	300/300
	native_search	21.10	19.50	F	300/300
	dimagent_v3	9.50	10.60	F	300/300
	dimagent_v3_native	22.20	28.70	E	300/300
gemini-3.5-flash	pure_qa	16.90	16.40	F	300/300
	native_search	16.40	16.50	F	113/300[‡]
	dimagent_v3	19.90	17.70	F	300/300
	dimagent_v3_native	19.70	17.60	F	300/300
kimi-k3	pure_qa	21.40	25.00	E	298/300[†]
	native_search	23.20	28.90	E	300/300
	dimagent_v3	24.00	33.40	E	281/300[†]
	dimagent_v3_native	24.70	30.10	E	300/300
Doubao-Seed-2.1-Pro	pure_qa	20.70	28.00	E	300/300
	native_search	22.50	28.60	E	300/300
	dimagent_v3	24.40	39.20	E	300/300
	dimagent_v3_native	25.50	44.00	D	300/300
DeepSeek-V4-Pro	pure_qa	19.10	30.20	E	300/300
	native_search	20.00	23.40	E	300/300
	dimagent_v3	24.00	34.50	E	300/300
	dimagent_v3_native	23.60	33.70	E	300/300

[†] Gateway issues: empty or non-JSON responses and HTTP 429 rate limiting; affected records were dropped from the metrics, not re-run. [‡] Mass failures: the 187 incomplete `native_search` runs span all six disciplines and never emitted a parseable scoring JSON. Rows with fewer than 300 successes are averaged over successful records only.

Systematic over- and under-estimation. In the tool-free mode a large positive bias appears: models tend to over-rate papers. Structured tool use reduces this bias and, in physics, can even make it slightly negative. The direction can still reverse—in the pure tool run of gpt-5.6-sol a strong negative bias occurred (-20.2) and the median per-paper relative error reached 100%, so more than half of the papers are severely under-rated in that mode (mechanism in Sec. 7.5). Scheme-1 scores overall remain low because the weight-averaged dimensional correlation (Part B) depends heavily on three challenging dimensions—conceptual_innovation, semantic_novelty, and disruptiveness—whose per-dimension correlation is near zero in every evaluation mode.

7.5 Engineering notes and the gpt-5.6-sol case

Several practical issues arose. **Gateway issues:** the gateway sometimes returned empty or non-JSON responses, or rate-limited requests until the retry budget was exhausted (HTTP 429), as for claude-fable-5 on a subset of biology, medicine, and chemistry titles and for kimi-k3 under `dimagent_v3`. Such records were dropped from the metrics rather than scored as zero, so the affected success counts remain under 300, and they are reported separately rather than treated as a model weakness. **Mass failures:** gemini-3.5-flash completed only 113/300 items under `native_search`, with failures spread across all six disciplines; in those runs the model never emitted a parseable scoring JSON, while its other three modes completed 300/300. **Parameter compatibility:** some models reject certain temperature, reasoning, and tool combinations, which the client handles per model, and Claude thinking required a larger `max_tokens` value to avoid truncation. **Resumption:** the harness supports `-resume` keyed on `bibcode`, plus a retry script for transient failures.

gpt-5.6-sol requires a caveat, and it is the same case as the strong negative bias reported in Sec. 7.4. A parameter-conflict bug (`temperature` + `reasoning_effort` + tools) forced a client-side fix and a full re-run, after which all four modes reached 300/300; `dimagent_v3` still underestimates systematically (Scheme-2 10.6, $\rho \approx 0.17$, bias -20.2, grade F in every discipline), whereas native search recovers it to 28.7 (E) but destabilizes the physics sub-run. This is a genuine model-level difference rather than a technical one: tool-calling configurations must be validated per model before large runs. Appendix F gives the full case study.

8 Limitations and Future Work

Since each discipline offers only 50 selected papers, this benchmark cannot cover every sub-discipline in depth, and some runs on biomedical and chemistry titles failed at the gateway, leaving gaps that cannot always be back-filled. The dataset is English-dominant and may under-represent non-English scholarship; because citation counts and journal indicators drift, the backup pool needs periodic refresh and re-sampling. The next steps are to expand to computer science, materials science, earth science, and the social sciences and to add human-expert annotations alongside the bibliometric ground truth.

A further limitation concerns the weight vector itself: the seven weights come from a single dual-target derivation trained on 998 quantum-physics papers with five regressors (including a random forest), yet they are shared across all six disciplines. We have not re-derived them per discipline, so the transferability of one common weight vector remains an assumption of the benchmark—one of its principal limitations—and per-discipline re-derivation and comparison are left to future work.

9 Conclusion

We present SciTaste-Bench, a six-discipline gold-standard benchmark for judging research innovativeness, constructed by (i) integrating metadata from ADS/OpenAlex/arXiv and scoring each paper with a seven-dimensional, machine-learning-weighted Academic Taste Index under a single weight vector shared across physics, mathematics, chemistry, biology, medicine, and finance; (ii) designing and implementing in-house a structured tool-calling agent, DimAgent-v3, with its nine bibliometric tools, two-round loop, and reviewer prompt; and (iii) evaluating models with correlation- and relative-error-based methods. Across ten large language models, structured tool-based evaluation generally outperforms tool-free and raw web-search modes, although the advantage is model-dependent, with gpt-5.6-sol's collapsed pure-tool run as the main counterexample. Physics is the easiest field, chemistry, biology, and medicine are the most difficult, and dimension-level evaluation remains the key bottleneck. The dataset, the harness, and the agent—all designed and built by the authors—are released to support reproducible research on scientific-innovation assessment. Beyond benchmarking, we view SciTaste-Bench as infrastructure for auto-research: autonomous systems that propose, evaluate, and prioritize scientific ideas require exactly this ability to judge the value potential of research papers.

Statement on AI Assistance

Large language models were used in preparing this manuscript solely to aid and polish the writing. A multi-hop LLM paraphrasing pipeline was applied segment-wise to the prose for language-level editing, style polishing, and readability, and an LLM assistant helped rephrase and reorganize author-written sentences and paragraphs—including the exposition of the results analysis—so that findings already established by the authors were expressed more clearly. In every case the underlying content, analytical points, tables, and figures were supplied by the authors. Large language models were not used to generate or label data, design experiments, run the evaluations, derive results, search the literature, or prove mathematical claims, and no model-generated text was accepted without human verification. All text was reviewed, corrected, and approved by the human authors, who take full responsibility for the content of this paper.

References

Alberto Accomazzi. Decades of transformation: Evolution of the NASA astrophysics data system's infrastructure. *arXiv preprint arXiv:2401.09685*, 2024.

Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

Kevin Funk and Jason Owen-Smith. A dynamic network measure of technological change. *Management Science*, 63(3):791–817, 2017.

Weixin Liang, Yuhui Zhang, et al. Can large language models provide useful feedback on research papers? a large-scale empirical analysis. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. arXiv:2310.01783.

Spyros Makridakis, Evangelos Spiliotis, and Vassilios Assimakopoulos. The M5 accuracy competition: Results, findings and conclusions. *International Journal of Forecasting*, 38(4):1346–1364, 2022.

Thi Huyen Nguyen and Zahra Ahmadi. LLM-based scientific peer review: Methods, benchmarks, and reliability challenges. *arXiv preprint arXiv:2606.25057*, 2026.

Michael Park, Erin Leahey, and Russell J. Funk. Papers and patents are becoming less disruptive over time. *Nature*, 613:138–144, 2023.

Jason Priem, Heather Piwowar, and Richard Orr. OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts. *arXiv preprint arXiv:2205.01833*, 2022.

Yujia Qin et al. ToolLLM: Facilitating large language models to master 16000+ real-world apis. In *International Conference on Learning Representations (ICLR)*, 2024. arXiv:2307.16789.

Chenglei Si, Diyi Yang, and Tatsunori Hashimoto. Can LLMs generate novel research ideas? a large-scale human study with 100 NLP researchers. In *International Conference on Learning Representations (ICLR)*, 2025. arXiv:2409.04109.

Mike Thelwall. Are the discretised lognormal and hooked power law distributions plausible for citation data? *Journal of Informetrics*, 2016. arXiv:1603.05078.

Lianmin Zheng, Wei-Lin Chiang, et al. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2023. arXiv:2306.05685.

A Sub-discipline taxonomy and the physics classification cascade

Table 4 lists the second-level sub-discipline taxonomy used for all six disciplines and Table 3 gives the four-stage cascade that assigns physics labels. Worked examples: `quant-ph` maps directly to quantum physics; at the journal stage, bibstems such as PRB or NATMAT indicate condensed

matter whereas APJ or MNRAS indicate astrophysics; at the keyword stage, two or more controlled-vocabulary hits, such as *superconductivity* and *graphene* for condensed matter, decide the label.

Table 3: Four-stage cascade for physics sub-discipline classification; the first stage that fires decides.

Priority	Basis	Rule
1	arXiv category	direct mapping of the subject class
2	journal inference	keyword match on the publication venue
3	keyword matching	≥ 2 controlled-keyword hits in title/abstract
4	title matching	≥ 2 keyword hits in the title

Table 4: Second-level sub-discipline taxonomy of the six disciplines in SciTaste-Bench.

Discipline	Second-level sub-disciplines
Physics (8)	classical physics; condensed matter; quantum physics; applied physics; high-energy physics; computational physics; interdisciplinary physics; astrophysics
Mathematics (8)	algebra and number theory; analysis and differential equations; geometry and topology; probability and statistics; computational mathematics; operations research and optimization; applied mathematics; logic and discrete mathematics
Chemistry (8)	organic chemistry; inorganic chemistry; physical chemistry; analytical chemistry; materials chemistry; computational and theoretical chemistry; chemical biology; polymers and soft matter
Biology (8)	molecular biology; cell biology; genetics and genomics; ecology and biodiversity; evolutionary biology; microbiology and infection; structural biology and biophysics; synthetic biology and bioengineering
Medicine (7)	clinical medicine; basic medicine; public health and epidemiology; pharmacy and drug development; imaging medicine and intervention; oncology; neuroscience
Finance (8)	asset pricing; corporate finance; market microstructure; risk management; financial econometrics; behavioral finance; derivatives and quantitative finance; macro-finance and financial technology

B DUAL-TARGET WEIGHT DERIVATION

Table 5 reports the per-dimension feature importances of the two Random-Forest analyses and the renormalized weights used throughout the paper.

The rationale for the dual-target design is that the weights should be assessed by the data rather than chosen by hand: citation count and journal impact are used as two independent supervision targets, so that each dimension's influence on these two community-sanctioned outcomes determines its weight. Specifically, we ran two independent Random-Forest analyses (Breiman, 2001) and interpreted feature importances as weights; both forests use 100 trees (seed 42) with an 80/20 train–test split. The fit of the citation model is good (test $R^2 = 0.74$), indicating that citation impact is largely predictable from the other six dimensions, whereas the journal model is much weaker ($R^2 = 0.17$), suggesting that journal impact carries relatively independent information. The two targets therefore provide complementary evidence for the final weight vector.

Table 5: Dual-target weight derivation. B–F average the two analyses; A and G are cross-assigned; the final column renormalizes the 128.86% total to 100%.

Dimension	Citation-analysis	Journal-analysis	Final (normalized)
A. Citation Impact	(target)	37.99%	29.5%
B. Knowledge Breadth	22.98%	7.52%	11.8%
C. Semantic Novelty	12.91%	16.87%	11.6%
D. Disruptiveness	18.14%	4.69%	8.9%
E. Conceptual Innovation	21.99%	11.20%	12.9%
F. Team Collaboration	4.24%	21.73%	10.1%
G. Journal Impact	19.73%	(target)	15.3%

C DIMENSION-TO-SCORE MAPPING

Each dimension's raw sub-parameters are mapped onto a 0–100 score.

A. Citation Dynamics (29.5%). From `total_citations` we blend a log-normalized score with an empirical percentile:

$$s_{\text{cit}} = 0.6 \underbrace{\frac{\log(1+c)}{\log(1+c_{\max})} \times 100}_{\text{log-normalized}} + 0.4 \underbrace{\frac{\#\{c_j < c\}}{N} \times 100}_{\text{percentile}}. \tag{3}$$

The logarithmic term handles the heavy-tailed citation distribution (Thelwall, 2016). The percentile term counts the fraction of papers that are strictly below the target, anchoring the score in the empirical distribution.

B–F. PCA + percentile rank. The set of secondary parameters is as follows. The breadth of knowledge uses the number of references, the diversity of reference fields, the diversity of reference journals, and the age range of references as indicators. Semantic novelty uses nearest neighbor and mean neighborhood semantic distances, number of preferred fields, field Shannon entropy, SPECTER distance to the reference list, SPECTER bridge score, and reference semantic dispersion. Disruptiveness uses the CD index and disruptive citation count. Conceptual Innovation uses core-concept score, concept count, core-topic score, and topic span. Team Collaboration is measured by the country count and the author count. Missing entries are imputed first by the column mean and then by zero. The resulting matrix is standardized and projected onto the first principal component. If PC1 is negatively correlated with the standardized variables, the sign is reversed. The projection is then mapped to $[0, 100]$ by mean-rank normalization:

$$r_{\text{norm}} = \frac{\text{rank}(z) - 1}{N_{\text{valid}} - 1} \times 100. \tag{4}$$

G. Journal Impact (15.3%). This score is a fixed-weight blend of the impact factor (70%) and two partition schemes (15% each):

$$s_G = 0.70\, s_{\text{IF}} + 0.15\, s_{\text{JCR}} + 0.15\, s_{\text{CAS}}. \tag{5}$$

The coefficients 0.70/0.15/0.15 were set from expert experience and validated through expert inspection of the final paper scores.

The impact factor score is the percentile rank of the journal's impact factor among observed values; missing entries receive 0 points. Both partition schemes enter the score: JCR (Journal Citation Reports) quartiles and CAS (Chinese Academy of Sciences zones) map Q1/zone-1→100, Q2/zone-2→75, Q3/zone-3→50, and Q4/zone-4→25, with missing partitions again scoring 0.

D DATASET STATISTICS AND RECORD FORMAT

Table 6 summarizes the released splits. Each record contains identifiers, text, year, and the ground-truth total plus seven per-dimension scores (Figure 2).

Table 6: SciTaste-Bench dataset statistics.

Metric	Value
Number of disciplines	6
Selected papers	300 (50 per discipline)
Backup papers	5,700 (950 per discipline)
Publication-year range	1996–2023
Mean composite score	53.33
Journal-tier balance (Q1/Q2/Q3/Q4)	78 / 73 / 76 / 73

D.1 DATASET RECORD FORMAT

Each record contains identifiers, text, year, and the ground-truth total plus seven per-dimension scores (2).


```
{
  "bibcode": "1998PhRvA..58.3983S",
  "title": "Nonadiabatic dynamics in evaporative cooling ...",
  "abstract": "Magnetically trapped neutral atoms can be cooled ...",
  "year": 1998,
  "ground_truth_score_100": 46.36,
  "ground_truth_dimensions": {
    "citation_impact": 51.55,
    "knowledge_breadth": 35.33,
    "semantic_novelty": 34.39,
    "disruptiveness": 79.22,
    "conceptual_innovation": 44.51,
    "team_collaboration": 13.77,
    "journal_impact": 57.58
  }
}
```


Figure 2: A single SciTaste-Bench record (physics).

E AGENT TOOLS, LOOP, AND EVALUATION MODES

Table 7 lists the nine bibliometric tools, Figure 3 shows the two-round DIMAGENT-V3 loop, and Table 8 lists the four evaluation modes with the capability each probes.

Table 7: The nine bibliometric tools exposed to DIMAGENT-V3.

Tool	Raw data returned	Supports
`get_citation_data`	citation count, reference count	citation_impact
`get_journal_info`	journal name, pub. date, DOI	journal_impact
`get_reference_list`	reference count + top-20 titles	knowledge_breadth
`get_concept_data`	keyword list (up to 15)	knowledge_breadth (aux)
`get_author_data`	author count, names, affiliations, countries	team_collaboration
`get_citation_network`	citing-paper list (up to 15)	disruptiveness (aux)
`get_topic_data`	topic keyword list	conceptual_innovation
`get_disruptiveness_data`	references + citation overlap for 5 citing papers	disruptiveness
`get_keyword_rarity`	corpus frequency + rarity score per keyword	semantic_novelty

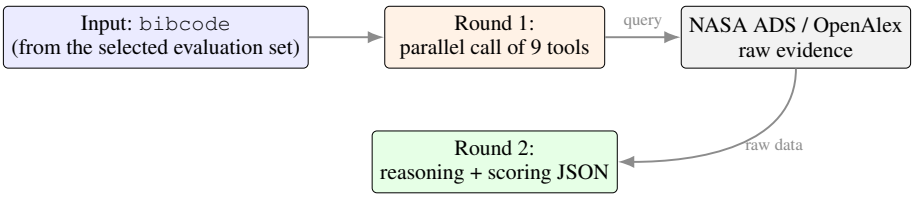

Figure 3: DIMAGENT-V3 agent loop: after one round of parallel tool calls, one round of scoring is performed.

Table 8: The four evaluation modes and the capability each probes.

Mode	Configuration	Capability probed
`pure_qa`	no tools, no search; internal knowledge only	prior knowledge & judgment
`native_search`	platform-native web search enabled	retrieval-augmented judgment
`dimagent_v3`	nine structured bibliometric tools	external-knowledge integration
`dimagent_v3_native`	nine tools + native search	structured + live retrieval

F THE GPT-5.6-SOL TOOL-CALLING CASE

gpt-5.6-sol requires a caveat; this is the same case as the strong negative bias reported in Section 7.4. The early tool-calling runs for this model were affected by a parameter-conflict bug. Specifically, combining `temperature` and `reasoning_effort` with tools triggered API errors. To resolve this, these two parameters were disabled for this model's tool calls on the client side, and a full re-run was performed. As a result, all four modes achieved 300/300. Five transient `pure_qa` failures during testing were back-filled by retry. The results of this re-run are included in Table 2. The `pure_qa` and `native_search` results for this model are broadly on par with other models, both receiving a grade of F, but `dimagent_v3` shows systematic severe underestimation. Scheme-2 scores were only 10.6, with $\rho{\approx}0.17$, bias value -20.2, and all disciplines received a grade of F. Enabling native search recovered performance to 28.7 (grade E), but the physics sub-run destabilized (ρ is undefined, Scheme-2 scores are 0.0). Therefore, this is not a technical issue but indicates a genuine model-level difference. The lesson is that tool-calling configurations must be validated per model before large runs, and a technically valid run does not guarantee behaviorally effective tool use.

G PER-DISCIPLINE RESULTS

These two tables back the cross-discipline analysis in the main text. Table 9 reports, for the `dimagent_v3` run of qwen-3.8-max, the per-discipline Spearman correlation, MAE, WAPE and Scheme-2 score with its grade. Table 10 reports the Scheme-2 score of all ten models in the same mode, discipline by discipline, which makes it possible to check whether the difficulty ordering of the six fields is a property of the benchmark or of each model's training mix.

Table 9: Results by discipline for `dimagent_v3` (qwen-3.8-max).

Discipline	Spearman ρ	MAE	WAPE	Scheme 2
Overall	0.5481	12.86	24.1%	51.80 [D]
Physics	0.7751	6.21	13.1%	70.40 [C]
Finance	0.6923	11.90	22.5%	55.00 [D]
Mathematics	0.5293	13.03	23.6%	52.70 [D]
Chemistry	0.5697	15.80	29.2%	41.70 [D]
Biology	0.4599	15.72	28.8%	42.50 [D]
Medicine	0.3099	14.51	26.1%	47.80 [D]

Table 10: Scheme-2 scores by discipline for `dimagent_v3` across all ten models. *The claude-opus-4-8 physics sub-run is anomalous ($\rho{=}-0.16$, WAPE 74.4%) and reflects an instability of that model under tool mode rather than the difficulty of the discipline.

Model	Overall	Physics	Finance	Math	Chem	Bio	Med
qwen-3.8-max	51.8	70.4	55.0	52.7	41.7	42.5	47.8
qwen3.7-plus	47.6	69.4	57.3	47.9	37.5	35.7	44.7
qwen3.7-max	47.7	70.7	50.4	46.0	38.3	38.3	45.7
claude-fable-5	47.0	66.9	48.5	51.4	38.2	37.0	42.7
claude-opus-4-8	34.5	6.3*	60.8	49.0	37.1	33.1	42.3
gpt-5.6-sol	10.6	10.2	13.0	9.5	10.5	10.2	13.6
gemini-3.5-flash	17.7	29.1	16.2	17.7	15.4	15.8	18.4
kimi-k3	33.4	46.9	35.6	37.5	29.8	28.8	24.7
Doubao-Seed-2.1-Pro	39.2	15.8	58.6	51.6	43.6	40.3	49.8
DeepSeek-V4-Pro	34.5	31.1	38.6	35.8	32.8	32.5	35.8

H CITATION-COUNT DISTRIBUTIONS

The distribution of citation counts is heavy-tailed. Two prevailing models are the discretised lognormal distribution and the hooked or shifted power-law distribution. The former is often favoured for the arts, humanities, social sciences, engineering fields, single disciplines, and papers published ≥ 3 years ago. The latter is more commonly favoured for the natural sciences, life sciences, recent papers, and datasets with many zero-cited items; this model naturally supports zero citations (Thelwall, 2016). There is no universal distribution. For cross-field comparison, percentile normalization (e.g., the FWCI indicator) combined with non-parametric tests is preferable. For modeling within a specific field, AIC/BIC criteria can be used to choose between candidate distributions. For regression analysis, $\log(c+1)$ combined with ordinary least squares is usually the most stable choice. All zero-cited papers should be included, heterogeneity across fields should be controlled by stratification, and goodness-of-fit should not be interpreted as evidence for a generative mechanism.

I ARTIFACT INVENTORY

The main outcomes published in this study include the merged dataset with its selected and backup splits, the identifier-to-OpenAlex alignment map, the report generator, the agent task configuration defining the nine tools and four evaluation modes, the scoring-metrics module implementing both scoring schemes, and the per-model run scripts, and the weight-derivation package containing the trained regression models (including the random forest) together with their hyperparameters and feature-importance outputs. All components are published together with the dataset, ensuring full replication. The detailed dataset and the full code will additionally be open-sourced on Hugging Face.

J EVALUATION PROMPTS AND A WORKED EXAMPLE

The original prompts were written in Chinese; the blocks below are condensed translations. All four modes share the same user template.

```
Please assess the innovativeness of the
following {discipline} paper.
Discipline: {discipline}
Title: {title}
Abstract: {abstract}
Identifier: {bibcode}
```

is a mode-specific termination instruction. The system prompt of `pure_qa` casts the model as a senior academic reviewer, requiring it to score the seven dimensions from its own scholarly knowledge, list the seven- dimension rubric using the shared weight vector, and conclude with "output the JSON directly, no explanation". Its user template closes with "score the seven dimensions from your scholarly knowledge and output JSON.". The system prompt of `native_search` retains the same skeleton, but requires the model to first retrieve the paper's citation count, venue reputation, authorship, and field information through native web search, and to ground every score in the retrieved

data. The DIMAGENT-V3 mode uses the system prompt summarized in Sec. 5: it consists of the nine ADS tools, the three-step parallel-call workflow, the rule that all nine tools must be called and that scores must be grounded in the returned payloads, and the same strict-JSON contract (composite score, seven dimension scores, reasoning trace, and self- reported confidence). Verbatim prompts ship with the released harness.

Worked example. Instantiating the template on a real item gives:

```
Discipline: physics
Title: Nonadiabatic dynamics in evaporative
  cooling of trapped atoms by a
  radio-frequency field
Abstract: Magnetically trapped neutral atoms can
  be cooled with the evaporation technique. ...
  [truncated]
Identifier: 1998PhRvA..58.3983S
```

The model must reply with the strict JSON; the ground truth of this item is

```
{"innovation_score": 46.36,
 "dimension_scores": {
  "citation_impact": 51.55,
  "journal_impact": 57.58,
  "conceptual_innovation": 44.51,
  "knowledge_breadth": 35.33,
  "semantic_novelty": 34.39,
  "team_collaboration": 13.77,
  "disruptiveness": 79.22}}
```