

Large Language Models Extend Known Research But Do Not Create New Research Directions

Zhihua Liu*

*Causality in Healthcare AI Hub (CHAI)
School of Engineering, University of Edinburgh*

zliu7@ed.ac.uk

Kurt Butler*

*Causality in Healthcare AI Hub (CHAI)
School of Engineering, University of Edinburgh*

kbutler2@ed.ac.uk

Damian Machlanski*

*Causality in Healthcare AI Hub (CHAI)
School of Engineering, University of Edinburgh*

d.machlanski@ed.ac.uk

Danilo Bueno

*Causality in Healthcare AI Hub (CHAI)
School of Engineering, University of Edinburgh*

dbueno@ed.ac.uk

Sotirios A. Tsafaris

*Causality in Healthcare AI Hub (CHAI)
School of Engineering, University of Edinburgh*

S.Tsafaris@ed.ac.uk

Abstract

The success of autonomous research systems relies on the quality of the scientific hypotheses they generate. Recent studies report that human reviewers often rate LLM-generated ideas as more novel than those from human experts (Si et al., 2025; 2026; Jiang, 2026), raising hope that AI can accelerate scientific discovery. Yet it is unclear whether these models can do more than extend existing knowledge (Bao et al., 2026). Can LLMs drive genuine innovation by revealing new research directions, or are they limited to repeating and recombining what science already knows? This unresolved question lies at the heart of discussions about AI and the future of science. To address this, we distinguish two types of novelty: distance, which measures how far a hypothesis lies from existing work, and direction, which captures whether the hypothesis pursues new lines of inquiry beyond those already present. These two measures diverge sharply for LLM-generated work. We represent scientific papers in a shared conceptual space using a pretrained scientific-paper embedding model (Singh et al., 2023). Within this space, we embed 26.9 million human-authored papers across 1,325 research topics in six natural-science fields and 524,308 LLM-generated hypotheses from eight models. Distance is measured using existing metrics Tang & Yang (2026). To quantify direction, we introduce the orthogonal novelty score (ONS). LLMs consistently generate ideas more distant from existing work than human-authored papers, up to 16% further when the LLM is given the papers farthest from the centre of the topic. However, LLM hypotheses have a 15–23% lower median ONS than published papers from the same topic. Thus, LLMs move further from what is already known, but remain within established directions. This pattern persists across changes in various configurations. As a result, LLMs excel at extending the frontier but the creation of new directions is still led by human scientists.

*Equal contribution.

Autonomous research agents are attracting substantial investment across academia and industry, driven by the promise to accelerate scientific discovery by automating hypothesis generation, validation and experimentation in a closed loop. Large language models (LLMs) have enabled a new generation of autonomous research systems, capable of performing open-ended ideation across scientific domains by leveraging the vast corpora of scientific literature in their training data (Gridach et al., 2025). This enthusiasm is fueled further by the availability of large, multimodal health databases (Liu et al., 2025; Yan et al., 2026). Recent studies into the feasibility of these systems have been encouraging (Si et al., 2025; Abdel-Rehim et al., 2025), showing that LLM-generated hypotheses can match or exceed humans in their novelty, although it remains unclear whether this reflects genuinely new lines of inquiry or existing literature restated in unfamiliar language.

In spite of this optimism, other research has identified reasons for caution. Hao et al. (2026) found that AI tools expand scientists’ impact but contract their focus, as evaluated through their bibliometric footprint. Tang & Yang (2026) show that LLM-generated ideas are more conceptually concentrated compared to human authored papers. Luyten & van der Schaar (2026) show theoretically that novel hypothesis generation is penalized by the need to optimize for correctness reduces the diversity of acceptable hypotheses. These findings create an apparent paradox: do LLM-based research agents broaden the horizon of science, or do they constrain it?

We argue that this contradiction stems from a failure to distinguish between two distinct dimensions of novelty. A scientific hypothesis can be distant from existing work while remaining entirely within the established conceptual directions. Normal science extends existing research directions, occasionally to a considerable distance, whereas conceptual change requires new directions (Kuhn & Hacking, 1970). Existing measures of novelty do not distinguish between these two notions, leaving open the question of whether the observed conservatism represents a superficial tuning problem or a fundamental structural constraint.

Here we separate distance and direction as distinct, measurable properties of scientific novelty, and we ask whether LLMs can match human researchers under either metric. We evaluate the performance of several state-of-the-art AI models against a large corpus of human authored papers across six natural-science fields. We find that direction and distance show differing patterns across experiments: LLMs are able to produce ideas that improve in distance-based novelty over human authors, but there is a deficit between LLMs and humans in direction-based novelty. This pattern holds across model families and is not affected by modifying the prompting strategy, context selection and agentic configuration. These results identify a specific and reproducible boundary on the contribution of LLMs to scientific discovery.

Overview of Methods and Findings

We embedded 26.9 million research papers (titles and abstracts) into a shared 768-dimensional embedding space (Singh et al., 2023). We refer to these published research papers as **seed-papers**, spanning 1,325 research topics in six natural-science fields and published between 2005 and 2026. We follow the hierarchy of OpenAlex in terms of organizing these seed-papers, which uses the following categorization: **field - sub-field - topic**. For each research topic, an LLM hypothesis generator received a set of published papers sampled from that topic as input, referred to as **focal-papers**, and generated hypotheses in the form of a paper title and abstract, which we refer to as **pseudo-papers**. All pseudo-papers were embedded in the same space as seed-papers and compared with approximately 1,000 papers published in the same topic that were not provided as input to the model (Fig. 1d). We evaluated all generated hypotheses using three metrics (Fig. 1b). Two of them measure the distance to the embedding centroid of the topic seed-papers, and the distance to the nearest published seed-paper embedding within the topic. The third metric is the orthogonal novelty score (ONS), which measures the proportion of the variance of a paper that lies outside the subspace spanned by the published literature of its topic.

We evaluated eight LLMs from four model families and generated 524,308 hypotheses in total. Through the large scale comparison using distance metrics and the direction metric, we found that LLM-generated hypotheses obtained lower ONS scores than published papers across all fields and models. The two distance measures exhibited behaviors distinct from ONS and from each other. Most models positioned hypotheses farther from the nearest published paper than real papers generally do, whereas their distance to the centroid of the topic remained mostly at or below that of published papers. Distance metrics responded to the manner

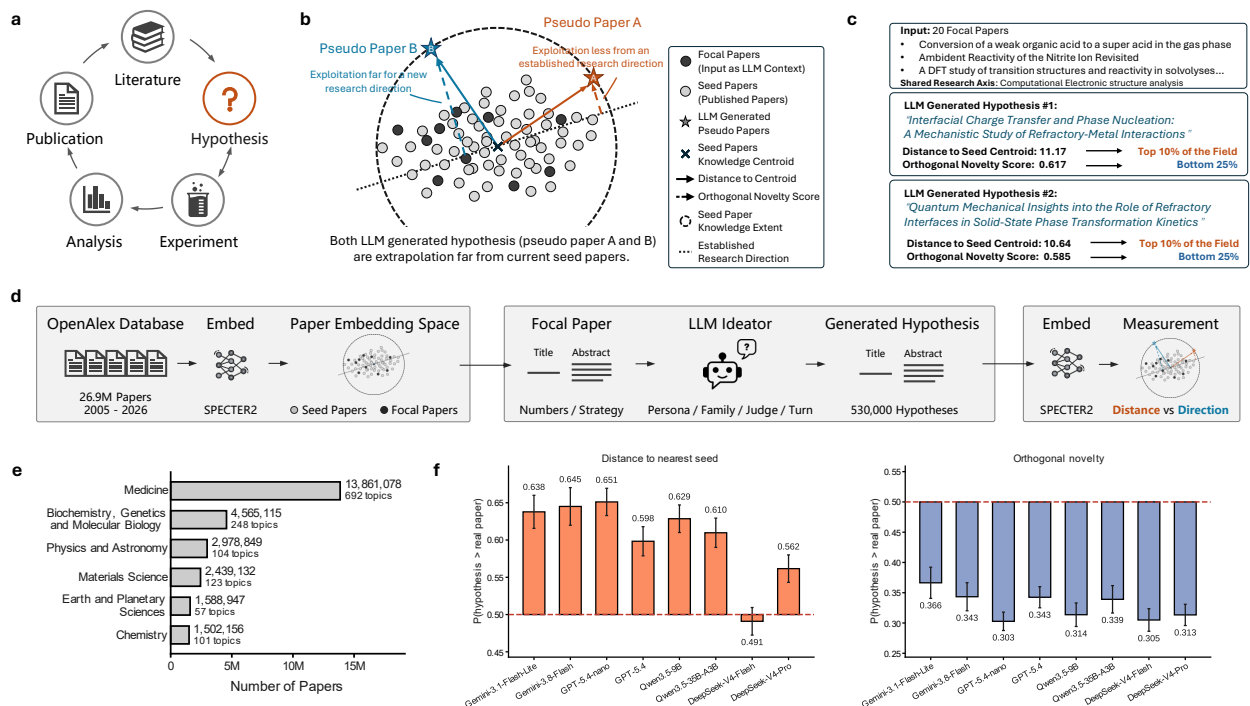


Figure 1: **LLM places hypotheses further from the published literature than real papers do, yet along fewer new research directions.** **a**, an illustration of hypothesis generation within an automated research loop. **b**, distinguish of distance and direction novelty of LLM generated hypothesis. **c**, an running example of LLM generated hypotheses fall within the top decile for distance extension and the bottom quartile for direction orthogonality. **d**, overview of our generation and evaluation pipeline. **e**, scale of the comparison corpus. **f**, comparison between generated hypotheses and human-authored papers across three evaluation metrics.

in which hypotheses were generated. Selecting focal papers from the far edge of a topic, or prompting the evaluator to favor unconventional ideas, shifted hypotheses well beyond published work. However, no intervention configuration achieved parity in terms of ONS: providing additional context, altering context selection, incorporating a critic loop, assigning a persona — neither of them closed the performance gap. This large-scale evaluation demonstrates that current LLMs function as engines of normal science: they extend established research directions, occasionally beyond published papers, but direct a smaller proportion of their content outside these existing paradigms.

Results

We compare each set of LLM generated hypotheses, formatted as title and abstract of pseudo-papers, with real papers from the same topic scored on the same three metrics. We further refer to pseudo-papers and real papers as **pseudo** and **real**, and calculate the win rate $P(\text{pseudo} > \text{real})$ over metrics, which is the probability that a pseudo-paper drawn at random from a topic scores higher than a real paper drawn at random from the same topic. Values below 0.5 mean that pseudo tend to score lower than real on a certain metric. We first present the results from a large-scale comparison across eight LLM models and six research fields (Fig. 2). We then test two categories of interventions that most systems using LLMs for automated research can control, including the context the LLM reads, through the number and the selection strategy of focal papers (Figs 3 and 4), and the process that refines LLM generated hypothesis, through a multi-turn dialogue with an LLM judge and the personas given to the two roles (Figs 5 and 6).

LLM Generated Hypotheses Open Fewer New Research Directions

By separating research distance from its direction, we aim to answer: *do LLM generated hypotheses fall short of published work, and does the answer depend on the field or on the model?* We conduct a large scale comparison across eight LLMs from four model families (Gemini, GPT, Qwen and DeepSeek). Each LLM generates roughly 50 pseudo-papers per research topic given 20 focal papers of that topic as context, yielding a total of 524,308 LLM generated hypotheses across 1,325 research topics.

Fig. 2 demonstrates the comparison results between LLM generated pseudo-papers and real papers over all metrics. Each comparison is reported as a win rate over a certain metric. Most LLMs generate hypotheses farther from their nearest published paper than real papers sit from theirs. By contrast, distance to the topic centroid is mostly shorter for pseudo-papers than for real papers, with median win rates across fields of 35–48% by model ($P < 0.05$). This result demonstrates that most LLMs move pseudo-papers away from their nearest neighbors without moving them further from the center of the topic.

The above evidence suggests that the generated hypotheses fill gaps within the topic rather than push its boundary outward. Indeed, pseudo-papers seem to be displaced less than real ones as they fall short with respect to direction more than in terms of distance across all models (median win rate 18–28% on ONS against 35–48% on centroid distance, $P < 0.001$). The median pseudo-paper has an ONS 15–23% lower than the median real paper and this contraction appears in all research fields and LLMs ($P < 0.001$). When a real paper and a pseudo-paper are drawn at random from the same topic, the real paper has the higher ONS in 72–82%. Moreover, the contraction is not driven by topic selection as for every tested LLM, it appears in 86–97% of the 1,325 topics and in at least 84 of the 86 sub-fields.

This observation remains consistent across all research fields. Pseudo-papers within the medicine field come the closest to real papers with respect to ONS for all models, yet they remain below real papers in all eight LLMs (median win rate 30–41%, $P < 0.001$).

Using larger models for the hypotheses ideator did not solve the issue as hypotheses generated by larger LLMs generally scored lower ONS as compared with smaller models: Gemini (win rate 26% to 22%), statistically unchanged for GPT (28% to 26%) and Qwen (28% to 26%), and raised it only for DeepSeek (18% to 25%), whose larger model still remains well below real papers. These findings show that the contraction in research directions is shared by every model and field we examined, whereas how far the hypotheses travel depends on both. As contraction appears to persist regardless of the model choice, our further investigation explored whether it depends on how models are used.

Context Moves Hypotheses Outward, Not Into New Directions

We first examine the LLM input, which comprises the focal papers a user selects and provides as context to the model. We select two LLMs, Gemini-3.1-Flash-Lite and GPT-5.4-nano, and varied both the number of focal papers and the selection strategy across 86 topics, taking from each of the 86 sub-fields the topic with the largest number of seed papers, and generating 50 pseudo-papers per topic in every setting.

Number of focal papers. A simple hypothesis for the contraction in large scale comparison is that the context (focal paper number $N = 20$) provide insufficient knowledge as input. We therefore varied the number of focal papers from 5 to 100 (Fig. 3, $N \in \{5, 10, 20, 50, 100\}$). None of the three metrics varied with the number of focal papers. On distance measurements, LLM generated hypothesis sat 3–5% further from their nearest published paper than real papers did at every N for both models (Fig. 3a,b, win rate 57–63%, $P < 0.05$), while their distance to the topic centroid stayed within 2% of that of real papers and did not differ significantly in any of the ten (win rate 46–52%, $P < 0.05$). On direction measurement, the median pseudo-paper had an ONS 16–18% lower than the median real paper at every N for both models (win rate 26–29%, $P < 0.05$), and this contraction stayed flat across the range of different focal paper input.

Selection strategy of focal papers. By setting $N = 20$, we provided the models with fixed number of focal papers selected by different strategies based on the euclidean distance between paper embeddings. These strategies reflect on the content difference of context holds, including 20 papers closest to the topic centroid, 20 papers drawn at random from the seed papers, or the 20 papers farthest from the centroid (Fig. 4). With the farthest papers as context, LLM generated hypotheses were located 11–12% further from the topic centroid and 16% further from their nearest published paper than real papers (Fig. 4a,b, win rate 77–84%, $P < 0.05$). With random selection, their distance to the centroid did not differ significantly from that of real papers for either model. With the closest papers as input, it did not differ for Gemini-3.1-Flash-Lite and was 3% shorter for GPT-5.4-nano (win rate 40%, $P < 0.05$). Distance to the nearest published paper was greater than that of real papers under all three strategies for both models ($P < 0.05$). However, the choice mattered far less on measurement of proposing new research directions. ONS was lower than that of real papers under all three strategies for both models, where The farthest papers brought ONS closest to that of real papers but did not reach it. GPT-5.4-nano remained significantly lower (win rate 39%, $P < 0.05$), and Gemini-3.1-Flash-Lite was lower but not significantly so (win rate 45%, $P < 0.05$).

Collectively, the number and the selection strategy of focal papers change how far pseudo-papers travel, but neither brings their ONS level with that of real papers. As changing the input does not remove the contraction in research directions, we next asked whether the generation process itself can.

Critique and Personas Change Hypotheses Without Opening New Directions

Autonomous research systems commonly pair an LLM that generates research hypothesis with an LLM as a judge that reviews them, and feed the reviews back to the LLM ideator for the next round of generation (Lu et al., 2026). We reproduced this design as a five-turn dialogue between an LLM ideator and an LLM judge, using the same two LLMs and the same 86 topics as in Figs. 3 and 4. At every turn, the ideator received the same 50 focal papers and generated 20 pseudo-papers per topic. From the second turn onwards, the judge first reviewed the previous batch as a whole, naming its weaknesses, its over-represented and missing themes and concrete suggestions for the next batch. The ideator then generated a new batch conditioned on the focal papers, the review comment and the titles of all earlier pseudo-papers, which it was told not to repeat. The two models judged each other’s pseudo-papers. Every topic completed all five turns in every configuration.

Ideator–judge dialogue. With both roles given default prompts, we compared the pseudo-papers of each turn with real papers (Fig. 5a,b) and with the ideator’s own first turn (Fig. 5c). Against real papers, the median pseudo-paper had an ONS 12–17% lower than the median real paper (Fig. 5a,b; win rate 24–34%, $P < 0.001$), while pseudo-papers sat 3–7% further from their nearest published paper than real papers did (win rate 57–69%, $P \leq 0.028$). Against the first turn, the dialogue changed the pseudo-papers of GPT-5.4-nano, whose generated hypotheses from later dialogue sat further from the topic centroid and from their nearest published paper, and further outside the directions spanned by the 50 focal papers it was given (Fig. 5c; win rate 58–71% over its first-turn pseudo-papers, $P \leq 0.017$). Those of Gemini-3.1-Flash-Lite did not change significantly on these measures (win rate 41–54%, $P \geq 0.07$). The ONS of both models rose only slightly over the first turn, by 5–8% for Gemini-3.1-Flash-Lite (win rate 56–61%, $P \leq 0.015$) and by 2–4% for GPT-5.4-nano (win rate 52–54%, $P < 0.05$), and stayed 12–17% below that of real papers at every turn, including hypotheses generated from the last dialogue turn.

Personas. The dialogue in Fig. 5 used default prompts, which asks for hypotheses that bridge the topic to neighboring areas or adjacent disciplines. As persona prompts change the behavior of LLMs (Kong et al., 2024) and are routinely used to assign roles in AI scientist systems (Lu et al., 2026), we repeated the five-turn dialogue with a persona given to one role at a time, either the ideator or the judge, while the other role kept its default prompt. The *conservative persona* asks for sharper and more rigorous hypotheses that extend the focal papers. The *wild persona* asks for the opposite, in terms close to what ONS measures that a hypothesis should bridge the topic with a distant discipline, invert an assumption the focal papers take for granted, attack a problem they treat as solved, or propose a method or framing the field has not yet seen. This setup yields four personas, including conservative ideator, wild ideator, conservative judge and wild judge (Fig. 6a–d).

No persona brought ONS level with that of real papers. Across all configuration, the median pseudo-paper had an ONS 2–18% lower than the median real paper (Fig. 6; win rate 24–47%, $P < 0.05$), while the distance to the nearest published paper was greater than that of real papers (win rate 54–77%, $P < 0.05$). Asking for new directions by using *wild persona* did not produce a hypotheses that inspires a new research direction. Compared directly at the same turn, the pseudo-papers of the wild ideator had a lower ONS than those of the conservative ideator at every turn for both models (win rate of wild over conservative 36–48%, $P < 0.05$).

The persona of the judge had the larger effect over distance. From the second turn when the judge first acts, the win rate on distance to the nearest published paper was 71–77% under the wild judge ($P < 0.001$) against 57–67% under the conservative judge ($P \leq 0.031$). Compared directly, the wild judge placed generated hypotheses further from their nearest published paper than the conservative judge at every turn for both models (win rate 61–77%, $P \leq 0.010$), whereas the two ideator personas differed far less and not always in the same direction (win rate 37–60%, $P < 0.05$). Yet from the second turn onwards, ONS was lower under the wild judge than under the conservative judge (Fig. 6c,d, win rate 42–54%, $P < 0.05$). A judge instructed to demand departures from the focal papers therefore moved pseudo-papers further from published work than a judge instructed to demand rigor, but gave them a lower ONS. Collectively, critique and personas changed where pseudo-papers landed, most strongly through the judge, but none brought their ONS level with that of real papers.

Discussion

Across six scientific domains and a broad range of experimental configurations, we find that LLMs underperform in generating new research directions relative to human researchers. On the other hand, LLMs excel in exploring and extending established research directions. This pattern held across model families, scientific domains, and every experimental configuration that we tested.

We hypothesize that the asymmetry between distance and direction points to a deeper property of how LLMs encode scientific knowledge. Trained on the existing literature, LLMs and their outputs contain a compressed representation of that corpus. This makes LLMs efficient at traversing the known conceptual space, but also penalizes the generation of scientific topics that might challenge or diverge what existing in the training corpus. The consistent deficit in ONS across all values of N (ranging from 5 to 100 focal papers) suggests that insufficient context is not the root cause. Also, the deficit persists across prompting strategies that encouraged lateral or conservative thinking, although distance was sensitive to this. One interpretation here is that orthogonal research directions are underrepresented in training data. Since the training corpus does not often show examples of genuine new pivots in research, the model has no gradient to follow to learn how to produce them on its own.

Our claims should be tempered by a few limitations of this methodology. We use distance and direction in the SPECTER2 embedding space as proxies for novelty. A high or low ONS paper is not necessarily novel or unoriginal, respectively. Further, we consider ablations for a number of key factors, but real AI workflows may diverge from our setup in ways that we did not anticipate. A key limitation is that we comment only on LLM-based systems. Many specialist AI systems that are not based on LLMs — including AlphaFold (Jumper et al., 2021) and Evo 2 (Brixi et al., 2026) — have proven to be uniquely useful and will likely continue to have great value for the scientific enterprise. LLMs-based systems are a first attempt at a generalist system and understanding their limitations is crucial for proper deployment.

For individual researchers using LLMs for ideation, our results are directly relevant. LLMs are demonstrably useful for generating hypotheses within or across existing topics. Where they are less useful is in the identification of new research directions. This suggests that LLMs can become powerful research engines when guided by humans, but those research agendas should be considered carefully and set by the user.

For companies that are building autonomous research systems at scale, our results identify specific limitations. Current LLM-based pipelines are well-suited to explore and ideate within a given domain. These pipelines may generate, refine and evaluate hypotheses at a scale and speed which is unmatched by human researchers. However, our results suggest that these pipelines cannot set their own directions. Repeated ideator-judge loops and multi-turn refinement don’t improve ONS. That said, embedding-based tools offer an efficient and practical tool for these pipelines: although a high ONS is not a necessary condition for novelty or utility, it proxies a useful signal for identifying which hypotheses can be pruned. Even highly cited papers are often conventional to a large degree, with novelty appearing only in a fraction of the work (Uzzi et al., 2013).

For policymakers, our results imply a careful assessment of how AI-led research is positioned with national and institutional research strategies. There is considerable optimism premised on the idea that LLMs can accelerate scientific discovery by autonomously generating new ideas, and several institutions have invested in the development of autonomous research platforms (Panfilov et al., 2026). Although across all of our experiments LLMs consistently failed to match humans in directional novelty, there still exist applications where this directional novelty is unneeded and more narrow autonomous research is valuable. What our work does suggest is that positioning LLMs as autonomous discovery agents, rather than tools that amplify human-directed research programs, risks misallocating our resources. Our current findings add to the growing body of research (Tang & Yang, 2026) that suggests that current LLMs contribute to science in real but bounded ways, and that matters for how we invest in and govern these systems.

Data availability

The OpenAlex dataset used in this study is publicly available at <https://openalex.org/>.

Code availability

The code to replicate the results presented in the Article and Supplementary Information is available via GitHub at <https://github.com/CHAI-UK>. The SPECTER 2.0 text embedding model is available at <https://hugging-face.co/allenai/specter2>.

References

- Abbi Abdel-Rehim, Hector Zenil, Oghenejokpeme Orhobor, Marie Fisher, Ross J Collins, Elizabeth Bourne, Gareth W Fearnley, Emma Tate, Holly X Smith, Larisa N Soldatova, et al. Scientific hypothesis generation by large language models: laboratory validation in breast cancer treatment. *Journal of the Royal Society Interface*, 22(227):20240674, 2025.
- Honglin Bao, Siyang Wu, Xiao Liu, Sida Li, Shiyun Cao, and James A Evans. Contemporary AI lacks the imagination to diverge or negate in science. *arXiv preprint arXiv:2606.08251*, 2026.
- Garyk Brixi, Matthew G Durrant, Jerome Ku, Mohsen Naghipourfar, Michael Poli, Gwanggyu Sun, Greg Brockman, Daniel Chang, Alison Fanton, Gabriel A Gonzalez, et al. Genome modelling and design across all domains of life with Evo 2. *Nature*, 652(8112):1349–1361, 2026.
- Mourad Gridach, Jay Nanavati, Khaldoun Zine El Abidine, Lenon Mendes, and Christina Mack. Agentic AI for scientific discovery: A survey of progress, challenges, and future directions. *arXiv preprint arXiv:2503.08979*, 2025.
- Qian Yue Hao, Fengli Xu, Yong Li, and James Evans. Artificial intelligence tools expand scientists’ impact but contract science’s focus. *Nature*, 649(8099):1237–1243, 2026.
- Bo Jiang. Hindsight: Evaluating LLM-generated research ideas via future impact. *arXiv preprint arXiv:2603.15164*, 2026.
- John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589, 2021.
- Aobo Kong, Shiwan Zhao, Hao Chen, Qicheng Li, Yong Qin, Ruiqi Sun, Xin Zhou, Enzhi Wang, and Xiaohang Dong. Better zero-shot reasoning with role-play prompting. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 4099–4113, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.228. URL <https://aclanthology.org/2024.naacl-long.228/>.
- Thomas S Kuhn and Ian Hacking. *The Structure of Scientific Revolutions*, volume 2. University of Chicago press Chicago, 1970.
- Tennison Liu, Silas Ruhrberg Estévez, David L Bentley, and Mihaela van der Schaar. Hypothesis hunting with evolving networks of autonomous scientific agents. *arXiv preprint arXiv:2510.08619*, 2025.
- Chris Lu, Cong Lu, Robert Tjarko Lange, Yutaro Yamada, Shengran Hu, Jakob Foerster, David Ha, and Jeff Clune. Towards end-to-end automation of ai research. *Nature*, 651(8107):914–919, 2026.
- Max Ruiz Luyten and Mihaela van der Schaar. The reasoning-creativity trade-off: Toward creativity-driven problem solving. *arXiv preprint arXiv:2601.00747*, 2026.
- Alexander Panfilov, Peter Romov, Igor Shilov, Yves-Alexandre de Montjoye, Jonas Geiping, and Maksym Andriushchenko. Claudini: Autoresearch discovers state-of-the-art adversarial attack algorithms for LLMs. *arXiv preprint arXiv:2603.24511*, 2026.
- 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*, volume 2025, pp. 94003–94092, 2025.
- Chenglei Si, Tatsunori Hashimoto, and Diyi Yang. The ideation-execution gap: Execution outcomes of llm-generated versus human research ideas. In *International conference on learning representations*, volume 2026, pp. 74195–74240, 2026.

Amanpreet Singh, Mike D’Arcy, Arman Cohan, Doug Downey, and Sergey Feldman. Scirepeval: A multi-format benchmark for scientific document representations. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pp. 5548–5566, 2023.

Yixuan Tang and Yi Yang. Ai research agents narrow scientific exploration. *arXiv preprint arXiv:2605.27905*, 2026.

Brian Uzzi, Satyam Mukherjee, Michael Stringer, and Ben Jones. Atypical combinations and scientific impact. *Science*, 342(6157):468–472, 2013.

Junyu Yan, Damian Machlanski, Kurt Butler, Panagiotis Dimitrakopoulos, Ewen M Harrison, Bruce Guthrie, and Sotirios A Tsafaris. Explainable AI for data-driven design of high-dimensional predictive studies. *medRxiv*, pp. 2026–05, 2026.

Acknowledgements

We acknowledge the support of the UKRI AI programme, and the Engineering and Physical Sciences Research Council (EPSRC), for the Causality in Healthcare AI Hub [grant number EP/Y028856/1].

Competing interests

The authors declare no competing interests.

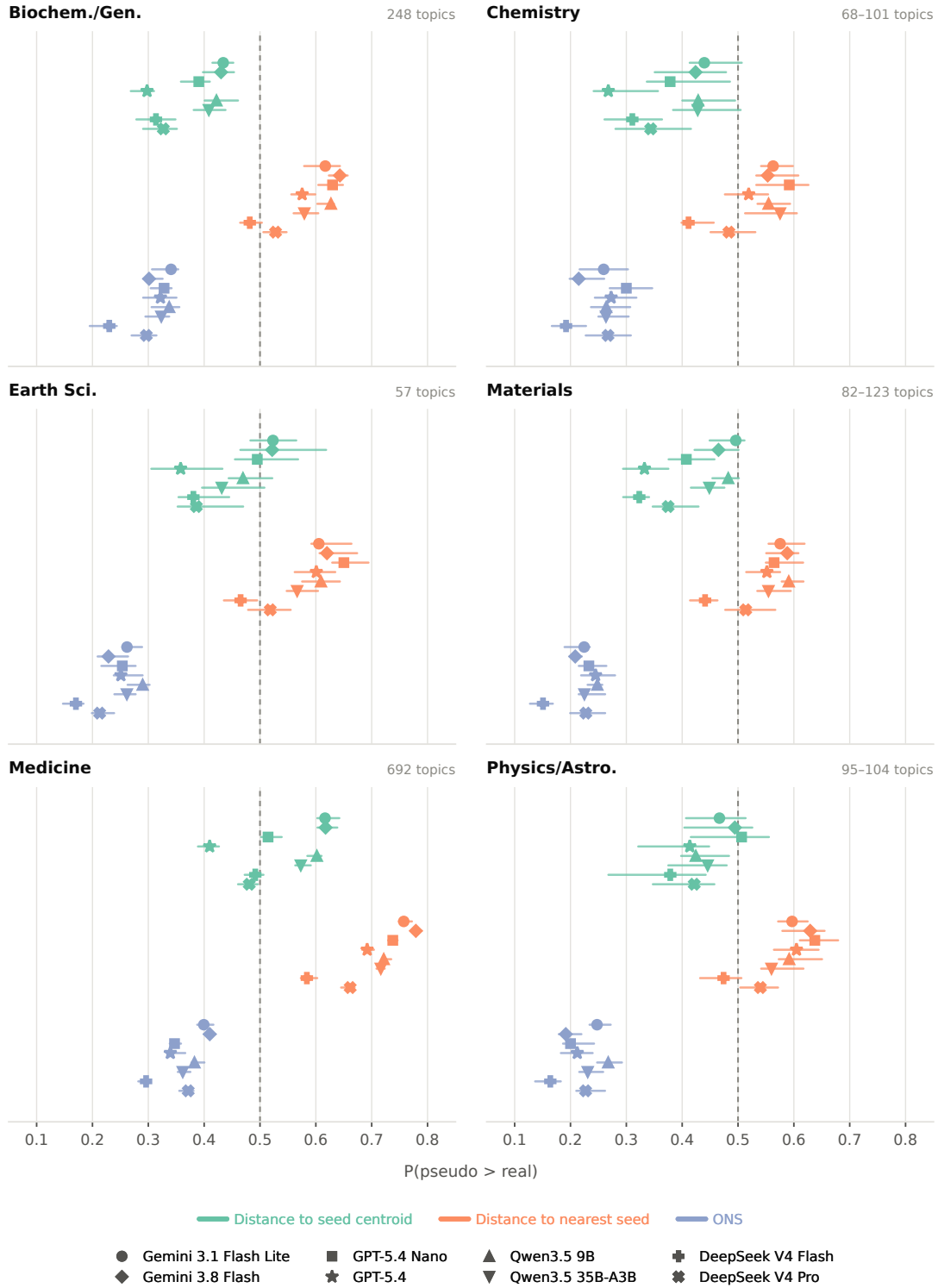


Figure 2: **LLM-generated hypotheses open fewer new research directions.** The ONS deficit is universal across models and fields. LLMs extend at least as far as real research does locally. The ordering is the same in every field and every model: nearest seed (orange) > centroid (green) > ONS (blue). Scale (i.e. topics) does not close the gap. Centroid displacement is at or below parity.

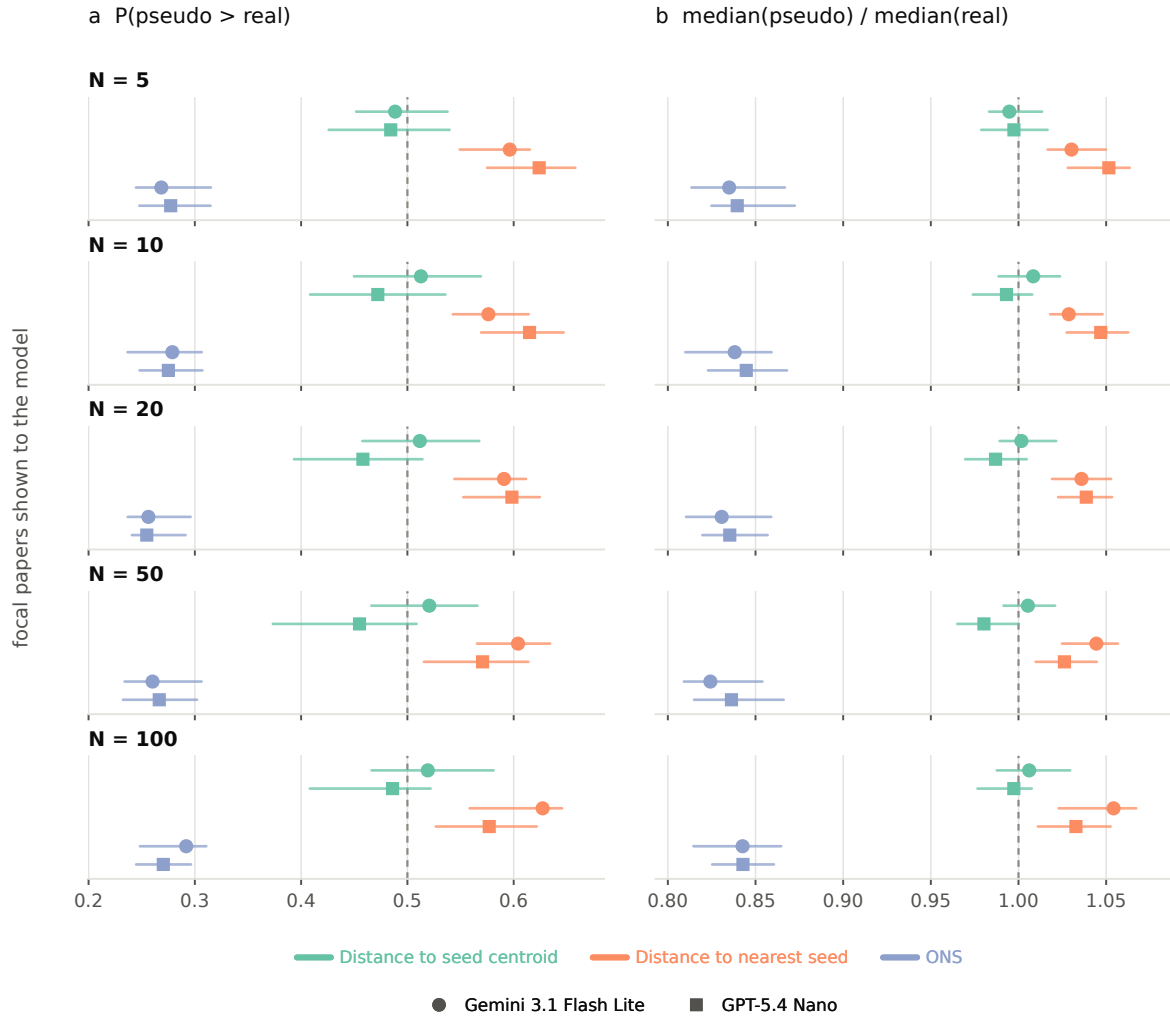


Figure 3: **Context size has no effect on generated hypotheses.** ONS is flat across a 20-fold increase in context. Displacement is also flat. In magnitude terms, LLM hypotheses reach about 84% of real papers' outside-span distance while matching or exceeding them on displacement. More reading does not buy new directions.

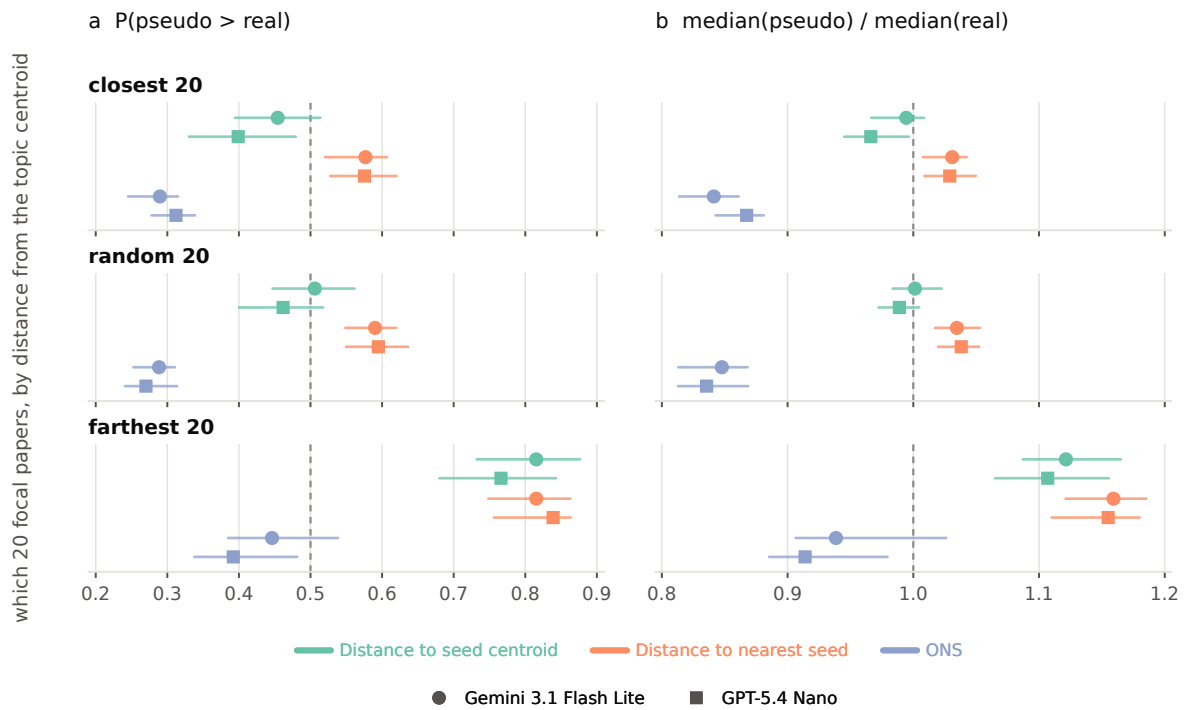


Figure 4: **Context selection pushes displacement and partially direction.** Selecting papers from the edge of the literature strongly amplifies displacement. ONS improves but stays below parity.

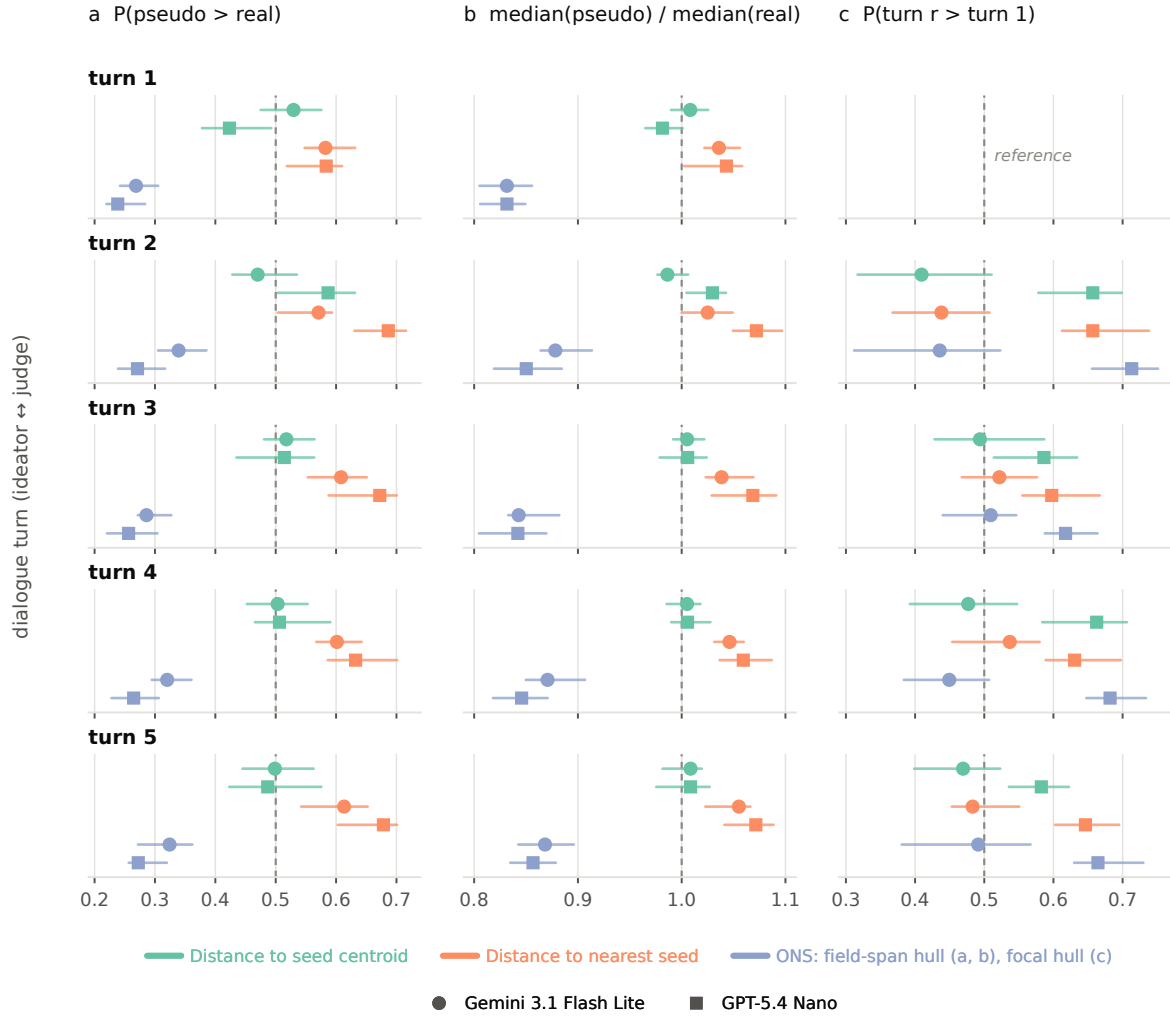


Figure 5: Dialogue increases displacement. Field-span ONS barely moves. Dialogue moves ideas away from where they started, but not out of the field. The effect depends on the model.

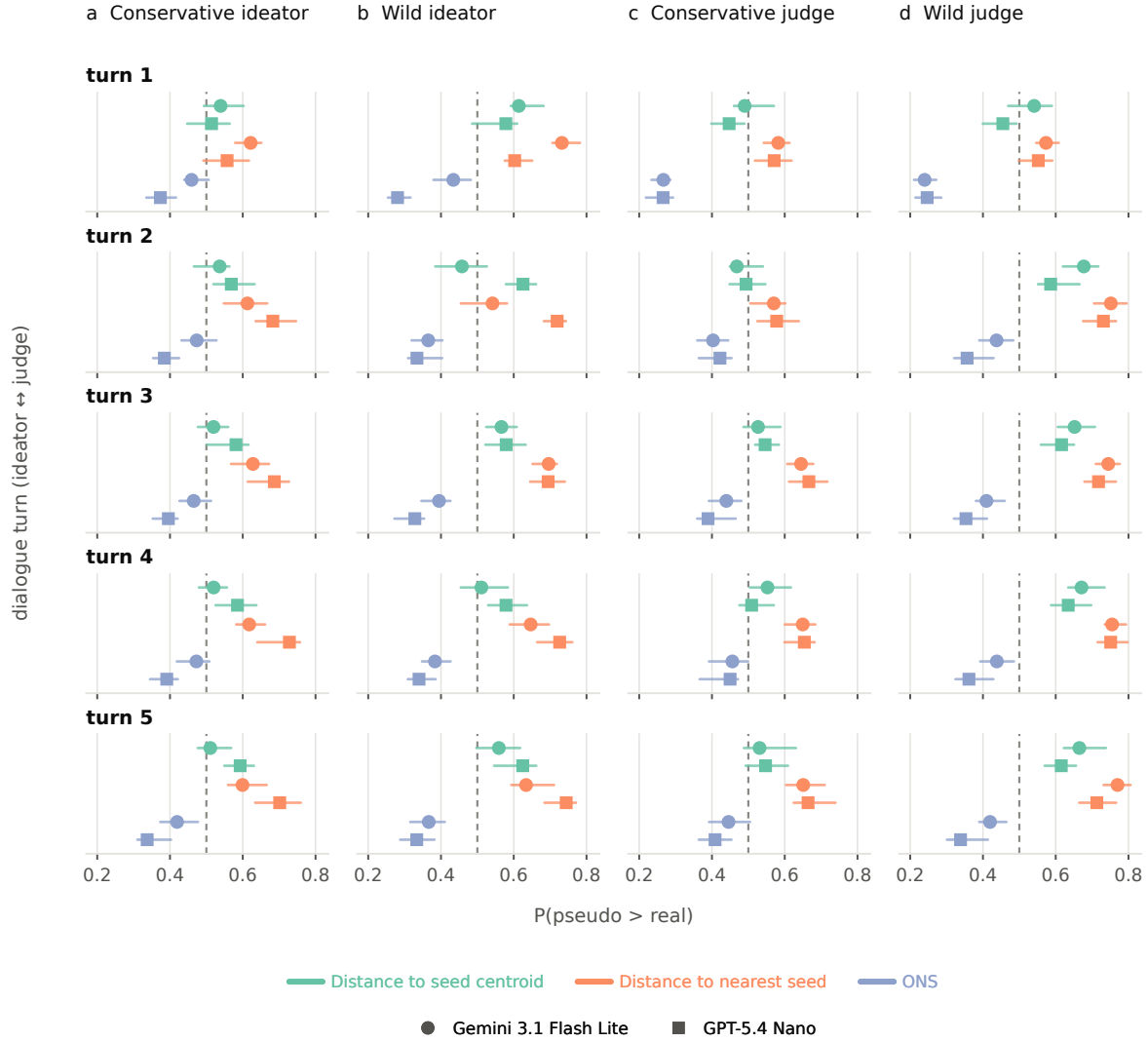


Figure 6: No persona brings ONS to parity. A wild judge is the strongest lever on displacement but not on direction. Asking for "wild" ideas doesn't produce more new directions than asking for conservative ones. Personas lift ONS modestly compared with the default dialogue in Figure 5.