Can large language models help young researchers develop new clinical research ideas?

Lancet Digit Health 2026
Published Online
https://doi.org/10.1016/j.landig.2026.100983

The foundation of science begins with defining the most important ideas.[1] Traditionally, this process of identifying new research questions (ideas or gaps) relies on a (often senior) scientist and their team to synthesise previous research, understand the direction of the field, consolidate feedback from colleagues and expert opinions, and build consensus through discussions and many meetings.[2] To address more complex research problems such as translational or interdisciplinary science, team-based interactive discussions extending beyond traditional single-specialty groups are needed.[3] For example, identifying clinical needs in clinical scenarios is an important pillar for conducting translational and applied biomedical research.[4] However, such approaches often require large teams of scientists, clinicians, and clinician–scientists; are time intensive and resource heavy; and are often conducted in established laboratories with advanced platforms such as academic medical centres or biomedical hubs. These requirements limit the pace and scale of research and innovation in less developed institutions and scientific communities, including younger scientists and clinicians from low-income and middle-income countries (LMICs), leading to dominance of research in developed high-income countries. In this context, identifying research knowledge gaps and novel ideas is a key yet challenging task for young researchers early in their careers, who often receive insufficient guidance from mentors. However, this demographic generally shows higher acceptance of internet-based tools and emerging technologies. Therefore, young researchers represent an ideal population to investigate whether large language models (LLMs) can serve as an effective tool to overcome experience-related constraints and enhance innovative research capabilities.

Development in LLMs, such as ChatGPT,[5] has provided opportunities to explore whether LLMs can generate and develop new scientific research ideas and directions, potentially transforming the entire design and conduct of research. If LLMs can develop sufficiently new and useful research ideas, they might improve the breadth and depth of novel hypothesis generation, enabling young researchers and small teams in less resourceful settings such as LMICs to start and conduct novel research. LLMs will also optimise the allocation of research time, allowing scientists to focus on the conduct and completion of research that is new and not replicative.

To test this hypothesis, we used ophthalmology as a case study. We conducted a comparative study of research idea generation and evaluation across three groups: (1) researchers only (n=3), (2) LLMs only (GPT-4o, Llama 3.1, and Gemma 2), and (3) researchers assisted by LLMs (appendix pp 1–5). In the LLM-assisted researcher group, researchers independently generated initial research ideas and subsequently used an LLM tool of their choice to enhance these ideas.

The three groups were tasked with generating research ideas focused on common ophthalmic conditions, including refractive errors, retinal diseases, and glaucoma, which represent the leading causes of vision impairment and blindness, across three scenarios: (1) proposing new ideas based on a given topic; (2) generating new ideas after reading a published scientific paper; and (3) developing ideas through group discussions, with all three groups evaluated under each scenario. In the first two tasks, each participant individually proposed five ideas, and in the third task, three participants worked together as a group to propose five ideas. To evaluate the ideas, a panel of 15 global ophthalmology expert reviewers with interdisciplinary backgrounds, masked to the source of the ideas, evaluated the quality of the research ideas using an overall evaluation (scored on a 10-point scale) and six evaluation categories (each on a 5-point scale): completeness, relevance, clarity, novelty, feasibility, and impact. Furthermore, a masked classification task was conducted in which reviewers were asked to judge whether each idea was generated by the researcher-only group, the LLM-only group, or the researchers-assisted-by-LLM group.

We used a t-test to assess differences in performance across the three groups, including overall evaluation scores, scores across the six evaluation categories, and stratified analyses of different tasks within each group. Cohen's Kappa statistic was used to determine reviewers' ability to distinguish between research ideas generated by researchers, LLMs, or LLM-assisted researchers. Data were analysed using SPSS (v.24.0) with a designated two-sided, significance level of 5%. This study was approved by the ethics committee of the Zhongshan Ophthalmic Center, Sun Yat-sen University (approval number: 2025KYPJ003),

See Online for appendix

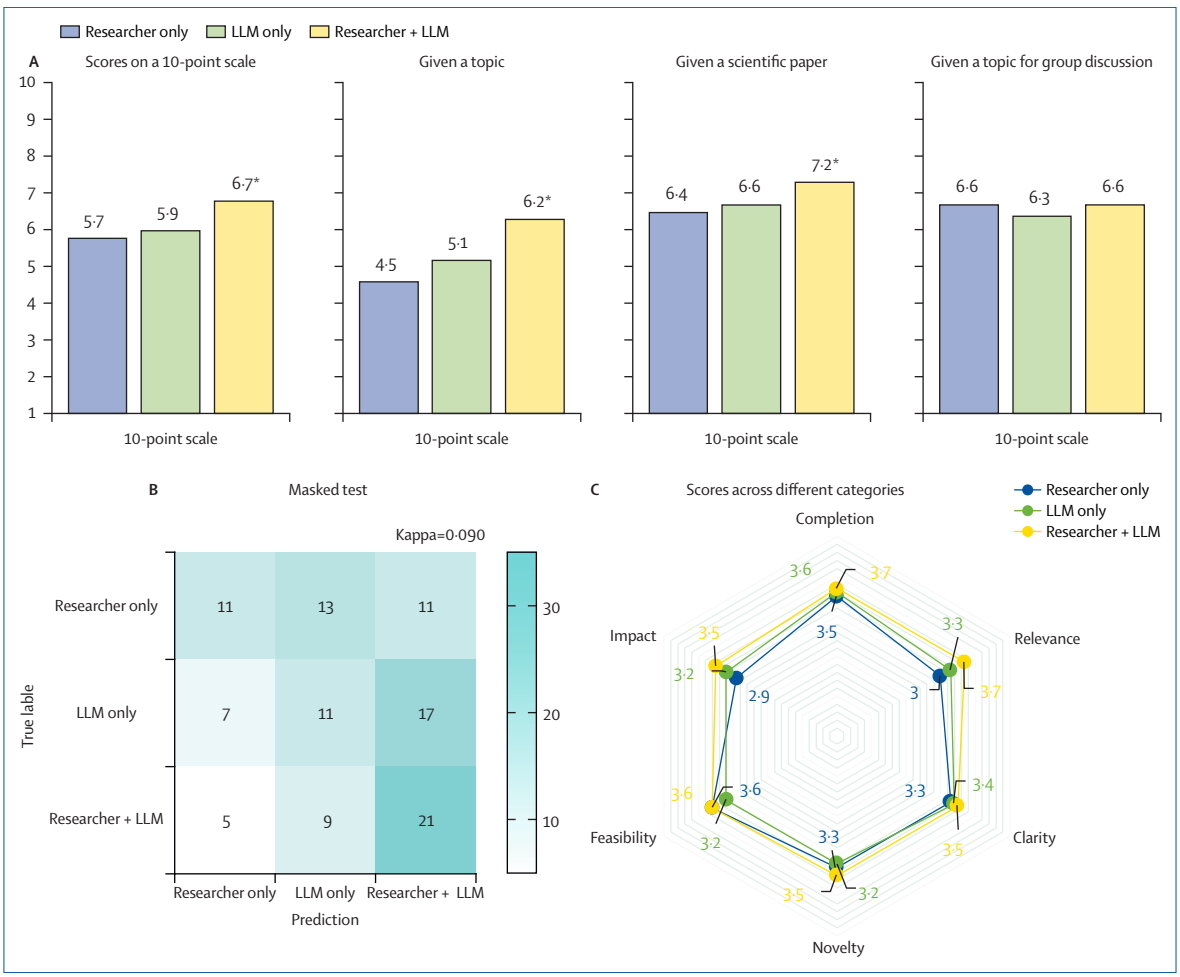

Figure: **Performance evaluation**
(A) Stratified analysis of performance between three different research groups (researcher-only, LLM-only, and LLM-assisted researcher [researcher + LLM]. (B) Masked test. (C) Performance across six evaluation categories. Asterisk indicates statistical significance at p values <0·05 compared with researcher-only group. LLM=large language model.

and registered on the Chinese Clinical Trial Register (ChiCTR2500097263). The participants (researchers) provided written informed consent.

In the overall evaluation using a 10-point scale, the LLM-assisted researcher group outperformed the researcher-only group (p<0·0001), and the performance of the LLM-only group was not significantly different compared with that of the researcher-only group (p=0·951; figure panel A, appendix p 6).

In the stratified analysis of different tasks, overall evaluation using the 10-point scale showed that the use of LLMs significantly improved researchers' performance when proposing research ideas for a given topic (appendix p 6) and when proposing ideas after reading a scientific paper. However, no significant improvement in performance was observed through group discussions on a given topic (figure panel A).

In the masked classification task, Cohen's Kappa value was 0·090 (figure panel B), indicating that reviewers had difficulty distinguishing among research ideas generated by researchers only, LLMs only, and researchers assisted by LLMs.

Finally, we investigated which group generated the highest-scoring research ideas across six evaluation categories. The research idea with the highest feasibility score originated from the researcher-only group. The most complete idea was generated by the LLM-only group. Notably, the ideas receiving the highest scores for novelty, impact, and relevance were produced by the LLM-assisted researcher group (figure panel C, appendix pp 6–7).

Our study provides evidence for the potential of using LLMs to generate novel research ideas and shows that researchers, when assisted by LLMs, showed markedly better performance in generating innovative and impactful ideas than researchers without assistance. This finding underscores a future in which LLMs function as powerful co-creators of ideas and collaborators in research idea generation to accelerate scientific discovery. We showed that this collaborative advantage was task dependent; the benefit was most prominent when generating ideas from a given topic or from previous knowledge (when given a paper) but was negligible in group discussions, wherein the interaction of humans, sharing experience and critical thinking, might yet be irreplaceable. A crucial limitation emerged when LLMs were used in isolation to generate novel ideas.

In terms of the potential impact of these findings, for a young graduate student or postdoctoral researcher, the process of scientific inquiry often starts from reading and learning from a copious amount of scientific (current state-of-the-art) literature and through discussing this literature with their scientific mentor. Through this process, the researcher identifies new research ideas (sometimes called gaps).[6] This process is long, often frustrating, and highly inefficient. In many countries, considerable amount of science is not novel, with many repetitive experiments and research that do not advance the field. Furthermore, because the best ideas[7] originate from large, advanced, well-structured laboratories and well-resourced teams in established biomedical hubs in high-income countries, research ideas and, thus, innovations tend to focus on problems eminent in these highly selected regions.[3] The potential of LLMs to generate new and relevant ideas could radically transform the initiation and conduct of research in other settings such as LMICs.

We also observed that, despite performing comparably to unassisted researchers, the outputs of LLMs showed a tendency towards homogeneity in certain tasks, echoing concerns that over-reliance on current models could reduce the diversity of ideas. This convergence might stem from LLMs being trained on existing literature, potentially perpetuating systemic biases and favouring safe ideas aligned with mainstream thought over high-risk, groundbreaking hypotheses. For instance, in the task generating research ideas based on a given paper, both GPT-4o and Llama 3.1 proposed ideas related to cost-effectiveness analysis of the DeepDR-LLM system (an AI system combining image analysis and language models for diabetic retinopathy screening and management) and its applicability across different types of diabetes patients. Therefore, to responsibly integrate these tools, researchers should prioritise originality. They should also transparently report LLM use in scientific papers, thus ensuring that the quest for efficiency does not compromise the diversity of scientific thought.[8–10]

Future efforts should also focus on exploring robust mechanisms to mitigate reproducibility and hallucination risks, including real-time knowledge grounding, fact-checking protocols, and fine-tuning domain-specific models with verifiable scientific content. These approaches will be crucial for enhancing the reliability of LLM-assisted research ideation while maintaining scientific rigour.

In conclusion, LLMs have shown promise in many areas of science. Our study shows that LLMs can generate high-quality and original research ideas, especially when assisting researchers, thereby enhancing feasibility, impact, and innovation. Our findings are substantiated by comprehensive evaluation that incorporated diverse research scenarios, several advanced LLMs, and masked tests to assess the distinguishability of LLM-generated research ideas. However, although ophthalmology experts with extensive professional experience were invited to conduct masked evaluations, human judgement remains inherently subjective. Additionally, challenges such as output homogeneity, model transparency, and ethical concerns such as academic integrity and data privacy necessitate clear guidelines and responsible integration. We suggest that researchers should strive to strike a balance between maximising the benefits of LLMs and mitigating potential risks. We suggest that LLMs can complement researchers in generating scientific ideas and questions, and this benefit is likely most useful for younger researchers, increasing the diversity of research, particularly in settings without mature research infrastructure or large teams.

CC-Y has received consulting fees from Medi-Whale. He is also a co-founder of Eye AI. PAK has received grants from the UK Research & Innovation Future Leaders Fellowship (MR/T019050/1) and The Rubin Foundation Charitable Trust. He has also received consulting fees from Retina Consultants of America, Roche, Boehringer-Ingelheim, and Topcon. Additionally, PAK has received payment or honoraria for lectures, presentations, and educational events from Zeiss (Boehringer-Ingelheim), Topcon (Apellis), and Novartis (Abbvie). He also holds active patents related to generalisable medical image analysis using segmentation and classification neural networks and has a pending patent for predicting disease progression from tissue images and tissue segmentation maps. PAK participates on advisory boards for Topcon, RetinAI, Bayer, Novartis, and Boehringer-Ingelheim. He has also received support for attending meetings or travel from Bayer, Topcon, and Roche. Furthermore, he holds stock in Cascader, Bitfount, and Big Picture Medical; TYW has received consulting fees from Abbvie Pte, Aldropika Therapeutics, Bayer, Boehringer-Ingelheim, Carl Zeiss, Genentech, Novartis,

Opthea, Plano, Quaerite Biopharm Research, Regeneron Pharmaceuticals Inc, Roche, Sanofi, and Shanghai Henlius. He is also a named inventor on patents related to retinal image analysis and artificial intelligence–based ophthalmic disease detection, and is the co-founder of start-up companies EyRiS and VISRE. The other authors declare that they have no conflict of interest. The authors declare support from the National Key R&D Program (Grant No: 2022YFC2502800 [TYW]), the National Natural Science Fund of China (Grant No: 82388101 [TYW], 82441003 [HL], and 82301265 [YY]), the Beijing Natural Science Foundation (Grant No: IS23096 [TYW]), the China Postdoctoral Science Foundation (Grant No: 2023M734047 [YY]), the Science and Technology Planning Project of Guangdong Province (Grant No: 2023A1111120011 [HL] and 2018B010109008 [HL]), and the Young Elite Scientists Sponsorship Program by China Association for Science and Technology (2024QNRC001 [YY]). YY and LZ had direct access to and verified the raw data in the study and take responsibility for the integrity of the data and the accuracy of the data analysis. TYW and HL conceived and designed the study. YY, DZ, and XC analysed the data. YY drafted the manuscript. DZ, XC, SB, ZL, YL, WL, ZZ, CYC, CKL, KHP, KOM, BS, SP, MA, ChYC, SMS, PK, ZhuZ, JBJ, and YW critically revised the manuscript for important intellectual content. All authors had access to the data, and the final version of the paper has been seen and approved by all the authors. All data generated or analysed during this study are included in this Comment and the appendix. YY, DZ, XC, and LZ are first co-authors.

*Yahan Yang, Danqi Zeng, Xinwei Chen, Lanqin Zhao, Shaowei Bi, Zhangkai Lian, You Li, Wangting Li, Zheming Zhang, Carol Y Cheung, Christopher Kai-shun Leung, Ki Ho Park, Kyoko Ohno-Matsui, Bin Sheng, Shamira Perera, Marcus Ang, Ching-Yu Cheng, Seang Mei Saw, Pearse A Keane, Zhuoting Zhu, Jost B Jonas, Yaxing Wang, *Haotian Lin, *Tien Yin Wong*

linht5@mail.sysu.edu.cn; wongtienyin@tsinghua.edu.cn

Zhongshan Ophthalmic Centre, Sun Yat-sen University, WHO Collaborating Centre for Eye Care and Vision, State Key Laboratory of Ophthalmology, Guangdong Provincial Key Laboratory of Ophthalmology and Visual Science, Guangdong Provincial Clinical Research Center for Ocular Diseases, Guangzhou, China (YY, DZ, XC, LZ, SB, ZL, YL, ZZ, HL, TYW); Shenzhen Eye Hospital, Shenzhen, China (WL); Department of Ophthalmology and Visual Sciences, The Chinese University of Hong Kong, Hong Kong SAR, China (CYC); Department of Ophthalmology, The University of Hong Kong, Hong Kong SAR, China (CKL); Seoul National University, Seoul, South Korea (KHP); Institute of Science Tokyo, Tokyo, Japan (KOM); School of Computer Science, Shanghai Jiao Tong University, Shanghai, China (BS); MOE Key Laboratory of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University, Shanghai, China (BS); Singapore Eye Research Institute, Singapore National Eye Centre, Singapore (SP, MA, ChYC, SMS, TYW); Centre for Innovation and Precision Eye Health, Yong Loo Lin School of Medicine, National University of Singapore, Singapore (ChYC); Department of Ophthalmology, Yong Loo Lin School of Medicine, National University of Singapore, Singapore (ChYC); Ophthalmology & Visual Sciences Academic Clinical Program (Eye ACP), Duke-NUS Medical School, Singapore (SMS); Saw Swee Hock School of Public Health, National University of Singapore, Singapore (SMS); Moorfields Eye Hospital NHS Foundation Trust, London, UK (PK); University College London (UCL), Institute of Ophthalmology, London, UK (PK); Centre for Eye Research Australia, Ophthalmology, University of Melbourne, Melbourne, VIC, Australia (ZhuZ); Rothschild Foundation Hospital, Institut Français de Myopie, Paris, France (JBJ); Beijing Visual Science and Translational Eye Research Institute, Beijing Tsinghua Changgung Hospital, Tsinghua Medicine, Tsinghua University, Beijing, China (JBJ, YW, TYW); Privatpraxis Prof Jonas und Dr Panda-Jonas, Heidelberg, Germany (JBJ); LV Prasad Eye Institute, Hyderabad, Telangana, India (JBJ); Hainan Eye Hospital and Key Laboratory of Ophthalmology, Zhongshan Ophthalmic Center, Sun Yat-sen University, Haikou, Hainan, China (HL); Department of Genetics and Biomedical Informatics, Zhongshan School of Medicine, Institute for Frontier Interdisciplinary Research in Health Sciences and Technology, Sun Yat-sen University, Guangzhou, Guangdong 510623, China (HL); School of Clinical Medicine, Tsinghua Medicine, Tsinghua University, Beijing 100084, China (TYW)

1 Seyferth A, Ratna A, Chung KC. The art of questioning. *Plast Reconstr Surg* 2022; **149:** 1031–35.
2 Vandenbroucke JP, Pearce N. From ideas to studies: how to get ideas and sharpen them into research questions. *Clin Epidemiol* 2018; **10:** 253–64.
3 Wuchty S, Jones BF, Uzzi B. The increasing dominance of teams in production of knowledge. *Science* 2007; **316:** 1036–39.
4 Seyhan AA. Lost in translation: the valley of death across preclinical and clinical divide–identification of problems and overcoming obstacles. *Transl Med Commun* 2019; **4:** 1–19.
5 van Dis EAM, Bollen J, Zuidema W, Van Rooij R, Bockting CL. ChatGPT: five priorities for research. *Nature* 2023; **614:** 224–26.
6 Ley TJ, Rosenberg LE. The physician-scientist career pipeline in 2005: build it, and they will come. *JAMA* 2005; **294:** 1343–51.
7 Naddaf M. 'Labour advantage' drives greater productivity at elite universities. *Nature* 2022; published online Nov 22. https://doi.org/10.1038/d41586-022-03784-4.
8 Padmakumar V, He H. Does writing with language models reduce content diversity?. In: International Conference on Learning Representations. 2023; 2309.05196.
9 Gallifant J, Afshar M, Ameen S, et al. The TRIPOD-LLM reporting guideline for studies using large language models. *Nat Med* 2025; **31:** 60–69.
10 Ibrahim H, Liu X, Denniston AK. Reporting guidelines for artificial intelligence in healthcare research. *Clin Exp Ophthalmol* 2021; **49:** 470–76.