

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our [Editorial Policies](#) and the [Editorial Policy Checklist](#).

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a Confirmed

- | | | |
|-------------------------------------|-------------------------------------|--|
| ☐ | ☒ | The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement |
| ☒ | ☐ | A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly |
| ☐ | ☒ | The statistical test(s) used AND whether they are one- or two-sided
<i>Only common tests should be described solely by name; describe more complex techniques in the Methods section.</i> |
| ☒ | ☐ | A description of all covariates tested |
| ☐ | ☒ | A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons |
| ☐ | ☒ | A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals) |
| ☐ | ☒ | For null hypothesis testing, the test statistic (e.g. F , t , r) with confidence intervals, effect sizes, degrees of freedom and P value noted
<i>Give P values as exact values whenever suitable.</i> |
| ☒ | ☐ | For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings |
| ☒ | ☐ | For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes |
| ☐ | ☒ | Estimates of effect sizes (e.g. Cohen's d , Pearson's r), indicating how they were calculated |

Our web collection on [statistics for biologists](#) contains articles on many of the points above.

Software and code

Policy information about [availability of computer code](#)

Data collection Python 3.9.16, Pandas 2.2.3, NumPy 2.0.1, transformers 4.49.0, sentence-transformers 3.4.1. Code is available at <https://doi.org/10.5281/zenodo.17707647>

Data analysis Python 3.9.16, Pandas 2.2.3, NumPy 2.0.1, Scikit-learn 1.6.0, SciPy 1.13.1. Code is available at <https://doi.org/10.5281/zenodo.17707647>

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio [guidelines for submitting code & software](#) for further information.

Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our [policy](#)

All data generated or analysed during this study have been deposited in <https://doi.org/10.5281/zenodo.17707879>

Research involving human participants, their data, or biological material

Policy information about studies with [human participants or human data](#). See also policy information about [sex, gender \(identity/presentation\), and sexual orientation](#) and [race, ethnicity and racism](#).

Reporting on sex and gender	Sex and gender were not considered in the study design. The research focuses on the technical evaluation of Large Language Model (LLM) outputs by domain experts. The biological sex or gender identity of the evaluators is not considered a relevant variable for the quality of technical assessment, and thus was not collected or analyzed.
Reporting on race, ethnicity, or other socially relevant groupings	Data on race, ethnicity, or other socially relevant groupings were not collected. The participants were selected solely based on their professional expertise relevant to the evaluation tasks, rather than demographic characteristics.
Population characteristics	The study involved 6 participants who are experts in the field of partial differential equations (PDEs). They were recruited from Renmin University of China. No clinical or sensitive personal data were collected.
Recruitment	Participants were recruited at Renmin University of China. All participants provided written informed consents on a voluntary basis without compensation.
Ethics oversight	The study was approved by the Scientific and Ethic Committee of Renmin University of China.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☒ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Behavioural & social sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	This is a quantitative study involving expert evaluation of Large Language Model (LLM) outputs. Data consists of scores/ratings provided by domain experts regarding the quality and accuracy of the model responses.
Research sample	The sample consisted of 6 domain experts from Renmin University of China. Participants were selected based on their specific expertise in PDEs to ensure accurate assessment of the model outputs. The sample is representative of qualified evaluators in this domain.
Sampling strategy	Participants were recruited through purposive sampling based on their professional qualifications. No statistical sample size calculation was performed. The sample size was determined to ensure that each model output received at least 6 independent evaluations to guarantee reliability.
Data collection	Data were collected using a digital questionnaire. Evaluators assessed the model outputs displayed on a computer screen. The study was blinded: evaluators were not aware of which specific model generated the responses they were evaluating to prevent bias.
Timing	Data collection was conducted during Feb to May 2025.
Data exclusions	No data were excluded from the analyses.
Non-participation	No participants dropped out or declined participation after providing consent.
Randomization	Participants were not allocated into experimental groups. All participants performed the same evaluation tasks.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Plants

Seed stocks

Report on the source of all seed stocks or other plant material used. If applicable, state the seed stock centre and catalogue number. If plant specimens were collected from the field, describe the collection location, date and sampling procedures.

Novel plant genotypes

Describe the methods by which all novel plant genotypes were produced. This includes those generated by transgenic approaches, gene editing, chemical/radiation-based mutagenesis and hybridization. For transgenic lines, describe the transformation method, the number of independent lines analyzed and the generation upon which experiments were performed. For gene-edited lines, describe the editor used, the endogenous sequence targeted for editing, the targeting guide RNA sequence (if applicable) and how the editor was applied.

Authentication

Describe any authentication procedures for each seed stock used or novel genotype generated. Describe any experiments used to assess the effect of a mutation and, where applicable, how potential secondary effects (e.g. second site T-DNA insertions, mosaicism, off-target gene editing) were examined.