nature portfolio

Corresponding author(s): Luca M. Schulze Buschoff

Last updated by author(s): Nov 7, 2024

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 *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	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 *Give P values as exact values whenever suitable.*
☒	☐	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	The open-source models (lama_adapter_v2_multimodal7b, Fuyu-8b, OTTER-Image-MPT7B) were installed per the instructions on their related Github or Huggingface repositories and evaluated on a Slurm-based cluster with a single A100. For the results reported as GPT-4V, we used the public ChatGPT interface and the OpenAI API, specifically the November 2023 release of gpt4-vision-preview model which is not available anymore via the completions endpoint. For CLAUDE-3 we used claude-3-opus-20240229 with the Anthropic API. Code for replicating our results is available on GitHub (github.com/lsbuschoff/multimodal).
Data analysis	All models were evaluated in Python using PyTorch (2.3.1 and py3.12_cuda12.1_cudnn8.9.2_0). Additional analyses were carried out using NumPy (1.26.4), Pandas (2.2.2), and SciPy (1.13.1). Matplotlib (3.8.4) and Seaborn (0.12.2) were used for plotting. Bayesian mixed effects models were computed using brms (2.21.0) in R.

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 used in our experiments is available on GitHub (github.com/lsbuschoff/multimodal). We have used subsets of openly available data sets from Lerer et al. (https://github.com/facebookarchive/UETorch/issues/25#issuecomment-235688223), Gerstenberg et al. (https://github.com/tobiasgerstenberg/eye_tracking_causality), Zhou et al. (https://github.com/ciclstanford/mental_jenga), Wu et al. (https://github.com/cicl-stanford/counterfactual_agents), and Jara-Ettinger et al. (https://osf.io/uzs8r/).

Human research participants

Policy information about studies involving human research participants and Sex and Gender in Research.

Reporting on sex and gender	107 subjects (55 female and 52 male), no sex or gender based analyses were performed
Population characteristics	native English speakers with a mean age of 27.73 (sd = 4.21)
Recruitment	Participants were recruited from Prolific with the constraint of requiring native English speakers
Ethics oversight	Ethics commission at the faculty of medicine, Eberhard Karls University Tübingen

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	Quantitative study: evaluation of vision language models on cognitive domains and comparison to human data
Research sample	Five state of the art multimodal large language models. This sample of models were chosen so that they adequately represent current SOTA models (large and small, open source and closed source).
Sampling strategy	No sample size calculation was performed. Samples were determined by the available data sets in the respective domains.
Data collection	The open-source models were evaluated on a Slurm-based cluster with a single A100. For GPT-4V, we used the public ChatGPT interface and the OpenAI API, specifically the November 2023 release of gpt4-vision-preview model. For CLAUDE-3 we used claude-3-opus-20240229 with the Anthropic API. Data collection code is available on GitHub (github.com/lsbuschoff/multimodal). Additional human data was collected on Prolific for one evaluation. Human subjects first saw an example trial, followed by 100 test images. In a 2AFC paradigm, subjects were asked if the block tower in a given image was stable or not stable. They were paid £1.5 and the median time they took to complete the experiment was 08:08 minutes, making the average base reward £11.07 per hour. Additionally, they received a bonus payment of up to £1 depending on their performance (1 cent for each correct answer). The researcher was not blinded to experimental condition or the study hypothesis.
Timing	Data collection mainly took place in October and November of 2023 and again in May and June of 2024.
Data exclusions	The raw answers from the large language models were parsed and non-intelligible answers were registered as N/A. Human subjects were excluded if they did not complete the experiment.
Non-participation	For the human experiment, 11 participants started the experiment but did not complete it.

Randomization	All models received the same stimuli in the same order. Models were input images together with each question, since gradients were enabled and the models queried in a new context for each question, randomization is not required. Human participants received the same stimuli in a randomized order.

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

Methods

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