

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 (<i>n</i>) 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. <i>F</i> , <i>t</i> , <i>r</i>) with confidence intervals, effect sizes, degrees of freedom and <i>P</i> 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 <i>d</i> , Pearson's <i>r</i>), 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	We used the open source package AnnDictionary v 0.3.65 to query large language models for cell type annotation through several APIs as detailed in the manuscript and repository documentation. The git repository benchmark_llms linked in the Code Availability section of the manuscript contains the full computational pipeline.
Data analysis	The code to run the LLM benchmarking pipeline and reproduce all figures and tables presented in this study is publicly available in the benchmark_llms repository at https://github.com/ggit12/benchmark_llms and is archived in Zenodo at https://doi.org/10.5281/zenodo.16816304. The AnnDictionary source code is publicly available in the anndictionary repository at https://github.com/ggit12/anndictionary. AnnDictionary is also available on PyPI for install via pip. For reproducibility, the specific version of the AnnDictionary associated with this publication is archived in Zenodo at https://doi.org/10.5281/zenodo.16815933. For general use of the package, install the latest version via pip. We have also included https://github.com/ggit12/benchmark_llms/blob/main/environment.yml in the benchmark_llms repository to show all package versions and included a SIF file which can be used as well. Below, we show the contents of the environment.yml, which contains all version numbers of software used. <pre>name: benchmark_llms channels: - conda-forge - defaults dependencies:</pre>

```
- _libgcc_mutex=0.1
- _openmp_mutex=5.1
- bzip2=1.0.8
- ca-certificates=2025.2.25
- expat=2.6.4
- firefox=134.0
- geckodriver=0.36.0
- ld_impl_linux-64=2.40
- libblas=3.9.0
- libcbblas=3.9.0
- libexpat=2.6.4
- libffi=3.4.4
- libgcc=14.2.0
- libgcc-ng=14.2.0
- libgfortran=14.2.0
- libgfortran5=14.2.0
- libgomp=14.2.0
- libhwloc=2.11.2
- libiconv=1.18
- liblapack=3.9.0
- libllvm15=15.0.7
- liblzma=5.6.4
- libnsl=2.0.1
- libopenblas=0.3.29
- libsqlite=3.45.2
- libstdcxx=14.2.0
- libstdcxx-ng=11.2.0
- libuuid=2.38.1
- libxcrypt=4.4.36
- libxml2=2.13.6
- libzlib=1.3.1
- llvmlite=0.44.0
- ncurses=6.4
- numba=0.61.0
- openssl=3.4.1
- pip=25.0
- python=3.12.2
- python_abi=3.12
- readline=8.2
- setuptools=75.8.0
- sqlite=3.45.2
- tbb=2022.0.0
- tk=8.6.13
- wheel=0.45.1
- xz=5.6.4
- zlib=1.3.1
- zstd=1.5.7
- pip:
  - adjusttext==1.3.0
  - aiohappyeyeballs==2.4.8
  - aiohttp==3.11.13
  - aiosignal==1.3.2
  - anndata==0.10.8
  - anndict==0.3.65
  - annotated-types==0.7.0
  - anthropic==0.49.0
  - anyio==4.8.0
  - appdirs==1.4.4
  - argon2-cffi==23.1.0
  - argon2-cffi-bindings==21.2.0
  - argparse-dataclass==2.0.0
  - array-api-compat==1.11.1
  - arrow==1.3.0
  - asttokens==3.0.0
  - async-lru==2.0.5
  - attrs==25.1.0
  - babel==2.17.0
  - beautifulsoup4==4.13.3
```

- bleach==6.2.0
- bokeh==3.4.3
- boto3==1.34.162
- botocore==1.34.162
- cachetools==5.5.2
- certifi==2025.1.31
- cffi==1.17.1
- charset-normalizer==3.4.1
- colorcet==3.1.0
- comm==0.2.2
- conda-inject==1.3.2
- configargparse==1.7
- connection-pool==0.0.3
- contourpy==1.3.1
- cycler==0.12.1
- dataclasses-json==0.6.7
- datrie==0.8.2
- debugpy==1.8.13
- decorator==5.2.1
- defusedxml==0.7.1
- distro==1.9.0
- docutils==0.21.2
- dpath==2.2.0
- executing==2.2.0
- fastjsonschema==2.21.1
- fonttools==4.56.0
- fqdn==1.5.1
- frozenlist==1.5.0
- gitdb==4.0.12
- gitpython==3.1.44
- google-ai-generativelanguage==0.6.6
- google-api-core==2.24.1
- google-api-python-client==2.162.0
- google-auth==2.38.0
- google-auth-http2==0.2.0
- google-generativeai==0.7.2
- googleapis-common-protos==1.69.0
- greenlet==3.1.1
- grpcio==1.71.0rc2
- grpcio-status==1.63.0rc1
- h11==0.14.0
- h5py==3.13.0
- harmonypy==0.0.10
- holoviews==1.19.1
- httpcore==1.0.7
- httplib2==0.22.0
- httpx==0.28.1
- humanfriendly==10.0
- idna==3.10
- igraph==0.11.8
- immutables==0.21
- ipykernel==6.29.5
- ipython==8.26.0
- ipywidgets==8.1.5
- isoduration==20.11.0
- jedi==0.19.2
- jinja2==3.1.5
- jiter==0.8.2
- jmespath==1.0.1
- joblib==1.4.2
- json5==0.10.0
- jsonpatch==1.33
- jsonpointer==3.0.0
- jsonschema==4.23.0
- jsonschema-specifications==2024.10.1
- jupyter==1.1.1
- jupyter-client==8.6.3
- jupyter-console==6.6.3

- jupyter-core==5.7.2
- jupyter-events==0.12.0
- jupyter-lsp==2.2.5
- jupyter-server==2.15.0
- jupyter-server-terminals==0.5.3
- jupyterlab==4.3.6
- jupyterlab-pygments==0.3.0
- jupyterlab-server==2.27.3
- jupyterlab-widgets==3.0.13
- kiwisolver==1.4.8
- krippendorff==0.7.0
- langchain==0.2.14
- langchain-anthropic==0.1.23
- langchain-aws==0.1.17
- langchain-community==0.2.12
- langchain-core==0.2.43
- langchain-google-genai==1.0.10
- langchain-openai==0.1.22
- langchain-text-splitters==0.2.4
- langsmith==0.1.147
- legacy-api-wrap==1.4.1
- leidenalg==0.10.2
- linkify-it-py==2.0.3
- markdown==3.7
- markdown-it-py==3.0.0
- markupsafe==3.0.2
- marshmallow==3.26.1
- matplotlib==3.9.2
- matplotlib-inline==0.1.7
- mdit-py-plugins==0.4.2
- mdurl==0.1.2
- mistune==3.1.3
- multidict==6.1.0
- mypy-extensions==1.0.0
- natsort==8.4.0
- nbclient==0.10.2
- nbconvert==7.16.6
- nbformat==5.10.4
- nest-asyncio==1.6.0
- networkx==3.4.2
- notebook==7.3.3
- notebook-shim==0.2.4
- numpy==1.26.4
- openai==1.65.3
- orjson==3.10.15
- outcome==1.3.0.post0
- overrides==7.7.0
- packaging==24.2
- pandas==2.2.2
- pandocfilters==1.5.1
- panel==1.4.5
- param==2.2.0
- parso==0.8.4
- patsy==1.0.1
- pexpect==4.9.0
- pillow==11.1.0
- plac==1.4.3
- platformdirs==4.3.6
- prometheus-client==0.21.1
- prompt-toolkit==3.0.50
- propcache==0.3.0
- proto-plus==1.26.0
- protobuf==4.25.6
- psutil==7.0.0
- ptyprocess==0.7.0
- pulp==2.9.0
- pure-eval==0.2.3
- pyasn1==0.6.1

- pyasn1-modules==0.4.1
- pycparser==2.22
- pydantic==2.10.6
- pydantic-core==2.27.2
- pygments==2.19.1
- pynndescent==0.5.13
- pyparsing==3.2.1
- pysocks==1.7.1
- python-dateutil==2.9.0.post0
- python-dotenv==1.0.1
- python-json-logger==3.3.0
- pytz==2025.1
- pyviz-comms==3.0.4
- pyyaml==6.0.2
- pyzmq==26.3.0
- referencing==0.36.2
- regex==2024.11.6
- requests==2.32.3
- requests-toolbelt==1.0.0
- reretry==0.11.8
- rfc3339-validator==0.1.4
- rfc3986-validator==0.1.1
- rpds-py==0.23.1
- rsa==4.9
- s3transfer==0.10.4
- scanpy==1.10.2
- scikit-learn==1.5.1
- scipy==1.14.1
- seaborn==0.13.2
- selenium==4.29.0
- send2trash==1.8.3
- session-info==1.0.0
- sigfig==1.3.19
- six==1.17.0
- smart-open==7.1.0
- smmap==5.0.2
- snakemake==8.29.0
- snakemake-executor-plugin-slurm==0.15.1
- snakemake-executor-plugin-slurm-jobstep==0.2.1
- snakemake-interface-common==1.17.4
- snakemake-interface-executor-plugins==9.3.3
- snakemake-interface-report-plugins==1.1.0
- snakemake-interface-storage-plugins==3.3.0
- sniffio==1.3.1
- sortedcontainers==2.4.0
- soupsieve==2.6
- sqlalchemy==2.0.38
- stack-data==0.6.3
- statsmodels==0.14.4
- stdlib-list==0.11.1
- tabulate==0.9.0
- tenacity==8.5.0
- terminado==0.18.1
- texttable==1.7.0
- threadpoolctl==3.5.0
- throttler==1.2.2
- tiktoken==0.9.0
- tinycss2==1.4.0
- tornado==6.4.2
- tqdm==4.67.1
- traitlets==5.14.3
- trio==0.29.0
- trio-websocket==0.12.2
- types-python-dateutil==2.9.0.20241206
- typing-extensions==4.12.2
- typing-inspect==0.9.0
- tzdata==2025.1
- uc-micro-py==1.0.3

```

- umap-learn==0.5.7
- uri-template==1.3.0
- uritemplate==4.1.1
- urllib3==2.3.0
- wcwidth==0.2.13
- webcolors==24.11.1
- webencodings==0.5.1
- websocket-client==1.8.0
- widgetsnbextension==4.0.13
- wrapt==1.17.2
- wsproto==1.2.0
- xyzservices==2025.1.0
- yarl==1.18.3
- yte==1.7.0

```

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](#)

This study used publicly available data. The gene expression data used in this study have been deposited in Figshare under accession code <https://doi.org/10.6084/m9.figshare.27921984.v1>. The raw fastq files are available under restricted access and access can be obtained by submitting a request at <https://tabula-sapiens.sf.czbiohub.org/whereisthedata>. All data processing can be reproduced via the provided code, see Code Availability. Source data are provided with this paper.

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

N/A

Reporting on race, ethnicity, or other socially relevant groupings

N/A

Population characteristics

N/A

Recruitment

N/A

Ethics oversight

N/A

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](https://www.nature.com/documents/nr-reporting-summary-flat.pdf)

Life sciences study design

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

Sample size

For cell type annotation, sample size was determined by using all available data in Tabula Sapiens V2, which includes 1,136,335 cells. For the biological process annotation study, we used 500 gene sets selected randomly from sets containing 3 to 100 genes from the human Gene Ontology Biological Process 2023 database, based on previous work: Hu, M., Alkhairy, S., Lee, I. et al. Evaluation of large language models for discovery of gene set function. Nat Methods 22, 82–91 (2025). <https://doi.org/10.1038/s41592-024-02525-x>

Data exclusions

Quality control filtering criteria for cells were previously applied by the Tabula Sapiens Consortium. For one control analysis, we removed

	pancreas due to low cell abundance.
Replication	All performance measurements were averaged across five replicates. All replicates were successful.
Randomization	This study collected data from large language models, which inherently produce stochastic output. This is why all measurements are reported across five replicates. This study also used a random sample of 500 gene sets selected randomly from sets containing 3 to 100 genes from the human Gene Ontology Biological Process 2023 database.
Blinding	Blinding was not relevant for this study because it is an automated benchmarking study with fixed datasets, automated metrics, and no human ratings.

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		Methods	
n/a	Involved in the study	n/a	Involved in the study
☒	☐ Antibodies	☒	☐ ChIP-seq
☒	☐ Eukaryotic cell lines	☒	☐ Flow cytometry
☒	☐ Palaeontology and archaeology	☒	☐ MRI-based neuroimaging
☒	☐ Animals and other organisms		
☒	☐ Clinical data		
☒	☐ Dual use research of concern		
☒	☐ Plants		

Plants

Seed stocks	N/A
Novel plant genotypes	N/A
Authentication	N/A