

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 have collected data using custom software - MyGPT, which uses various open-source components such as LLMs embedding models. We have mentioned all the versions of LLMs and embedding models in the manuscript. We used Ollama v0.1.48 to host various LLMs used in the study. In addition, we used Chroma vector database version 0.5.7 and reported them in the manuscript. A fully open-source code for MyGPT is available in GitHub at https://github.com/stjude/MyGPT and via Zenodo at https://doi.org/10.5281/zenodo.22165249
Data analysis	We performed data analysis using custom Python (v3.10) and JavaScript (Node v20.17.0) programs. The Python scripts used packages such as Pandas v2.1.4, scikit-learn v1.6.1, seaborn v0.13.2, and matplotlib v3.9.4, and JavaScript used D3.js v7.8.3 for statistical analysis.

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](#)

The BioASQ evaluation corpus used in this study was curated as a collection of PDF documents and has been deposited in Zenodo under <https://doi.org/10.5281/>

zenodo.22165150. The PubMedQA and Open RAG Benchmark datasets, as well as the kinase literature use case dataset, are based on published articles obtained through institutional subscriptions and licensing agreements. Although the corresponding PDF files cannot be redistributed, the PubMed IDs (PMIDs) and Digital Object Identifiers (DOIs) required to retrieve the source publications are provided in the above Zenodo repository. The BMTCT Standard Operating Procedure (SOP) documents are not publicly available as they contain confidential institutional information that is subject to institutional legal and administrative requirements. The authors do not have permission to redistribute these materials. PDF of health policy documents analyzed in the multilingual study are publicly available and have been deposited in the above Zenodo repository. Reviewer-validated assessments of generated multilingual policy analyses are not publicly available but may be obtained from the corresponding author for academic research purposes, subject to any applicable data-sharing requirements. Evaluation results for all other experiments presented in the manuscript are provided in Supplementary Data 1. For all datasets, the scripts used to reproduce the results are available at: <https://github.com/stjude/MyGPT-evaluations> and <https://doi.org/10.5281/zenodo.22165150>.

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

Use the terms sex (biological attribute) and gender (shaped by social and cultural circumstances) carefully in order to avoid confusing both terms. Indicate if findings apply to only one sex or gender; describe whether sex and gender were considered in study design; whether sex and/or gender was determined based on self-reporting or assigned and methods used. Provide in the source data disaggregated sex and gender data, where this information has been collected, and if consent has been obtained for sharing of individual-level data; provide overall numbers in this Reporting Summary. Please state if this information has not been collected. Report sex- and gender-based analyses where performed, justify reasons for lack of sex- and gender-based analysis.

Reporting on race, ethnicity, or other socially relevant groupings

Please specify the socially constructed or socially relevant categorization variable(s) used in your manuscript and explain why they were used. Please note that such variables should not be used as proxies for other socially constructed/relevant variables (for example, race or ethnicity should not be used as a proxy for socioeconomic status). Provide clear definitions of the relevant terms used, how they were provided (by the participants/respondents, the researchers, or third parties), and the method(s) used to classify people into the different categories (e.g. self-report, census or administrative data, social media data, etc.) Please provide details about how you controlled for confounding variables in your analyses.

Population characteristics

Describe the covariate-relevant population characteristics of the human research participants (e.g. age, genotypic information, past and current diagnosis and treatment categories). If you filled out the behavioural & social sciences study design questions and have nothing to add here, write "See above."

Recruitment

Describe how participants were recruited. Outline any potential self-selection bias or other biases that may be present and how these are likely to impact results.

Ethics oversight

Identify the organization(s) that approved the study protocol.

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 study presents and evaluates MyGPT, an open-source, privacy-preserving retrieval-augmented generation (RAG) platform. The study combines software engineering, benchmarking, and quantitative evaluation methods. Performance was assessed using three public question-answering datasets (PubMedQA, BioASQ, and Open RAG Benchmark) and three real-world use cases involving biomedical literature, clinical standard operating procedures (SOPs), and multilingual health policy documents. Outcomes included answer correctness, retrieval performance, and confidence metrics (Question Relevance Score, Answer Relevance Score, and Hallucination Index).

Research sample

The research sample consisted of document collections and associated question-answer datasets rather than human participants. Public benchmarking datasets included 294 PubMedQA articles and questions, 163 BioASQ questions with 210 curated PDF documents, and 170 Open RAG Benchmark questions with 165 associated documents. Real-world evaluations included 83 biomedical publications with 282 expert-generated question-answer pairs, 68 BMTCT SOP documents with 136 questions, and 20 multilingual health policy documents evaluated using 320 questions across Spanish and French policy collections. Expert reviewers and domain specialists assessed answer correctness against predefined reference answers.

Sampling strategy

Public benchmark datasets were selected because they are widely used for evaluating question-answering and RAG systems. Subsets were used when full datasets could not be evaluated because of article-access limitations or the extensive manual effort required for validation. Sample sizes were guided by power analyses that accounted for paired evaluations and clustering within documents. For

	real-world use cases, question-answer datasets were generated by domain experts to reflect practical information-seeking tasks relevant to clinical SOPs, biomedical literature, and global health policies.
Data collection	Data were collected using publicly available benchmark datasets, institutionally licensed scientific literature, confidential clinical SOP documents, and publicly available health policy documents. MyGPT was used to generate answers to predefined questions. Correctness labels were obtained through independent manual review by researchers or domain experts using predefined reference answers and structured evaluation procedures. For public benchmarks, evaluators were blinded to the identity of the RAG system that generated each answer. Data collection and analysis were conducted using custom Python and JavaScript software.
Timing	Benchmarking, validation, and data collection were performed during software development and manuscript preparation. The study used existing public datasets, institutional SOP libraries, biomedical literature collections, and health policy repositories available at the time of analysis. No longitudinal follow-up or repeated collection periods were involved. The manuscript does not rely on participant enrollment over a defined time interval. Data collection and analyses were conducted between 2024 and 2026.
Data exclusions	No predefined exclusions were applied to the final evaluation datasets beyond limitations imposed by document availability, institutional access to published articles, and data completeness requirements for specific analyses. For machine-learning analyses, question-answer records with missing semantic, keyword, reranking, or correctness-label information were excluded prior to model training.
Non-participation	The study did not involve recruitment of human research participants. Consequently, there were no participant dropouts, withdrawals, refusals, or non-response rates. Expert reviewers and domain specialists completed assessment tasks as part of the evaluation workflow.
Randomization	No participant randomization or allocation to experimental groups was performed. All benchmark datasets and document collections were evaluated directly using predefined study protocols. For machine-learning analyses, datasets were split into training (80%) and test (20%) sets using stratified sampling, and 10 independent replications with different random seeds were performed to assess robustness. Evaluators of public benchmark datasets were blinded to the identity of the RAG system that generated each answer.

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	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.