

Democratizing Reliable Knowledge-Seeking with MyGPT: A Privacy-First, Open-Source Retrieval-Augmented Generation Platform

Corresponding Author: Dr M. Madan Babu

This file contains all reviewer reports in order by version, followed by all author rebuttals in order by version.

Version 0:

Reviewer comments:

Reviewer #1

(Remarks to the Author)

The authors present MyGPT, a privacy-first, open-source Retrieval-Augmented Generation (RAG) platform designed to operate locally. The proposed system integrates various components of a RAG pipeline, such as document chunking, organization, and LLM inference, into a end-to-end workflow intended to allow researchers to query private documents without data leakage. The system is benchmarked against a standalone LLM using a curated set of biomedical questions.

While the paper presents a functional tool, it more resembles an engineering technical report rather as compared to a traditional paper in this venue. In addition, I have five main concerns regarding the scientific novelty and validation of the work:

1. Lack of Differentiation from Existing Solutions.

Although the authors mention other open-source solutions around lines 289–295 and in Extended Data Table 5, the manuscript does not demonstrate how MyGPT substantially differs from previous work. The comparison offers only qualitative claims regarding flexibility. Furthermore, the comparison in Table 5 is incomplete; it omits numerous prominent open-source RAG frameworks, such as RAGFlow (<https://github.com/infiniflow/ragflow>) and Kotaemon (<https://github.com/Cinnamon/kotaemon>), both of which offer simple deployment solutions and user interface support than the proposed. The discussion in Extended Data Figure 1 should also include these open-source baselines to better establish novelty, rather than limiting the comparison to commercial models. Ultimately, the manuscript does not articulate how the architecture advances the field of Information Retrieval or NLP beyond providing an interface wrapper around existing libraries.

2. Insufficient Benchmarking Baselines.

The included benchmark results only compare MyGPT against Llama2 alone. It is well-established in the literature that RAG outperforms standalone LLMs in domain-specific, knowledge-intensive tasks. Consequently, proving that MyGPT outperforms a model without context is quite a trivial result. There is no experiment demonstrating the superior ability (or even parity) of the proposed MyGPT against other state-of-the-art, open-source RAG frameworks.

3. Questionable Validity of the Hallucination Index

The proposed Hallucination Index is questionable. It is calculated as a weighted combination of the Question Relevance Score (QRS) and Answer Relevance Score (ARS), which are derived from the embedding similarity between the query/context and context/answer. This approach relies on semantic distance, which is not a proxy for factual correctness. For example, the statement "The drug is safe" and "The drug is not safe" may share high semantic similarity in vector space, yet represent opposite factual claims. A high similarity score could therefore coexist with a factually incorrect answer, rendering the index unreliable.

4. Flawed Experimental Design and Sample Size.

The benchmarking methodology is problematic. First, the evaluation relies on only 40 questions derived from the manuscript itself. This sample size is (unnecessarily) too small to draw statistically significant conclusions regarding the robustness of the system or the utility of the Hallucination Index. Second, the baseline setting is methodologically flawed because the

standalone LLM was not trained on the private manuscript; therefore, a 0% accuracy rate is entirely expected. This comparison only confirms that having context is helpful, which is already known. A correct evaluation would provide the same document chunks to both the baseline (e.g., a standard RAG implementation) and MyGPT to demonstrate that MyGPT offers better retrieval or higher-quality context processing.

5. Structural and Organizational Deficiencies.

Finally, the manuscript requires substantial structural revision to meet the standards of a technical research paper. Currently, the text transitions directly from the Introduction to the Results, omitting key transitions in the main body. Consequently, crucial details regarding the architectural design and pipeline implementation are fragmented or relegated to the later sections at the end of the document and supplementary.

Conclusion:

Although MyGPT provides a complete, reproducible, end-to-end framework for RAG, I cannot recommend acceptance in its current state. The work acts primarily as a wrapper for existing tools without offering significant algorithmic novelty or rigorous comparative benchmarking. It is believed that significant modification, including comparison against proper RAG baselines, is required before resubmission.

(Remarks on code availability)

I believe the source code could still provide value to the community, particularly because it includes detailed instructions for using and deploying MyGPT. That said, I want to note that the proposed MyGPT framework does not appear to differ substantially from existing RAG solutions—such as those built with LangChain—beyond its emphasis on cost effectiveness and its integration with scholarly sources like Zotero.

Moreover, even within the open-source ecosystem, there already exist extensive community-maintained resources (e.g., the comprehensive collection at https://github.com/NirDiamant/RAG_Techniques). This raises some questions about the extent to which the contribution meaningfully advances or enriches the existing landscape.

Reviewer #2

(Remarks to the Author)

I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

(Remarks on code availability)

I believe the source code could still provide value to the community, particularly because it includes detailed instructions for using and deploying MyGPT. That said, I want to note that the proposed MyGPT framework does not appear to differ substantially from existing RAG solutions—such as those built with LangChain—beyond its emphasis on cost effectiveness and its integration with scholarly sources like Zotero.

Moreover, even within the open-source ecosystem, there already exist extensive community-maintained resources (e.g., the comprehensive collection at https://github.com/NirDiamant/RAG_Techniques). This raises some questions about the extent to which the contribution meaningfully advances or enriches the existing landscape.

Reviewer #3

(Remarks to the Author)

The paper proposes MyGPT, an open-source system that lets users run retrieval-augmented generation (RAG) on their own documents without sending data to commercial AI services. MyGPT works by converting documents into chunks and then vector embeddings, retrieving the most relevant text chunks for each question, and giving the model only those chunks to answer with. The system highlights the exact document sources used and introduces three confidence metrics, including Question Relevance Score, Answer Relevance Score, and a Hallucination Index to help users judge reliability.

Details comments are as below:

+ The paper focuses on open-source models, which addresses privacy concerns and allows users to deploy RAG systems locally.

+ The system has a clear modular structure, visible source citations, and built-in relevance and hallucination scores, which help users understand where answers come from and check their reliability. The interface is simple and supports customization of models, chunking, and search settings, making it flexible while still keeping the core workflow easy to use.

- The novelty of the work seems to be limited. The components (document chunking, embeddings, relevance computing) and the framework are very normal and standard. It is not clear what are the technical challenges that are solved by the system, or the new perspective of approaching the problem.

- The document chunking and embedding process could be the bottleneck of the performance of RAG systems, since embedding-based similarity search may not always return useful contents. Simple relevance scores will not help much. This work could develop more advanced techniques.

- The paper conducted very detailed experimental analysis on the efficiency of the proposed system. However, evaluation of accuracy or effectiveness on downstream tasks is very limited.

- Although preserving privacy is good, it seems more like a by-product of using open-source models, not the unique contribution of the work.

(Remarks on code availability)

Reviewer #4

(Remarks to the Author)

- The manuscript presents an open-source and locally deployable RAG platform that can operate on private document collections without the need for cloud services. The system aims to produce more reliable answers and reduce hallucinations. It introduces three confidence metrics, the question relevance score, the answer relevance score and the Hallucination Index. These metrics help users judge how well the model used the provided context. The platform offers traceable source references, supports several embedding models and LLMs, and is simple to install. The authors also demonstrate its use in scientific literature analysis, clinical SOP documents and multilingual global health texts. The fact that the system can run on small devices adds practical value. These strengths make the tool immediately relevant in settings with strict privacy requirements. The work is not only a technical demonstration but a functioning tool with modular components, a clear user interface and flexible configuration options. Earlier open systems such as PrivateGPT or BioChatter (Lobentanzer et al., 2024) pursue related goals, but MyGPT distinguishes itself by combining an integrated pipeline with explicit confidence metrics and more comprehensive user control.
- The overall importance of the work is clear, especially in clinical or institutional environments where cloud-based systems cannot be used. The authors claim that MyGPT offers advantages over alternatives such as PrivateGPT, ChatPDF, BioChatter or LlamaIndex. This may well be true, but the manuscript does not provide evidence. Competing systems are listed, yet no direct comparison is given. The paper does not report accuracy, speed, resource usage or hallucination rate in a head-to-head format. This is a major gap. A fair comparison using the same dataset is needed. At least one competing RAG system should be included. Without this, it is difficult to judge the real performance impact of MyGPT.
- The evaluation of the confidence metrics is too narrow. All results rely on one question set that was created by domain experts from the manuscript itself. This makes the test circular and limits the ability to generalize the metrics. There is no evidence that the questions reflect real user behavior. There is no blinded evaluation or agreement between independent reviewers. The p-values that are reported come from a very small sample size of forty questions, which reduces their meaning. To support claims about metric robustness, the authors need to evaluate the metrics on public benchmarks. Suggested datasets include MedQA or BioASQ for biomedical content and Natural Questions or HotpotQA for general domains. Without this, the metrics cannot be considered validated beyond the manuscript's own content.
- The formula of the Hallucination Index also raises concerns. The choice of weights is not explained. There is no ablation or sensitivity study. The authors mention the option for users to change the values, but they do not offer guidance on how to do so or why. They should test the stability of the formula and provide threshold ranges that signal unreliable answers. The number of context chunks is also not studied. The authors state that three to five chunks are usually enough, but they do not show the impact of this parameter. A quantitative investigation is needed, because context selection is central to RAG performance.
- The structured chunking option does not have supporting evidence. The authors include this method, but they do not show measurable improvement compared to fixed-size chunks. They also state that more advanced chunking strategies do not improve results. This creates a contradiction. If structure matters, then the evidence must be shown. If it does not, then the emphasis should be reduced.
- The clinical and multilingual use cases are interesting but anecdotal. The GPCR example contains one complex question, but it does not include accuracy data or metric values. The SOP and global health examples do not include any quantitative results. Claims about real-world utility cannot rely only on descriptive examples. A broader user study is expected.
- The multilingual workflow is another area that needs clarification. It is not clear whether non-English questions are translated or whether the embedding model handles all languages directly. The effect of translation on retrieval accuracy is not tested. The authors should explain the pipeline and show the impact of different approaches.
- The cost discussion is also incomplete. The authors claim the platform is cheaper than commercial solutions, but they provide no real numbers beyond a general cloud GPU price. They do not consider installation, maintenance or staff time. These costs can be significant. While the Jetson Orin result is promising, the system requirements can still be too high for many groups. A more complete analysis is needed.
- The paper is dense at several points, and the structure makes it hard to follow. The HI formula should be included in the main text. The roles of QRS, ARS and HI should be explained with clear visuals. The extended data tables, especially those about benchmarking, are important but not well integrated. Bringing key information into the main text would improve clarity.
- Despite these issues, the paper addresses an important problem and offers a promising tool for private, transparent and controllable RAG workflows. Its engineering strengths and its practical potential make it a valuable contribution. However, the claims made about performance, generalizability and usability require more evidence. I expect the authors to address the points listed above with concrete additions, broader evaluation and clearer presentation. Once these improvements are made, I would recommend the manuscript for publication.

(Remarks on code availability)

I successfully installed and ran MyGPT on a personal macOS computer (Apple Silicon). The code is well-structured, modular, and production-ready.

Strengths:

- Clear, tiered installation guide for multiple deployment scenarios (personal computer, server/VM, cloud)
- Platform-specific instructions (macOS, Linux, Windows) with dedicated README files
- Well-organized project structure with modular components (/frontend, /backend, /ollama)
- Docker Compose for reproducible deployments (CPU and GPU)

- GitHub Actions CI/CD pipelines demonstrate mature software engineering practices
- Installation completed without errors, documentation is accessible for moderate technical skills

Minor limitation:

- Main README lacks explicit version specifications for key dependencies (Python, Node.js, Django, React). Pinning exact versions in requirements.txt and package.json would strengthen reproducibility

The code is a usable resource for the community, particularly for researchers and clinicians seeking a local, privacy-preserving RAG system. The modular architecture supports future contributions. No critical issues prohibit publication. Minor enhancements to version specification would strengthen the reproducibility narrative.

Version 1:

Reviewer comments:

Reviewer #3

(Remarks to the Author)

This paper presents MyGPT, an open-source, locally deployable RAG platform for question answering over private document libraries. The system integrates vector-based semantic search, BM25 keyword search, and cross-encoder reranking, and introduces three confidence metrics (QRS, ARS, and HI) to help users assess the reliability of generated answers. The authors benchmark MyGPT against four open-source RAG systems on three public datasets and validate it in three real-world scenarios at St. Jude Children's Research Hospital.

Strengths:

This work addresses a genuine need for a document question answering tool that does not rely on commercial cloud services in privacy-sensitive clinical and research settings. The system is well-engineered, with a modular design that allows users to flexibly swap individual components, support for edge device deployment, and source highlighting that enhances transparency. The validation across three real-world use cases, particularly the clinical SOP scenario (94.85% accuracy) and the efficiency comparison for kinase literature (2 hours vs. 30–40 hours manually), demonstrates practical value.

Major concerns:

1. Limited methodological novelty. The system is primarily composed of existing open-source components, and the core contribution lies more in system integration and application validation. QRS, ARS, and the original HI are mainly weighted combinations of existing retrieval and relevance scores. Their methodological novelty and incremental value over established RAG evaluation metrics require further clarification. As it stands, this work is closer to a well-validated engineering systems paper than a research paper proposing novel retrieval or generation methods.

2. The claimed performance advantage over RAGFlow lacks sufficient statistical support. The accuracy gaps between the two systems across the three datasets are modest (e.g., 91.4% vs. 90.2% on BioASQ), and the authors' own power analysis indicates that demonstrating this difference would require sample sizes far exceeding those actually used (5,106 documents needed for BioASQ vs. 210 used). The available data can only show that MyGPT's point estimate of accuracy is higher than RAGFlow's; it cannot yet establish that MyGPT's performance is statistically superior.

3. The confidence metrics should not be interpreted as direct measures of factual correctness. QRS, ARS, and HI primarily measure the relevance and support relationships among the question, retrieved text, and answer, and are better viewed as measures of contextual support. Even when an answer is highly consistent with the documents, it may still be factually incorrect due to errors, outdated information, or incompleteness in the source documents, or due to inaccuracies in numbers, negations, or qualifying conditions in the answer. Furthermore, the effect sizes for distinguishing correct from incorrect answers on the public datasets are uniformly small ($r = 0.11$ – 0.17), and the random forest achieves a test ROC-AUC of approximately 74%. In the optimal linear configuration across the three combined public datasets, the textual entailment term in ARS receives a weight of zero, indicating that it does not provide additional contribution under this setting.

(Remarks on code availability)

Reviewer #4

(Remarks to the Author)

I thank the authors for the substantially revised manuscript. The paper presents MyGPT, an open-source, privacy-first retrieval-augmented generation platform designed for local or secure institutional deployment. The revised version is considerably stronger than the earlier version. In particular, the addition of public benchmarking datasets, head-to-head comparisons with other open-source RAG systems, real-world use cases in clinical SOPs, biomedical literature and multilingual health policy documents, power-based sample size considerations, and a clearer description of the code and deployment model have significantly improved the manuscript. The main strengths of the work are its practical orientation, privacy-preserving local deployment model, modularity, source surfacing, and the attempt to provide user-facing confidence metrics for RAG-based question answering. The manuscript now evaluates MyGPT on PubMedQA, BioASQ and Open RAG

Benchmark datasets, reporting higher accuracy than RAGFlow, Dify, PrivateGPT and Kotaemon in the tested subsets. The clinical SOP use case is also compelling, with 129/136 correct answers and 94.85% accuracy reported in a private Bone Marrow Transplantation SOP library. The biomedical literature and multilingual policy examples further demonstrate that the platform has relevance beyond a single narrow domain. Overall, I find the manuscript timely, relevant and potentially useful to researchers, clinical institutions and other organizations that need to query private document collections without sending sensitive material to commercial cloud-based LLM services. The revision has addressed most of the major conceptual and methodological concerns. I support publication after minor revisions, mainly to improve precision, reproducibility and calibration of claims.

Specific comments:

1. The QRS, ARS and HI framework is a useful contribution, but the manuscript should be careful not to overstate the strength of these metrics. In the public benchmark datasets, QRS, ARS and HI distinguish correct from incorrect answers statistically, but the reported effect sizes are small for QRS and ARS/HI in the combined public datasets. The language around these metrics should therefore emphasize that they are useful risk indicators and triage signals, not definitive hallucination detectors. This is partly acknowledged in the limitations section, but the Results and Discussion should use similarly cautious wording.
2. The new power-analysis section is very helpful. However, the authors explicitly state that substantially larger sample sizes would be required for adequately powered comparisons with RAGFlow in PubMedQA, BioASQ and Open RAG Benchmark, but these larger evaluations were not performed. The manuscript should make this limitation visible in the Results or Discussion, not only in the Methods/Supplementary material. Claims that MyGPT outperforms RAGFlow should be softened accordingly, whereas the stronger claims against Dify, PrivateGPT and Kotaemon appear better supported by the available sample sizes.
3. The revised manuscript now states that correctness labels for public datasets were provided by two independent researchers and that private datasets were evaluated by domain experts. However, the manuscript should provide additional detail on whether evaluators were blinded to the tool identity, how disagreements were resolved, and whether inter-rater agreement was measured. This is especially important because the comparisons involve manually evaluated answers and because the use cases include private institutional datasets.
4. There appear to be inconsistencies between the main text, Methods and Supplementary Figure 4. The main text reports 282 Q&A pairs, 206/282 correct answers, and 227/282 correct answers after using 15 chunks. The Methods section, however, describes 297 Q&A pairs for the kinase library, and Supplementary Figure 4 reports 238/297 correct answers with $n=15$. These numbers should be reconciled before publication.
5. The multilingual health-policy use case is valuable, but the Methods state that the reranker was omitted in this evaluation because of limited foreign-language expertise and manual validation constraints, and that future evaluation of reranker suitability is needed. This should be clearly reflected in the Results and Discussion. The authors should avoid implying that the full MyGPT confidence-metric pipeline has been equally validated in multilingual settings if the pipeline differs materially from the English-language evaluations.
6. The manuscript states that the code is currently private, will be made public upon acceptance, and is provided as supplementary material for review. Before publication, the authors should ensure that the public repository includes a clear README, installation instructions, dependency versions, example data, instructions for reproducing the benchmark evaluations where possible, and a statement on how private datasets were handled or substituted. Given the centrality of MyGPT as a software resource, this is essential for reproducibility.
7. Several typographical errors and wording issues remain and should be corrected during revision. Examples include "populat" in the confidence metric definitions, "comiplied" and "evauate" in the multilingual policy section, "cusromize" in Figure 6, and "MyGTP" in the Figure 6 legend. These do not affect the scientific contribution but should be corrected before publication.

Additional assessment of Reviewer #1's concerns

I have also reviewed Reviewer #1's original comments together with the authors' point-by-point responses and the corresponding revisions to the manuscript. In my assessment, the authors have substantially addressed the major concerns raised by Reviewer #1. In particular, the addition of direct benchmarking against multiple open-source RAG systems across three public datasets substantially strengthens the evidence base and addresses the earlier concern that comparison with a standalone LLM was insufficient. The revised architecture and workflow descriptions, the addition of BM25 retrieval and reranking, the expanded real-world evaluations, and the clearer comparison with existing RAG solutions also substantially improve the differentiation and scientific framing of MyGPT. The concern regarding the Hallucination Index has been addressed in part through a more sophisticated formulation incorporating semantic, keyword and entailment-based information, sensitivity analysis, and additional validation. However, I agree that these metrics should still be interpreted as indicators of potential answer unreliability rather than definitive measures of factual correctness, as noted in Comment 1 above. Similarly, while the comparative benchmarking is a major improvement, the authors' own power calculations indicate that the comparisons with RAGFlow are underpowered to establish superiority, and the corresponding claims should therefore be moderated, as noted in Comment 2. Overall, I consider Reviewer #1's concerns to have been substantially, though not completely, addressed; the remaining issues appear suitable for minor revision.

(Remarks on code availability)

Open Access This Peer Review File is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

In cases where reviewers are anonymous, credit should be given to 'Anonymous Referee' and the source.

The images or other third party material in this Peer Review File are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

To view a copy of this license, visit <https://creativecommons.org/licenses/by/4.0/>

Note to All Referees: Summary of Major Revisions and Additions

We are grateful to the referees for their careful and constructive feedback. In response to their comments, we have substantially revised the manuscript to strengthen the methodological rigor, expand benchmarking and validation, and improve clarity and transparency throughout. Below we summarize the key changes made in the revised version.

1. Substantially Expanded Benchmarking Against Other RAG Systems

- a. New head-to-head benchmarking against four widely used open-source RAG solutions (RAGFlow, Dify, PrivateGPT, and Kotaemon) using three public datasets: PubMedQA, BioASQ, and the Open RAG Benchmark (new **Results** section; **Fig. 4; Methods**).
- b. Demonstrated that MyGPT consistently outperforms all comparators in accuracy, document processing, and source highlighting, achieving >80–90% accuracy across datasets depending on benchmark.
- c. Confidence metrics (QRS, ARS, HI) are shown to robustly distinguish correct from incorrect answers across all benchmarks, with statistical testing and effect sizes reported.
- d. Detailed feature comparison of open-source and commercial RAG solutions, showing that transparency and validation features are not provided by all RAG frameworks (updated **Figure 6b**).

We hope that these additional results directly address concerns about comparative performance, generalizability, and whether MyGPT offers unique advantages beyond existing RAG tools.

2. Formalization and Strengthening of Confidence Metrics

- a. The definitions of Question Relevance Score (QRS), Answer Relevance Score (ARS), and Hallucination Index (HI) have been fully formalized mathematically, with explicit equations, normalization schemes, and rationale (revised **Results** section).
- b. Global sensitivity analysis is now used to determine optimal weights for QRS, ARS, and HI across public datasets (new **Methods** section; **Supplementary Fig. 8**) with QRS, ARS and HI distribution provided for both public datasets (**Fig. 4**) and St. Jude use-cases (**Fig. 5, Supplementary Fig. 4**).
- c. A machine learning-based hallucination predictor (random forest) trained on >1,000 Q&A pairs from public and private datasets is introduced, significantly improving discrimination between correct and incorrect answers (new **Results** section; **Methods, Supplementary Fig. 6**).

We hope that these additional approaches demonstrate that the confidence metrics are now quantitatively grounded, validated, and extensible.

3. Enhanced Retrieval Architecture: Keyword Search and Reranking

- a. The retrieval pipeline (updated **Fig. 1**) now explicitly combines:
 - Semantic search using embedding models and vector database
 - Keyword search using BM25 to capture domain-specific terminology
 - Reranking of retrieved chunks with cross-encoder (QNLI-based) to select a final “golden context.”
- b. The Reranker’s contribution to improved accuracy is demonstrated empirically, including direct comparison of two Rerankers (**Results; Fig. 5**).

We hope that this addition addresses concerns regarding missed keywords, hallucinations caused by weak retrieval, and reproducibility of context selection.

4. Expanded and Quantified Real-World Use Cases

- a. Three application domains are now systematically validated with quantitative evaluation by experts in the field:
 - Private clinical SOPs (Bone Marrow Transplantation and Cell Therapy at St. Jude): 94.85% accuracy with strong separation by confidence metrics (updated **Results**; **Fig. 5**; **Methods**).
 - Biomedical literature mining (protein kinases): large-scale expert-curated Q&A with performance gains vs manual curation (new **Results**; **Supplementary Fig. 4**; **Methods**).
 - Multilingual global health policies: validated Spanish and French policy libraries with 80–88% accuracy (updated **Results**; **Supplementary Fig. 5**; **Methods**) with detailed pipeline modification for multilingual Q&A.
- b. Each use case includes precise customization settings (embedding model, chunking, reranking, LLM) and corresponding performance metrics.

We hope that these additions demonstrate real-world utility beyond toy examples, including privacy-sensitive and multilingual settings.

5. Greatly Expanded Technical Transparency and Reproducibility

Methods and supplementary data now include:

- Detailed descriptions of chunking strategies (fixed vs structure-preserving recursive chunking) with results showing performance of both methods (**Supplementary Fig. 4**).
- Explicit criteria for determining the number of retrieved chunks.
- Hosting requirements and costs for local deployment (**Supplementary Table 6**), cloud deployment (updated **Supplementary Table 3**), and cost analyses for MyGPT vs other similar solutions (**Methods**, **Supplementary Table 7**) showcasing resource usage, performance trade-offs, and deployment scenarios.
- Rigorous description of sample sizes for benchmarking comparisons using formal power analysis (new **Methods** and **Supplementary Table 5**).

We hope that these additional data addresses reproducibility, feasibility, and operational concerns often raised for AI systems.

6. Clarified Scope, Limitations, and Claims

- a. Limitations have been expanded and made explicit, including:
 - Inability to parse non-textual elements (tables, figures, equations).
 - Dependence on library completeness.
 - Contexts where advanced chunking offers little benefit.
- b. Claims regarding generalizability and hallucination reduction are now carefully bounded by empirical evidence.

We thank the referees again for their constructive comments. We believe these changes fully address the referees' concerns and substantially enhance the manuscript's robustness and value to the community. We hope that the revised manuscript meets the expectations of the reviewer for publication in *Nature Communications*.

Sincerely,

Jaimin Patel and M. Madan Babu
On behalf of all authors

Author response to reviewer comments:

Reviewer comments are in *black*.

Responses are in *blue*.

Reviewer #1 (Remarks to the Author):

The authors present MyGPT, a privacy-first, open-source Retrieval-Augmented Generation (RAG) platform designed to operate locally. The proposed system integrates various components of a RAG pipeline, such as document chunking, organization, and LLM inference, into a end-to-end workflow intended to allow researchers to query private documents without data leakage. The system is benchmarked against a standalone LLM using a curated set of biomedical questions.

While the paper presents a functional tool, it more resembles an engineering technical report rather as compared to a traditional paper in this venue. In addition, I have five main concerns regarding the scientific novelty and validation of the work:

We thank the reviewer for their careful evaluation of our manuscript and very insightful comments, which were instrumental in guiding our revision. As described in the note to all referees above, in the revised manuscript, we have strengthened both the novelty and validation. Specifically, MyGPT is no longer benchmarked only against a standalone LLM, but is now rigorously evaluated against multiple state-of-the-art open-source RAG frameworks across three public datasets. We further formalize and validate the proposed relevance and hallucination metrics, including mathematical definitions, statistical testing, and power analysis-based sample size justification. In addition, we expand quantitative evaluations to multiple real-world use cases (private clinical documents, biomedical literature, and multilingual global health policies), demonstrating generality beyond a single curated task. Together, these additions establish our manuscript as a methodologically grounded and broadly validated RAG framework rather than an engineering technical report. We hope the revised version addresses all the concerns constructively and that the referee would find the revised manuscript suitable for publication in *Nature Communications*.

1. Lack of Differentiation from Existing Solutions.

Although the authors mention other open-source solutions around lines 289–295 and in Extended Data Table 5, the manuscript does not demonstrate how MyGPT substantially differs from previous work. The comparison offers only qualitative claims regarding flexibility. Furthermore, the comparison in Table 5 is incomplete; it omits numerous prominent open-source RAG frameworks, such as RAGFlow (<https://github.com/infiniflow/ragflow>) and Kotaemon (<https://github.com/Cinnamon/kotaemon>), both of which offer simple deployment solutions and user interface support than the proposed. The discussion in Extended Data Figure 1 should also include these open-source baselines to better establish novelty, rather than limiting the comparison to commercial models. Ultimately, the manuscript does not articulate how the architecture advances the field of Information Retrieval or NLP beyond providing an interface wrapper around existing libraries.

We appreciate the critique from the reviewer and have extensively revised the manuscript to differentiate MyGPT from other solutions. Specifically, we include the following in the revised manuscript:

- New **Figure 4** shows a comparative evaluation of the performance of MyGPT and four other RAG applications, including RAGFlow and Kotaemon, directly addressing the reviewer's concern about missing prominent open-source baselines. Please see our response to the reviewer's second concern immediately below.
- New **Figure 6b** presents a comparative summary of the features and architectural scope of MyGPT and nine current open-source and commercial GPT solutions, including Kotaemon and

RAGFlow, in the main text. Prior **Extended Data Table 5** has been removed to avoid an incomplete or qualitative-only comparison.

- **Extended Data Figure 1** was revised to more broadly compare MyGPT and open-source end-to-end RAG solutions with other information-seeking tasks, rather than limiting the comparison to commercial models. It is now presented as **Supplementary Figure 1**.

In addition, the revised manuscript explicitly clarifies how the MyGPT architecture goes beyond an interface wrapper around existing libraries by formally defining the system-level abstractions for document organization, retrieval supervision, and hallucination-aware evaluation, and by quantitatively validating these components across multiple datasets and use cases. These clarifications are presented in the **Methods** and **Results** sections to articulate the methodological contribution to information retrieval-driven RAG systems.

2. Insufficient Benchmarking Baselines.

The included benchmark results only compare MyGPT against Llama2 alone. It is well-established in the literature that RAG outperforms standalone LLMs in domain-specific, knowledge-intensive tasks. Consequently, proving that MyGPT outperforms a model without context is quite a trivial result. There is no experiment demonstrating the superior ability (or even parity) of the proposed MyGPT against other state-of-the-art, open-source RAG frameworks.

We thank the reviewer for these constructive comments. We fully agree that demonstrating improvement over a standalone LLM is insufficient, as the benefit of RAG over context-free generation is well established. Accordingly, in the revised manuscript we have added new analyses that benchmark MyGPT against multiple state-of-the-art open-source RAG frameworks, rather than solely against Llama2. Specifically, we now evaluate MyGPT and its confidence metrics using three publicly available datasets (PubMedQA, BioASQ, and Open-RAG Benchmark) and quantitatively compare its performance with four open-source RAG systems: PrivateGPT, Kotaemon, RAGFlow, and Dify. These results demonstrate that MyGPT achieves performance that is superior to, or at minimum on par with, these RAG frameworks, as described in the **Results** section (pp. 7–8) and shown in new **Figure 4** (“*Benchmarking MyGPT with other open-source RAG solutions using public datasets*”). To avoid over-emphasizing a straight-forward baseline, the original comparison of MyGPT with Llama2 alone has now been moved to the **Supplementary Information**, where it is retained for completeness and to demonstrate that MyGPT works as expected in such a use case.

3. Questionable Validity of the Hallucination Index

The proposed Hallucination Index is questionable. It is calculated as a weighted combination of the Question Relevance Score (QRS) and Answer Relevance Score (ARS), which are derived from the embedding similarity between the query/context and context/answer. This approach relies on semantic distance, which is not a proxy for factual correctness. For example, the statement “The drug is safe” and “The drug is not safe” may share high semantic similarity in vector space, yet represent opposite factual claims. A high similarity score could therefore coexist with a factually incorrect answer, rendering the index unreliable.

We thank the referee for this helpful and important comment. We agree that semantic similarity alone is not a proxy for factual correctness, as statements with opposite factual meaning can indeed be close in embedding space. To address this concern as far as possible within the scope of a RAG framework, the confidence metrics in the revised manuscript were substantially strengthened in several ways. First, the relevance metrics were extended beyond embedding similarity to incorporate BM25 matching and cross-encoder reranking scores, which evaluates whether retrieved context supports or contradicts a given answer (**Results**, pp. 5–7). Second, we introduced a machine learning-based approach to predict answer correctness, using Question Relevance Score, Answer Relevance Score, and additional retrieval features

as input (**Results**, p. 9: *Machine learning-based optimization of hallucination index*; **Supplementary Fig. 6**). This model was trained on diverse datasets and demonstrates statistically significant separation between correct and incorrect responses across all evaluated benchmarks. In addition, we performed a global sensitivity analysis to determine the default weights used to compute QRS, ARS, and HI (**Methods**, p. 26: *Finding the optimal setting of QRS, ARS, and HI*; **Supplementary Fig. 8**), reducing arbitrariness in metric construction.

Importantly, MyGPT also explicitly surfaces the exact source text used to generate each response, enabling users to directly validate answers against the underlying documents. Accordingly, we now clarify in the revised manuscript that the Hallucination Index is intended as a guidance and triage metric, rather than a definitive measure of factual correctness. The scope and limitations of the confidence metrics are explicitly discussed in the **Results** section (p. 10).

4. Flawed Experimental Design and Sample Size.

The benchmarking methodology is problematic. First, the evaluation relies on only 40 questions derived from the manuscript itself. This sample size is (unnecessarily) too small to draw statistically significant conclusions regarding the robustness of the system or the utility of the Hallucination Index. Second, the baseline setting is methodologically flawed because the standalone LLM was not trained on the private manuscript; therefore, a 0% accuracy rate is entirely expected. This comparison only confirms that having context is helpful, which is already known. A correct evaluation would provide the same document chunks to both the baseline (e.g., a standard RAG implementation) and MyGPT to demonstrate that MyGPT offers better retrieval or higher-quality context processing.

We thank the referee for this comment. As noted above in our reply to Point 2, the revised manuscript now presents new analyses shown in **Figure 4**: “*Benchmarking MyGPT with other open-source RAG solutions using public datasets*”, which directly address the limitations of the original evaluation. Specifically, rather than relying on 40 manuscript-derived questions or comparing against a standalone LLM, the revised benchmarking evaluates MyGPT against multiple state-of-the-art open-source RAG frameworks using shared public datasets. The optimal sample size for performance comparisons of different RAG applications was determined through a formal power analysis using the approach described in pp. 24–25 of the **Methods** (“*Sample size calculations using power analysis*”). The number of documents and Q&A pairs used to benchmark MyGPT (shown in **Supplementary Table 5**: “*Sample size calculations for MyGPT vs other RAG solutions*”) meet or exceed the sample size required by the power analysis, ensuring statistically meaningful comparisons. These benchmarks compare MyGPT directly with other RAG implementations under equivalent retrieval settings, rather than against a context-free LLM baseline. The original evaluation comparing MyGPT with a standalone LLM on a private manuscript has been moved to the **Supplementary Information** and is retained for completeness. We hope that the new results and benchmarking data presented in the revised manuscript now adequately address the concerns raised by the referee regarding experimental design and sample size.

5. Structural and Organizational Deficiencies.

Finally, the manuscript requires substantial structural revision to meet the standards of a technical research paper. Currently, the text transitions directly from the Introduction to the Results, omitting key transitions in the main body. Consequently, crucial details regarding the architectural design and pipeline implementation are fragmented or relegated to the later sections at the end of the document and supplementary.

We appreciate the feedback and have extensively restructured the manuscript to address these concerns. In the revision, we have explicitly corrected the lack of structure and transitions identified by the referee. Specifically, we now present the rationale of our approach, design goals and overall strategy for the MyGPT

framework within the **Introduction** (p. 2), followed by a dedicated, logically ordered section in the main body of the paper detailing the system design and workflow (“*Outline of MyGPT architecture and workflow*,” pp. 3-4), rather than deferring these elements to later sections or the Supplementary Information. These passages are marked with track changes in revised manuscript that has been submitted as an additional file. A new **Figure 1** provides an integrated overview of the end-to-end MyGPT workflow. **Figure 2** presents the MyGPT user interface for real-time Q&A and source inspection, while a new **Figure 3** presents MyGPT pipeline customizations, with the functions and selection of individual components described in detail in pp. 4–5 of the Results (“*MyGPT workflow customization options*”). The manuscript then proceeds in a standard research-paper structure, presenting benchmarking results (**Figure 4**), followed by multiple application use cases (**Figure 5**), and concluding with considerations, limitations, and comparison to existing solutions (**Figure 6**). We believe these revisions resolve the organizational and structural deficiencies noted by the referee.

Conclusion:

Although MyGPT provides a complete, reproducible, end-to-end framework for RAG, I cannot recommend acceptance in its current state. The work acts primarily as a wrapper for existing tools without offering significant algorithmic novelty or rigorous comparative benchmarking. It is believed that significant modification, including comparison against proper RAG baselines, is required before resubmission.

We appreciate the constructive criticism. In response, we have made extensive revisions to the manuscript. As detailed above and in the note to all referees, the revised version now includes rigorous and comprehensive comparative benchmarking against multiple state-of-the-art open-source RAG frameworks, rather than standalone LLM baselines, as well as formalization, sensitivity analysis, and statistical validation of the proposed confidence metrics. We have also added new architectural components and capabilities, including BM25 search, reranking of chunks, multilingual support, and expanded quantitative evaluations across several real-world use cases spanning clinical documents, biomedical literature, and global health policies. In addition, the manuscript has been substantially reorganized to more clearly articulate the design rationale and methodological contributions beyond an interface wrapper. We believe that these revisions directly address the reviewer’s concerns and hope that the new manuscript will meet their expectations for publication in *Nature Communications*.

Reviewer #1 (Remarks on code availability):

I believe the source code could still provide value to the community, particularly because it includes detailed instructions for using and deploying MyGPT. That said, I want to note that the proposed MyGPT framework does not appear to differ substantially from existing RAG solutions—such as those built with LangChain—beyond its emphasis on cost effectiveness and its integration with scholarly sources like Zotero.

Moreover, even within the open-source ecosystem, there already exist extensive community-maintained resources (e.g., the comprehensive collection at https://github.com/NirDiamant/RAG_Techniques). This raises some questions about the extent to which the contribution meaningfully advances or enriches the existing landscape.

We thank the referee for their positive remarks about the value of the source code to the community. For a more detailed response, we request this referee to see our response to the comment of Reviewer 2, below (who co-reviewed the paper with Reviewer 1).

Reviewer #2 (Remarks to the Author):

I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

We thank the reviewer for their constructive evaluation of our work.

Reviewer #2 (Remarks on code availability):

I believe the source code could still provide value to the community, particularly because it includes detailed instructions for using and deploying MyGPT. That said, I want to note that the proposed MyGPT framework does not appear to differ substantially from existing RAG solutions—such as those built with LangChain—beyond its emphasis on cost effectiveness and its integration with scholarly sources like Zotero. Moreover, even within the open-source ecosystem, there already exist extensive community-maintained resources (e.g., the comprehensive collection at https://github.com/NirDiamant/RAG_Techniques). This raises some questions about the extent to which the contribution meaningfully advances or enriches the existing landscape.

We thank the reviewer for these comments. We have extensively revised the manuscript text and figures to emphasize the unique features of the MyGPT framework (**Fig. 6b**: “*MyGPT applications and comparison to current GPT solutions*”), and we demonstrate its improved performance compared to other RAG solutions (**Fig. 4**: “*Benchmarking MyGPT with other open-source RAG solutions using public datasets*”). We have also explained the rationale underlying MyGPT’s design (first section of the **Results**, “*Outline of MyGPT architecture and workflow*,” p. 3) and emphasize how it fills a gap in the current landscape of LLM-based knowledge-seeking tools (bottom of the **Introduction**, p. 2, and summarized in the **Discussion**, pp. 10-11). We believe that our extensive revisions, including rigorous and comprehensive comparative benchmarking, new features (e.g. BM25, re-ranking), multi-lingual capabilities, several use cases across domains with performance evaluation (private clinical documents, biomedical literature, and global health policies), and additional analyses (power analysis, sensitivity analysis) demonstrate how MyGPT meaningfully advances and enriches the existing landscape.

We acknowledge the existence of extensive community-maintained RAG resources; however, these collections serve largely as independent reference methods and techniques rather than validated, citable, end-to-end systems. Also, some of these methods will require the use of API keys from commercial components, such as Contextual AI, OpenAI, and Cohere, making them unusable for privacy-focused air-gapped RAG solutions. The revised manuscript clarifies how MyGPT complements this ecosystem by providing a fully integrated and benchmarked framework with explicit evaluation and confidence metrics. We hope these revisions address the reviewer’s concerns regarding suitability for publication in *Nature Communications*.

Reviewer #3 (Remarks to the Author):

The paper proposes MyGPT, an open-source system that lets users run retrieval-augmented generation (RAG) on their own documents without sending data to commercial AI services. MyGPT works by converting documents into chunks and then vector embeddings, retrieving the most relevant text chunks for each question, and giving the model only those chunks to answer with. The system highlights the exact document sources used and introduces three confidence metrics, including Question Relevance Score, Answer Relevance Score, and a Hallucination Index to help users judge reliability.

Details comments are as below:

- + The paper focuses on open-source models, which addresses privacy concerns and allows users to deploy RAG systems locally.
- + The system has a clear modular structure, visible source citations, and built-in relevance and hallucination scores, which help users understand where answers come from and check their reliability. The interface is simple and supports customization of models, chunking, and search settings, making it flexible while still keeping the core workflow easy to use.
- The novelty of the work seems to be limited. The components (document chunking, embeddings, relevance computing) and the framework are very normal and standard. It is not clear what are the technical challenges that are solved by the system, or the new perspective of approaching the problem.

We thank the reviewer for the constructive feedback, which we have used to carefully revise the manuscript. In the revised version, we articulate more clearly the design goals, technical motivations, and system-level challenges addressed by MyGPT, particularly the need for a privacy-preserving, inspectable, and deployable end-to-end RAG framework with built-in evaluation and transparency. These aspects are now described explicitly in the “*Outline of MyGPT architecture and workflow*” section of the **Results** (p. 3), at the end of the **Introduction** (p. 2), and summarized in the **Discussion** (pp. 10–11).

While individual components such as document chunking and embeddings are standard, MyGPT’s contribution is the integration of evaluation, answer traceability, and deployment of the core RAG concepts into an end-to-end, efficient and reliable workflow, enabling users to systematically assess reliability and verify outputs rather than relying on ad-hoc inspection. We further emphasize the distinguishing features of the MyGPT framework beyond standard RAG pipelines, including integrated confidence metrics and unified evaluation across retrieval, generation, and hallucination detection (**Fig. 6b**: “*MyGPT applications and comparison to current GPT solutions*”). In addition, we demonstrate quantitatively improved performance relative to other open-source RAG systems through rigorous benchmarking on public datasets (**Fig. 4**) and validate MyGPT’s multilingual and cross-domain applicability using real-world use cases spanning private clinical documents, biomedical literature, and global health policies (**Results**, p. 8; **Supplementary Fig. 5**). We hope these revisions clarify the technical perspective and contributions of the work and directly address the reviewer’s concerns regarding novelty.

- The document chunking and embedding process could be the bottleneck of the performance of RAG systems, since embedding-based similarity search may not always return useful contents. Simple relevance scores will not help much. This work could develop more advanced techniques.

We appreciate the reviewer’s criticism. To address this concern, we have expanded MyGPT to include the BM25-based retrieval and cross-encoder reranking, which mitigate known limitations of embedding-only similarity search. We have also added a machine-learning-based model to predict answer correctness. These additions are described in pp. 4 and 9 of the revised manuscript and lead to improved retrieval quality and robustness, as illustrated in **Supplementary Figure 4** (“*Q&A with a large library of biomedical*”).

literature") and in the **Results** (p. 8). We hope these changes directly address the identified bottleneck in retrieval and relevance estimation.

- The paper conducted very detailed experimental analysis on the efficiency of the proposed system. However, evaluation of accuracy or effectiveness on downstream tasks is very limited.

Thank you for raising this point. In response, we have expanded the evaluation of accuracy and effectiveness of the downstream tasks in the revised manuscript. With the involvement of multiple domain experts, we could assess the responses for the downstream tasks rigorously. Accuracy, relevance scores, and HI are now reported for all three real-world use cases presented in the revision (**Results** pp.8-9): Private clinical documents (**Figure 5**), Biomedical literature (**Supplementary Figure 4**) and Multi-lingual global health policies (**Supplementary Figure 5**). We believe that these additions substantially strengthen the assessment of MyGPT's effectiveness beyond efficiency assessment alone.

- Although preserving privacy is good, it seems more like a by-product of using open-source models, not the unique contribution of the work.

The reviewer's point is well taken. While privacy can be enabled by open-source models, in the revised manuscript we clarify that privacy preservation is treated as an explicit design constraint rather than a by-product, particularly for deployment in clinical and institutional settings where data cannot leave secure environments. MyGPT is therefore designed as a local, firewall-protected, end-to-end RAG system with built-in evaluation and traceability, addressing practical barriers that limit the use of commercial or cloud-based solutions in such contexts.

Reviewer #4 (Remarks to the Author):

- The manuscript presents an open-source and locally deployable RAG platform that can operate on private document collections without the need for cloud services. The system aims to produce more reliable answers and reduce hallucinations. It introduces three confidence metrics, the question relevance score, the answer relevance score and the Hallucination Index. These metrics help users judge how well the model used the provided context. The platform offers traceable source references, supports several embedding models and LLMs, and is simple to install. The authors also demonstrate its use in scientific literature analysis, clinical SOP documents and multilingual global health texts. The fact that the system can run on small devices adds practical value. These strengths make the tool immediately relevant in settings with strict privacy requirements. The work is not only a technical demonstration but a functioning tool with modular components, a clear user interface and flexible configuration options. Earlier open systems such as PrivateGPT or BioChatter (Lobentanzer et al., 2024) pursue related goals, but MyGPT distinguishes itself by combining an integrated pipeline with explicit confidence metrics and more comprehensive user control.

We thank the reviewer for these supportive and thoughtful comments. We are glad that the reviewer recognizes the practical relevance, modular design, transparency, and privacy-preserving nature of MyGPT, as well as its distinction from related open-source systems through the integration of an end-to-end pipeline with explicit confidence metrics and user control.

- The overall importance of the work is clear, especially in clinical or institutional environments where cloud-based systems cannot be used. The authors claim that MyGPT offers advantages over alternatives such as PrivateGPT, ChatPDF, BioChatter or LlamaIndex. This may well be true, but the manuscript does not provide evidence. Competing systems are listed, yet no direct comparison is given. The paper does not report accuracy, speed, resource usage or hallucination rate in a head-to-head format. This is a major gap. A fair comparison using the same dataset is needed. At least one competing RAG system should be included. Without this, it is difficult to judge the real performance impact of MyGPT.

We appreciate the reviewer's constructive criticism. To address this important gap in our work, we have added rigorous head-to-head and comprehensive comparative benchmarking of MyGPT against multiple open-source RAG systems using the same public datasets (PubMedQA, BioASQ, and Open-RAG Benchmark). We now quantitatively compare MyGPT with PrivateGPT, Kotaemon, RAGFlow, and Dify with the same LLMs and RAG settings, and resources where possible, in terms of accuracy and confidence metrics, as shown in **Figure 4** ("*Benchmarking MyGPT with other open-source RAG solutions using public datasets*") and described in **Results** (pp. 7–8). Although it is beyond the scope of this manuscript to benchmark all RAG steps across other open-source RAG solutions, we now emphasize the distinguishing features of MyGPT compared to other open-source RAG frameworks in **Fig. 6b**: "*MyGPT applications and comparison to current GPT solutions*". We hope that these additions provide a fair and comprehensive comparison requested by the reviewer.

- The evaluation of the confidence metrics is too narrow. All results rely on one question set that was created by domain experts from the manuscript itself. This makes the test circular and limits the ability to generalize the metrics. There is no evidence that the questions reflect real user behavior. There is no blinded evaluation or agreement between independent reviewers. The p-values that are reported come from a very small sample size of forty questions, which reduces their meaning. To support claims about metric robustness, the authors need to evaluate the metrics on public benchmarks. Suggested datasets include MedQA or BioASQ for biomedical content and Natural Questions or HotpotQA for general domains. Without this, the metrics cannot be considered validated beyond the manuscript's own content.

Thank you for these suggestions. As noted in our response to the point above, MyGPT is now rigorously evaluated against four state-of-the-art open-source RAG frameworks across three public datasets. We further formalize and validate the proposed relevance and hallucination metrics, including mathematical definitions, statistical testing, and power analysis-based sample size justification. In addition, we expand quantitative evaluations to multiple real-world use cases (private clinical documents, biomedical literature, and multilingual global health policies), demonstrating generality beyond a single curated task. We hope that these additions, as well as those that we describe in our response to the point below, address concerns regarding metric robustness.

- The formula of the Hallucination Index also raises concerns. The choice of weights is not explained. There is no ablation or sensitivity study. The authors mention the option for users to change the values, but they do not offer guidance on how to do so or why. They should test the stability of the formula and provide threshold ranges that signal unreliable answers. The number of context chunks is also not studied. The authors state that three to five chunks are usually enough, but they do not show the impact of this parameter. A quantitative investigation is needed, because context selection is central to RAG performance.

We thank the reviewer for these comments. The confidence metrics have been updated to include the BM25 algorithm and reranking scores (**Results** pp5-7), and a machine-learning-based approach was added to predict answer correctness (**Results** p9: '*Machine learning-based optimization of hallucination index*' and **Supplementary Figure 6**: '*Random forest model for generalizable HI predictor*'). To address the reviewer's concerns regarding the formulation of the Hallucination Index (HI), we performed a global sensitivity analysis (**Methods** p26: '*Finding the optimal setting of QRS, ARS, and HI*'; **Supplementary Figure 8**) to determine default weights for each metric and assess their stability across parameter settings. Considerations, guidance on usage, and limitations of the confidence metrics, including how users may interpret, are explicitly discussed on p10 of **Results**.

The effect of number of context chunks is quantitatively examined in the '*Biomedical literature*' use case (**Supplementary Figure 4**) described on p. 8 of **Results**. Using the same setup involving 282 questions, we show that increasing the number of chunks improves accuracy (80.5% with 15 chunks versus 73.0% with the default 5 chunks; $p = 7 \times 10^{-2}$; effect size = 0.34 (phi coefficient Φ , medium); McNemar chi-square test). We also evaluated structure-preserving chunking (default of 5 chunks; 75.8% accuracy), which did not yield a significant improvement over fixed chunking. We clarify that this feature may be useful when users wish to focus on a specific section or avoid specific sections while generating a response.

- The structured chunking option does not have supporting evidence. The authors include this method, but they do not show measurable improvement compared to fixed-size chunks. They also state that more advanced chunking strategies do not improve results. This creates a contradiction. If structure matters, then the evidence must be shown. If it does not, then the emphasis should be reduced.

We thank the referee for this comment. The effect of chunking strategies is examined in the "*Biomedical literature*" use case (**Supplementary Fig. 4**). With fixed-size chunking, we report 73.0% (206/282) accuracy. When the experiment was repeated using structure-preserving chunking with 5 chunks, accuracy increased to 75.8% (214/282), which did not represent a statistically significant improvement ($p = 0.27$; effect size = 0.15 (Φ , small); McNemar test). We therefore do not claim general performance gains from structure-preserving chunking. In the '*Considerations and limitations*' section of **Results** (p10), we note that the third chunking method, advanced semantic chunking, also does not provide a substantial performance improvement, as reported by Qu, R., *et al.*; Findings of the Association for Computational Linguistics: NAACL 2025, pp2155-2177 and cited in ref. 31. As this method is computationally expensive and does not offer an advantage over fixed chunking, further investigation is required before it can be incorporated into MyGPT. We have accordingly reduced their emphasis and instead present structured chunking as an

optional feature that may be useful when users wish to either exclude or focus on specific document sections rather than as a general performance optimization.

- The clinical and multilingual use cases are interesting but anecdotal. The GPCR example contains one complex question, but it does not include accuracy data or metric values. The SOP and global health examples do not include any quantitative results. Claims about real-world utility cannot rely only on descriptive examples. A broader user study is expected.

To address the reviewer's concern, we performed searches of Private Clinical SOP documents (**Figure 5**: "Standard Operating Procedures (SOPs) for Bone Marrow Transplantation") and of the Biomedical Literature (**Supplementary Figure 4**: "Q&A with large library of biomedical literature") to quantitatively assess the performance of MyGPT. We now report the accuracy and relevance scores of MyGPT's answers, as well as the time required for the pipeline to complete the task, in the 'Real-world use cases for MyGPT' section of **Results** (pp.8-9). We hope that these controlled evaluations across diverse real-world datasets provide quantitative evidence of downstream task performance and support the claims regarding MyGPT's practical utility. Please note that the reviewer's query regarding the multilingual use case is addressed in the next point.

- The multilingual workflow is another area that needs clarification. It is not clear whether non-English questions are translated or whether the embedding model handles all languages directly. The effect of translation on retrieval accuracy is not tested. The authors should explain the pipeline and show the impact of different approaches.

Thank you for this comment. We have clarified the multilingual workflow in the revised manuscript by explicitly describing the roles of the translation model (TranslateGemma), the embedding model, and the BM25 retrieval algorithm in the "Multilingual global health policies" use case (**Results** pp.8-9; **Supplementary Fig. 5**). We clarify when non-English queries are translated versus when multilingual embeddings are used directly, and we describe the available language-specific options for BM25 and reranking that users can select through the UI (**Results** p. 4: "MyGPT workflow customization options"). In addition, we have added a detailed description and quantitative analysis of multilingual pipeline configurations to the **Methods**, including their impact on retrieval and answer quality.

- The cost discussion is also incomplete. The authors claim the platform is cheaper than commercial solutions, but they provide no real numbers beyond a general cloud GPU price. They do not consider installation, maintenance or staff time. These costs can be significant. While the Jetson Orin result is promising, the system requirements can still be too high for many groups. A more complete analysis is needed.

In addition to cloud hosting costs reported in **Supplementary Table 3**, the revised manuscript now includes a more comprehensive cost analysis. Specifically, **Supplementary Table 6** ("Cost of hosting MyGPT on commodity hardware for group or department usage") provides estimates that account for local hardware requirements, deployment scale, and operational considerations, while **Supplementary Table 7** ("Cost comparison with various commercial and open-source RAG solutions") presents a side-by-side comparison with alternative systems. These analyses, described in the "Cost and resource usage" section of the **Methods** (p. 26), clarify the trade-offs between cloud-based and local deployment, including scenarios where resource requirements may remain prohibitive for some groups, while illustrating the conditions under which MyGPT offers a cost-effective solution.

- The paper is dense at several points, and the structure makes it hard to follow. The HI formula should be included in the main text. The roles of QRS, ARS and HI should be explained with clear visuals. The

extended data tables, especially those about benchmarking, are important but not well integrated. Bringing key information into the main text would improve clarity.

We thank the reviewer for the valuable suggestions, which we have incorporated in the revised manuscript. The Hallucination Index (HI) formula and the roles of QRS, ARS, and HI are now included in the main text and are explicitly defined and visually illustrated in **Fig. 4b-d**, with full formulations presented in the “*Confidence metrics*” section of the **Results** (pp. 5-7). We have also revised the manuscript structure to improve clarity and narrative flow and better integrated key **Supplementary Tables** (particularly those related to benchmarking) into the main **Results**, with explicit references and discussion where they are introduced.

- Despite these issues, the paper addresses an important problem and offers a promising tool for private, transparent and controllable RAG workflows. Its engineering strengths and its practical potential make it a valuable contribution. However, the claims made about performance, generalizability and usability require more evidence. I expect the authors to address the points listed above with concrete additions, broader evaluation and clearer presentation. Once these improvements are made, I would recommend the manuscript for publication.

We thank the reviewer for this constructive assessment. As detailed above, we have addressed all points raised through concrete additions, including broader and fair benchmarking against multiple RAG baselines, quantitative evaluation across real-world and multilingual use cases, clarified methodology, and improved presentation and structure. We hope that the additional analyses, improved data presentation, and the revised *Discussion* now fully address the reviewer's concerns for publication in *Nature Communications*.

Reviewer #4 (Remarks on code availability):

I successfully installed and ran MyGPT on a personal macOS computer (Apple Silicon). The code is well-structured, modular, and production-ready.

Strengths:

- Clear, tiered installation guide for multiple deployment scenarios (personal computer, server/VM, cloud)
- Platform-specific instructions (macOS, Linux, Windows) with dedicated README files
- Well-organized project structure with modular components (/frontend, /backend, /ollama)
- Docker Compose for reproducible deployments (CPU and GPU)
- GitHub Actions CI/CD pipelines demonstrate mature software engineering practices
- Installation completed without errors, documentation is accessible for moderate technical skills

We thank the reviewer for the careful evaluation of the performance of the MyGPT platform.

Minor limitation:

- Main README lacks explicit version specifications for key dependencies (Python, Node.js, Django, React). Pinning exact versions in requirements.txt and package.json would strengthen reproducibility. The code is a usable resource for the community, particularly for researchers and clinicians seeking a local, privacy-preserving RAG system. The modular architecture supports future contributions. No critical issues prohibit publication. Minor enhancements to version specification would strengthen the reproducibility narrative.

We appreciate the suggestions: detailed documentation has been added in the source code at "MyGPT/dependencies.md". Dependency files for the frontend and backend are provided in "frontend/package.json" and "backend/requirements.txt", respectively.

Response to Reviewer's Comments

Reviewer #3 (Remarks to the Author)

This paper presents MyGPT, an open-source, locally deployable RAG platform for question answering over private document libraries. The system integrates vector-based semantic search, BM25 keyword search, and cross-encoder reranking, and introduces three confidence metrics (QRS, ARS, and HI) to help users assess the reliability of generated answers. The authors benchmark MyGPT against four open-source RAG systems on three public datasets and validate it in three real-world scenarios at St. Jude Children's Research Hospital.

Strengths: This work addresses a genuine need for a document question answering tool that does not rely on commercial cloud services in privacy-sensitive clinical and research settings. The system is well-engineered, with a modular design that allows users to flexibly swap individual components, support for edge device deployment, and source highlighting that enhances transparency. The validation across three real-world use cases, particularly the clinical SOP scenario (94.85% accuracy) and the efficiency comparison for kinase literature (2 hours vs. 30–40 hours manually), demonstrates practical value.

We thank the reviewer for the thoughtful assessment of our work and for recognizing its strengths. We are pleased that the reviewer found the privacy-preserving design, modular architecture, local deployment capabilities, source-surfacing functionality, and real-world validations to be valuable contributions. We also appreciate the recognition of the practical utility demonstrated across the clinical SOP, biomedical literature, and global health policy use cases. The reviewer's comments have been helpful in refining the presentation of the manuscript, and we have addressed the points raised below.

Major concerns:

1. Limited methodological novelty. The system is primarily composed of existing open-source components, and the core contribution lies more in system integration and application validation. QRS, ARS, and the original HI are mainly weighted combinations of existing retrieval and relevance scores. Their methodological novelty and incremental value over established RAG evaluation metrics require further clarification. As it stands, this work is closer to a well-validated engineering systems paper than a research paper proposing novel retrieval or generation methods.

We appreciate the reviewer's comments. We agree that the primary contribution of this work lies in the development, integration, and validation of a privacy-preserving RAG platform rather than in proposing fundamentally new retrieval or generation methods. We further acknowledge that QRS, ARS, and HI build upon existing retrieval, relevance, and entailment signals to provide user-facing confidence indicators within a unified framework. Given the rapid adoption of RAG systems, there remains a significant need for an end-to-end, locally deployable solution that supports sensitive institutional data and provides users with interpretable confidence signals to assess answer reliability in real-time interactions. We have revised the manuscript to better clarify the scope of our contribution, emphasize the engineering and validation aspects of the work, and moderate statements that could be interpreted as claims of methodological novelty.

2. The claimed performance advantage over RAGFlow lacks sufficient statistical support. The accuracy gaps between the two systems across the three datasets are modest (e.g., 91.4% vs. 90.2% on BioASQ), and the authors' own power analysis indicates that demonstrating this difference would require sample sizes far exceeding those actually used (5,106 documents needed for BioASQ vs. 210 used). The available data can only show that MyGPT's point estimate of accuracy is higher than RAGFlow's; it cannot yet establish that MyGPT's performance is statistically superior.

We thank the reviewer for this helpful and important comment. We agree that the observed accuracy differences between RAGFlow and MyGPT are modest and that the available sample sizes are insufficient to establish statistical superiority. We have therefore revised the manuscript to soften our claims and clarify that the current results demonstrate a higher point estimate of accuracy for MyGPT, but do not provide sufficient statistical evidence to conclude that MyGPT performs better than RAGFlow. We now discuss this explicitly in the Results and Discussion sections in addition to the Methods.

3. The confidence metrics should not be interpreted as direct measures of factual correctness. QRS, ARS, and HI primarily measure the relevance and support relationships among the question, retrieved text, and answer, and are better viewed as measures of contextual support. Even when an answer is highly consistent with the documents, it may still be factually incorrect due to errors, outdated information, or incompleteness in the source documents, or due to inaccuracies in numbers, negations, or qualifying conditions in the answer. Furthermore, the effect sizes for distinguishing correct from incorrect answers on the public datasets are uniformly small ($r = 0.11\text{--}0.17$), and the random forest achieves a test ROC-AUC of approximately 74%. In the optimal linear configuration across the three combined public datasets, the textual entailment term in ARS receives a weight of zero, indicating that it does not provide additional contribution under this setting.

We thank the reviewer for this constructive comment. We agree that the confidence metrics should be presented as indicators of contextual support rather than direct measures of factual correctness. These metrics quantify the relationships among the user question, retrieved context, and generated answer, and should therefore be interpreted as signals of answer support rather than definitive measures of correctness. In addition to the statement in the Considerations and Limitations section, we have revised the Results, Discussion, and Limitations sections to more clearly communicate the intended role and limitations of QRS, ARS, and HI (including the relatively modest effect sizes observed in the public benchmark datasets), and the importance of human validation when interpreting the generated responses.

Reviewer #4 (Remarks to the Author)

I thank the authors for the substantially revised manuscript. The paper presents MyGPT, an open-source, privacy-first retrieval-augmented generation platform designed for local or secure institutional deployment. The revised version is considerably stronger than the earlier version. In particular, the addition of public benchmarking datasets, head-to-head comparisons with other open-source RAG systems, real-world use cases in clinical SOPs, biomedical literature and multilingual health policy documents, power-based sample size considerations, and a clearer description of the code and deployment model have significantly improved the manuscript. The main strengths of the work are its practical orientation, privacy-preserving local deployment model, modularity, source surfacing, and the attempt to provide user-facing confidence metrics for RAG-based question answering. The manuscript now evaluates MyGPT on PubMedQA, BioASQ and Open RAG Benchmark datasets, reporting higher accuracy than RAGFlow, Dify, PrivateGPT and Kotaemon in the tested subsets. The clinical SOP use case is also compelling, with 129/136 correct answers and 94.85% accuracy reported in a private Bone Marrow Transplantation SOP library. The biomedical literature and multilingual policy examples further demonstrate that the platform has relevance beyond a single narrow domain. Overall, I find the manuscript timely, relevant and potentially useful to researchers, clinical institutions, and other organizations that need to query private document collections without sending sensitive material to commercial cloud-based LLM services. The revision has addressed most of the major conceptual and methodological concerns. I support publication after minor revisions, mainly to improve precision, reproducibility, and calibration of claims.

We thank the reviewer for the thoughtful and constructive assessment of our revised manuscript. We are grateful for the recognition that the revisions, including the addition of public benchmarking datasets, head-to-head comparisons with other open-source RAG systems, real-world validation studies, power-based sample size considerations, and improved descriptions of the code and deployment model, have substantially strengthened the work. We also appreciate the reviewer's positive assessment of the privacy-preserving deployment model, modular architecture, source-surfacing capabilities, and user-facing confidence metrics. We are encouraged that the reviewer finds the platform timely, relevant, and potentially useful across research, clinical, and other privacy-sensitive settings. The reviewer's remaining suggestions regarding precision, reproducibility, and calibration of claims were extremely helpful, and we have addressed each point in detail below.

Specific comments:

1. The QRS, ARS and HI framework is a useful contribution, but the manuscript should be careful not to overstate the strength of these metrics. In the public benchmark datasets, QRS, ARS, and HI distinguish correct from incorrect answers statistically, but the reported effect sizes are small for QRS and ARS/HI in the combined public datasets. The language around these metrics should therefore emphasize that they are useful risk indicators and triage signals, not definitive hallucination detectors. This is partly acknowledged in the limitations section, but the Results and Discussion should use similarly cautious wording.

We thank the reviewer for this constructive feedback. We agree that QRS, ARS, and HI should be presented as risk indicators and triage signals rather than definitive hallucination detectors or direct measures of factual correctness. We further agree that the relatively small effect sizes observed in the combined public benchmark datasets warrant careful interpretation of these metrics. While these measures can help users assess the degree of contextual support for a generated response and identify potentially unreliable answers, they do not replace human validation. Accordingly, we have revised the Abstract, Results, Discussion, and Considerations and Limitations sections to use more cautious language and to clearly communicate the intended scope and limitations of these metrics.

2. The new power-analysis section is very helpful. However, the authors explicitly state that substantially larger sample sizes would be required for adequately powered comparisons with RAGFlow in PubMedQA, BioASQ and Open RAG Benchmark, but these larger evaluations were not performed. The manuscript should make this limitation visible in the Results or Discussion, not only in the Methods/Supplementary material. Claims that MyGPT outperforms RAGFlow should be softened accordingly, whereas the stronger claims against Dify, PrivateGPT and Kotaemon appear better supported by the available sample sizes.

We thank the reviewer for this helpful comment. We agree that the power analysis highlights an important limitation of the current comparison with RAGFlow and that the available sample sizes are insufficient to establish statistical superiority. We have therefore revised the Results and Discussion sections to make this limitation more explicit and softened claims regarding MyGPT's performance relative to RAGFlow, emphasizing that the current data support higher point estimates of accuracy but not definitive evidence of superior performance. We have also clarified that the comparative findings against Dify, PrivateGPT, and Kotaemon are better supported by the available sample sizes.

3. The revised manuscript now states that correctness labels for public datasets were provided by two independent researchers and that private datasets were evaluated by domain experts. However, the manuscript should provide additional detail on whether evaluators were blinded to the tool identity, how disagreements were resolved, and whether inter-rater agreement was measured. This is especially

important because the comparisons involve manually evaluated answers and because the use cases include private institutional datasets.

We thank the reviewer for these constructive comments and for identifying the need for additional details regarding the evaluation methodology. The evaluators for the public datasets were blinded to the identity of the RAG system that generated each answer. For PubMedQA and BioASQ, the presence of clearly defined ground-truth answers and supporting context enabled straightforward evaluation, and no significant disagreements requiring a formal adjudication process occurred between the two researchers. For the Open RAG Benchmark, we first used GPT-OSS-120B as an LLM-as-a-judge to generate preliminary labels (correct, incorrect, or ambiguous) together with supporting reasoning. Two independent researchers then reviewed and verified these labels, documenting instances in which their assessments differed from the LLM-generated labels. Because disagreements between the two human evaluators were infrequent and readily resolved through discussion, we did not perform a formal inter-rater agreement analysis. For the private datasets, only MyGPT was evaluated using different system customizations, and domain experts were blinded to the specific configuration that generated each answer. Ground-truth responses were established before evaluation, and experts were asked to assign correctness labels and provide supporting comments. We have updated the Methods section accordingly.

4. There appear to be inconsistencies between the main text, Methods and Supplementary Figure 4. The main text reports 282 Q&A pairs, 206/282 correct answers, and 227/282 correct answers after using 15 chunks. The Methods section, however, describes 297 Q&A pairs for the kinase library, and Supplementary Figure 4 reports 238/297 correct answers with $n=15$. These numbers should be reconciled before publication.

We thank the reviewer for identifying this inconsistency, which arose from an error in reporting. The correct dataset size for the kinase literature evaluation is 282 Q&A pairs, and the values reported in the Methods section and Supplementary Figure 4 were not updated following a revision of the dataset. We have reconciled these discrepancies and revised the Methods, Results, and Supplementary Figure 4 to consistently report 282 total Q&A pairs and the corresponding accuracy values for each chunking configuration. We appreciate the reviewer's careful attention to detail and have verified that all related values are now internally consistent throughout the manuscript.

5. The multilingual health-policy use case is valuable, but the Methods state that the reranker was omitted in this evaluation because of limited foreign-language expertise and manual validation constraints, and that future evaluation of reranker suitability is needed. This should be clearly reflected in the Results and Discussion. The authors should avoid implying that the full MyGPT confidence-metric pipeline has been equally validated in multilingual settings if the pipeline differs materially from the English-language evaluations.

We thank the reviewer for this constructive assessment. We agree that the multilingual health-policy use case differs from our English-language evaluations and that the full MyGPT confidence-metric pipeline was not evaluated under the same conditions. We have revised the Results and Discussion sections to make this distinction more explicit and to clearly state that further evaluation of reranking strategies and confidence metrics is needed in multilingual settings. This revision ensures that our claims accurately reflect the scope of the validation performed and avoids implying that multilingual performance has been validated to the same extent as the English-language evaluations. We also highlight this as an area for future work.

6. The manuscript states that the code is currently private, will be made public upon acceptance, and is provided as supplementary material for review. Before publication, the authors should ensure that the public repository includes a clear README, installation instructions, dependency versions, example data,

instructions for reproducing the benchmark evaluations where possible, and a statement on how private datasets were handled or substituted. Given the centrality of MyGPT as a software resource, this is essential for reproducibility.

We thank the reviewer for these valuable suggestions regarding reproducibility. The source code provided for review already includes an extensive README with installation instructions for multiple operating systems, virtual machines, and cloud platforms, as well as dependency requirements, FAQs, evaluation datasets, and benchmark results. We agree that comprehensive documentation is essential given the central role of MyGPT as a software resource. Accordingly, for the public release we have expanded the repository to include detailed instructions for reproducing benchmark evaluations where possible, example datasets and workflows, and a clarified Data Availability statement describing how private institutional datasets were handled and why they cannot be publicly distributed. These additions will further improve the transparency, reproducibility, and usability of MyGPT for the broader research community.

7. Several typographical errors and wording issues remain and should be corrected during revision. Examples include “populat” in the confidence metric definitions, “comipled” and “evauate” in the multilingual policy section, “cusromize” in Figure 6, and “MyGTP” in the Figure 6 legend. These do not affect the scientific contribution but should be corrected before publication.

We thank the reviewer for pointing this out. We have corrected these typographical errors in the manuscript.

Additional assessment of Reviewer #1’s concerns

I have also reviewed Reviewer #1’s original comments together with the authors’ point-by-point responses and the corresponding revisions to the manuscript. In my assessment, the authors have substantially addressed the major concerns raised by Reviewer #1. In particular, the addition of direct benchmarking against multiple open-source RAG systems across three public datasets substantially strengthens the evidence base and addresses the earlier concern that comparison with a standalone LLM was insufficient. The revised architecture and workflow descriptions, the addition of BM25 retrieval and reranking, the expanded real-world evaluations, and the clearer comparison with existing RAG solutions also substantially improve the differentiation and scientific framing of MyGPT. The concern regarding the Hallucination Index has been addressed in part through a more sophisticated formulation incorporating semantic, keyword and entailment-based information, sensitivity analysis, and additional validation. However, I agree that these metrics should still be interpreted as indicators of potential answer unreliability rather than definitive measures of factual correctness, as noted in Comment 1 above. Similarly, while the comparative benchmarking is a major improvement, the authors’ own power calculations indicate that the comparisons with RAGFlow are underpowered to establish superiority, and the corresponding claims should therefore be moderated, as noted in Comment 2. Overall, I consider Reviewer #1’s concerns to have been substantially, though not completely, addressed; the remaining issues appear suitable for minor revision.

We thank the reviewer for carefully considering Reviewer #1’s original concerns, our responses, and for concluding that the major issues have been substantially addressed through the revisions to the manuscript. We agree with the reviewer’s remaining observations regarding the interpretation of the confidence metrics and the limitations of the current comparison with RAGFlow. As described in our responses to Comments 1 and 2 above, we have revised the Abstract, Results, Discussion, and Limitations sections to use more cautious language regarding QRS, ARS, and HI, and to more clearly acknowledge the limited statistical power of the current RAGFlow comparison and the resulting need to moderate claims of superiority.