Supplementary Information for Patel *et al. NCOMMS-25-82619A*

SUPPLEMENTARY TABLES (1-7)

Supplementary Table 1: Example LLMs that can be used with MyGPT via Ollama.

LLM	Parameters	Size	Developed by	Note	Ideal hosting environment	Minimum requirements
Llama2:7b	6.7B	3.8 GB	Meta	Released on 18th July 2023	PC/Laptop, Cloud service	1 GPU, 8 GB RAM
Llama3:8b	8.03B	4.7 GB			PC/Laptop, Cloud service	1 GPU, 8 GB RAM
Llama3:70b	70.6B	40 GB			VM, server, Cloud service	1 GPU, 64 GB RAM
Llama3.1:8b	8.03B	4.9 GB		Released on 23rd July, 2024	PC/Laptop, Cloud service	1 GPU, 8 GB RAM
Llama3.2:1b	1.24B	1.3 GB		Released on 25th September, 2024	Edge AI device, PC/Laptop, Cloud service	1 GPU, 8 GB RAM
Llama3.3:70b	70.6B	43 GB		Released on 7th December, 2024	VM, server, Cloud service	1 GPU, 64 GB RAM
Gemma:7b	8.54B	5 GB	Google DeepMind	Released on 21st February 2024	PC/Laptop, Cloud service	1 GPU, 8 GB RAM
Gemma2:9b	9.24B	5.4 GB		Released on 27th June 2024	PC/Laptop, Cloud service	1 GPU, 8 GB RAM
Gemma2:2b	2.61B	1.6 GB			Edge AI device, PC/Laptop, Cloud service	1 GPU, 8 GB RAM
Gemma3:1b, 4b	1B, 4.3B	0.8GB, 3.3GB		Released on 12th March 2025	Edge AI device, PC/Laptop, Cloud service	1 GPU, 8 GB RAM
Gemma3:12b, 27b	12.2B, 27.4B	8.1GB, 17GB		Released on 12th March 2025	PC/Laptop, Cloud service	1 GPU, 32 GB RAM
Gemma4:26b, 31b	25.8B, 31.3B	18GB, 20GB		Released on 2nd April 2026	PC/Laptop, Cloud service	1 GPU, 32 GB RAM
Gpt-oss:20b	20.9B	14GB	OpenAI	Released on 5th August 2025 (Thinking model)	PC/Laptop, Cloud service	1 GPU, 32 GB RAM
DeepSeek-R1:8b, 14b	8.19B, 14.8B	5.2GB, 9GB	DeepSeek	Released on January 20, 2025 (Thinking model)	PC/Laptop, Cloud service	1 GPU, 32 GB RAM
Qwen3:8b, 14b	8.19B, 14.8B	5.2GB, 9.3GB	Alibaba	Released on April 29th, 2025 (Thinking model)	PC/Laptop, Cloud service	1 GPU, 16 GB RAM
Mistral:7b	7.2B	4.1 GB	Mistral AI	Released on 27th September 2023	PC/Laptop, Cloud service	1 GPU, 8 GB RAM
Vicuna:7b	6.74B	3.8 GB	LMSYS Org, UC Berkeley	Finetuned Llama2	PC/Laptop, Cloud service	1 GPU, 8 GB RAM

Supplementary Table 2: LLM performance for different environments.

LLM size category	LLM	Host environment	Infrastructure	GPU	Avg. Speed (seconds)
Mid-size LLMs (most common)	**Llama2:7b**	Google Cloud Platform	2 vCPUs, 83 GB RAM	1 Nvidia A100 Tensor Core	3.80
		Google Cloud Platform	2 vCPUs, 12 GB RAM	1 Nvidia T4 Tensor Core	6.46
		St. Jude internal server	64 CPUs, 2 TB RAM	1 Nvidia A100 Tensor Core	2.89
		Apple Mac m1	8 CPUs, 64 GB RAM	1 Apple M1 Metal	9.26
		Apple Mac m3	8 CPUs, 64 GB RAM	1 Apple M3 Metal	5.21
	Llama3:8b	Google Cloud Platform	2 vCPUs, 83 GB RAM	1 Nvidia A100 Tensor Core	3.34
		Google Cloud Platform	2 vCPUs, 12 GB RAM	1 Nvidia T4 Tensor Core	4.96
		St. Jude internal server	64 CPUs, 2 TB RAM	1 Nvidia A100 Tensor Core	3.40
		Apple Mac M1	8 CPUs, 64 GB RAM	1 Apple M1 Metal	7.47
		Apple Mac M3	8 CPUs, 64 GB RAM	1 Apple M3 Metal	6.00
	Llama3.1:8b	Google Cloud Platform	2 vCPUs, 83 GB RAM	1 Nvidia A100 Tensor Core	2.49
		Google Cloud Platform	2 vCPUs, 12 GB RAM	1 Nvidia T4 Tensor Core	6.31
		St. Jude internal server	64 CPUs, 2 TB RAM	1 Nvidia A100 Tensor Core	2.05
		Apple Mac M1	8 CPUs, 64 GB RAM	1 Apple M1 Metal	7.05
		Apple Mac M3	8 CPUs, 64 GB RAM	1 Apple M3 Metal	4.45
	Gemma:7b	Google Cloud Platform	2 vCPUs, 83 GB RAM	1 Nvidia A100 Tensor Core	2.74
		Google Cloud Platform	2 vCPUs, 12 GB RAM	1 Nvidia T4 Tensor Core	4.38
		St. Jude internal server	64 CPUs, 2 TB RAM	1 Nvidia A100 Tensor Core	1.61
		Apple Mac M1	8 CPUs, 64 GB RAM	1 Apple M1 Metal	5.52
		Apple Mac M3	8 CPUs, 64 GB RAM	1 Apple M3 Metal	2.80
	Gemma2:9b	Google Cloud Platform	2 vCPUs, 83 GB RAM	1 Nvidia A100 Tensor Core	2.72
		Google Cloud Platform	2 vCPUs, 12 GB RAM	1 Nvidia T4 Tensor Core	4.57
		St. Jude internal server	64 CPUs, 2 TB RAM	1 Nvidia A100 Tensor Core	1.46
		Apple Mac M1	8 CPUs, 64 GB RAM	1 Apple M1 Metal	6.01
		Apple Mac M3	8 CPUs, 64 GB RAM	1 Apple M3 Metal	3.28
	Mistral:7b	Google Cloud Platform	2 vCPUs, 83 GB RAM	1 Nvidia A100 Tensor Core	3.37
		Google Cloud Platform	2 vCPUs, 12 GB RAM	1 Nvidia T4 Tensor Core	5.69
		St. Jude internal server	64 CPUs, 2 TB RAM	1 Nvidia A100 Tensor Core	2.67
		Apple Mac M1	8 CPUs, 64 GB RAM	1 Apple M1 Metal	7.87
		Apple Mac M3	8 CPUs, 64 GB RAM	1 Apple M3 Metal	5.28
	Vicuna:7b	Google Cloud Platform	2 vCPUs, 83 GB RAM	1 Nvidia A100 Tensor Core	3.96
		Google Cloud Platform	2 vCPUs, 12 GB RAM	1 Nvidia T4 Tensor Core	8.73
		St. Jude internal server	64 CPUs, 2 TB RAM	1 Nvidia A100 Tensor Core	3.56
		Apple Mac M1	8 CPUs, 64 GB RAM	1 Apple M1 Metal	9.51
		Apple Mac M3	8 CPUs, 64 GB RAM	1 Apple M3 Metal	8.93
Small LLMs	**Llama3.2:3b**	Apple Mac M1	8 CPUs, 64 GB RAM	1 Apple M1 Metal	3.79
		Apple Mac M3	8 CPUs, 64 GB RAM	1 Apple M3 Metal	2.28
		Nvidia Orin-nano	1 CPU, 8 GB RAM	1 1024-core NVIDIA Ampere	19.88
	Llama3.2:1b	Apple Mac M1	8 CPUs, 64 GB RAM	1 Apple M1 Metal	2.15
		Apple Mac M3	8 CPUs, 64 GB RAM	1 Apple M3 Metal	1.52
		Nvidia Orin-nano	1 CPU, 8 GB RAM	1 1024-core NVIDIA Ampere	9.82
	Gemma:2b	Apple Mac M1	8 CPUs, 64 GB RAM	1 Apple M1 Metal	1.25
		Apple Mac M3	8 CPUs, 64 GB RAM	1 Apple M3 Metal	0.80
		Nvidia Orin-nano	1 CPU, 8 GB RAM	1 1024-core NVIDIA Ampere	5.26
Large LLMs	**Llama3:70b**	Google Cloud Platform	2 vCPUs, 83 GB RAM	1 Nvidia A100 Tensor Core	11.17
		St. Jude internal server	64 CPUs, 2 TB RAM	1 Nvidia A100 Tensor Core	31.11
		Apple Mac M1	8 CPUs, 64 GB RAM	1 Apple M1 Metal	60.27
		Apple Mac M3	8 CPUs, 64 GB RAM	1 Apple M3 Metal	49.53

Supplementary Table 3. Cost of hosting LLMs on Cloud infrastructure.

Cloud service provider	Infrastructure	GPU	Avg. cost per month (USD)
Azure	24 vCPUs, 220 GB RAM	1 Nvidia A100 Tensor Core	$2,681.29
Azure	8 vCPUs, 56 GB RAM	1 Nvidia T4 Tensor Core	$548.96
Google	12 vCPUs, 85 GB RAM	1 Nvidia A100 Tensor Core	$2,682.57
Google	8 vCPUs, 30 GB RAM	1 Nvidia T4 Tensor Core	$374.03
Google	4 vCPUs, 16 GB RAM	1 Nvidia L4 Tensor Core	$516.99

Supplementary Table 4. Benchmarking of various MyGPT steps for different library sizes.

Device	Pages	PDF count	Chunks count (fixed chunk size)	Upload time	Retrieval time (1st)	Retrieval time (avg)	Generation time (1st) llama3:7b	Generation time (avg)	end-to-end time (avg)
Mac Laptop (M1 chip)	10	1	70	2m10s	1.28s	1.06s	6.39s	2.99s	11s
	100	2	235	2m33s	1.18s	0.62s	4.16s	6.03s	17s
	1000	4	3211	5m53s	0.73s	0.59s	3.16s	3.9s	12s
	10000	791	58869	95m19s	1.94s	0.64s	5.52s	5.82s	17s
Mac Laptop (M4 chip)	10	1	70	2.05s	0.38s	0.225s	0.84s	1.205s	4.5s
	100	2	235	4.9s	0.1s	0.0775s	2.04s	2.295s	4.5s
	1000	4	3211	1m18s	0.14s	0.11333s	1.9s	2.049s	4.2s
	10000	791	58869	23m51s	0.69s	0.651s	1.86s	2.124s	5.0s
SJ VM	10	1	70	11s	5.63s	0.80s	11.22s	0.51s	11s
	100	2	235	13s	0.77s	0.79s	1.2s	1.07s	6.75s
	1000	4	3211	2m12s	0.68s	0.73s	1.16s	1.49s	12.6s
	10000	791	58869	46m2s	0.77s	0.85s	1.19s	1.39s	8s

Supplementary Table 5: Sample size calculations for MyGPT vs other RAG solutions.

Dataset	MyGPT correction rate (p_1)	Other RAG Tools	Correction rate ($p2$)	MyGPT-vs-RAG-tool corrections correlation (ρ)	Required number of docs (c)
PubMedQA Q&A=294, docs used=294)	0.8163	RAGFlow	0.7755	0.608	806
		PrivateGPT	0.5102	0.062	43
		Dify	0.6803	0.146	179
		Kotaemon	0.4558	0.134	28
BioASQ (Q&A=163, docs used=210)	0.9141	RAGFlow	0.9018	0.561	5106
		PrivateGPT	0.7485	0.327	72
		Dify	0.6564	0.009	48
		Kotaemon	0.7607	0.085	108
Open RAG Benchmark (Q&A=170, docs used =165)	0.9294	RAGFlow	0.9	0.444	1063
		PrivateGPT	0.7824	0.299	84
		Dify	0.7882	0.306	89
		Kotaemon	0.5588	0.218	20

Supplementary Table 6: Cost of hosting MyGPT on commodity hardware for group or department usage.

Hardware	Configuration	GPU	Price	Best LLM to host
Mac Mini	64 GB RAM, 512 GB Storage	M4 Pro chip	$1,839	Gpt-os:20b, Gemma4:31b
Framework Desktop	64 GB RAM, 512 GB Storage	AMD Ryzen™ AI Max+ 395	$2,213	Gpt-os:20b, Gemma4:31b
Mac Studio	128 GB, 512 GB storage	M4 Max chip	$3,149	Gpt-oss:120b, Llama3.3:70b
Framework Desktop	64 GB RAM, 512 GB Storage	AMD Ryzen™ AI Max+ 395	$3,253	Gpt-oss:120b, Llama3.3:70b
Nvidia DGX Spark	128 GB, 1 TB storage	Nvidia Blackwell Architecture	$4,000	Gpt-oss:120b, Llama3.3:70b

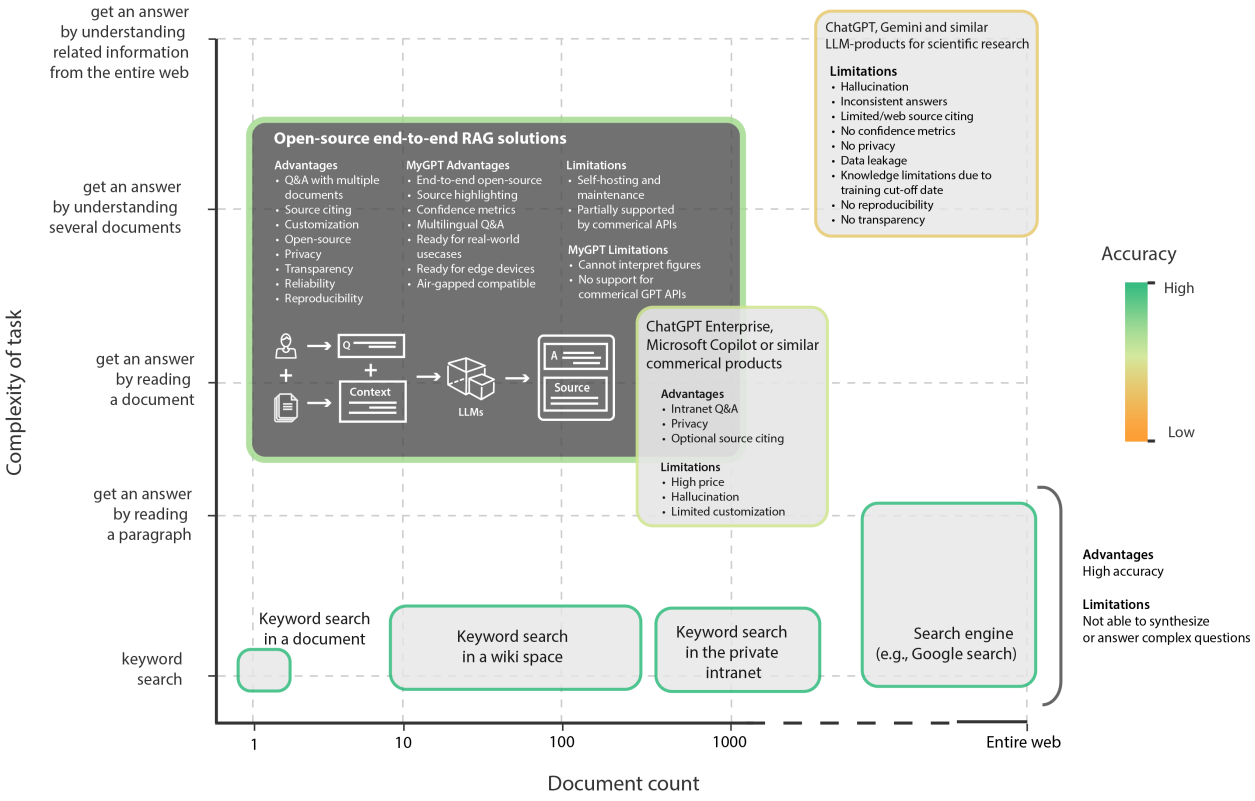
Supplementary Table 7: Cost comparison with various commercial and open-source RAG solutions.

Tool	Type	Pricing structure	Cost	Notes	Sources
Microsoft 365 Copilot (Business)	Commercial, SaaS	Per user/month	$18/user/mo (through Jun 2026), then $21/user/mo	Requires a qualifying M365 license on top of the Copilot fee. RAG grounded in M365 Graph data.	[1]
Microsoft 365 Copilot (Enterprise)	Commercial, SaaS	Per user/month	$30/user/mo on top of M365 E3 ($39) or E5 ($60)	True all-in cost ranges from $66 - $87/user/month when base license is included.	[2, 3]
Gemini for Google Workspace	Commercial, SaaS	Per user/month	Now bundled into Workspace plans ($7--$22/user/mo after 17--22% price increase)	Previously sold as a $20 - $30/user/mo add-on; standalone Gemini add-ons discontinued for new purchases. Vertex AI API access is separate.	[4, 5]
Gemini Enterprise (Google Cloud)	Commercial, SaaS	Per user/month + consump-tion	$30/user/mo as Workspace add-on for existing subscribers; volume discounts at 500+ and 1000+ seats. Vertex AI: $1.88/M input tokens, $7.50/M output tokens.	Includes agentic platform, compliance controls (HIPAA, FedRAMP).	[6, 7]
ChatGPT Business	Commercial, SaaS	Per user/month	$25/user/mo (annual) $30/user/mo (monthly)	Includes admin controls, SOC 2, SAML SSO. Data not used for training.	[8, 9]
ChatGPT Enterprise	Commercial, SaaS	Custom quote	Negotiated	Adds custom data retention, volume licensing, dedicated support.	[8]
Glean	Commercial, SaaS	Per user/month (quote-based)	$45-$50+ user/mo base; AI add-on ~$15/user/mo extra	100-seat minimum; enterprise contracts often exceed $200K/year. No public pricing. Renewal increases of 7--12% reported.	[10, 11]
Vectara	Commercial, SaaS/VPC	Usage-based (credits)	Enterprise plans: Small $100K/yr, Medium $250K/yr, Large $500K/yr	30-day free trial. Pricing based on index size and query volume, not per-seat. VPC and on-prem options available.	[12, 13]
PrivateGPT (Abstracta)	Commercial, self-hosted (AWS)	Custom quote	No public list price; request private offer via AWS Marketplace	Self-hosted on EKS/ECS. Includes SSO, per-user spend caps, RAG pipelines, usage analytics.	[14]
Kotaemon	Open source (Apache 2.0)	Free (self-hosted)	$0 (software)	Self-hosted via Docker. Supports OpenAI, Azure, Ollama, local LLMs. Ongoing costs are infra + LLM API tokens.	[15, 16]
LlamaIndex / LangChain	Open source	Free	$0 (framework)	Developer frameworks, not turnkey products. Total cost depends on hosting, vector DB, and LLM API spend.	[17]
Meilisearch	Open source + Cloud	Freemium / usage-based	Build plan $30/mo (50K searches, 100K docs); self-hosted is free	Search engine with RAG integration, not a full copilot/assistant.	[17]

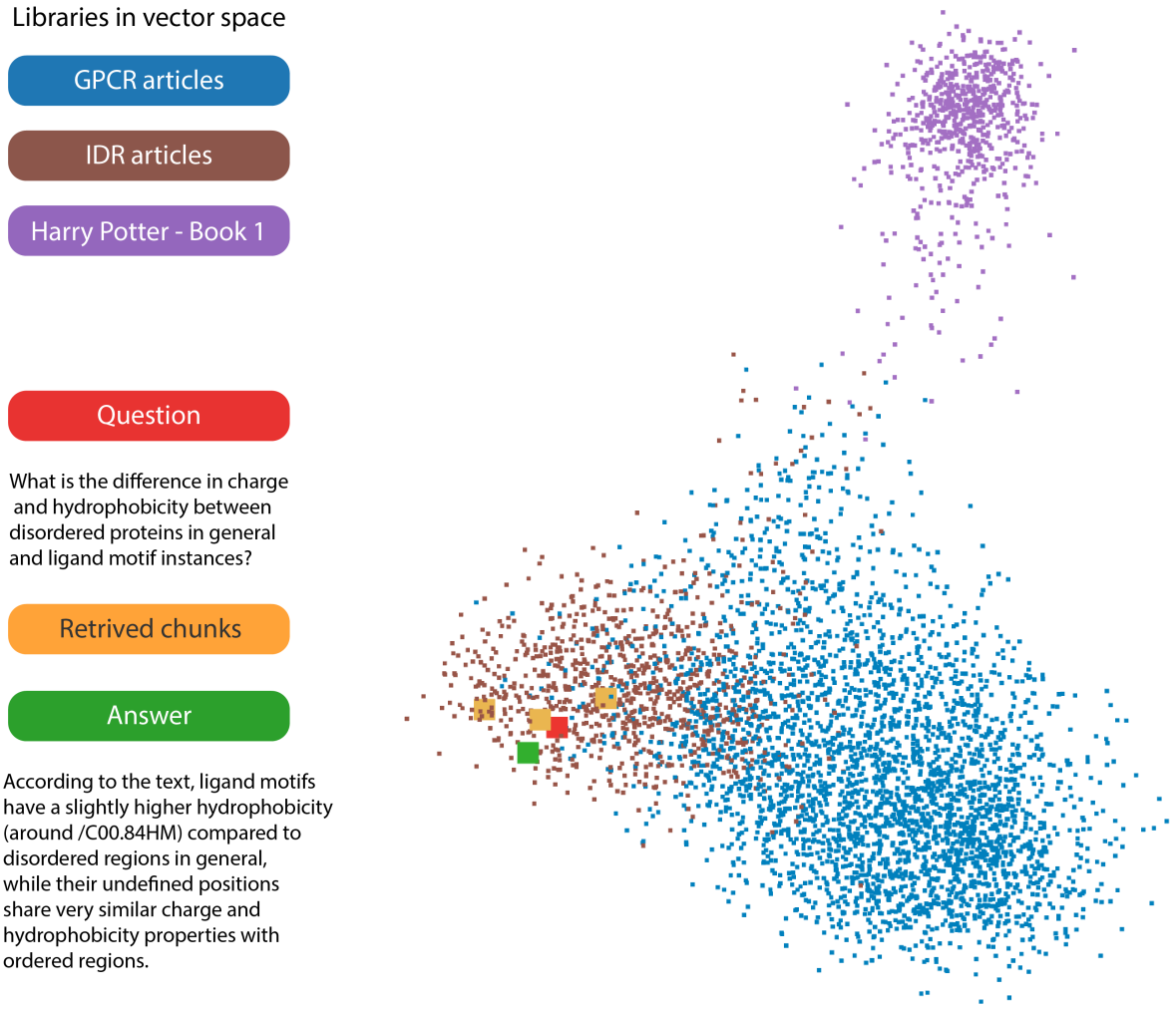
Key takeaways

- **Per-user tools carry hidden costs: Microsoft Copilot requires a qualifying M365 base license, so the real per-seat cost is roughly double the headline add-on price. Glean layers on support fees (~10% of contract value) and reported renewal increases.**
- **Open-source options shift cost to operations. Kotaemon, LlamaIndex, and LangChain have zero licensing fees but require engineering time, infrastructure, and LLM API bills.**
- **Gemini pricing restructured in 2025. Google folded Gemini into base Workspace plans with a 17--22% price increase, eliminating the separate add-on SKU for new customers. Existing subscribers on legacy Gemini add-ons should evaluate whether they are double-paying for capabilities also available through Vertex AI.**
- **Vectara is the outlier in pricing model. It charges by usage credits (index size + queries) rather than seats, which can benefit organizations with many documents but relatively few heavy users.**

SUPPLEMENTARY FIGURES (1-14)

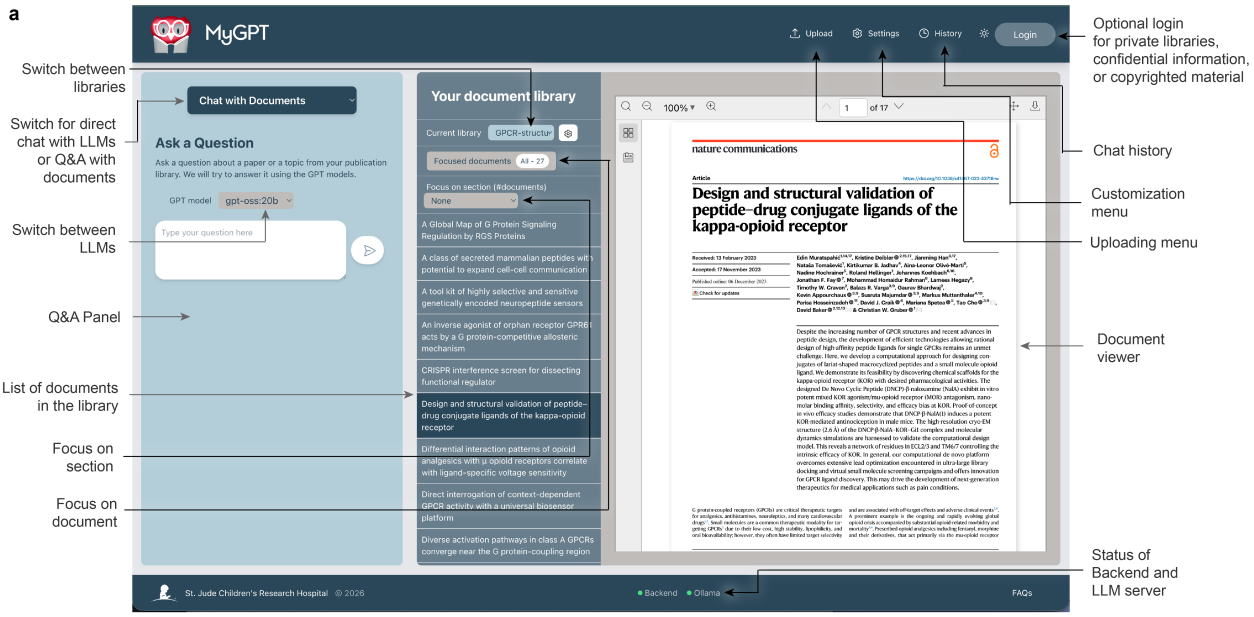


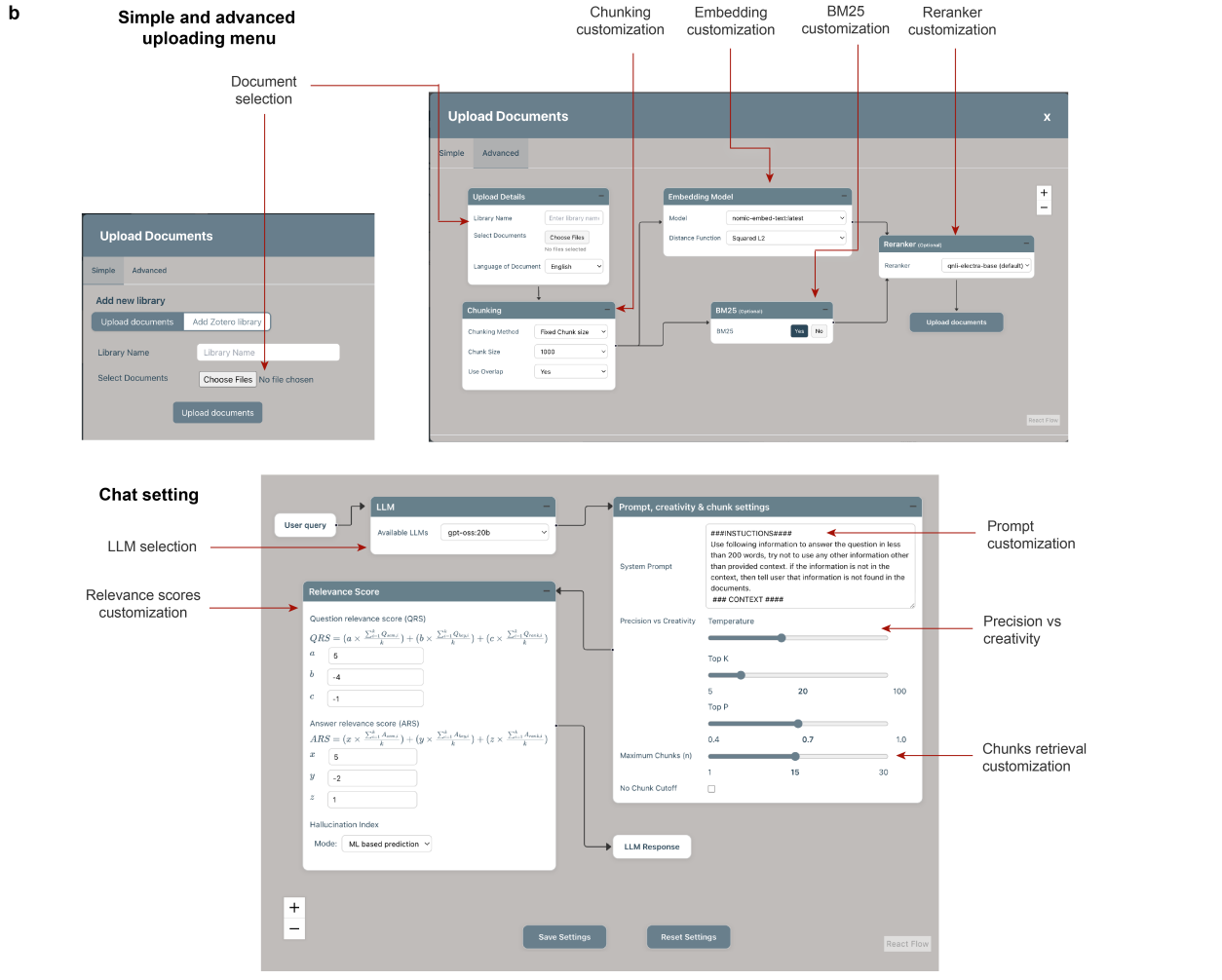
Supplementary Figure 1. Information-seeking tasks and solutions.
Keyword searches from a single document, multiple documents, a private intranet, or the entire web are routine tasks for most knowledge workers that have been optimized over several decades to provide accurate information. With ChatGPT and similar products, users can get answers to complex questions requiring processing of related material, but accuracy and factuality are lower than those of typical search solutions. MyGPT and other RAG solutions fill a void in the landscape of available discovery tools. The orange-to-green color of the framed boxes indicates the level of accuracy of the information obtained from the indicated task.



Supplementary Figure 2. Visualization of chunks from library of documents in the vector space of embeddings.

The PCA plot visualizes vector embeddings of chunks from two different libraries of publicly available biomedical articles on (a) G protein-coupled receptors (GPCRs; 27 articles) in blue, (b) intrinsically disordered regions (IDRs; 8 articles) in brown, and text from (c) Harry Potter-Book 1 in purple (**Source Data**). Each dot in the plot denotes a chunk from the three different libraries. A question related to the IDR library is converted into a vector (red square). MyGPT then identifies the chunks (orange squares) that are closest to the question and provides them as context for the LLM to generate a response. The generated answer (green square), when converted into a vector using the same embedding model, is closer to the context, highlighting similarity.


a
MyGPT
Upload
Settings
History
Login
Optional login for private libraries, confidential information, or copyrighted material
Switch between libraries
Chat with Documents
Switch for direct chat with LLMs or Q&A with documents
Ask a Question
Ask a question about a paper or a topic from your publication library. We will try to answer it using the GPT models.
GPT model
gpt-oss:20b
Switch between LLMs
Type your question here
Q&A Panel
List of documents in the library
Focus on section
Focus on document
Your document library
Current library
GPCR-structur
Focused documents
All - 27
Focus on section (#documents)
None
A Global Map of G Protein Signaling Regulation by RGS Proteins
A class of secreted mammalian peptides with potential to expand cell-cell communication
A tool kit of highly selective and sensitive genetically encoded neuropeptide sensors
An inverse agonist of orphan receptor GPR6 acts by a G protein-competitive allosteric mechanism
CRISPR interference screen for dissecting functional regulator
Design and structural validation of peptide–drug conjugate ligands of the kappa-opioid receptor
Differential interaction patterns of opioid analgesics with μ opioid receptors correlate with ligand-specific voltage sensitivity
Direct interrogation of context-dependent GPCR activity with a universal biosensor platform
Diverse activation pathways in class A GPCRs converge near the G protein–coupling region
100%
1
of 17
Chat history
Customization menu
Uploading menu
nature communications
Article
Design and structural validation of peptide–drug conjugate ligands of the kappa-opioid receptor
Document viewer
St. Jude Children's Research Hospital © 2026
Backend
Ollama
FAQs
Status of Backend and LLM server


b
Simple and advanced uploading menu
Chunking customization
Embedding customization
BM25 customization
Reranker customization
Document selection
Upload Documents
Simple
Advanced
Add new library
Upload documents
Add Zotero library
Library Name
Select Documents
Choose Files
No file chosen
Upload documents
Upload Details
Embedding Model
Model
nomic-embed-text:latest
Distance Function
Squared L2
Reranker (Optional)
Reranker
qnli-electra-base (default)
Language of Document
English
Chunking
Chunking Method
Fixed Chunk size
Chunk Size
1000
Use Overlap
Yes
BM25 (Optional)
BM25
Yes
No
Upload documents
Chat setting
User query
LLM
Available LLMs
gpt-oss:20b
LLM selection
Prompt, creativity & chunk settings
Prompt customization
System Prompt
###INSTUCTIONS####
Use following information to answer the question in less than 200 words, try not to use any other information other than provided context. if the information is not in the context, then tell user that information is not found in the documents.
CONTEXT
Precision vs Creativity
Temperature
Precision vs creativity
Top K
5
20
100
Top P
0.4
0.7
1.0
Maximum Chunks (n)
1
15
30
Chunks retrieval customization
No Chunk Cutoff
Relevance scores customization
Relevance Score
Question relevance score (QRS)
a
5
b
-4
c
-1
Answer relevance score (ARS)
x
5
y
-2
z
1
Hallucination Index
Mode:
ML based prediction
LLM Response
Save Settings
Reset Settings
React Flow

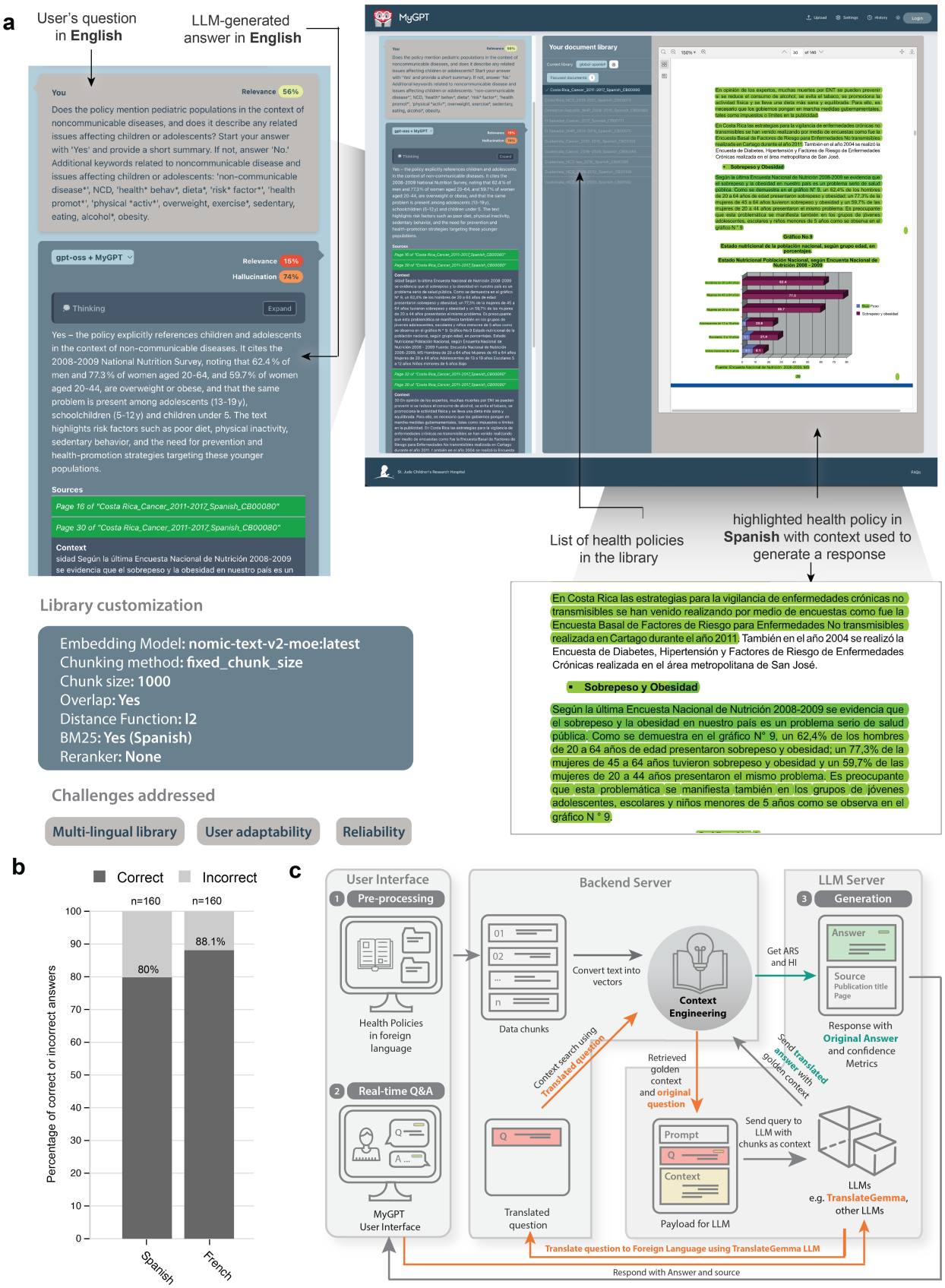
Supplementary Figure 3. User interface for MyGPT

a MyGPT UI home screen. The home screen displays a chat box (left) with options to switch between pre-loaded document libraries and LLMs, and a dropdown menu that lets users chat directly with the LLM or with documents. The “Your document library” panel (middle) displays a list of documents in the library, with the selected document loaded and visible in the document viewer (right). The top navigation bar displays icons to enable user login to private libraries (e.g., group- or department-level access). Users can also check chat history for various libraries and access the settings menu from this area. Users can verify whether the Backend and LLM server are properly connected and functional from the page footer. **b** MyGPT upload and chat menu. From the Upload menu, users can upload documents and create a library; each preprocessed library is accessible from the MyGPT home page. When uploading, users can either select a Simple upload with default customization pre-selected or can go to Advanced settings to customize various components such as chunk size, chunk method, chunk overlap, embedding model, distance function, BM25, and Reranker. In the chat settings, users can change the LLM, prompt, precision vs. creativity parameters, chunk retrieval customization, and relevance score customization.



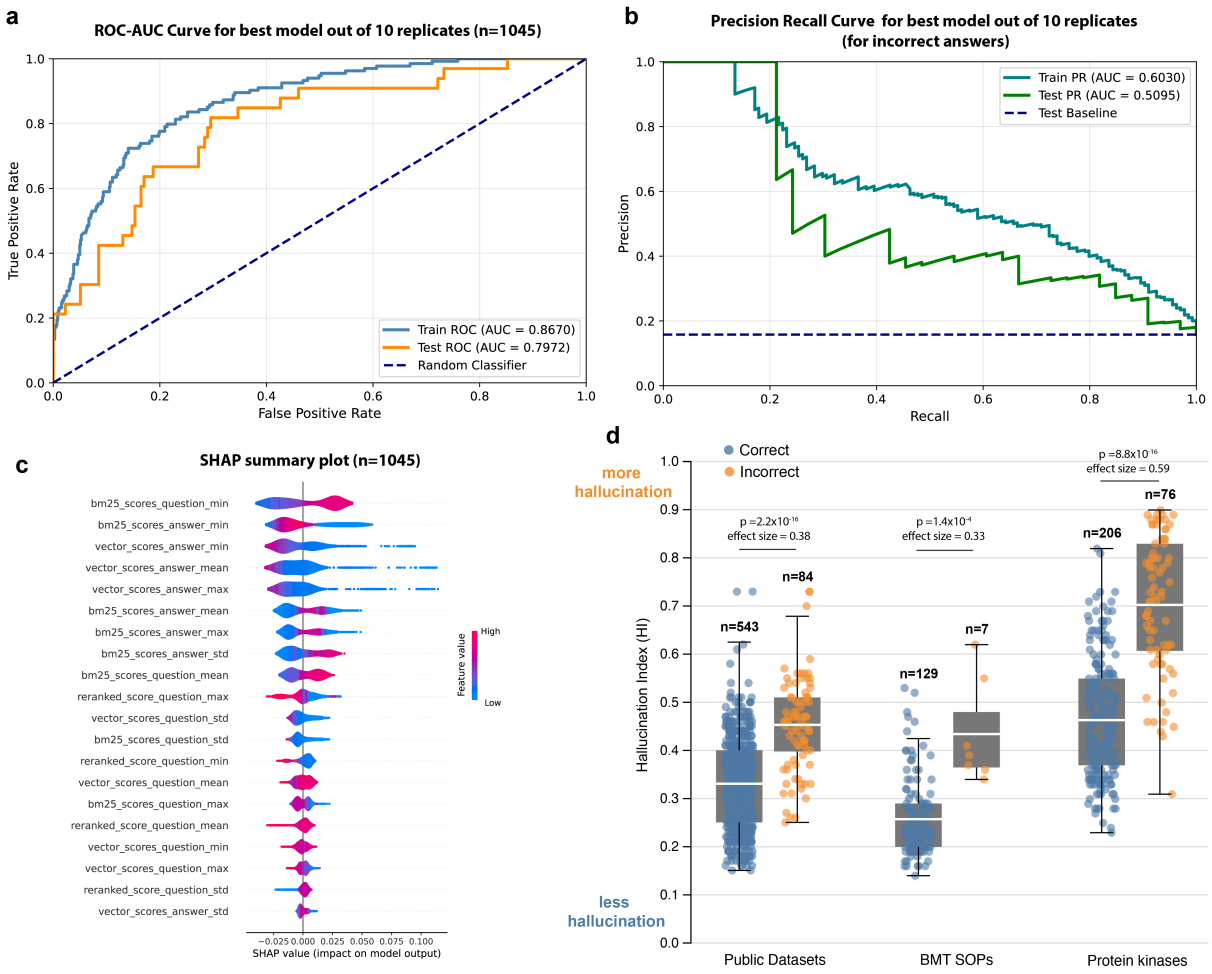
Supplementary Figure 4. Use case 1: Q&A with large library of biomedical literature.
a Example of a complex question and answer for the library of Biomedical literature on protein kinases. The user's question is shown along with the highlighted section of the articles used by the LLM (gpt-oss:20b) to generate the answer. This use case addresses the challenge of retrieving related information from multiple sections of an article and synthesizing an answer after analyzing the data. The library was created using

the nomic-embed-text embedding model, which utilizes the reasoning capabilities of the LLM (gpt-oss:20b) to generate an answer. **b** 282 questions (with ground truth) were collected from 83 documents for this evaluation. MyGPT was able to correctly answer 206 questions with 73% accuracy. The same experiment performed with 15 chunks per question (n=15, maximum chunks per semantic and keyword search) rather than the default 5 chunks yielded more correct answers with 80.5% (227/282) accuracy, showing significant improvement (p = $0.7x10^{-2}$; effect size = 0.34 (Φ, medium); McNemar chi-square test). Also, changing the chunking strategy from fixed to structure-preserving (with n=5) yielded 75.8% (214/282) accuracy, which did not provide a significant improvement (p = 0.27; effect size=0.15 (Φ, small); McNemar chi-square test) over fixed chunking with 5 chunks. Source data are provided as a Source Data file. **c** MyGPT's Question relevance score (QRS) distribution for questions with correct or incorrect answers (p = 0.13; effect size = 0.08 (r, small); Mann Whitney U test: two-sided) is not statistically different, but the Answer relevancy score (ARS) distribution for correct and incorrect answers (p = $5x10^{-12}$; effect size = 0.41 (r, medium); Mann Whitney U test: two-sided) shows higher scores for the correct answers. MyGPT's Default Hallucination Index (HI) for correct and incorrect answers (p = $6x10^{-13}$; effect size = 0.43 (r, medium); Mann Whitney U test: two-sided) shows significantly higher hallucination index values for incorrect answers. Box-plots with individual data points: center line, mean; box, 25th-75th percentiles (IQR); whiskers, most extreme values within 1.5 x IQR (Tukey); points beyond whiskers are outliers. All graph data are provided in **Source Data.**

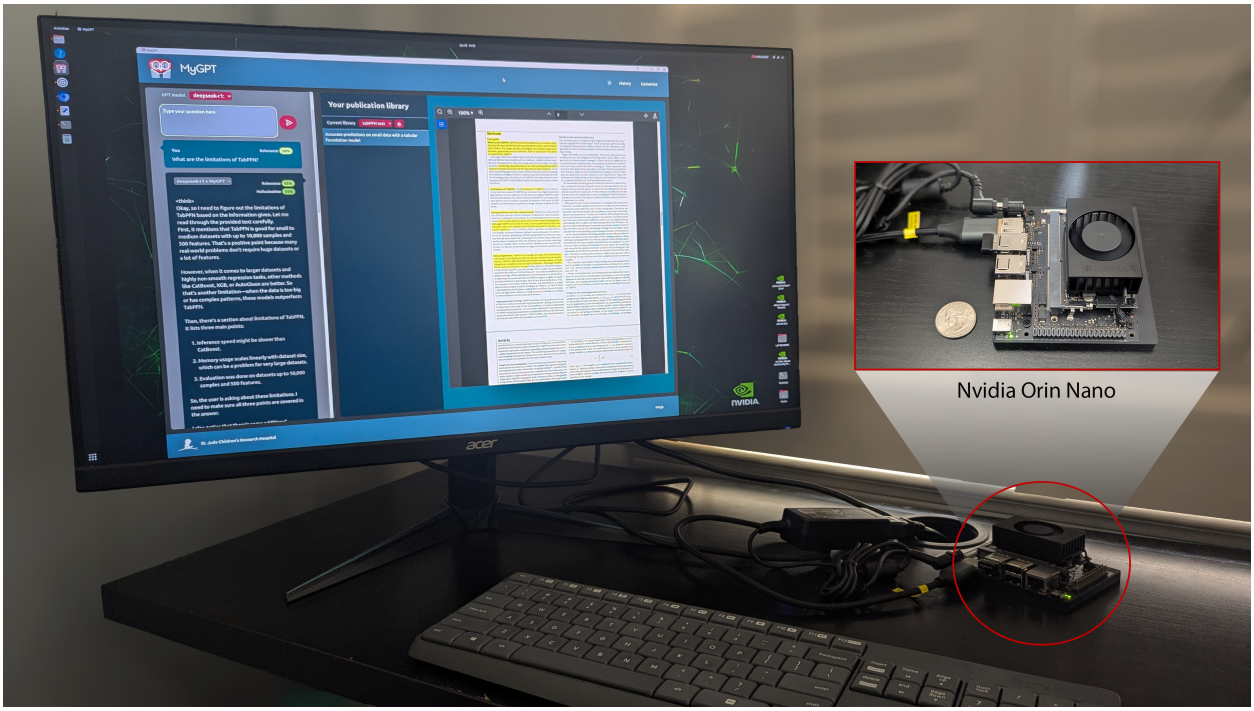


Supplementary Figure 5. Use case 2: Global health policies in multiple languages.
a Example of a question and answer for global health policies with documents in Spanish from the Policy Monitor, a global set of national cancer-related Health Policies library curated by St. Jude Global. A user's

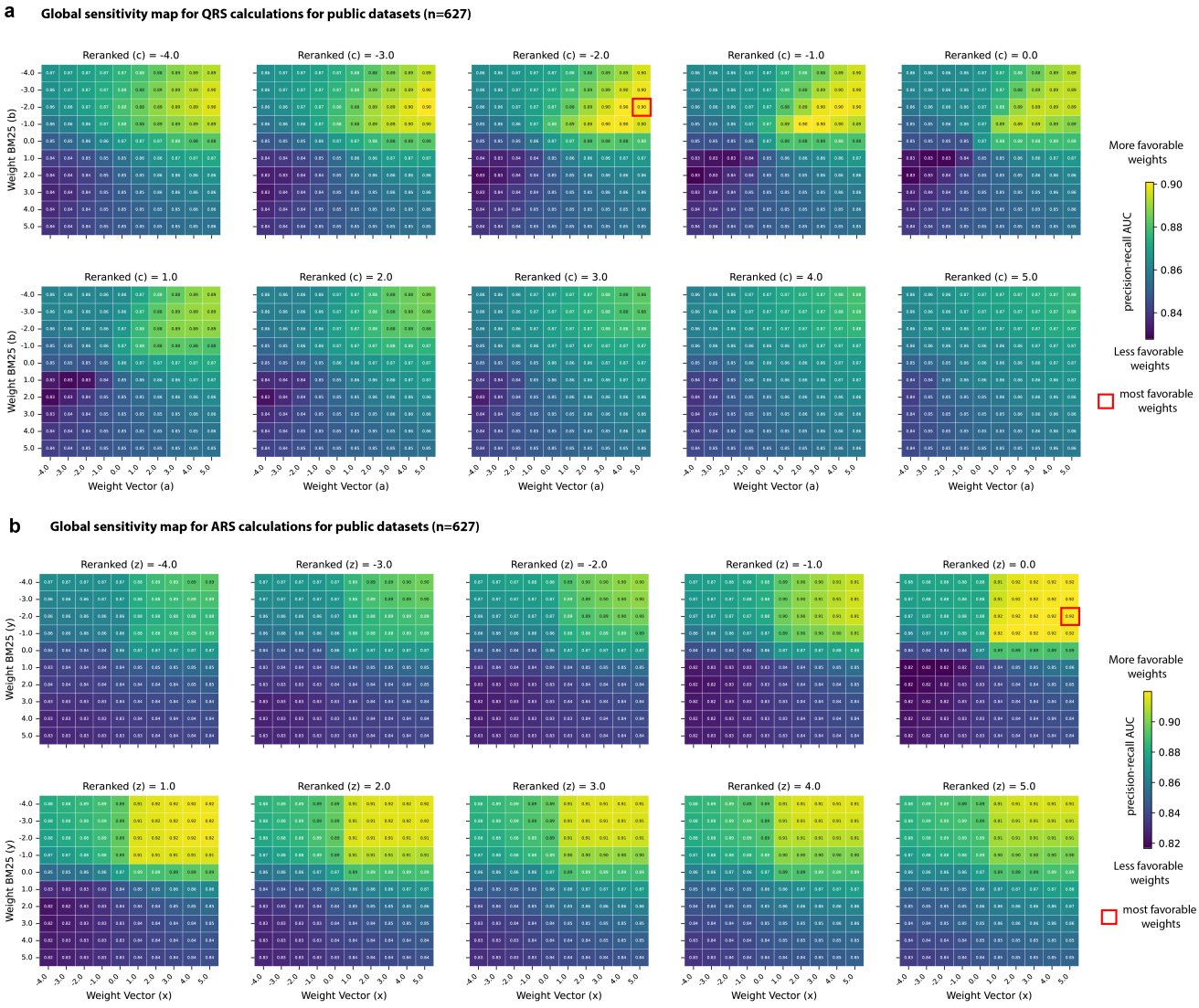
question, posed in English, was compared across various document sections using an embedding model with multilingual support (nomic-text-v2-moe:latest). The context in Spanish was utilized by the language model (GPT-oss:20b) to generate the answer in English. **b** 128 and 160 questions (with ground truth) were collected from 10 health policies in Spanish or French, respectively, providing 320 total Q&A for this evaluation. MyGPT correctly answered 128 (80%) of questions in Spanish and 141 (88.1%) of questions in French from the library. Graph data are provided in **Source Data**. **c** MyGPT uses the user-provided document language to select the appropriate embedding model and BM25 version. User questions were translated into the document's language using the TranslateGemma LLM before performing semantic and keyword searches for context retrieval. Using this context, the user-selected LLM generates a response in the user's language, and MyGPT uses translated answers from TranslateGemma to compute ARS and HI. Modifications to MyGPT's workflow for this use case are indicated in orange (backend) or green (LLM) bold font.



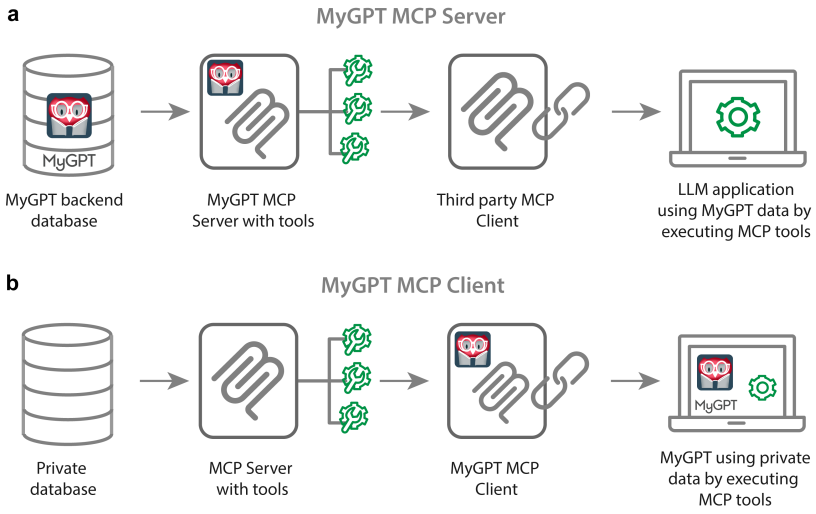
Supplementary Figure 6. Random forest model for generalizable HI predictor.
a Performance of 10 replicates of the random forest model trained and tested on 1045 Q&A from public and private datasets with an 80:20 split to predict incorrect answers. For the best-performing replicate out of 10, the ROC-AUC is 86.70% for training and 79.72% for testing. **b** The precision-recall curve of the same best-performing replicate of the model shows 60.3% for training and 50.95% for testing. The baseline of 17% represents the proportion of incorrect answers for 1045 samples, which will be the performance of a random predictor. Precision-recall of 50.95% for the random forest model signifies a 3-fold better performance than that of a random predictor. **c** SHAP summary plot of 20 features used to train and test the random forest model to predict incorrect answers. Features are ranked by mean absolute SHAP value, with positive SHAP values pushing the model toward predicting "incorrect" and negative SHAP values toward predicting "correct" answers. The high feature values are marked in pink and the low feature values in blue. **d** The HI values of all 1045 Q&A using the random forest model. The 3 public datasets combined (p = 2.2×10^{-16}; effect size = 0.38 (r, medium); Mann Whitney U test: two-sided), BMT (Bone Marrow Transplantation) SOPs (p = 1.4×10^{-4}; effect size = 0.33 (r, medium); Mann Whitney U test: two-sided) and Protein kinases (p = 8.8×10^{-16}; effect size = 0.59 (r, large); Mann Whitney U test: two-sided) datasets display significant differences in HI scores between correct and incorrect Q&A pairs with large effect size. Box-plots with individual data points: center line, mean; box, 25th-75th percentiles (IQR); whiskers, most extreme values within 1.5 x IQR (Tukey); points beyond whiskers are outliers. All graph data are provided in **Source Data**.



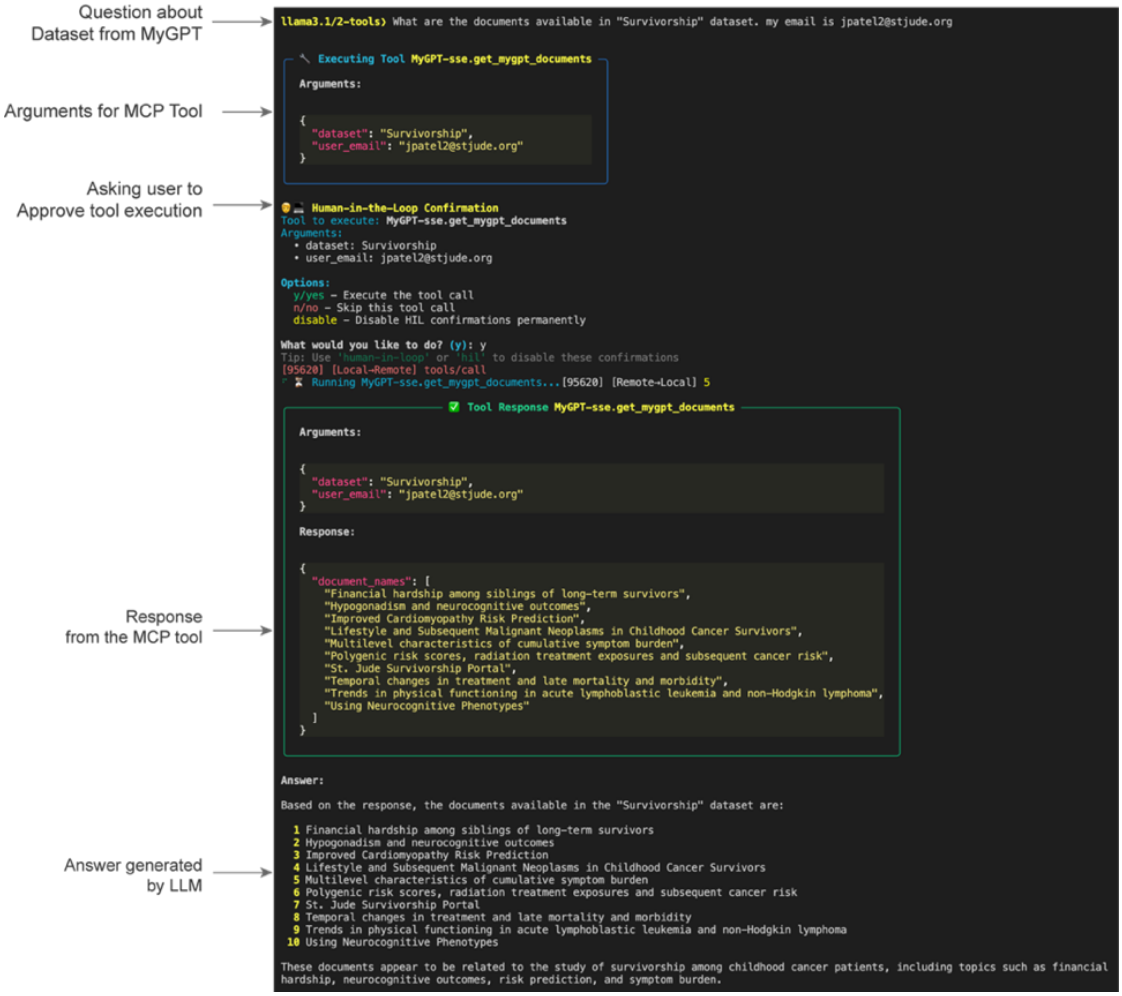
Supplementary Figure 7. MyGPT running on Nvidia Orin Nano with small-size LLMs.
Small LLMs include Llama3.2:3b, Llama3.2:1b, Gemma2:2b, and DeepSeek-R1:1.5b that are ideal for remote locations with limited or no internet access or for air-gapped deployments.



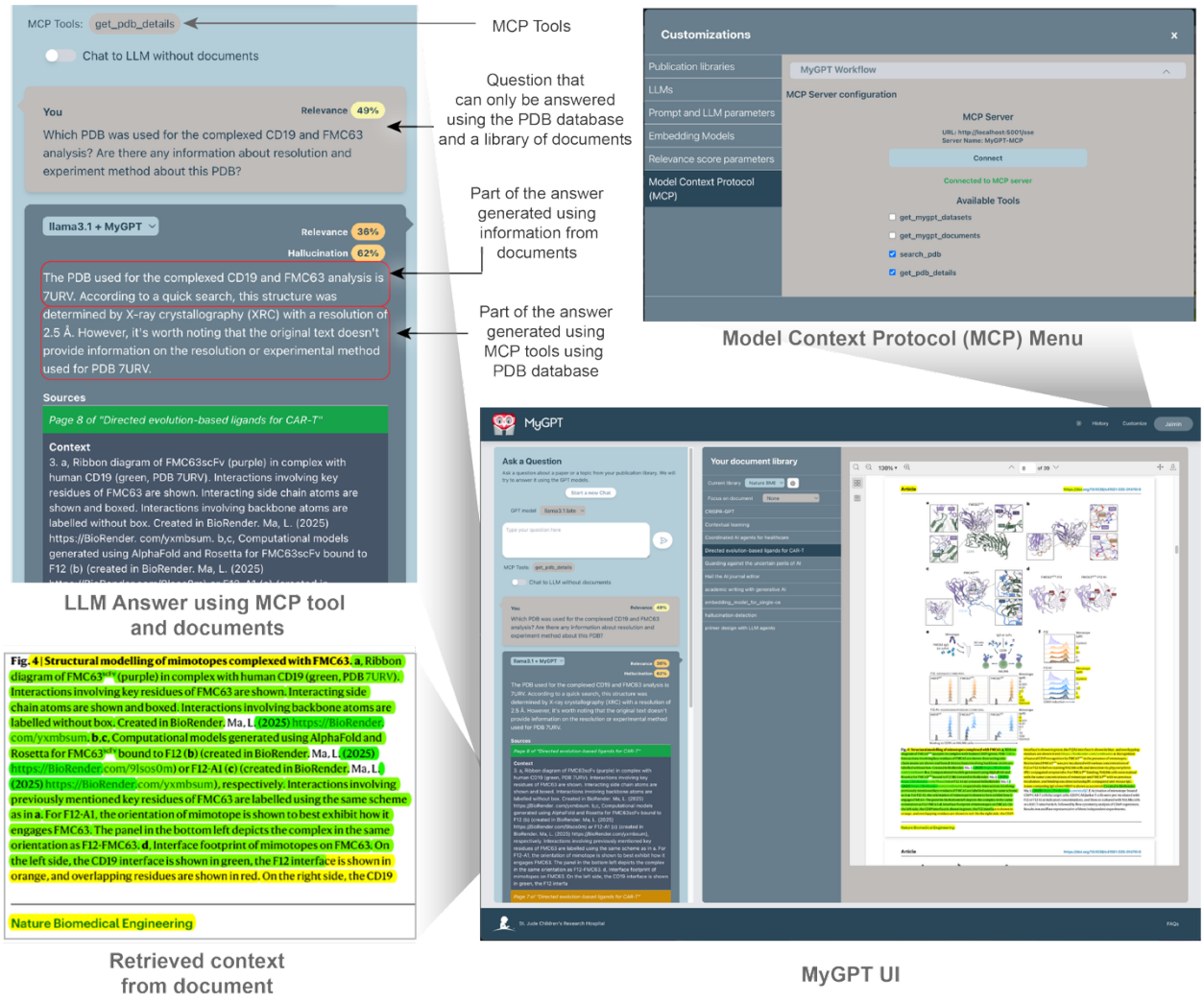
Supplementary Figure 8. Global sensitivity analysis to calculate QRS and ARS for public datasets. **a** Heatmaps showing classification performance (AUC) across combinations of weight parameters used in QRS computation for a range of vector, BM25, and reranking weight cutoffs. The weights marked in red box are the optimal weights for QRS calculations, 5 for vector, -2 for BM25 and -2 for Reranker. **b** Corresponding sensitivity maps for ARS across weight configurations for a range of vector, BM25, and reranking weight cutoffs. Color scale indicates AUC, highlighting the robustness and dependence of QRS and ARS on parameter choices. The weights marked in red box are the optimal weights for ARS calculations, 5 for vector, -2 for BM25 and 0 for Reranker.



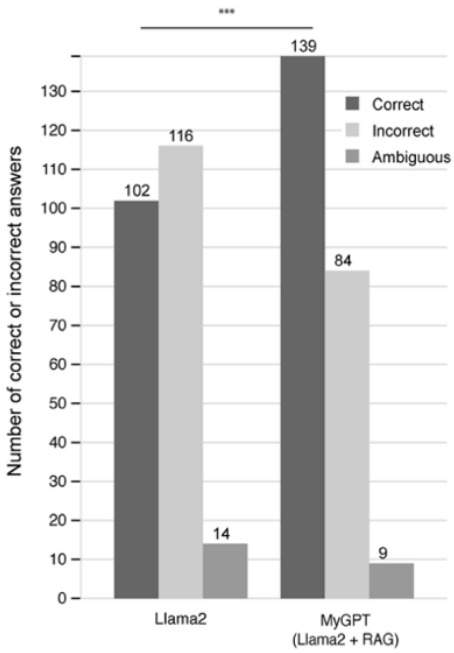
Supplementary Figure 9 MCP server and client capabilities in MyGPT.
a MCP server and **b** MCP client workflow with MyGPT.



Supplementary Figure 10. MyGPT MCP server exposing information using MCP tools.
The MCP server exposes data from the MyGPT backend to the third-party MCP Client for Ollama.

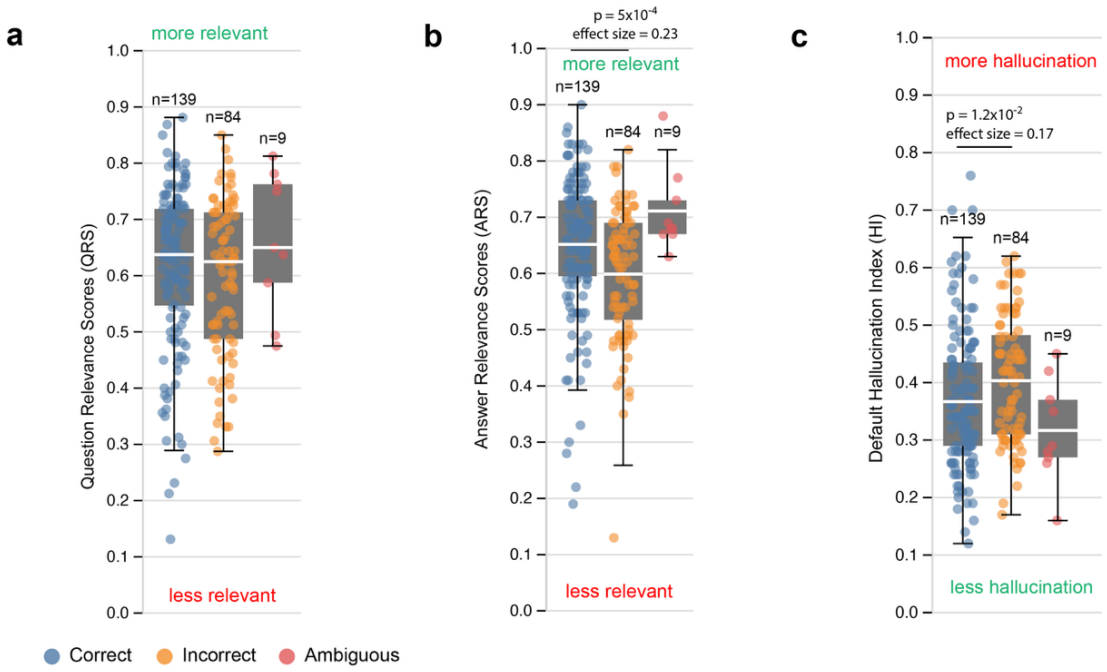


Supplementary Figure 11. MyGPT MCP Client User Interface (UI) and Customization Menu. Illustration of the MyGPT user interface showing example question and answer using the MCP tools from the MCP Client. Enlarged images display the user customization menu (top) and answer generated from the query document and the PDB database (left).



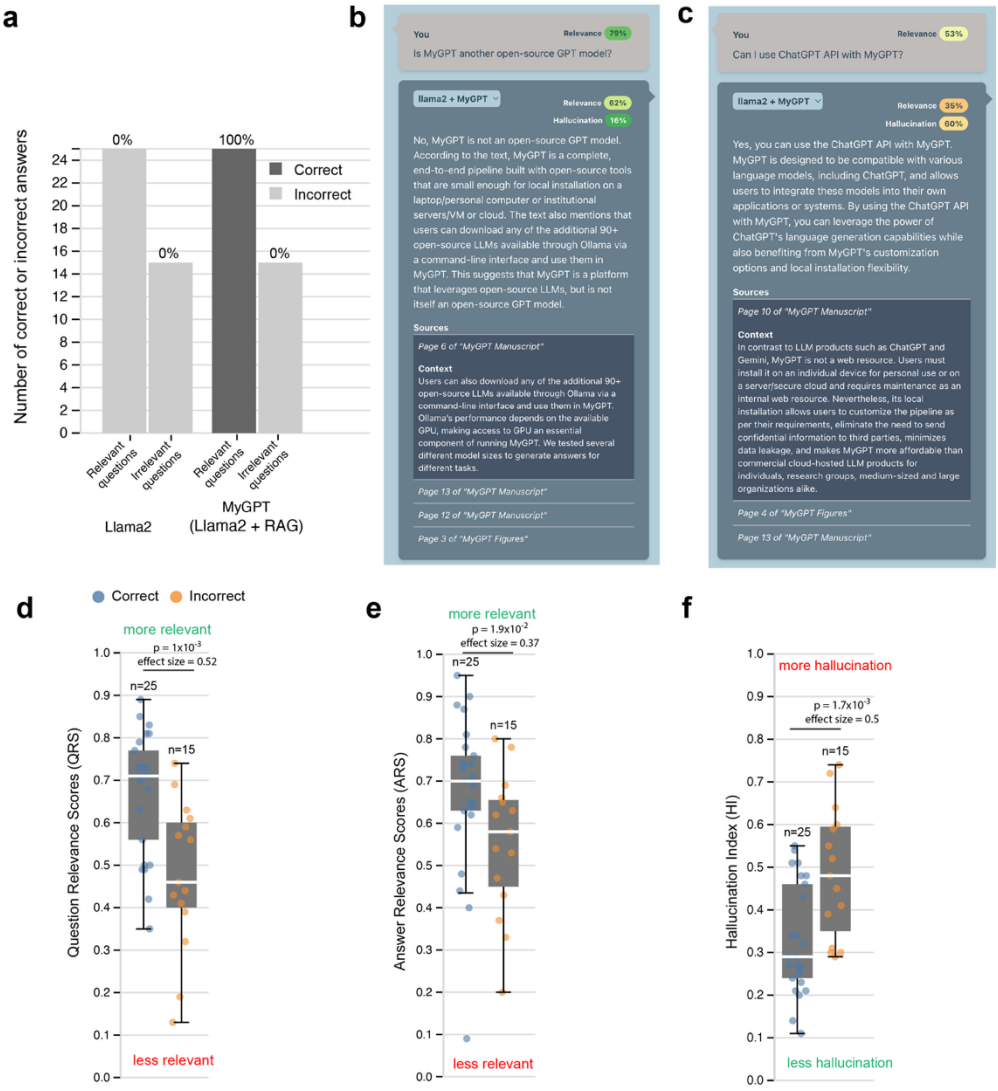
Supplementary Figure 12. Verification of MyGPT-generated answers on the dataset of public biomedical articles.

232 questions were collected from 10 domain experts from 5 publication libraries. Answers were generated using Llama2 (no context) and MyGPT (Llama2:7b + RAG). Frequency of answer correctness by standalone LLM and MyGPT. Users were blinded and asked if the generated answers were correct or not, and they provided feedback with “Yes” (dark grey), indicating the answer generated was correct, and “No” (light grey), indicating the answer generated was incorrect. Answers were considered “Ambiguous” if domain experts felt the generated answer was too vague to answer the intended question or if the answer was only partially correct. Graph data are provided in **Source Data**.



Supplementary Figure 13. QRS, ARS and HI distribution of correct vs incorrect Q&A for biomedical dataset.

a Question relevance score (QRS) distribution for correct and incorrect answers using Semantic scores (Q_{sem}). Box-plots with individual data points: center line, mean; box, 25th-75th percentiles (IQR); whiskers, most extreme values within 1.5 x IQR (Tukey); points beyond whiskers are outliers. **b** Answer relevance score (ARS) distribution using Semantic scores (A_{sem}) for correct and incorrect answers (p = 5´10^{-4}; effect size = 0.23, Wilcoxon rank-sum test). **c** Hallucination Index for correct and incorrect answers using default parameters (p = 1.2´10-2; effect size=0.17; Wilcoxon rank-sum test). Graph data are provided in **Source Data**.



Supplementary Figure 14. Verifying MyGPT (Llama2 + RAG) generated answers for a private document.

a Frequency of answer correctness with Llama2:7b alone and MyGPT (Llama2:7b + RAG). Users were asked if the generated answers were correct or not, and they provided feedback with "Yes" (dark grey), indicating the answer generated was correct, and "No" (light grey), indicating the answer generated was incorrect. The 40 questions were designed to have 25 questions where the necessary information to answer was in the private document and 15 irrelevant questions (i.e., questions lacking the necessary information in the document to answer them correctly). **b** Example of relevant question and correct answer, indicated by high QRS, ARS, and low HI. **c** Example of irrelevant questions and incorrect answers, indicated by low QRS, ARS, and high HI. **d** Question relevancy score (QRS, using Semantic scores (Q_{sem})) distribution for questions of correct and incorrect answers (p=1x10^{-3}; effect size=0.52; Wilcoxon rank-sum test). Box-plots with individual data points: center line, mean; box, 25th-75th percentiles (IQR); whiskers, most extreme values within 1.5 x IQR (Tukey); points beyond whiskers are outliers. **e** Answer relevancy score (ARS, using Semantic scores (A_{sem})) distribution for correct and incorrect answers (p=1.9x10^{-2}; effect size=0.37; Wilcoxon rank-sum test). **f** Default Hallucination Index (HI) for correct and incorrect answers using arithmetic means between QRS and ARS (p=1.7x10^{-3}; effect size=0.5; Wilcoxon rank-sum test). All graph data are provided in **Source Data**.

SUPPLEMENTARY NOTES (1-3)

Supplementary Note 1. Model Context Protocol

Model Context Protocol (MCP)[18] is an open standard developed by Anthropic that enables the integration of language model (LLM) applications with various external tools and data sources, both public and private. MCP operates on a client-server architecture. This means that information from external applications, local databases, or local files can be made available through an MCP server, and that LLMs from an AI application can, in turn, access this information using the MCP client.

MyGPT provides a basic implementation of the MCP server and client, thus facilitating integration with other private LLM applications and data sources within the same AI ecosystem. The MyGPT MCP server exposes data from the MyGPT backend using MCP tools, which any third-party MCP client can utilize. Conversely, the MyGPT MCP client enables connections to the MCP server, whether it is running locally or in a private network (**Supplementary Fig. 9**). This feature allows the client to use private database information as context in addition to information from documents.

The MyGPT server is built using a simple Python application that leverages the FastMCP[19] library, which is available as a Docker image for quick setup like all components of the MyGPT pipeline. For prototyping purposes, we have developed two MCP tools: *get_mygpt_datasets* and *get_mygpt_documents*. To demonstrate how to use these tools, we employed the private third-party *MCP Client for Ollama (ollmcp)*[20]. As shown in **Supplementary Fig. 10**, users can inquire about the available libraries and documents from MyGPT. The LLM first executes the tool and then uses the information generated to answer the user's questions.

The MyGPT MCP Client allows users to connect to either locally running or privately hosted MCP servers. Users follow instructions on the MyGPT GitHub page to establish a connection and once linked, the customization menu displays details about the server and lists the available tools. Users can select tools and compatible LLMs such as Llama3.1, Gpt-oss, or Qwen3 to prepare for real-time Q&A. For example, an MCP server was created using MCP tools to utilize the Protein Data Bank (PDB). After connecting to the server and selecting Llama3.1:8b, users posed a complex question that required data from both the document and the PDB. As shown in **Supplementary Fig. 11**, MyGPT successfully collected the PDB ID from the document, used that information to retrieve additional data from the PDB database, and finally generated an answer, illustrating its capability to integrate various data sources to enhance answers.

Supplementary Note 2. Verifying MyGPT performance on biomedical Q&A using public literature

To ensure that MyGPT works as expected, we curated a dataset consisting of questions and correct answers (ground truths) from domain experts, along with a relevant set of publications to answer the questions. To do so, we asked 10 participants to provide us with at least 20 questions from 8-56 publications in their area of expertise. We thus collected 232 questions from 5 publication libraries (see **Source Data** for full dataset) covering different scientific domains within the area of drug discovery (G protein-coupled receptors (GPRCs), intrinsically disordered regions (IDRs), Chimeric Antigen Receptor T Cells (CAR-T), protein kinases, and Post-Translational Modifications (PTMs)). This evaluation was conducted from February to April 2024, during which time Llama2:7b was the latest available open-source LLM. Of 120 articles considered for the evaluation, 66 were published before the training cut-off date of 2022 for llama2:7b, while 54 were published afterward, thus assuring that not all the questions could be answered with LLM alone and would need RAG to work. This study design ensured that (a) not all answers to questions or information on the broad topic were expected to be found within the library (e.g., given the library size of documents is small); (b) the LLM models may have been partially trained on information in these documents; and (c) the LLM models alone would be unable to answer certain questions that rely on information from articles published after their training cut-off date.

The expert-generated questions were answered by MyGPT using Llama2 with the relevant library and Llama2:7b (alone, without context). We used the all-MiniLM-L6-v2 embedding model and chunk size of 1000 characters without overlaps. The squared Euclidean distance metric was used to find the distance

between the chunks and the question and Ollama's default setting for temperate (0.8), top-P (0.9) and top-K (40) to generate the answers. Participants were blinded to the response and were asked to provide feedback on whether the individual answers generated from MyGPT and Llama2 were correct, incorrect or ambiguous. We observed a 31.7% improvement in the number of correct answers generated by Llama2 when used within the MyGPT pipeline compared to Llama2 on its own for our dataset (**Supplementary Fig. 12**).

We then analyzed the QRS, ARS, and the default HI distribution for correct and incorrect answers (**Supplementary Fig. 13**). For our dataset of 232 questions from 5 publication libraries, we observed that the QRS scores were comparable for the questions with correct and incorrect answers. This suggested that for our dataset, the library of documents was generally relevant for the kinds of questions asked. In contrast, we observed that the ARS was higher for correct answers compared to incorrect answers ($p = 5\times10^{-4}$), suggesting that correct responses tended to incorporate specific material from the library of documents while generating the response. Consistently, we also observed that the HI (default) was lower for correct compared to incorrect answers ($p = 1.2\times10^{-2}$). The relationship between QRS, ARS, and HI for each question in the dataset shows how either higher QRS, ARS, or both result in lower HI for correct compared to incorrect answers. We wish to clarify that we only used Semantic scores (Q_{sem}, A_{sem}) for QRS and ARS calculations in this exercise to validate whether the RAG concepts[21] proposed by Lewis et al. work as expected. Specifically, we excluded the Keyword and Re-rank score-based hybrid retrieval methods that are introduced in the manuscript as advanced RAG approaches.

Taken together, these observations suggest that the QRS, ARS, and HI provide useful indicators of contextual support that can help users prioritize responses for review. However, these metrics do not directly establish factual correctness and should be interpreted together with source inspection and human validation. To enhance trustworthiness in AI-powered tools, users can manually evaluate their datasets, optimize the HI scoring function, and use statistical learning approaches to leverage QRS, ARS, and HI scores to predict and classify responses as correct or incorrect. HI optimization and classification require generating a library, domain-specific questions, and labels indicating whether LLM-generated answers are correct or incorrect. By leveraging QRS, ARS, and HI, users can significantly improve the reliability and accuracy of answers generated by LLMs. We want to emphasize that correctness labels for LLM-generated answers are subjective. A user's preference for what constitutes a correct answer may result in varying ratios of correct to incorrect labels.

Supplementary Note 3 Verifying MyGPT performance on a private document

There is growing concern and evidence that LLMs are trained on public data, including research articles[22], making it challenging to determine which specific publications were included in the training datasets for the various LLMs. To circumvent this concern, we used the previous version of MyGPT manuscript (which could not have been included in the LLM training dataset) as a private document, allowing us to test how the RAG answers questions regarding a private document.

We curated a dataset of 40 questions related to the MyGPT manuscript from 2 domain experts: 25 were related to information present in the manuscript (relevant questions), and 15 were related to the topic but the manuscript lacked the information to answer them (irrelevant questions). MyGPT answered all 25 relevant questions correctly and failed to answer the 15 irrelevant questions (**Supplementary Fig. 14a-c**). In contrast, the LLM alone generated incorrect (or hallucinated) answers for all 40 questions. The confidence metrics (ARS, QRS, and HI) obtained using MyGPT could differentiate between correct and incorrect answers and provided signals of contextual support that were associated with answer correctness. (**Supplementary Fig. 14d-f**). Like the previous study, we have used only Semantic scores (Q_{sem}, A_{sem}) for QRS and ARS calculations. We observed that QRS ($p = 1\text{x}10^{-3}$) and ARS ($p = 1.9\text{x}10^{-2}$) were significantly higher for relevant questions and correct answers, and HI ($p = 1.7\text{x}10^{-3}$) were significantly higher for irrelevant questions and incorrect answers. These findings demonstrate that MyGPT (and the confidence metrics) works as expected and generates more correct answers than the standalone LLMs when using private documents.

Supplementary References

1 Microsoft Copilot Pricing Explained: Plans, Cost & Licensing (2026 Guide), https://copilot-experts.com/microsoft-copilot-pricing-guide/ (2026).
2 Copilot is a commercial, run it that way, https://redresscompliance.com/microsoft-copilot-licensing-guide-2026 (2026).
3 Microsoft to increase Office suite prices across the board starting July 2026, https://www.directionsonmicrosoft.com/microsoft-to-increase-office-suite-prices-across-the-board-starting-july-2026 (2025).
4 Google Workspace Pricing 2026 | Plans, Features & Costs, https://www.lineserve.net/blog/google-workspace-pricing-2026 (2026).
5 Google Workspace: Gemini AI Features and Pricing Changes, https://www.cumulusglobal.com/google-workspace-gemini-ai-features-and-pricing-changes/ (2025).
6 Google Gemini enterprise pricing & licensing, https://atonementlicensing.com/blog/gemini-enterprise/ (2025).
7 Google workspace pricing, https://workspace.google.com/pricing (2026).
8 ChatGPT Business pricing, https://chatgpt.com/pricing/?type=team (2026).
9 ChatGPT Pricing: My Honest Take on the 2026 Plans (Free to Pro), https://tldv.io/blog/chatgpt-pricing/ (2026).
10 Glean Review 2026: Features, Pricing, and My Honest Experience, https://fritz.ai/glean-review/ (2026).
11 Glean Pricing Explained — And Why Buyers Want More Transparency, https://www.gosearch.ai/blog/glean-pricing-explained/ (2026).
12 Vectara's offerings, https://www.vectara.com/pricing (2026).
13 Vectara Analysis: $25M Raised, https://www.extruct.ai/hub/vectara-com-funding/ (2025).
14 PrivateGPT - AWS private offer, https://aws.amazon.com/marketplace/pp/prodview-j3dxmeroxeqju (2026).
15 Kotaemon - GitHub, https://github.com/Cinnamon/kotaemon (2026).
16 Kotaemon: An Open-Source RAG-based Tool for Chatting with Your Documents, https://www.marktechpost.com/2024/09/01/kotaemon-an-open-source-rag-based-tool-for-chatting-with-your-documents/ (2024).
17 10 best RAG tools and platforms: full comparison [2026], https://www.meilisearch.com/blog/rag-tools (2026).
18 Model Context Protocol (MCP) (GitHub https://github.com/modelcontextprotocol, 2024).
19 FastMCP (GitHub https://github.com/jlowin/fastmcp, 2024).
20 MCP Client for Ollama (ollmcp) (GitHub https://github.com/jonigl/mcp-client-for-ollama, 2025).
21 Lewis, P. et al. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. ArXiv **abs/2005.11401** (2020).
22 Gibney, E. Has your paper been used to train an AI model? Almost certainly. Nature **632**, 715–716 (2024). https://doi.org//10.1038/d41586-024-02599-9