Perplexity-Inspired Metasearch-Based Alternatives to FAIR GPT: Open-source AI Consultants for Research Data Management

Thomas Schmidt [1,*], Renat Shigapov [1], Jan Kamlah [1], Irene Schumm [1]

[1] University Library, University of Mannheim, Germany
[*] Corresponding author: thomas.schmidt@uni-mannheim.de

Chatbots and virtual assistants are becoming increasingly popular for user questions and support. With FAIR GPT, the Mannheim University Library released a virtual assistant for research data management (RDM) in 2024, designed to help researchers and institutions in making their data FAIR (Findable, Accessible, Interoperable, Reusable). FAIR GPT provides various RDM services, e.g. metadata enhancement, repository selection and FAIR assessment. However, FAIR GPT has numerous disadvantages: As a 'Custom GPT' of OpenAI, it is proprietary software that only outputs sources for the generated answers if it uses its internal web search tool (which cannot be controlled by the user) and therefore lacks transparency. Reliance on external cloud-based services leads to privacy concerns when dealing with sensitive (meta)data and the chatbot is still prone to hallucinations, thus reducing its trustworthiness.

These issues led us to explore alternative open-source solutions. We searched for open-source alternatives to *Perplexity.ai*, a system known for its ability to provide citations for the information it retrieves through web searches. We identified three candidates available on GitHub: *Perplexica*, *sensei*, and *farfalle*. These tools use local instances of the metasearch engine SearXNG to perform internet searches, using the results as input for Large Language Models (LLMs). We modified these tools to focus specifically on RDM tasks, releasing the new versions on GitHub openly under the names *FAIRplexica*, *FAIR-sensei* and *FAIR-farfalle*.

Keywords: Research Data Management, LLMs, Chatbot, RDM Assistants, FAIR data

Published in: Vincent Heuveline, Philipp Kling, Florian Heuschkel, Sophie G. Habinger, and Cora F. Krömer (Hrsg.): E-Science-Tage 2025. Research Data Management: Challenges in a Changing World. Heidelberg: heiBOOKS, 2025. DOI: https://doi.org/10.11588/heibooks.1652.c23938 (CC BY-SA 4.0).

1 AI consultants for research data management

Chatbots such as ChatGPT provide intuitive access to information through graphical user interfaces and conversations in natural language. Users can ask individual questions and get answers tailored to their needs. Due to their intuitive usability, such virtual assistants can be implemented particularly well for support tasks: instead of tediously navigating through extensive documentation or FAQ collections, users receive targeted answers to individual questions, even if the Large Language Models (LLMs) on which these systems are based on have disadvantages, such as hallucinations or a lack of transparency regarding the sources on which the response is built, if they are not connected to tools that are able to perform web searches (Fecher et al. 2023).

Regarding RDM, such virtual assistants could support "level one" advisory services as defined in the RISE framework (Rans and Whyte 2017) or complement the service areas "communication and training" as defined in RISE-DE (Hartmann, Jacob, and Weiß 2019). They could answer common questions from researchers in a targeted manner: What are the FAIR principles and why should I publish my research data FAIR? What are best practices for documenting my dataset and which repository is best suited for my research? Virtual assistants could be used to answer such entry-level RDM questions more easily, quickly and flexibly, provided that the response is reliable, robust and transparent.

2 FAIR GPT

A first step towards a virtual AI assistant for RDM is FAIR GPT (Shigapov and Schumm 2024), which has been released as an OpenAI 'Custom GPT', "that combines instructions, extra knowledge, and any combination of skills" (OpenAI 2023). To reduce hallucinations and improve accuracy for certain tasks, FAIR GPT uses external APIs (Institut Français de Bioinformatique 2025; MaastrichtU-IDS 2025; TIB – Leibniz Information Centre for Science and Technology 2025; re3data.org 2013) and uploaded RDM resources (Horizon 2020 guidelines and UB Mannheim 2024a). Its functionalities include metadata enhancement, dataset organization, repository selection, FAIRness assessment, license recommendations, and generating documentation such as data management plans, README files, and codebooks.

However, FAIR GPT has limitations. It does not always provide sources for its answers (depending if it is using an internal web search tool or not), which reduces transparency and trust in its outputs. As part of OpenAI's Custom GPTs, FAIR GPT is not open source, which limits customization, and it lacks an API for integration into existing RDM workflows. Reliance on external cloud-based services leads to privacy concerns when dealing with sensitive (meta)data.

3 Open-source alternatives to FAIR GPT

To mitigate these limitations open-source alternatives to FAIR GPT were sought, with Perplexity.ai being taken as an inspiration. Like a search engine, Perplexity.ai generates answers to user questions by consulting internet sources. The sources used are referenced in the LLM-generated answers by means of footnotes. Users are thus able to scrutinize the LLM's answers for truthfulness, which emphasizes the robustness and reliability of the information and at the same time improves the disadvantages in terms of transparency that apply to FAIR GPT.

Three open-source alternatives to FAIR GPT that follow the Perplexity.ai model were found on GitHub: Perplexica (ItzCrazyKns 2025), sensei (jjleng 2024) and farfalle (rashadphz 2024). Although slight differences in the tech stack of each tool exist, the architectures are comparable:

1. A client-side chat interface provides an input for user questions.
2. Based on the user questions web searches with the open-source metasearch engine SearXNG (2025) are performed and relevant sources retrieved.
3. These sources are scraped, parsed and post-processed using LLMs (API services like OpenAI, Groq and Anthropic and local LLMs via HuggingFace or Ollama are implemented).
4. The generated answer is streamed to the client-side chat interface; Footnotes are included for sentences/text chunks that are based on the content of a specific web source; The footnotes link directly to the corresponding source.

A key point in identifying suitable FAIR GPT alternatives was the open-source nature of all its components. Perplexica as well as sensei and farfalle are published under open-source licenses and are based on the open-source metasearch engine SearXNG. In addition to the use of proprietary API providers such as OpenAI, Anthropic or Google, all tools also implement the use of local open-source LLMs, which offer major advantages in terms of data protection issues that arise when using cloud-based API providers (cf. Table 1).

After identifying and evaluating all three tools, the forks FAIRplexica (UB Mannheim 2025b), FAIR-sensei (UB Mannheim 2024b) and FAIR-farfalle (UB Mannheim 2025a) were created. For FAIR-sensei and FAIR-farfalle the changes to the code base mainly focus on updating the web search by introducing a configurable global context variable that is appended to the search query, which is then processed by SearXNG. Those changes as well as smaller fixes and UI changes are documented in the commit history of the respective repositories (cf. UB Mannheim 2024b, 2025a).

Table 1: Components and additional criteria for the assessment of FAIR GPT alternatives.

	Perplexica	**sensei**	**farfalle**
Components			
LLM API Providers	OpenAI, Anthropic, Google, Groq	OpenAI, Anthropic	OpenAI, Groq
Local LLMs	Ollama, custom OpenAI API endpoints	HuggingFace	Ollama, litellm
Web search	SearXNG	SearXNG	SearXNG
Additional criteria			
License	MIT	Apache 2.0	Apache 2.0
Last Commit	March 2025	October 2024	September 2024

4 FAIRplexica

FAIRplexica was selected for further development based on two criteria: Firstly, the original Perplexica repository was actively maintained, which led to a continuous development activity with numerous bug fixes and updates to the code base until March 2025. In contrast, the other two tools identified had not had any development updates since September (farfalle) or October 2024 (sensei). Secondly, Perplexica's MIT license offered the greatest possible openness in the further development of the software.

The adjustments made by FAIRplexica relate to these areas (cf. to the commit history of UB Mannheim 2025b for further details):

1. Restricting the web search with SearXNG to RDM-specific sources by appending a customisable global context variable to each query, ensuring domain-relevant results.
2. Implementation of an admin dashboard to make settings for LLMs and APIs in a password-protected admin area. The changes introduce a credential-based authentication with NextAuth (an open-source authentication solution for Next.js applications), a new "admin" role as well as several adaptations to the UI. These changes allow the instance to be used directly as an easy-to-use pubic chatbot without the need for a new wrapper.
3. UI customizations to the CI/CD of the research data center.

5 Conclusions

The search for open-source alternatives to FAIR GPT resulted in three tools freely available on GitHub that could be adapted for the purposes of an AI assistant for RDM. Perplexica proved to be the most promising tool and was therefore developed further under the name of FAIRplexica.

FAIRplexica eliminates numerous disadvantages of FAIR GPT: it enables the use of an AI assistant whose components are completely open-source, uses and cites specific RDM sources from the Internet to counteract hallucinations and enable the verifiability of sources and (when using local LLMs, e.g. in a university network) also prevents data protection problems.

Despite these promising results, some limitations remain. Running FAIRplexica on a server with local LLMs requires a suitable, often costly GPU. Despite these technical limitations a next step should focus on evaluating FAIRplexica to assess the quality of its generated responses. Identifying and testing suitable evaluation procedures (as well as relevant data sets) are therefore necessary tasks. Frameworks such as DeepEval (Confident AI 2025) may provide a foundation for the evaluation, but other evaluation strategies will be considered as well. We anticipate that a thorough evaluation will generate added insights and value into LLM driven assistants for research data management.

Acknowledgements

We wish to thank the anonymous peer reviewers for their valuable feedback. The project on which this report is based was partially funded by the Federal Ministry of Education and Research (BMBF) under the funding code 13IHS264B TransforMA.

Authorship Contributions

- Thomas Schmidt: Writing – original draft
- Renat Shigapov: Writing – review & editing
- Jan Kamlah: Writing – review & editing
- Irene Schumm: Writing – review & editing

Conflict of Interest

The authors are not aware of any potential conflicts of interest.

Bibliography

Confident AI. 2025. *DeepEval – The LLM Evaluation Framework.* GitHub Repository. Visited on May 26, 2025. https://github.com/confident-ai/deepeval.

Fecher, Benedikt, Marcel Hebing, Melissa Laufer, Jörg Pohle, and Fabian Sofsky. 2023. "Friend or Foe? Exploring the Implications of Large Language Models on the Science System". *arXiv,* https://doi.org/10.48550/ARXIV.2306.09928.

Hartmann, Niklas K., Boris Jacob, and Nadin Weiß. 2019. *RISE-DE – Referenzmodell für Strategieprozesse im institutionellen Forschungsdatenmanagement.* Zenodo. https://doi.org/10.5281/zenodo.3585556.

Institut Français de Bioinformatique. 2025. *FAIR-Checker: Assessing FAIR Principles of Web Resources.* Visited on March 23, 2025. https://fair-checker.france-bioinformatique.fr/.

ItzCrazyKns. 2025. *Perplexica: An AI-Powered Search Engine.* GitHub Repository. Visited on March 23, 2025. https://github.com/ItzCrazyKns/Perplexica.

jjleng. 2024. *Sensei: An Open-Source AI-Powered Answer Engine.* GitHub Repository. Visited on March 23, 2025. https://github.com/jjleng/sensei.

MaastrichtU-IDS. 2025. *FAIR Enough API: Evaluating FAIRness of Digital Resources.* Visited on March 23, 2025. https://api.fair-enough.semanticscience.org/.

OpenAI. 2023. *Introducing GPTs.* Visited on March 23, 2025. https://openai.com/index/introducing-gpts/.

Rans, Jonathan, and Angus Whyte. 2017. *Using RISE, the Research Infrastructure Self-Evaluation Framework.* Digital Curation Centre. Visited on March 23, 2025. http://www.dcc.ac.uk/sites/default/files/documents/publications/UsingRISE_v1_1.pdf.

rashadphz. 2024. *Farfalle: AI-Powered Search Engine.* GitHub Repository. Visited on March 23, 2025. https://github.com/rashadphz/farfalle.

re3data.org. 2013. *Registry of Research Data Repositories.* https://doi.org/10.17616/R3D.

SearXNG. 2025. *SearXNG: A Free Internet Metasearch Engine.* GitHub Repository. Visited on March 23, 2025. https://github.com/searxng/searxng..

Shigapov, Renat, and Irene Schumm. 2024. “FAIR GPT: A virtual consultant for research data management in ChatGPT”. *arXiv,* https://doi.org/10.48550/ARXIV.2410.07108.

TIB – Leibniz Information Centre for Science and Technology. 2025. *TIB Terminology Service: Access to Multidisciplinary Ontologies.* Visited on March 23, 2025. https://terminology.tib.eu/ts.

UB Mannheim. 2024a. *Awesome RDM: A Curated List of Research Data Management Resources.* GitHub Repository. Visited on March 23, 2025. https://github.com/UB-Mannheim/awesome-RDM.

———. 2024b. *FAIR-sensei: An Open-Source Perplexity Analogue for Research Data Management.* GitHub Repository. Visited on March 23, 2025. https://github.com/UB-Mannheim/FAIR-sensei.

———. 2025a. *FAIR-farfalle: An Open-Source AI-Powered Search Engine for Research Data Management.* GitHub Repository. Visited on March 23, 2025. https://github.com/UB-Mannheim/FAIR-farfalle.

———. 2025b. *FAIRplexica: An Open-Source AI Assistant for Research Data Management.* GitHub Repository. Visited on March 23, 2025. https://github.com/UB-Mannheim/FAIRplexica.