Open Source Large Language Models in Action: A Bioinformatics Chatbot for PRIDE database

Jingwen Bai[1], Selvakumar Kamatchinathan[1], Deepti J Kundu[1], Chakradhar Bandla[1], Juan Antonio Vizcaino[2], and Yasset Perez Riverol[1]

[1]European Bioinformatic Institute
[2]EMBL-EBI

March 12, 2024

Abstract

We here present a chatbot assistant infrastructure (https://www.ebi.ac.uk/pride/chatbot/) that simplifies user interactions with the PRIDE database, the most popular proteomics data repository. Our system utilizes two advanced Large Language Models (LLM), llama2-13b and chatglm2-6b, and includes a web service API (Application Programming Interface), web interface, and sophisticated algorithms. We have developed a novel approach to construct vector-based representations for enabling the LLM responses, featuring a curated version and a comprehensive database of relevant links and paragraphs for each generated response. An important part of the framework is a benchmark component based on an Elo-ranking system, providing a scalable method for evaluating not only the performance of llama2-13b and chatglm2-6b but also, of any other available and future open-source LLMs. Throughout the benchmarking process, the PRIDE documentation for external users was refined to enhance the clarity and efficacy in addressing user queries. Importantly, while our infrastructure is exemplified through its application in the PRIDE database context, the modular and adaptable nature of our approach positions it as a valuable tool for improving user experiences across a spectrum of bioinformatics and proteomics tools and resources, among other domains. The integration of advanced LLMs, innovative vector-based construction, the benchmarking framework, and optimized documentation collectively form a robust and transferable chatbot assistant infrastructure.

TECHNICAL BRIEF

Open Source Large Language Models in Action: A Bioinformatics Chatbot for PRIDE database

Jingwen Bai [a,&], Selvakumar Kamatchinathan[a,&], Deepti J Kundu [a], Chakradhar Bandla [a], Juan Antonio Vizcaíno [a], Yasset Perez-Riverol [a, *]

[a] European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK.

[&] These authors contributed equally to this work.

Word count: 3712

Keywords: Large language models, proteomics, documentation, training, public data, bioinformatics, software architectures.

Abstract :

We here present a chatbot assistant infrastructure (*https://www.ebi.ac.uk/pride/chatbot/*) that simplifies user interactions with the PRIDE database, the most popular proteomics data repository. Our system utilizes two advanced Large Language Models (LLM), llama2-13b and chatglm2-6b, and includes a web service

API (Application Programming Interface), web interface, and sophisticated algorithms. We have developed a novel approach to construct vector-based representations for enabling the LLM responses, featuring a curated version and a comprehensive database of relevant links and paragraphs for each generated response. An important part of the framework is a benchmark component based on an Elo-ranking system, providing a scalable method for evaluating not only the performance of llama2-13b and chatglm2-6b but also, of any other available and future open-source LLMs. Throughout the benchmarking process, the PRIDE documentation for external users was refined to enhance the clarity and efficacy in addressing user queries. Importantly, while our infrastructure is exemplified through its application in the PRIDE database context, the modular and adaptable nature of our approach positions it as a valuable tool for improving user experiences across a spectrum of bioinformatics and proteomics tools and resources, among other domains. The integration of advanced LLMs, innovative vector-based construction, the benchmarking framework, and optimized documentation collectively form a robust and transferable chatbot assistant infrastructure.

Main

One of the most important aspects of any popular bioinformatics and/or proteomics tool is its public documentation and training materials [1, 2]. As tools, frameworks, and bioinformatics resources evolve, so do the corresponding functionalities, algorithms, and the documentation accompanying them. However, creating documentation that can be used by both new and expert users poses many challenges [3]. Traditional tools in proteomics and bioinformatics predominantly rely on static formats such as Word/PDF documents, training websites, or in more advanced cases, video tutorials [2]. However, the use of more advanced methods such as chatbots, or AI (Artificial Intelligence)-powered automatic assistants remains mainly unexplored.

AI and deep learning algorithms are transforming the field of bioinformatics and proteomics [4, 5]. Large Language Models (LLMs) in particular, have been explored with success for example in functional protein design [6-8], and in *de novo* peptide identification algorithms [9]. Open-source LLMs are gaining momentum, with models such as llama2-13b [10] and chatglm2-6b (*https://huggingface.co/THUDM*) demonstrating an increased stability and scalability (search models available here:*https://huggingface.co/models*). However, these models, unlike their commercial counterparts such as ChatGPT (OpenAI), require a comprehensive framework encompassing API (Application Programming Interface) integration, web interfaces, and database construction to optimize their functionality.

The PRoteomics IDEntifications database (PRIDE -*https://www.ebi.ac.uk/pride/*) [11] is the most popular data repository for mass spectrometry-based proteomics data, and it is one of the founding members of the global ProteomeXchange consortium [12]. The PRIDE ecosystem provides software tools, services, APIs, libraries, and file formats that enable the submission, dissemination, and reuse of public proteomics data. However, the extent and complexity of PRIDE documentation can make it difficult for new and expert users to understand specific topics. This often results in time-consuming communications with the PRIDE helpdesk. To address this issue, new approaches are needed to make PRIDE documentation easier to navigate, search, and understand. This will help users to optimize their time and enable PRIDE staff to focus on more complex issues.

Here, we present an infrastructure featuring a web service API, a user-friendly web interface, and specialized algorithms for the open-source LLMs llama2-13b and chatglm2-6b. Our approach involves the creation of vector-based representations for LLM responses, the integration of enhanced documentation and the construction of a comprehensive database. We also evaluated the LLMs using an Elo-ranking benchmarking system [13], triggering the refinement of the PRIDE documentation. While deployed on the PRIDE infrastructure (*https://www.ebi.ac.uk/pride/chatbot/*), our open-source framework is adaptable to other bioinformatics resources and omics domains, combining advanced LLMs, innovative vector construction, a benchmarking framework, and optimized documentation into a robust and versatile chatbot assistant infrastructure.

In this manuscript, first of all, we will describe the developed infrastructure and some of the technical details related to the implementation and deployment of the chatbot, by using two LLMs: llama2-13b and chatglm2-6b. Second, we will describe the benchmarking framework developed for assessing the performance of the

LLMs, and the results of our benchmarking. Third, we will describe the functionality of the chatbot web interface, and finally, we will highlight some conclusions and future perspectives.

Commercial and open-source LLMs frequently fail to provide accurate responses about highly specialized subdomains/topics due to the generalist nature of the content of the databases that are used for training them. This limitation in target knowledge can be resolved by integrating private knowledge (e.g. documentation files) with LLMs [14]. A pragmatic strategy entails extracting relevant knowledge as fragments from training/documentation files, establishing an index database for enabling user queries, and integrating this acquired knowledge with the existing knowledge base of the model (**Figure 1**). Next, we describe in detail the developed PRIDE chatbot infrastructure and some technical details involving the two open source LLMs.

Knowledge database indexes creation: To create the vector databases from Markdown-formatted private documents (the PRIDE documentation in this concrete case), the following steps were followed (Figure 1). The documents from PRIDE's user documentation were segmented, converted into vectors, and stored in two twin vector databases. The first database contains text chunks without Markdown formatting, enabling similarity searches for user queries. The second database aligns with the knowledge segments identified in the first database, allowing users to navigate through the PRIDE documentation. By using this approach, chatbot users can ask questions against the cleaned segmented database, and at the same time get access to the original links and paragraphs from where the information was extracted. We could name the way to create the vector database based on PRIDE documents as vectorization or tokenization. The terms come from the concept of embedding which is defined as a method for mapping discrete entities (e.g. words, images) into a dense, high-dimensional feature space, effectively transforming them into continuous vectors [15, 16].

Database Management Capability: A new database management component allows developers and maintainers of the chatbot to manage (add, delete and update) the content of the two vector knowledge databases (**Figure 1**). The addition function enables the upload of new files, integrating them into the vector database, whereas the delete function enables the removal of specific files from both databases. The viewing function provides content visualization and document segmentation structure. Users can inspect all uploaded files and the segmentation in a tree-diagram format.

Chatbot component : As shown in step three (**Figure 1**), upon the user's query input, the query is first transformed into a vector using an embedding model, followed by performing a similarity search within the database. The top three vectors, in terms of similarity, are converted back into text and combined with the user's query in a prompt template, before being processed by the LLM to produce a response.

Currently, the most popular prompt frameworks of LLMs consist of the following elements (*https://github.com/mattnigh/ChatGPT3-Free-Prompt-List)*: (i) Capacity and Role (CR): this element defines the desired role one wants the chatbot to assume; (ii) Insight (I): this involves background information and context relevant to the interaction; (iii) Statement (S): this specifies what one expects the chatbot to accomplish; and (iv) Personality (P): this dictates the style or manner in which one wants the chatbot to respond. Based on the above frameworks, we summarized the most proper prompt template for PRIDE documentation, using llama2-13b and chatglm2-6b (**Box 1**).

More specifically, each pair of *<s>* and*</s>* encapsulates a complete round of dialogue. *<s>* is used to indicate the beginning of a sentence or input, and </s> signifies the end. This helps the model to understand where an input starts and stops. *[INST]* is placed at the beginning of the instruction for LLM, and *[/INST]* is used to mark the end. The content between *[INST]* and *[/INST]* has two parts: system prompt and user prompt. The system prompt is encapsulated by*<<SYS>>* and*<</SYS>>* tags, signifying the configuration of the model's capabilities and roles, and indicating the preferred style of response, as mentioned earlier, comprising CR and P. The user prompt represents the questions provided by the users, which should be placed between*</SYS>* and*</INST>* [*https://llama-2.ai/prompting-llama-2/*].

As shown in Box 1, we have provided the LLMs with an identity and set of the tasks and requirements to be accomplished. In this context, we have inserted the relevant chunk retrieved from the vector database, and in the question part, the user query is included. The open-source llama2-13b and chatglm2-6b models will

generate an output based on the above two prompt templates. In addition to the LLM's output, the system also returns the original, Markdown-formatted text from the alternate database for user reference.

Tracing and monitoring: To assess the performance of the model, each user input (query), the content extracted from the knowledge base, and the model's inference results (single or multi-turn) are stored in a separate database. This database, distinct from the vector knowledge databases, is an SQL database dedicated to storing plain text. It is exclusively used for the backend data collection needed for user model evaluation and does not grant management access to users.

Document processing algorithm: Extracted document fragments are segmented using \n#, \n##, and \n### as delimiters, creating document fragments that are duplicated and stripped of Markdown markers. These segments are then input into vector databases managed by ChromaDB (*https://www.trychroma.com/*). Each database entry includes the document file path, a unique ID, and a URL to the original document in the PRIDE help pages.

LLMs Selection: The primary function of the chatbot is to answer user queries based on sections that are split from private documents. This task imposes several requirements on LLMs:

- (i) The LLM should be deployable in local infrastructures to ensure the security of the knowledge base documents.
- (ii) The LLM should be able to process long input texts, due to the varying lengths of knowledge fragments retrieved through similarity searches from private documents.
- (iii) The LLM should balance performance and cost.

Considering the requirements of our local deployment environment, we chose two specific LLMs: llama2-13b and chatglm2-6b. llama2-13b is an open-source predictive model trained with 2T tokens, featuring variants with up to 70 billion (B) parameters. llama2-13b was selected as the optimal LLM due to better performance than llama2-7B and lower GPU costs than llama2-70B. chatglm2-6b is a low-resource open-source predictive model developed in China with excellent performance. The latest version, chatglm2-6b, has enhanced context length capabilities of up to 32 K and requires minimal VRAM, making it suitable for consumer-level hardware.

The proposed framework allows for the deployment, testing and benchmarking of other available and future LLMs. Before benchmarking llama2-13b and chatglm2-6b, we tested two additional open-source LLMs: the GPT4ALL (https://github.com/nomic-ai/gpt4all) and mpt-7b-chat (https://huggingface.co/mosaicml/mpt-7b-chat) models. However, they were excluded from the benchmark due to poor performance on simple responses (data not shown). The chatglm2-6b model and its corresponding parameters were obtained from HuggingFace, whereas the llama2-13b parameters were obtained from Meta's official website.

We next describe the developed benchmark framework for evaluating the responses of the chatbot. We used the Elo-ranking system as a benchmark to evaluate the performance of LLMs [13, 17]. The Elo-ranking system was originally designed to rank chess players based on head-to-head comparisons. This method has been previously used to evaluate and benchmark LLM methods (*https://arena.lmsys.org/*). We leveraged this system to compare the llama2-13b and chatglm2-6b models by providing questions to them and ranking them based on their responses. We conducted three rounds (iterations) of comparisons, and three PRIDE staff members were involved in the process. The summary of the results is indicated in **Table 1** . Using a database, all responses were traced, and a JSON API endpoint was enabled to retrieve both the given response and the PRIDE staff evaluations (**Supplementary Note 1**).

We selected 31 frequently asked questions to the PRIDE support team for each iteration. These questions covered different levels of complexity, being targeted to different types of users, including both new and more advanced PRIDE users (**Supplementary Note 2**). During the first evaluation, both models failed to produce correct responses to an 11% of the questions that were evaluated. After carefully improving the documentation (by including topics that could be used to respond to the provided questions and by also removing old, outdated documents), the percentage of incorrect responses decreased from 11% to 1.8%.

Interestingly, the llama2-13b model in the first iteration achieved an Elo of 1206, compared to 794 for

chatglm2-6b (a difference of 412 points). However, after carefully improving the documentation and adjusting the model templates, the difference between the two models was reduced to 90 Elo points, after the third iteration (Table 1). This indicates that with improvements of the documentation and adjustments to the model templates, the model's performance can be significantly improved.

Web interface : The developed framework provides a chatbot assistant interface as the main entry point for the application (*https://www.ebi.ac.uk/pride/chatbot/*). Users can select the desired LLM and pose questions in the input box. After the similar results obtained after the last iteration of the benchmark, we decided to enable both LLMs, but the default model is llama2-13b. The program searches the database for relevant documents based on the input and feeds them into the LLM to generate responses. A hyperlink in the lower right corner of the model's response provides access to several documents from the database involved in the Q&A session (**Figure 2**).

In addition, the framework offers several other components to manage the knowledge databases (**Supplementary Note 3**). The administration visualization component allows users to display the graph structure of the documentation including the visualization of each section for each documentation topic as structured in the markdown. Users can also manually administer the content of the knowledge database by adding, deleting or updating every document used to construct the database (**Supplementary Note 3 - Figure 1**). The proposed framework was deployed at the European Bioinformatics Institute (EMBL-EBI) cluster using the GPU capabilities. The GPU node capability has 48 cores, 100 GB of memory and an A100 GPU - NVIDIA.

The chatbot assistant framework holds significant potential for future applications in proteomics and bioinformatics training. By integrating advanced LLMs, vector-based construction, benchmarking, and optimized documentation, the framework enhances user experiences with bioinformatics resources. The integration of open-source LLMs, such as llama2-13b and chatglm2-6b, provides a flexible and adaptable solution, and the benchmarking framework creates a basis for assessing and implementing future LLMs.

In this work, we introduce a novel chatbot assistant infrastructure using open-source LLMs for bioinformatics and proteomics training and support deployed on top of the PRIDE database infrastructure. The system employs the llama2-13b and chatglm2-6b LLMs, accompanied by a web service API, web interface, and sophisticated algorithms. The PRIDE chatbot incorporates a unique approach to construct vector-based representations for LLM responses, enhancing user interactions with bioinformatics resources. LLMs are susceptible to a well-known issue known as the hallucinations [18], wherein the models generate inaccurate or misleading responses. However, by benchmarking, and refining the PRIDE documentation, we reduced the number of wrong and hallucinating responses from the chatbot. Although we aimed to evaluate and benchmark the LLMs more analytically, using metrics such as word error rate and semantic metrics, we ultimately relied on a simple human evaluation framework based on Elo ranking to identify hallucinations and bad responses. We plan to further improve our framework in the future by measuring the accuracy of each model using other quantification methods.

The developed framework, deployed at EMBL-EBI, can be potentially used by diverse bioinformatics resources from different domains, offering an adaptable and robust chatbot assistant infrastructure beyond the PRIDE database. The proposed solution could be adapted to other proteomics tools and services such as search engines, workflow environments or proteomics tools and Consortia such as Mascot, MaxQuant [19], quantms [20], HUPO-PSI [21] or ProteomeXchange [12]. Additionally, we have refined the PRIDE documentation, which has led to an improvement in the performance of the models. In the near future, we plan to work with other bioinformatics resources at EMBL-EBI to deploy training and documentation materials.

Acknowledgements

The authors of the manuscript acknowledge funding from Wellcome [grant number 223745/Z/21/Z], BBSRC grants 'PTMeXchange' [BB/S01781X/1], 'GRAPPA' [BB/T019670/1]; '3D-Proteomics' [BB/V018779/1], and 'DIA-eXchange' [BB/X001911/1] and EMBL core funding.

References

[1] Karimzadeh, M., Hoffman, M. M., Top considerations for creating bioinformatics software documentation. *Brief Bioinform* 2018,*19* , 693-699.

[2] Perez-Riverol, Y., Wang, R., Hermjakob, H., Muller, M.*, et al.* , Open source libraries and frameworks for mass spectrometry based proteomics: a developer's perspective. *Biochim Biophys Acta* 2014,*1844* , 63-76.

[3] Williams, J. J., Teal, T. K., A vision for collaborative training infrastructure for bioinformatics. *Ann N Y Acad Sci*2017, *1387* , 54-60.

[4] Qin, C., Luo, X., Deng, C., Shu, K.*, et al.* , Deep learning embedder method and tool for mass spectra similarity search.*J Proteomics* 2021, *232* , 104070.

[5] Rehfeldt, T., Gabriels, R., Bouwmeester, R., Gessulat, S.*, et al.* , ProteomicsML: An Online Platform for Community-Curated Datasets and Tutorials for Machine Learning in Proteomics. 2022.

[6] Ferruz, N., Schmidt, S., Hocker, B., ProtGPT2 is a deep unsupervised language model for protein design. *Nat Commun* 2022,*13* , 4348.

[7] Madani, A., Krause, B., Greene, E. R., Subramanian, S.*, et al.* , Large language models generate functional protein sequences across diverse families. *Nat Biotechnol* 2023, *41* , 1099-1106.

[8] Huang, T., Li, Y., Current progress, challenges, and future perspectives of language models for protein representation and protein design. *Innovation (Camb)* 2023, *4* , 100446.

[9] Yilmaz, M., Fondrie, W. E., Bittremieux, W., Nelson, R.*, et al.* , Sequence-to-sequence translation from mass spectra to peptides with a transformer model. *Biorxiv* 2023, 2023.2001. 2003.522621.

[10] Touvron, H., Martin, L., Stone, K., Albert, P.*, et al.* , Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288* 2023.

[11] Perez-Riverol, Y., Bai, J., Bandla, C., Garcia-Seisdedos, D.*, et al.* , The PRIDE database resources in 2022: a hub for mass spectrometry-based proteomics evidences. *Nucleic Acids Res* 2022,*50* , D543-D552.

[12] Deutsch, E. W., Bandeira, N., Perez-Riverol, Y., Sharma, V.*, et al.* , The ProteomeXchange consortium at 10 years: 2023 update. *Nucleic Acids Res* 2023, *51* , D1539-D1548.

[13] Li, R., Patel, T., Du, X., Prd: Peer rank and discussion improve large language model based evaluations. *arXiv preprint arXiv:2307.02762* 2023.

[14] Song, C. H., Wu, J., Washington, C., Sadler, B. M.*, et al.* , *Proceedings of the IEEE/CVF International Conference on Computer Vision* 2023, pp. 2998-3009.

[15] Cellucci, C. J., Albano, A. M., Rapp, P., Comparative study of embedding methods. *Physical Review E* 2003, *67* , 066210.

[16] Mikolov, T., Chen, K., Corrado, G., Dean, J., Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* 2013.

[17] Dettmers, T., Pagnoni, A., Holtzman, A., Zettlemoyer, L., Qlora: Efficient finetuning of quantized llms. *arXiv preprint arXiv:2305.14314* 2023.

[18] Wang, J., Zhou, Y., Xu, G., Shi, P.*, et al.* , Evaluation and analysis of hallucination in large vision-language models.*arXiv preprint arXiv:2308.15126* 2023.

[19] Cox, J., Mann, M., MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat Biotechnol* 2008, *26* , 1367-1372.

[20] Dai, C., Pfeuffer, J., Wang, H., Sachsenberg, T.*, et al.* , quantms: A cloud-based pipeline for proteomics reanalysis enables the quantification of 17521 proteins in 9,502 human samples. 2023.

[21] Deutsch, E. W., Vizcaino, J. A., Jones, A. R., Binz, P. A., *et al.* , Proteomics Standards Initiative at Twenty Years: Current Activities and Future Work. *J Proteome Res* 2023,*22* , 287-301.

Figure 1: Graphical representation of the implementation of the PRIDE chatbot assistant framework. **(1)** The first component involves the creation of a vector/index database, utilizing the PRIDE documentation in Markdown format. This process results in the generation of two knowledge databases: one preserves the text without Markdown formatting, facilitates similarity searches with user queries, and supplies the identified segments to the LLM. The second database aligns with the knowledge segments identified in the first, presenting their original documents and the links to the PRIDE documentation.**(2)** The framework incorporates a dedicated component to manage the vector database, designed for groups without a bioinformatics background. This component allows for the deletion of existing databases, the addition of new documentation files, and the visualization of the document database structure. **(3)** The chatbot interface facilitates the interaction between the users and the LLM and includes the storage of the user queries. This stored information contributes to the accumulation of questions and answers provided to the model and can be used to enhance the existing documentation.

Box 1 : Prompt templates for llama2-13b and chatglm2-6b. llama2-13b uses special tags such as <s>, </s>, <<SYS>>, <</SYS>>, [INST], and [/INST], which are commonly used in the context of machine learning and are distinct markers indicating various parts in the llama2-13b prompt template. The chatglm2-6b template is not like the llama2-13b one It is simpler and is composed only by the context part and the question part. There are no special tags to indicate different parts or contents in the template. We only need to describe what the content is, which is enough for chatglm2-6b to understand.

Table 1 : Summary of the results of the benchmark iterations using the Elo ranking method. During each iteration, we focused on improving the PRIDE documentation in the following ways: including more information to address PRIDE staff and users' questions, removing outdated concepts and deprecated functionalities, and formatting data to enhance vectorization and improve the final LLMs responses.

Model	Iteration 1 Elo	Iteration 1 % wrong responses	Iteration 2 Elo	Iteration 2 % wrong responses	Iteration 3 Elo	Iteration 3 % wrong responses
llama2-13b	1206	11%	1168	5.4%	1045	1.8%
chatglm2-6b	794		832		955	

Figure 2 : Screenshot of the PRIDE chatbot web interface. Users can choose one of the two models llama2-13b or chatglm2-6b. Each time a user asks a question, the chatbot response includes links or sections from the documentation that were used to create the final response. This helps users to find the relevant information they seek more efficiently.