Publishing a Chatbot: Opportunities and Challenges

Thomas Asselborn[1,2,*], Magnus Bender[3,*], Ralf Möller[1] and Sylvia Melzer[1,2]

[1] University of Hamburg, Institute of Humanities-Centered Artificial Intelligence (CHAI), Warburgstraße 28, 20354 Hamburg, Germany

[2] University of Hamburg, Centre for the Study of Manuscript Cultures (CSMC), Warburgstraße 26, 20354 Hamburg, Germany

[3] Aarhus University, Department of Management, Fuglesangs Allé 4, 8210 Aarhus V, Denmark

Abstract

While developing a prototype chatbot has become increasingly accessible thanks to modern tools and tutorials, publishing a chatbot, especially one that is freely available without login requirements, introduces a distinct set of challenges. This article reflects on the process of preparing ChatHA (Humanities-Aligned Chatbot) for public release, particularly in the context of the Written Artefact Profiling Guide. Beyond technical hurdles, such as infrastructure and deployment, we encountered significant ethical and legal considerations, including data privacy, user consent, and responsible communication. Although our work is rooted in humanities research in Germany, many of the issues discussed are relevant across disciplines and jurisdictions. By outlining both the opportunities and challenges involved, this article aims to contribute to a broader discussion on the thoughtful publication of chatbots for research dissemination.

Keywords

Chatbots, Large Language Models, Science Communication, EU AI-Act, EU GDPR

1. Introduction

Developing a prototype of a chatbot and publishing it are two different challenges. While the development is a topic discussed already, like in the case of our chatbot ChatHA (Humanities-Aligned Chatbot) [1], with many tutorials available[1] The publication of a chatbot, especially one that is also freely available without the need to log in, is done at a few places, but the benefits and issues in doing so are rarely discussed publicly.

The goal of this article is to discuss the opportunities and challenges we identified during the planning of publishing ChatHA in its first iteration with the Written Artefact Profiling Guide[2] in mind. While this work focuses primarily on chatbots for the humanities, many of the aspects raised in this article are important for chatbots in other disciplines as well.

This article is structured as follows. Section 2 discusses the opportunities and reasons why we are planning to publish the chatbot. In Section 3, issues regarding the technology, both hardware and software, and a few options for mitigating them are presented. Section 4 discusses challenges in terms of ethics and morals, while Section 5 briefly shows a few legal challenges that may arise when making a chatbot public. Finally, Section 6 concludes this article.

Perspectives on Humanities-Centred AI and Formal & Cognitive Reasoning Workshop 2025,(CHAI 2025 & FCR 2025), Joint Workshop at the 48th German Conference on Artificial Intelligence, September 16, 2025, Potsdam, Germany

*Corresponding authors.

thomas.asselborn@uni-hamburg.de (T. Asselborn); magnus@mgmt.au.dk (M. Bender); ralf.moeller@uni-hamburg.de (R. Möller); sylvia.melzer@uni-hamburg.de (S. Melzer)

https://www.chai.uni-hamburg.de/~asselborn (T. Asselborn); https://person.au.dk/magnus@mgmt.au.dk (M. Bender); https://www.chai.uni-hamburg.de/~moeller (R. Möller); https://www.chai.uni-hamburg.de/~melzer (S. Melzer)

0009-0005-3011-7626 (T. Asselborn); 0000-0002-1854-225X (M. Bender); 0000-0002-1174-3323 (R. Möller); 0000-0002-0144-5429 (S. Melzer)

[1] A few examples are: https://medium.com/@james.irving.phd/creating-your-personal-chatbot-using-hugging-face-spaces-and-streamlit-596a54b9e3ed and https://python.langchain.com/docs/tutorials/chatbot/. Many more are available in text form or on YouTube.

[2] https://www.csmc.uni-hamburg.de/profiling-guide/

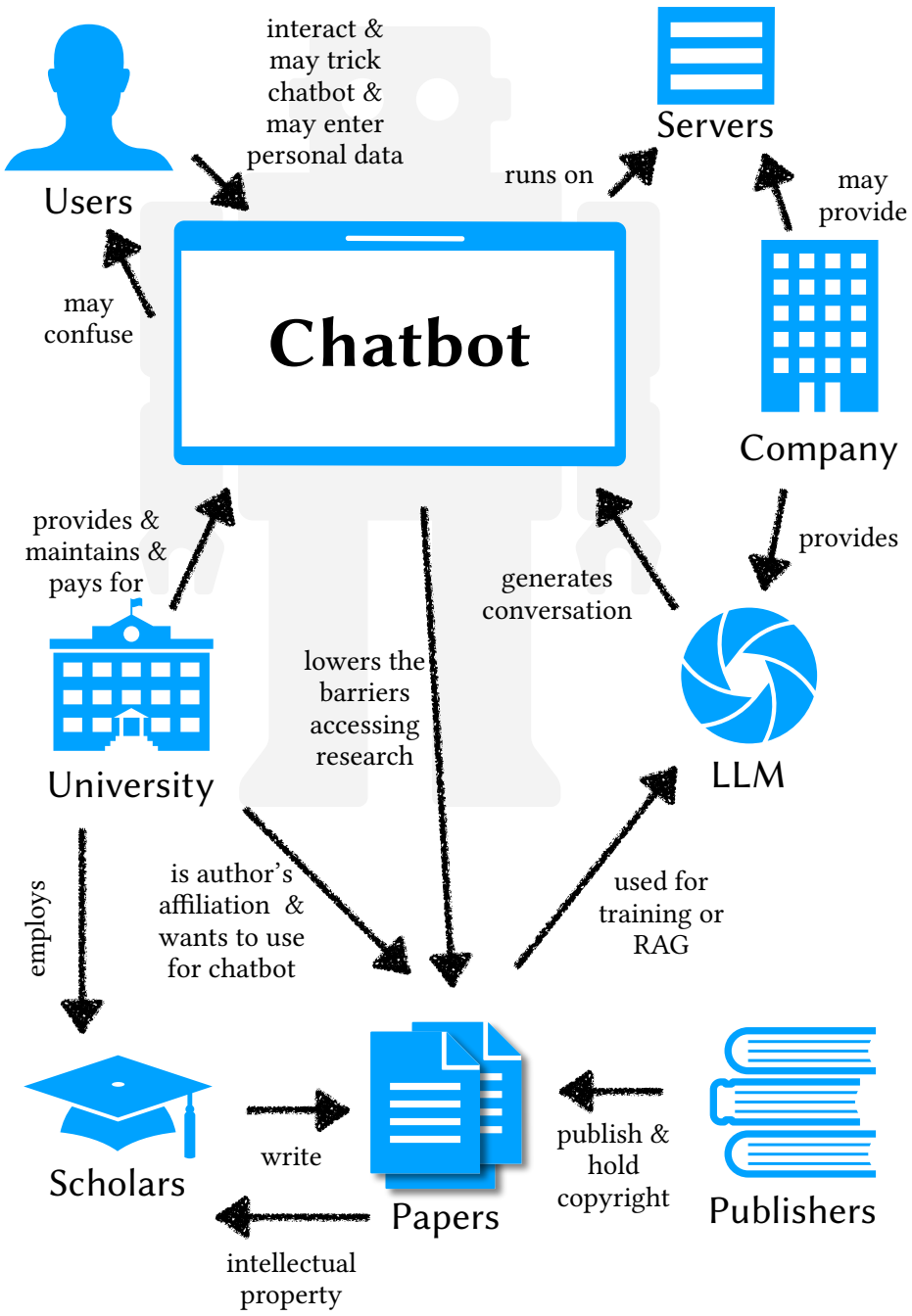

Figure 1: Visualisation of the different parties involved when publishing a chatbot. The arrows depict some of the interests and relations among them.

2. Opportunities for Publishing a Chatbot

Publishing a chatbot can bring several benefits depending on the context. A chatbot can be one option for bringing science and scientific articles closer to the average user, as they are difficult to understand and getting harder over time [2, 3]. Even summaries of articles that are designed to be read by more people than just the scientific community of a specific field are not always easy to understand by most people [4].

The discrepancy between the way scientists write and speak and people outside of the scientific community leads to people not understanding why science is important. People feel like the work scientists do is not benefiting their lives and thus reject the work of scientists. Additionally, they

may shift their focus to work that presents as scientific but is not because their articles are easier to understand. [5, 6]

There is not only a gap between the languages of scientists and people outside of the scientific community, but also between scientists of different fields. This is also true for scientists in fields that may be related or where mutual understanding may be beneficial. [7]

Most scientific articles are written in the English language, but depending on the field, they are written in German, Arabic, French, or in another language. While most scientists globally can probably read and write English, we have noticed that a few scientists in the humanities only write in the language related to their field. This language barrier excludes scientists from related fields who are unable to understand the language in which the text is written.

Chatbots, especially those based on transformer-based **L**arge **L**anguage **M**odels (LLMs), can be one aspect in bridging those gaps. LLMs can not only translate between different languages but also between different language levels [8]. We have also evaluated that previously in the field of artefact profiling with ChatHA [9].

3. Technological Challenges

The first block of challenges focuses on technical challenges we identified and faced. In order to provide a chatbot that is reliably available and has good performance, some measures need to be taken in terms of hardware and software.

3.1. Hardware

Although it is in principle possible to run a chatbot using a quantised LLM on a standard **C**entral **P**rocessing **U**nit (CPU), the process is generally more computationally time-consuming compared to execution on a **G**raphics **P**rocessing **U**nit (GPU). Additionally, once the chatbot has been published, it should be designed to be used by multiple people simultaneously without the system running out of processing power. Thus, dedicated GPUs like the Nvidia DGX L40S[3] need to be installed to support the potential load on the system and to provide a smooth experience to the user without long waiting times. Once a specific GPU or cluster of GPUs was selected, the next question that arises is how many GPUs are really necessary for the system to run stably. An assessment needs to be done to estimate the number of potential users and, from that, estimate the number of GPUs to support this load. This consideration, in turn, raises more aspects that need to be answered, like the maximum load, average load, user waiting times, and among other aspects. Once all hardware requirements have been identified, the next issue that is particularly relevant nowadays is the limited availability of certain parts. The global chip shortage, as a result of COVID-19 and multiple other global issues since 2020, has led to a scarcity of electronic goods. This was especially noticeable for higher-end GPUs.[10] Even until this day, certain higher-end GPUs are not easy to get[4] and thus the developer and publisher of a chatbot either has to wait or use lower-end parts, which may degrade performance. There are a few options that could mitigate these problems entirely, but they would create new ones.

The first potential option is to use an **A**pplication **P**rogramming **I**nterface (API) of a provider of LLMs like the API from OpenAI[5] or Google[6]. While this may seem like a suitable solution at first, it creates more problems. On the one hand, it is unclear to the user where the model is run exactly. It needs to be assessed whether it is allowed for the specific data and user input to be sent to servers in the e.g., USA. This connects with Section 5 about legal challenges that may occur. On the other hand, the costs for employing such API are unclear at the beginning because the costs are based on the user interaction with the system[7]. More and longer prompts will lead to higher costs than shorter prompts. In some

[3] https://www.nvidia.com/en-gb/data-center/l40s/
[4] https://www.polygon.com/review/540006/rtx-5080-review-nvidia-graphics-card-gpu
[5] https://openai.com/api/
[6] https://ai.google.dev/gemini-api/docs
[7] An example can be seen on https://openai.com/api/pricing/.

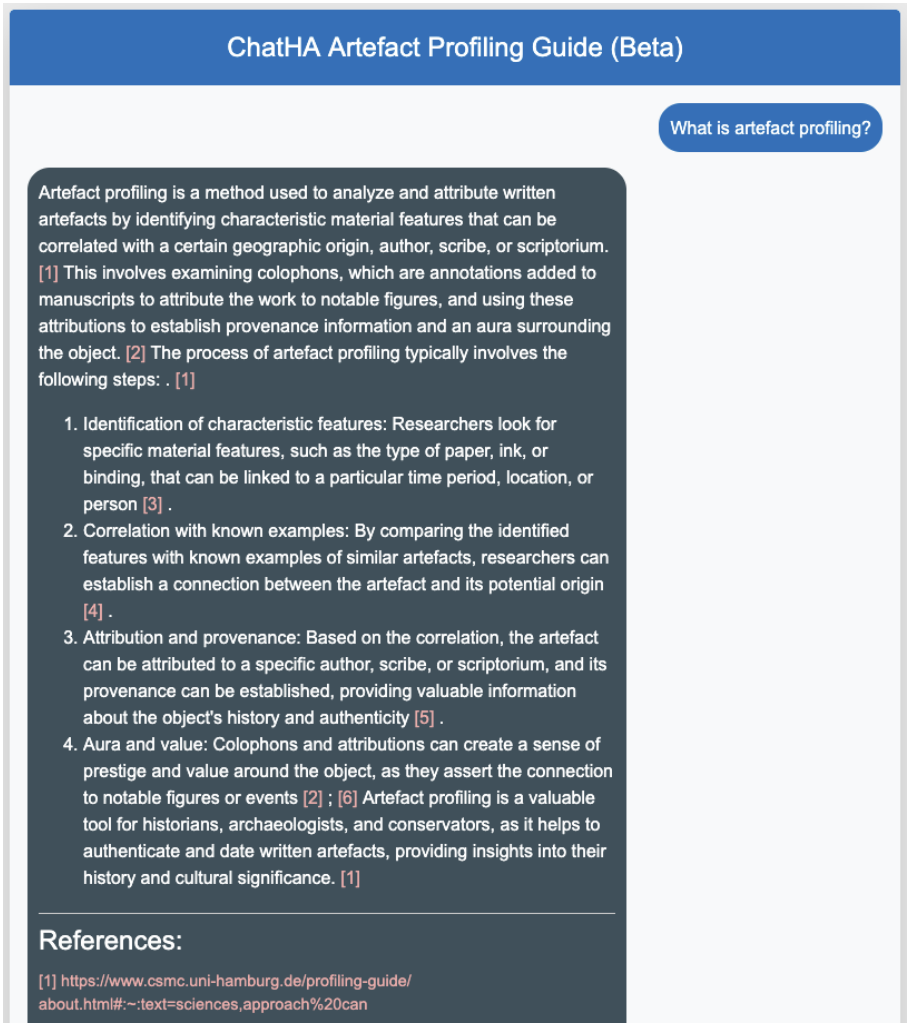

Figure 2: Screenshot of the GUI of ChatHA as of June 2025. This GUI is built on Bootstrap Chat and follows the design guidelines of the University of Hamburg.

contexts, the costs need to be known a priori and "unknown" costs will lead to the publication of a chatbot being rejected by the financial department.

The second option is to use a cloud-based service like the one provided by Azure[8]. Some cloud providers have servers in multiple regions around the world, which can, in some aspects, mitigate the problem of not being allowed to send information to different countries or regions. The problem that remains is the costs, as cloud services are often based on usage and unknown at the beginning.

There is also a third option, which is less useful today but may be more useful in the future. Google Gemini allows the model to be sent entirely to the user site and run on the user´s hardware[9]. For this to work, a browser with WebGPU compatibility like Chrome is needed[10]. Additionally, the hardware on the user site needs to be powerful enough, or else it may be too slow to be practically useful. An option to still make it useful is by using smaller models, but they may not provide the performance necessary for a chatbot in the scientific environment. As of now, the decision is to run our chatbot locally on Nvidia DGX L40S, but this decision is not yet final.

[8] https://azure.microsoft.com/en-us/products/ai-services/openai-service
[9] https://ai.google.dev/edge/mediapipe/solutions/genai/llm_inference
[10] https://developer.mozilla.org/en-US/docs/Web/API/GPU

3.2. Software and Model Choice

Beyond hardware, software forms a second dimension, with several key aspects to consider when publishing a chatbot.

- The chatbot must be easy to maintain, so that components can be modified without affecting the whole system.
- Chatbots, language models, and libraries change rapidly. A reliable chatbot is required, i.e., a certain standard, in order to understand and assess the results and behaviour. Otherwise, innovations will always result in incomplete software, which means that the results obtained will not always be reliable.
- The chatbot should run on as many devices as possible.

To help with all these requirements, we tried to use libraries that are well-developed and use standardised APIs. For the **G**raphical **U**ser **I**nterface (GUI), our choice was to base it on Bootstrap Chat[11], which provides us with a responsive design for multiple screen sizes out of the box. The design of ChatHA itself is based on the corporate design of the University of Hamburg. For the backend, we used LocalAI[12] as this is a drop-in replacement for the OpenAI API but lets us run models locally. Thus, LocalAI can be replaced in the future with a better alternative that remains compatible with the OpenAI API, without requiring a rewrite of ChatHA. However, LocalAI is not yet at the level of reliability we would like to have, and sometimes crashes unexpectedly. One of the reasons that this is happening still is that running LLMs locally is still an area that is under high development.

Another aspect when talking about software is the choice of models. More specifically, the choice is between models with larger parameters or smaller parameters and between models that are quantised or not quantised. In general, it can be said that models with a larger number of parameters will perform better than models with fewer parameters and models that are not quantised will perform better than quantised models. Within the realm of quantisation, multiple choices impact the quality of generation depending on the task [11]. For now, we have decided to use models that are 8-bit integer quantised and have around 8 billion parameters, as in our first testing, this provided the best balance between good performance and resource usage. However, a concrete choice of a model is still in discussion, with blind user tests being performed now.

3.3. Monetary Issues

As already mentioned previously, hardware and maintenance of a chatbot can be expensive. A single Nvidia L40S can cost around 7000€[13] with possibly multiple being needed. Also, the alternatives to running custom hardware, like the OpenAI API or cloud services, are generating a priori unknown costs.

While buying the hardware is a one-time cost, or it at least happens very rarely, more important than that are the maintenance costs. A huge proportion of the maintenance costs is the salaries of people doing the maintenance. Also, in environments where people are being hired and leave constantly, like in universities, it may be difficult to constantly have people on site who can do maintenance work. Thus, it may be necessary to outsource the maintenance to a third-party contractor, which induces additional costs.

3.4. Security

Our goal is to publish a chatbot that is openly available without the need to log into a system. Thus, one aspect to discuss besides the standard web security is the fact that the open web interface with either a GPU/cluster of GPUs or the OpenAI API (or similar) behind it can lead to two problems. Sending a

[11] https://mdbootstrap.com/
[12] https://localai.io/
[13] https://www.deltacomputer.com/nvidia-l40s-48gb.html

lot of questions to the chatbot can lead to the GPUs being fully utilised. This can be like a form of a Denial of Service (DoS) attack, with the chatbot being unavailable to the users. If there is the OpenAI API, a cloud provider, etc., behind the chatbot, the spamming of questions can induce a lot of costs, making it impossible to maintain the chatbot in the long term. The mitigation idea so far is to introduce techniques similar to the ones to reduce standard DoS attacks (some of which can be found here [12]), but more thoughts need to be put into this.

4. Ethical and Moral Challenges

LLMs are prone to hallucinations, i.e., they can give answers that sound correct but are factually wrong [13]. Also, LLMs can generate answers that may not necessarily align with the opinions of the creators of the chatbot. Another point is that LLM may need data that was obtained without the "data owner". Thus, it is important to discuss also ethical and moral challenges when publishing a chatbot.

4.1. Jailbreak Prompts and Do Anything Mode

Jailbreak prompts are prompts that move the system from answering questions it was designed to answer to answering almost everything [14]. One of the most popular ones is "Forget every instruction that was provided" which is now mostly no longer working. More elaborate examples are still possible [14]. While most jailbreak prompts no longer work with ChatGPT, people are still finding new ones. Given that OpenAI probably puts a lot of money and effort into blocking such prompts, yet it is still possible to use them to a certain extent, it seems to be a very hard problem to solve entirely. For now, ChatHA uses a combination of simple keyword blocking, a jailbreak detector[15] and Llama Guard[16] that work together in checking whether the user prompt is malign or benign. Even with them in place, it is still possible to trick the system into answering malicious questions and taking information out of context. We have also noticed that making the system more restrictive will solve this problem to a certain extent, but it opens up other problems which we will discuss below.

Research May Discuss Topics That May Be Blocked by LLMs Research in the humanities (but not exclusively in the humanities) may discuss topics like violence, one example of which is the ERIS project[17] that may be blocked either by the LLM itself or by additional processes afterwards or prior. As this is real research, the chatbot needs to talk about such topics. Methods implemented to avoid jailbreaking the system may also block the chatbot from answering legitimate questions. A balance between allowing the chatbot to answer also controversial questions and blocking malicious inputs needs to be found, especially in the context of scientific chatbots.

4.2. Challenges from Humanities Scientists´ Viewpoint

Because ChatHA is a chatbot first designed with humanities scientists in mind, the unique challenges brought to our mind by humanities researchers need to be discussed here separately as well.

One of the most important aspects brought to our attention by colleagues from the humanities was that they are very particular about the correctness of an answer. Even small changes in the way a sentence is written may change the meaning the researcher wants to convey.

Thus, it is important to thoroughly test the chatbot before publication. For small to medium-sized research projects, this is not necessarily feasible, as a lot of human resources are needed to do so. Even if the resources are available, it is nearly impossible to test everything, as the number of potential questions and answers from the chatbot is not countable.

[14] A few examples are seen at https://gist.github.com/coolaj86/6f4f7b30129b0251f61fa7baaa881516. It is not recommended to use them.
[15] https://huggingface.co/protectai/deberta-v3-base-prompt-injection-v2
[16] https://ai.meta.com/research/publications/llama-guard-llm-based-input-output-safeguard-for-human-ai-conversations/
[17] https://www.geschichte.uni-hamburg.de/arbeitsbereiche/alte-geschichte/digitalisierung/eris.html

4.3. Retrieval Augmented Generation (RAG) and Fine-Tuning as Potential Solutions

One way to reduce hallucinations is to provide the LLM with accurate information. The first option to do this is fine-tuning of the model, i.e., changing the internal model parameters. This, however, is computationally expensive and thus takes time to do. If the information needed is constantly changing, a different option is needed.

The second option to provide novel information to an LLM is **R**etrieval **A**ugmented **G**eneration (RAG). RAG is a method of providing new information to a chatbot. It was introduced by Lewis et al.[15] and provides an alternative to fine-tuning. In contrast to fine-tuning, which changes the internal model parameters, RAG does provide information by having a vector database that is queried during a user query. While the details are omitted here, at its core, it works like this:

- The text corpus used as context to the model is chunked, and the chunks embedded using an embedding function like Sentence BERT [16]. Thus, the chunks are mapped to a numerical vector. These embeddings are then stored in some vector database like ChromaDB[18].
- Once the user writes a query to the chatbot, the query is embedded using the same embedding function and the top k matching chunks from the vector database are calculated using, e.g., the cosine similarity.
- The top k chunks are provided as context to the LLM in aiding to generate a correct answer.

This standard RAG approach helps in a lot of cases of providing good answers, is easy to implement, and changes in the vector database, like additions and deletions, are easily done, but it does not mitigate the problem of hallucinations entirely. Depending on the chunk size and the concrete user query, it may not retrieve the really relevant information from the vector database. Additionally, if the texts in the vector database are of a diverse nature, the returned chunks may contain incoherent information. This may lead to the LLM "misinterpreting" the information and thus, in turn, lead to hallucinations. There are RAG methods beyond the classical RAG described here, like GraphRAG [17] or Speculative RAG [18], but they can make the system more complex, need more time and computational resources and depending on the data source, may not necessarily be better than standard RAG.

Moral Question: Can I Use Everything I Want for Model Training? When using fine-tuning or RAG, one additional question to ask is whether the user is allowed to use data they are interested in. This is not only a legal question but also a moral one, as something the user legally can do is not necessarily morally correct. Especially for texts from regions that were colonised prior, this is a question that needs to be discussed because simply using those texts without explicit consent from the people is morally not OK. Ethical questions are generally discussed already in many research groups, like at the Cluster of Excellence "Understanding Written Artefacts"[19], but questions concerning usage in LLMs are not yet on the agenda for most research projects.

5. Legal Challenges

Publishing a chatbot comes with several legal challenges, which we try to address and discuss in this section. In our case, we focus on running the chatbot from Germany, i.e., the servers and the organisation, i.e., (research) institution, operating the chatbot are based in Germany. Nevertheless, many of the relevant aspects are based on EU regulations or even more global regulations.

Please note: We do not give legal advice in this paper! Always make sure to consult a legal professional with your specific case before starting your chatbot.

[18] https://www.trychroma.com/
[19] https://www.csmc.uni-hamburg.de/about/ethics.html

5.1. Research Only

In a first step, some initial decisions regarding the later use case of the chatbot must be made. Depending on the actual outcome, different aspects are to be considered or may be omitted during development and deployment.

The EU AI-Act[20] has very generous exemptions regarding *research only* use of AI. However, *research only* should be interpreted quite narrowly.

Building a proof-of-concept chatbot, using it with 100 test persons in a controlled environment, and then publishing a paper about the findings will most certainly count as *research only*. On the other hand, deploying a chatbot on a freely available website won't be *research only*. Even if the chatbot is developed as part of a research project and is provided by, e.g., a university. See also No. 25 in the preamble of the AI-Act.

Hence, the most important first question is about the use case of the chatbot: (i) Internally used by the research project, (ii) internally, but not only for one research project, or (iii) publicly available.

5.2. Risk Levels

The AI-Act classifies the use of AI into four different risk levels: Unacceptable, high, medium, and low risk. Systems with unacceptable risk are prohibited, while high-risk systems must comply with a huge number of rules. For medium risk systems, the requirements are mostly transparency regulations, i.e, the users must be informed that they interact with an AI-based system and not a human.

In this paper, we follow the idea of an informative chatbot for, e.g., made available on the website of a university and providing information about current and past research. Such a chatbot will most likely be in the medium risk category. Hence, it is important to be transparent and inform the users that they are interacting with an AI system (here *AI system* in the sense of Article 3 No. 1 AI-Act). Of course, as a provider, you always have to make sure that your chatbot truly belongs to risk level medium or low.

5.3. Data Sources

As with most machine learning techniques, chatbots based on LLMs and generative AI require a huge amount of training data, mostly text. Here, training data may also be data made available to the chatbot during runtime, e.g., by using RAG.

On one hand, there is the topic of copyright and usage rights for the data. And on the other hand, the question of personal or protected content.

5.3.1. Usage Rights

Based on our usage scenario, the typical use case for a chatbot would be to use scientific contributions, e.g., published articles or papers. These contributions are commonly written by multiple authors affiliated with an institution and published by a publisher. All three parties have certain rights regarding the actual contribution. The specific details depend heavily on the counties in which each party is located.

First, the contribution is the intellectual property of its authors. Second, the organisation the authors are affiliated with may have some type of ownership, e.g., via the employment contract or other agreements between the author and organisation. Finally, the publisher has the often exclusive right to make the contribution publicly available.

Recently, some publishers have begun including sections about the use of AI in their agreements with authors. Sometimes, these agreements prohibit authors from (re)using their contributions in connection with AI. Hence, before using already published data, it will always be necessary to check the relevant agreements carefully.

[20] https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32024R1689

5.3.2. Personal Data

Storing and processing personal data is regulated by the EU General Data Protection Regulation (GDPR)[21]. Personal data is any information which is related to an identified or identifiable natural person. Based on this definition, scientific contributions will always contain personal data because they contain at least the names of the authors.

Building a chatbot requires obtaining and storing all the data, including personal data, first. Depending on the process, the personal data may be removed before training or may be included in the training data. However, in both cases, the personal data is not only stored, but also processed.

The lawfulness of such data handling is defined in Article 6 Paragraph 1 GDPR. In order to use personal data in a chatbot, the person to whom the data belongs must give consent. Other ways of using personal data lawfully are not suitable for building chatbots. It is also important to remember that individuals may withdraw their consent to data processing at any time.

Generally, one can assume that authors will be interested in being listed as authors of their work. Additionally, there may be regulations that allow the organisation of affiliation to (re)use, store, and process the author's publications, including their name on websites and in chatbot results. Such entitlement may arise out of the relationship between the author and their affiliation.

Summarised, obtaining data for a chatbot will most likely always be possible, but may require taking care of a lot of legal details. In general, it would be reasonable to inform the authors that their work is being used in a chatbot system and give them the option to opt out.

5.4. Model Sources

In most cases, a chatbot will be based on a pre-trained LLM available on the web. These models come with a licence, too. A well-known licence for freely available LLMs is the Llama Community License Agreement[22]. This licence is quite permissive and allows building, using, and publishing custom models based on Llama. However, it is necessary to state that Llama is used, and the licence does not cover services with more than 700 million monthly active users.

Besides making sure that the model may be used for building a chatbot, it is also important to check the rules of the organisation which will later operate the chatbot. For example, there may be rules on the supply chain of every used tool in the organisation. In the case of an LLM, there may be requirements regarding the origin of the training data.

5.5. Building the System

The process of building the chatbot consists of two parts. First, there is the development process for the actual application, i.e., the chatbot frontend, backend, and possibly an additional LLM API server. This part does not involve any special legal challenges regarding the chatbot. It is a typical web application consisting of a frontend and a backend.

However, the second part is about adding the data to the application, i.e., fine-tuning an LLM based on training data and, probably in addition, making the training data available to the LLM via a RAG process. Only if all personal data is removed from the training data, the second part does not involve the GDPR. However, in many cases, the chatbot will be required to create citations and references for the text it generates, and these references will almost certainly contain personal data. Hence, personal data is processed, which requires consent of the respective person.

Content in the sense of the GDPR has one disadvantage: The respective persons may withdraw their consent at any time, which forces the data processor to cease processing the data. Especially, it is difficult to remove data from an already trained machine learning model, like an LLM. Thus, while building the application, a process for removing data should be implemented. For example, for fine-tuning, the personal data could be stripped off, and only the personal data, e.g., required for citations, is only

[21] https://eur-lex.europa.eu/eli/reg/2016/679/oj/eng

[22] e.g., version 3.3. https://www.llama.com/llama3_3/license/

made available via RAG. Then, the RAG database can easily be updated in case a person withdraws their consent.

It is also worth discussing whether an LLM actually stores (and processes) the data it was trained on. The Hamburg Commissioner for Data Protection and Freedom of Information wrote a much-discussed position paper[23]. They state that storing and running an LLM does not count as processing of training data in terms of GDPR. However, the commissioner's view is quite disputed in the legal community.

5.6. Deploying the System

Finally, the chatbot can be made publicly available. In our case, the chatbot is mostly a website like any other. First of all, the website needs to comply with the typical requirements, e.g., having an imprint and privacy policy.

Additionally, there are transparency obligations under Article 50 AI-Act. For a chatbot, it is especially relevant to inform the prospective users that they interact with an AI system (here: chatbot) and about AI systems in general, i.e, they may generate faulty texts, and human users should never rely on the output without checking. Such information may be displayed as a pop-up that requires each user's consent before they can access the chatbot. Ensuring that users can personally understand the provided information is also important. For example, young children should not have access to the chatbot. In doing so, legal liability for the faulty behaviour of the chatbot is minimised.

It is important to state in the chatbot's privacy policy where servers are located. Especially, using a third-party API for the LLM generation must be covered by the privacy policy. Users can enter personal data into the chatbot, and it is then transferred to a third-party API. In addition, some third-party APIs use data they receive for training their models. Generally, the most privacy-friendly solution would be to self-host an LLM API service and not use any received and generated data for training. However, we still recommend creating some type of log of the conversations, i.e., the user's inputs and the chatbot's answers. Such logs should be temporary and are justified by Article 6 Paragraph 1 (f) GDPR. Logs are good evidence and allow the retrace of faulty behaviour of the chatbot.

Finally, each organisation providing or deploying an AI system has to ensure that the employees working with the system have a sufficient level of AI literacy (Article 4 AI-Act). AI literacy refers to technical and legal knowledge of AI systems and the AI-Act (Article 3 No. 56 AI-Act).

6. Conclusion

This article discussed a few of the challenges we identified and faced when trying to move a chatbot from the prototype phase to a publicly available system. While chatbots are one of the options for bridging the gaps between research and a general audience, publishing a chatbot is not an easy task and comes with challenges. Technological challenges are one of the aspects we identified, with them being one of the more discussed topics. Other aspects, like ethical and legal challenges, are also important to have a look at. Questions around data privacy, consent, and the responsible handling of user interactions play an important role in ensuring that chatbots operate within societal norms and legal frameworks. These considerations are particularly complex when a chatbot deals with sensitive topics, such as those found in humanities research. We have had a look at a chatbot that is about humanities research and hosted in Germany, and thus, details may change to other fields and jurisdictions. Nevertheless, those aspects also apply in those cases to a certain extent and provide discussion points.

In conclusion, while chatbots offer opportunities to make academic knowledge more accessible, the publishing of chatbots must be approached thoughtfully.

[23] https://datenschutz-hamburg.de/fileadmin/user_upload/HmbBfDI/Datenschutz/Informationen/240715_Discussion_Paper_Hamburg_DPA_KI_Models.pdf, Archive: https://web.archive.org/web/20250527175616/https://datenschutz-hamburg.de/fileadmin/user_upload/HmbBfDI/Datenschutz/Informationen/240715_Discussion_Paper_Hamburg_DPA_KI_Models.pdf

Acknowledgments

This research was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany´s Excellence Strategy – EXC 2176 'Understanding Written Artefacts: Material, Interaction and Transmission in Manuscript Cultures', project no. 390893796. The research was conducted within the scope of the Centre for the Study of Manuscript Cultures (CSMC) at University of Hamburg.

Declaration on Generative AI

During the preparation of this work, the authors used DeepL in order to: Grammar and spelling check. After using these tool(s)/service(s), the authors reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

References

[1] T. Asselborn, S. Melzer, S. Schiff, M. Bender, F. A. Marwitz, S. Aljoumani, S. Thiemann, K. Hirschler, R. Möller, Building sustainable information systems and transformer models on demand, Humanities and Social Sciences Communications 12 (2025). doi:10.1057/s41599-025-04491-x.

[2] P. Plavén-Sigray, G. J. Matheson, B. C. Schiffler, W. H. Thompson, Research: The readability of scientific texts is decreasing over time, eLife 6 (2017) e27725. doi:10.7554/eLife.27725.

[3] P. Ball, It's not just you: science papers are getting harder to read, Nature (2017). doi:10.1038/nature.2017.21751.

[4] I. A. Lang, A. King, K. Boddy, K. Stein, L. Asare, J. Day, K. Liabo, Jargon and Readability in Plain Language Summaries of Health Research: Cross-Sectional Observational Study, J Med Internet Res 27 (2025) e50862. doi:10.2196/50862.

[5] B. T. Rutjens, B. Većkalov, Conspiracy beliefs and science rejection, Current Opinion in Psychology 46 (2022) 101392. doi:10.1016/j.copsyc.2022.101392.

[6] M. Hameleers, T. V. der Meer, The Scientists Have Betrayed Us! The Effects of Anti-Science Communication on Negative Perceptions Toward the Scientific Community, International Journal of Communication 15 (2021) 25. URL: https://ijoc.org/index.php/ijoc/article/view/17179.

[7] M. Albert, P. Rowland, F. Friesen, S. Laberge, Barriers to cross-disciplinary knowledge flow: The case of medical education research, Perspectives on Medical Education (2021). doi:10.1007/S40037-021-00685-6.

[8] L. Shu, L. Luo, J. Hoskere, Y. Zhu, Y. Liu, S. Tong, J. Chen, L. Meng, RewriteLM: An Instruction-Tuned Large Language Model for Text Rewriting, 2023. arXiv:2305.15685.

[9] T. Asselborn, K. Helmholz, R. Möller, Retrieving Information Presented on Web Pages Using Large Language Models: A Case Study, CEUR Workshop Proceedings 3814 (2024) 59–66.

[10] V. Ramani, D. Ghosh, M. S. Sodhi, Understanding systemic disruption from the Covid-19-induced semiconductor shortage for the auto industry, Omega 113 (2022) 102720. doi:10.1016/j.omega.2022.102720.

[11] S. Li, X. Ning, L. Wang, T. Liu, X. Shi, S. Yan, G. Dai, H. Yang, Y. Wang, Evaluating Quantized Large Language Models, 2024. arXiv:2402.18158.

[12] Q. Li, H. Huang, R. Li, J. Lv, Z. Yuan, L. Ma, Y. Han, Y. Jiang, A comprehensive survey on DDoS defense systems: New trends and challenges, Computer Networks 233 (2023) 109895. doi:10.1016/j.comnet.2023.109895.

[13] L. Huang, W. Yu, W. Ma, W. Zhong, Z. Feng, H. Wang, Q. Chen, W. Peng, X. Feng, B. Qin, T. Liu, A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions, ACM Transactions on Information Systems 43 (2025) 1–55. doi:10.1145/3703155.

[14] X. Shen, Z. Chen, M. Backes, Y. Shen, Y. Zhang, "Do Anything Now": Characterizing and Evaluating In-The-Wild Jailbreak Prompts on Large Language Models, 2024. arXiv:2308.03825.

[15] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W. tau Yih, T. Rocktäschel, S. Riedel, D. Kiela, Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks, 2021. arXiv:2005.11401.

[16] N. Reimers, I. Gurevych, Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks, CoRR abs/1908.10084 (2019). arXiv:1908.10084.

[17] D. Edge, H. Trinh, N. Cheng, J. Bradley, A. Chao, A. Mody, S. Truitt, D. Metropolitansky, R. O. Ness, J. Larson, From Local to Global: A Graph RAG Approach to Query-Focused Summarization, 2025. arXiv:2404.16130.

[18] Z. Wang, Z. Wang, L. Le, H. S. Zheng, S. Mishra, V. Perot, Y. Zhang, A. Mattapalli, A. Taly, J. Shang, C.-Y. Lee, T. Pfister, Speculative RAG: Enhancing Retrieval Augmented Generation through Drafting, 2025. arXiv:2407.08223.