# Retrieving Information Presented on Web Pages Using Large Language Models: A Case Study

Thomas Asselborn[1,2,3], Karsten Helmholz[2,3] and Ralf Möller[1]

[1] *Universität Hamburg, Institute of Humanities-Centered Artificial Intelligence, Warburgstraße 28, 20354 Hamburg, Germany*

[2] *Universität Hamburg, Centre for the Study of Manuscript Cultures, Warburgstraße 26, 20354 Hamburg, Germany*

[3] *University of Hamburg, Cluster of Excellence 'Understanding Written Artefacts' (UWA), Warburgstraße 26, 20354 Hamburg, Germany*

**Abstract**

Developing web pages is a task that requires constant updating of new information. Additionally, multiple web pages with the same information must be developed but compiled differently if multiple user groups are targeted. Thus, we introduce a new approach that uses LLMs (**L**arge **L**anguage **M**odels), RAG (**R**etrieval **A**ugmented **G**eneration) and SCDs (**S**ubjective **C**ontent **D**escriptions) to query the information on a web page and also provide sources to the original data. Since the LLM can rewrite the response based on the target user group, this reduces the need to make multiple web pages. Based on the example of the Artefact Profiling Guide, this approach reduces the need to provide user-specific web pages. Additionally, this method reduces the need for an expert in web programming and designing by offloading the task of presenting new data to the LLM. The prototype system has shown promising results so far. It provides the correct answers grounded by the source and written appropriately for the target group.

**Keywords**

Artefact Profiling, Large Language Models (LLMs), Generative Pre-Trained Transformer (GPT), Retrieval Augmented Generation (RAG), Web Pages

## 1. Introduction

Web pages, like those for universities or corporations, are typically segmented into parts that offer information targeted to a specific user group. A university web page is usually divided into sections specifically curated to provide information for students, people interested in studying, employees, other researchers, the general public, etc. While this approach works fine, as shown in practice, there are specific scenarios where one may want to offer a single web page targeted to many user groups that may not even be known a priori. One example is that the information provided may interest a large group of people with different needs, be it pupils or scientific researchers. In this case, offering different web pages with duplicate content would generally be necessary, each curated for the respective audience. Alternatively, one has the option of offering a single web page written in such a way that it can address all target groups at once. However, web pages designed to be an all-in-one solution run the risk of making them too complicated for some yet too easy for others, thereby disappointing most if not all, users.

LLMs utilising the transformer architecture [1] like GPT (**G**enerative **P**re-trained **T**ransformer) [2] have several beneficial properties for making them an ideal candidate to be used on web pages with different and diverse user groups. On the one hand, they can be tailored to a specific downstream task,

*Humanities-Centred AI (CHAI), 4th Workshop at the 47th German Conference on Artificial Intelligence, September 23, 2024, Würzburg, Germany*

thomas.asselborn@uni-hamburg.de (T. Asselborn); karsten.helmholz@uni-hamburg.de (K. Helmholz); ralf.moeller@uni-hamburg.de (R. Möller)

https://www.philosophie.uni-hamburg.de/chai/personen/asselborn.html (T. Asselborn); https://www.csmc.uni-hamburg.de/about/people/helmholz.html (K. Helmholz); https://www.philosophie.uni-hamburg.de/chai/personen/moeller.html (R. Möller)

0009-0005-3011-7626 (T. Asselborn); 0000-0002-1174-3323 (R. Möller)

allowing them to acquire specific knowledge. Additionally, they can adapt their language to the specific user writing queries, e.g., by telling the user that the query should be explained to a 10-year-old child.

In this paper, we thus propose a different method (Section 3) of dealing with the problem of generating multiple distinct web pages by using LLMs to return the content a user may ask about and to provide citations to the original resource. It aims to investigate whether it is possible to replace web pages with a ChatGPT-like chatbot (Section 4) and which benefits and potential problems this approach may have (Sections 4 and 5). Before going into detail, Section 2 introduces the Artefact Profiling Guide.

### 1.1. Related Work

Working with LLMs and RAG is an actively researched topic. Gao et al. [3] are comparing different methods of applying the general principles of RAG. Ramesh et al. [4] have proposed context tuning for RAG to improve retrieval.

RAG is a method that several companies also use in practice to increase productivity. Examples include but are not limited to Telescope (a sales automation platform), Assembly (a human resources platform) and Causal (a financial planning tool). [5] Also, bigger companies like IBM, Google, NVIDIA, and Microsoft are using RAG for various tasks. [6]

Using LLMs to generate web pages is a feature that Perplexity introduced a few months ago, which they called Perplexity Pages [1]. There, Perplexity queries and results can be displayed automatically on a web page. This helps to produce web pages without the need to know web development and design. However, their approach focuses on making web pages with information already present on the internet and compiling them into a single view. In contrast, our approach is focused on retrieving information that would typically be displayed on a web page with the help of LLMs.

## 2. The Artefact Profiling Guide

The Artefact Profiling Guide is an online guide written by members of various lab teams of CSMC (**C**entre for the **S**tudy of **M**anuscript **C**ultures). Its goal is to explain the scientific methods and analytical means used to analyse written artefacts to foster interdisciplinary work between the humanities, the natural sciences and computer science. It is meant as a broad overview with more detailed explanations provided as links to dedicated sources.

While the guide is primarily written for a target group of researchers and specialists, it is also meant to provide information to other potential user groups, such as libraries, museums, and other collections that do not necessarily have the expertise or the equipment to do analyses. Another potential user group may be private collectors who want more information about an artefact in their collection. Beyond those specific user groups, the guide is also meant to provide the general public with information about some aspects of the research carried out at CSMC. The general public is a blurry group that could range from young pupils to older people wanting to broaden their knowledge and anyone in between. This variety of target groups makes it difficult to provide the desired explanation level and find a language appropriate for all users.

*Artefact profiling* is an interdisciplinary field combining analytical means from various fields to understand written artefacts. It combines, among others not mentioned here, multiple so-called “omics” approaches:

- **Genomics:** This studies the genetic material in the (written) artefact.
- **Proteomics:** Analysation of proteins and peptides.
- **Metabolomics:** Metabolites and small molecules are examined.
- **Metallomics:** This deals with the metals and their distribution.
- **Isotopolomics:** Isotope ratios in the written artefacts are studied.

[1] https://www.perplexity.ai/de/hub/blog/perplexity-pages

Artefact profiling aids in dating, locating and authenticating written artefacts.[7]

Artefact profiling is a dynamic field of research with new approaches emerging and new devices being developed to implement them. (see [2]). Additionally, researchers are switching universities and positions frequently. Thus, the Artefact Profiling Guide needs to be updated regularly.

## 3. Method

The data in the artefact profiling guide is constantly evolving and probably also not part of most LLMs pre-training. Thus, a way to ensure the LLM can understand new information must be used. Currently, there are two main methods of making sure that the LLM works better on a specific downstream task:

1. **Fine-tuning:** In fine-tuning, one takes a so-called pre-trained model, like Llama 3 [8], which is typically trained in a self-supervised[3] way. The model is typically fine-tuned in a supervised way using a smaller task-specific data set. With fine-tuning, the model´s parameters are changed. While it is far less demanding regarding data set size and hardware resources than pre-training and is thus more feasible for smaller companies or research groups, fine-tuning still needs a GPU to be done efficiently.
2. **RAG [9]:** RAG is another method of making sure that the LLM works better on a specific downstream task. In contrast to fine-tuning, the internal parameters of the models are not changed; instead, context is provided using automatic prompt augmentation. Typically, the data the user wants to add to the LLM is encoded in a vector database using, e.g., BERT [10] embeddings. When the user sends a query, it is first encoded using the same encoding, the k best results from the vector database are returned and then automatically appended to the query, which is then sent to the LLM. This works because LLMs have a property called in-context learning [11], allowing the LLM to answer queries about a topic that was never seen in the pre-training data set.

We have decided to use RAG for our approach as this is less demanding in terms of hardware requirements, i.e., while fine-tuning requires a dedicated GPU, RAG also works with CPUs. Additionally, this is necessary because the data in the Artefact Profiling Guide is dynamically changing.

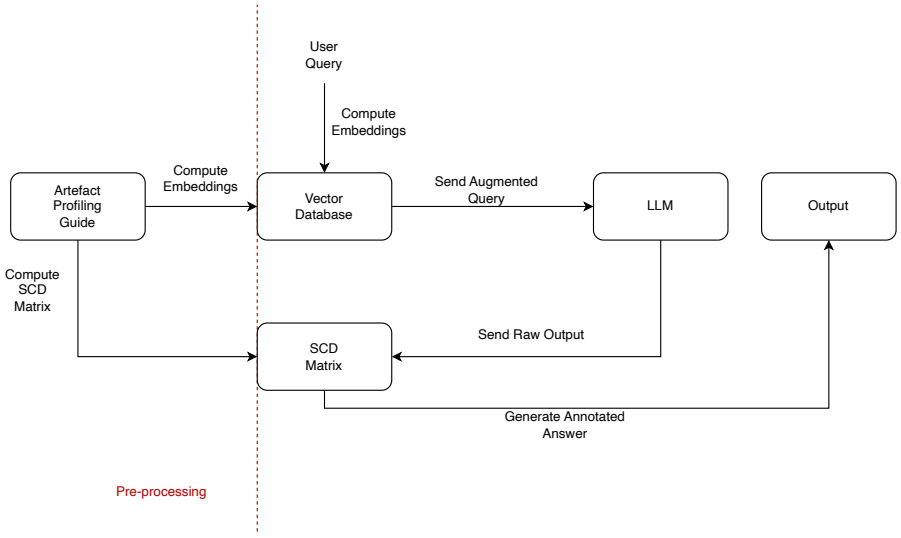



**Figure 1:** Broad overview of the underlying process. Everything on the left of the dotted, red line is part of the pre-processing. This only needs to be done once per change of the content of the Artefact Profiling Guide. Everything on the right of the line is executed after every user query.

[2] https://www.csmc.uni-hamburg.de/publications/blog/2024-04-30-enci-inauguration.html
[3] https://www.ibm.com/topics/self-supervised-learning

Figure 1 shows a broad overview of the process. It is divided into two sections. On the left of the dotted red line is the pre-processing. This only gets executed when a change to the data source, in this case, the Artefact Profiling Guide, has happened. From there, the embeddings to be stored in the vector database and the SCDs[12] are computed. SCDs contain additional data that is attached to locations in a text document. They can be descriptions, links, or labels and can be automatically generated. Thus, both pre-processing steps are completely automated without the need for human intervention.

The process on the right of the dotted, red line is executed at every user query. It can be split up into the following steps. It follows the same structure that was already introduced in [13] as ChatHA.

1. The user enters a query and selects a user group. A few groups and example queries may be predefined (see Figure 2a).
2. After the user has entered its query, it is embedded using the same algorithm that was used in the pre-processing, e.g., Sentence BERT[4]. This is then used to retrieve the n most similar entries from the vector database using, e.g., using the cosine similarity or Euclidean distance [14]. The results are then used to augment the original user query. Additionally, the query is also augmented by the user group that was selected prior.
3. Having the augmented query, this is then used to send the query to the LLM. This could be the OpenAI API using GPT4 [15] or a local, open-source LLM like Llama 3 [8].
4. The output from the LLM, which we call raw output, is then annotated with links to the original data source. Using the Most Probably Suited SCD (MPS$^2$CD) algorithm [16], the most suitable SCDs from the set of known SCDs (stored in the SCD matrix) are computed for the raw output. This then provides the citation with links to the source. Figure 2b shows how this may look practically
5. The final output with the citations is shown to the user.

## 4. Experiment

A few tests have been performed to test this approach. First, performance for the pre-processing step was evaluated, i.e., the time it took to calculate the embeddings for the vector database and the SCDs. Everything was computed on a Macbook Pro with an M3 processor and 16GB of RAM, which is no special hardware. Thus, it is reasonable to assume that it should take roughly the same time. Since both processes are independent, they can be executed in parallel, which was also done for this test. We used ChromaDB[5] with the default Sentence BERT embeddings for the vector database. The results are as follows:

- Total number of words in the document: 67,438
- Total time to calculate and store embeddings: 10.5 min.
- Total time to calculate and store SCDs: 26.5 min.

All the numbers are in a range that makes it possible to be rerun regularly, e.g., once per week.

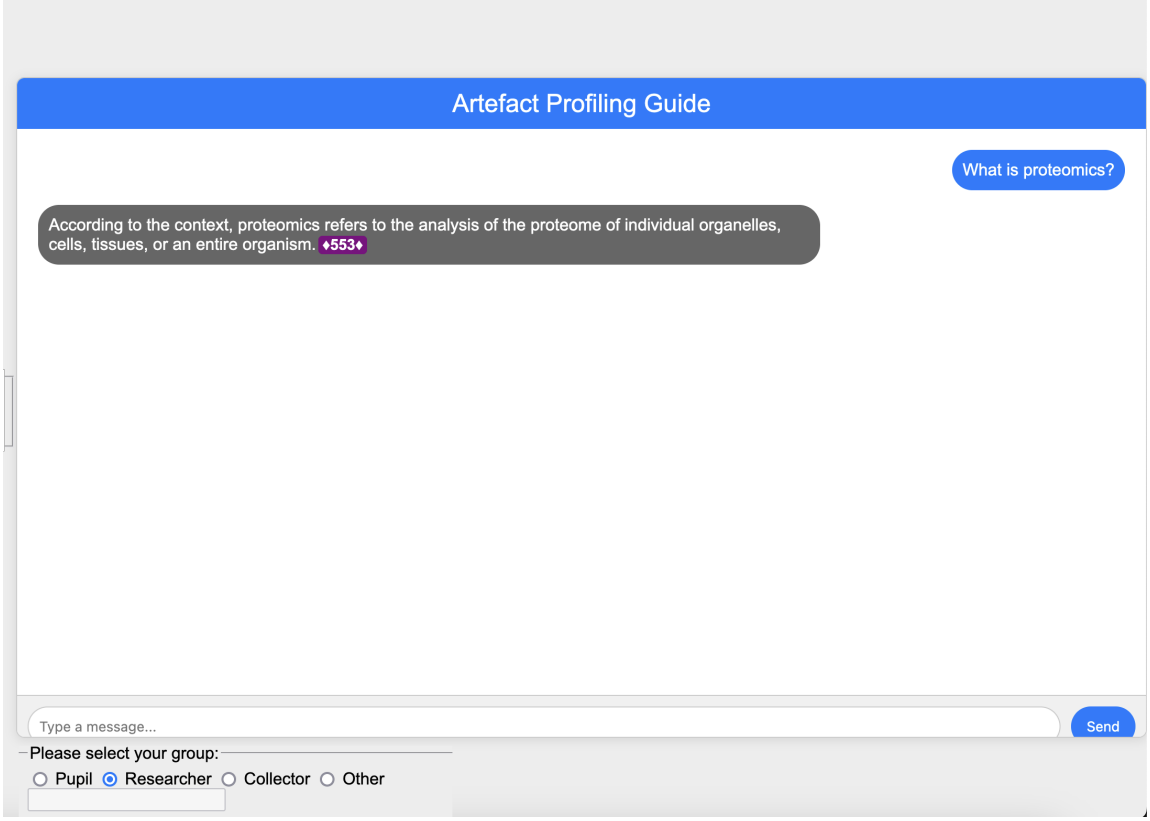
As a second experiment, a few test questions were asked to the system. This is then evaluated qualitatively. The key results are the following.

### RAG does work in our test example

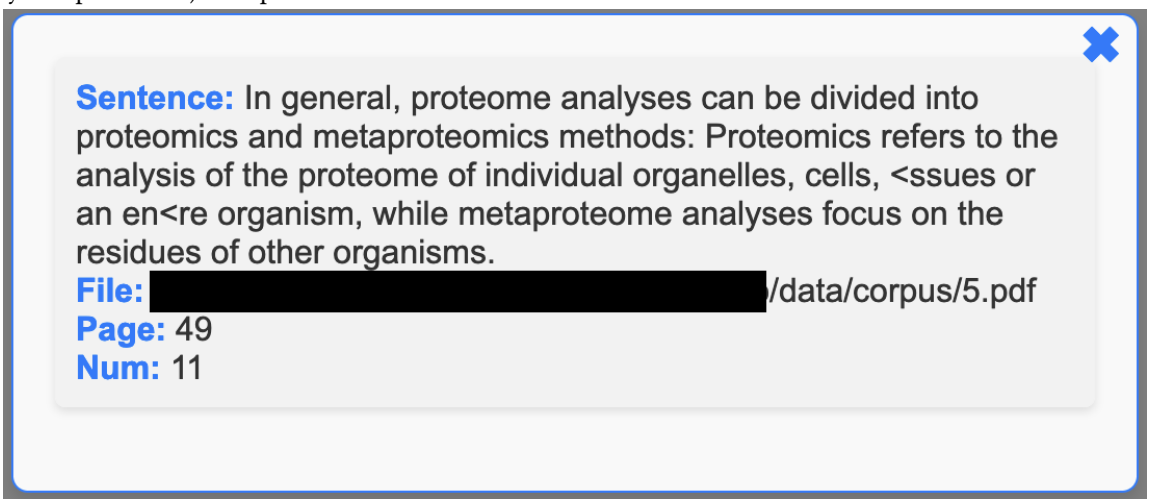
Mostly, the answers returned by the system are correct and factually backed by the data source. Also, this helped to answer questions that a standard LLM like Llama 3 could not answer. One such example is:

- Question: “What is ENCI?”

[4]https://sbert.net/
[5]https://www.trychroma.com/




(a) Mockup of how the web page may look like. User can select their group or a free input to target generation for their specific needs. Additionally, one can consider adding a few example queries as a starting point (not yet implemented). The pink field with the number inside indicates references the user can follow.




(b) Example, how citations may look like after clicking on the pink field as seen in Figure 2a. In this example, information came from a PDF file but can potentially come from any type of text data.

**Figure 2:** Prototype implementation of the web page replacement. The top shows the question and answer, while the bottom shows what one citation may look like.

- Answer: "Based on the provided context, ENCI is mentioned as a type of equipment used at the CSMC (Center for the Study of Manuscript Cultures) for non-destructive 3D material characterisation. The equipment is described as being able to study surface topography and mechanical properties, and its mobility is considered to be stationary. It is also mentioned that ENCI can be used for in-situ measurements with portable equipment. It appears that ENCI is a specialised tool used for the analysis of materials and artifacts, particularly in the field of

manuscript cultures and studies."

Standard Llama 3 was unable to answer that question at all.

### System is capable of automatically changing the "language" used to answer

The system aims to give answers in a language appropriate to the specific user group. The following example illustrates this:

- Question: "What is Raman spectroscopy?"
- Answer for 10-year old: "Raman spectroscopy is a special way to look at things without touching them or hurting them. It's like taking a picture of the things, but instead of using a camera, it uses special light to see what's inside the things. And it's very good for looking at old and special things, like art and treasures, because it doesn't hurt them at all!"
- Answer for a private collector: "Dear fellow collector, Raman spectroscopy is a fantastic technique that allows us to examine and analyse our beloved artefacts without causing any harm to them. It's a non-invasive and nondestructive method that uses special light to "see" what's inside our objects, allowing us to learn more about their composition, materials, and even their age. It's been particularly useful in the field of Cultural Heritage, where preserving our treasures is of utmost importance."
- Answer for a researcher: "As a researcher, I can tell you that Raman spectroscopy is a non-invasive and non-destructive analytical technique that relies on the inelastic scattering of monochromatic light by atomic vibrations in the visible, near-infrared, and near-ultraviolet range. This technique has gained significant popularity in the field of Cultural Heritage over the past few decades, as it allows us to analyse samples without causing any harm. The advantages of Raman spectroscopy include its ability to be truly non-destructive."

### Citation system with SCDs has issues when the language used in answer differs significantly from original language

While the system can provide the appropriate language depending on the target group, this leads to problems with generating citations. Depending on how the LLM rewrites the answer, the MPS²CD algorithm may only sometimes be able to find the correct citations. This needs to be investigated in the future in more detail.

## 5. Problems and Challenges

Working with LLMs still has some issues that may be faced also in the context of the Artefact Profiling Guide. The first observation was that the automatic prompt extension influenced the results significantly. If the user is interested in information only grounded by the provided documents, it should be clearly stated in the prompt extension. Otherwise, the LLM also tends to give general information in the prompt. Depending on the task, this may or may not be wanted.

While RAG is a method that aims at reducing hallucinations, the LLM may still give (partially) incorrect answers. One example is the following:

- Question: "What is the CSMC?"
- Answer (using Llama 3 Instruct 8B): "I apologise for not providing context earlier! The CSMC likely refers to the Center for Systems Medicine and Computational Biology, which appears to be a research institution or organisation. However, without more information, it's difficult to determine the specific details about this center.
  If you could provide more context or clarify what the CSMC is, I'd be happy to help with your original question!"

This incorrect answer was provided even though the documents provided had the necessary information to give a correct answer.

Opening an LLM to be used by the public also adds problems that cannot be solved by using RAG or even fine-tuning. They mainly arise because the system should be openly available without logging into it. A few of these issues are discussed here.

One issue that may occur is that opening the LLM without user registration (registration is probably not what the provider of a public web page wants) opens the possibility for Denial-of-Service (DoS) attacks, i.e., the input is flooded with queries that use up all of the hardware resources on the server making the web page unreachable for all legitimate users. Since LLMs have a higher demand for hardware than standard web pages, the limit will likely hit sooner. Thus, mitigation methods need to be thought of thoroughly.

Additionally, user protection must be incorporated into the system, especially for children. One of the texts in the data set has a medieval Arabic medicine book as an example containing medical exams and treatments. This also includes descriptions and illnesses that may not be suitable for children, yet it would still be technically correct for the LLM to return them as an answer. Some ways must be introduced to avoid giving inappropriate but technically correct answers.

## 6. Conclusion and Outlook

In this paper, we introduced a new method of getting and providing information on a web page using LLM and RAG. We showed that it is feasible, but we also pointed out some potential issues that needed to be mitigated, such as DoS attacks or child safety issues. Additionally, changing the concrete prompt used for automatic prompt augmentation makes a difference in how the LLM will answer.

Future focus should be on how citations are handled when the answer uses significantly different words while still giving the correct answer, i.e., when the LLM answers in simple words while the original data source is written in a scientific language. Additionally, user tests with a more extensive user group may provide more insights into evaluating the introduced method. Currently, the method only works with text data, with future work investigating methods of incorporating images and videos that are integrated into the RAG pipeline and in the citation with the SCDs. It would also be beneficial to extend the system to understand the user automatically without telling it which user group the user belongs to.

## Acknowledgments

The research for this contribution 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 Universität Hamburg.

## References

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, I. Polosukhin, Attention is all you need, in: I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (Eds.), Advances in Neural Information Processing Systems, volume 30, Curran Associates, Inc., 2017. URL: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

[2] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever, Improving language understanding by generative pre-training (2018). URL: https://openai.com/research/language-unsupervised.

[3] Y. Gao, Retrieval-augmented generation for large language models: A survey, arXiv preprint arXiv:2312.10997 (2023).

[4] R. A. Ramesh, T. Bethi, D. Vodianik, S. V. Chappidi, Context tuning for retrieval augmented generation, in: EACL Workshop, 2023. URL: https://arxiv.org/abs/2312.05708.
[5] J. Gitlin, 7 examples of retrieval-augmented generation (rag), 2024. URL: https://www.merge.dev/blog/rag-examples, accessed: 2024-06-30.
[6] NVIDIA, What is retrieval-augmented generation aka rag?, 2024. URL: https://blogs.nvidia.com/blog/what-is-retrieval-augmented-generation/, accessed: 2024-06-30.
[7] M. Creydt, M. Fischer, Artefact profiling: Panomics approaches for understanding the materiality of written artefacts, Molecules 28 (2023) 4872. doi:10.3390/molecules28124872.
[8] M. AI, Introducing meta llama 3: The most capable openly available llm to date, 2024. URL: https://ai.meta.com/blog/meta-llama-3/, accessed: 2024-05-23.
[9] 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.
[10] J. Devlin, M. Chang, K. Lee, K. Toutanova, BERT: pre-training of deep bidirectional transformers for language understanding, CoRR abs/1810.04805 (2018). URL: http://arxiv.org/abs/1810.04805. arXiv:1810.04805.
[11] S. Min, X. Lyu, A. Holtzman, M. Artetxe, M. Lewis, H. Hajishirzi, L. Zettlemoyer, Rethinking the role of demonstrations: What makes in-context learning work?, 2022. arXiv:2202.12837.
[12] F. Kuhr, T. Braun, M. Bender, R. Möller, To Extend or not to Extend? Context-specific Corpus Enrichment, Proceedings of AI: Advances in Artificial Intelligence (2019) 357–368. doi:10.1007/978-3-030-35288-2_29.
[13] T. Asselborn, S. Melzer, S. Aljoumani, M. Bender, F. A. Marwitz, K. Hirschler, R. Möller, Fine-tuning bert models on demand for information systems explained using training data from pre-modern arabic, in: Proceedings of the Workshop on Humanities-Centred Artificial Intelligence (CHAI 2023), CEUR Workshop Proceedings, 2023, pp. 38–51. URL: https://ceur-ws.org/Vol-3580/paper5.pdf.
[14] A. Singhal, Modern information retrieval: A brief overview, 2001. URL: http://singhal.info/ieee2001.pdf, accessed: 2024-06-28.
[15] OpenAI, Gpt-4 technical report, 2023. arXiv:2303.08774.
[16] F. Kuhr, M. Bender, T. Braun, R. Möller, Augmenting and automating corpus enrichment, Int. J. Semantic Computing 14 (2020) 173–197. doi:10.1142/S1793351X20400061.