Digitalisation Workflows in the Age of Transformer Models: A Case Study in Digital Cultural Heritage

Mahsa Vafaie*[1,2], Mary Ann Tan*[1,2] and Harald Sack[1,2]

[1]*FIZ Karlsruhe – Leibniz Institute for Information Infrastructure, Hermann-von-Helmholtz-Platz 1, 76344 Eggenstein-Leopoldshafen, Germany*

[2]*Applied Informatics and Formal Description Methods (AIFB), Karlsruhe Institute of Technology (KIT), Kaiserstraße 89, 76133 Karlsruhe, Germany*

Abstract

The advent of transformer architecture revolutionised the field of Artificial Intelligence (AI) and its various applications. It is only recently that digitalisation of cultural heritage data has become heavily dependent on AI solutions. This paper presents case studies from two different projects, to explore the integration of transformer-based technologies into digitalisation workflows for cultural heritage data. The transformative effects of these models on such workflows are showcased and the benefits and drawbacks of this paradigm shift are briefly discussed.

Keywords

Transformer-based technologies, Large Language Models (LLMs), Transformer based Optical Character Recognition, Digital Cultural Heritage, Digital Humanities, Digital Libraries.

1. Introduction

In the diverse landscape of Cultural Heritage (CH), technological advances have been reshaping the way we perceive and interact with our collective legacy. By employing state-of-the-art tools and techniques from different subfields of Artificial Intelligence (AI), digitalisation initiatives have been revolutionising the restoration, preservation and accessibility of cultural resources. This transformative approach not only safeguards artifacts and collections for future generations [1], but also paves the way for a more democratic and transnational access to knowledge by liberating it from the confines of physical museums, archives, and libraries [2]. As technology continues to integrate with CH, new horizons open up for scholars and researchers, empowering them by the emergence of new data collections, metadata sources, and Linked Open Data [3], as well as novel methods for analysis and collaboration.

CH data encompasses a diverse variety of information sources, spanning from unstructured data to structured data and metadata. Unstructured data, such as raw scans and photographed documents within archives, encapsulates the unprocessed narratives, historical documents, and creative expressions that form the essence of our cultural heritage. One example of a

* These authors contributed equally to this work.

digitalisation project that works with unstructured data in Germany is the project *"Themenportal zur Wiedergutmachung nationalsozialistischen Unrechts"* [1] [Thematic Portal for Compensation for National Socialist Injustice]. Also referred to as *"Themenportal Wiedergutmachung"*, this project focuses on *Wiedergutmachung* records, which contain documents related to compensation for injustices that occurred during the National Socialist era.

Metadata on the other hand, provides a structured lens through which to navigate the cultural resources. Exemplified by library catalogues, metadata contains essential details about artifacts, such as authorship, date of creation, and thematic categorisation. The *"Deutsche Digitale Biblothek"* [2] [German Digital Library], as a national aggregator to the Europeana [4], is an example of a digitalisation effort that collects, transforms, and publishes metadata representing tangible and intangible CH objects across Germany. These metadata are provided by hundreds of different CH institutions including, but not limited to, libraries, archives, museums, media libraries, and historical site preservation.

The advent of transformer-based technologies marks a historic juncture in the realm of AI. These cutting-edge technologies, exemplified by models such as BERT [5] and GPT-4 [6], both categorised as Large Language Models (LLMs), demonstrate unrivaled capabilities in automatic processing and generation of complex information without the need for domain-specific fine-tuning [7]. In the context of digitalisation workflows, the incorporation of transformer models promises a paradigm shift by enhancing both the efficiency and quality of the digitalisation process. In Computer Vision, transformers excel in text and image recognition tasks, enabling more accurate and context-aware digitisation of visual artifacts [8, 9]. In Natural Language Processing, these models facilitate nuanced understanding of textual content, aiding in the extraction of valuable information from diverse sources. The attention mechanisms inherent in transformer architectures [10] enable capturing global dependencies instead of considering only local contexts (as in RNNs), potentially leading to improved digitalisation outcomes. This synergy between transformer-based technologies, Computer Vision, and Natural Language Processing, has the potential to reshape workflows in the digitalisation projects when used in its best capacities, while introducing potential challenges that need careful consideration, such as compromised accuracy in capturing information and breaching data privacy and security measures by transmitting sensitive data to external servers.

In this position paper, we showcase the power of transformer models to reshape digitalisation workflows through increased efficiency. Our study is focused on workflows for the transformation of CH data into knowledge graphs, namely, *"Themenportal Wiedergutmachung"* and *"DDB"*. Section 2 outlines potential applications of transformer-based models relevant to CH digitalisation pipelines, and provides an overview of such endeavours and their outcomes. Section 3 introduces the aforementioned projects in more detail, and examines the capability of transformer-based technologies to influence workflows in these projects. In Section 4 the implications of transformer-based technologies for CH digitalisation projects are discussed, exploring both their benefits and drawbacks. Section 5 concludes the paper by emphasising the crucial aspects of utilisation of transformer-based technologies for digitalisation of CH data.

---

[1]https://is.gd/bundesfinanzministeriumwgm
[2]https://www.deutsche-digitale-bibliothek.de/

2. Related Work

The release of BERT [5] in 2018 caused a stir due to its groundbreaking performance in a wide variety of NLP tasks during that period. Its success can be attributed to the transformer architecture, which makes use of the attention mechanism, and the novel approach of training a language model bidirectionally to leverage the left and the right context of a word in the sentence. Since then, transformer-based models have been applied in several areas of research.

However, it was not until the release of ChatGPT that the term "LLM" reached public awareness. Early adopters used ChatGPT for frivolous fun, but for some, the implications became dire when Europe's top selling newspaper informed their editors that they will be replaced by AI [11]. Lund et al. [12] discussed the impact of ChatGPT in academia, scholarly research and publishing, while Meyer et al. [13] enumerated the benefits and limitations of using ChatGPT in academic writing, education, and programming.

During a briefing at the European Parliament in the context of museums, Pasikowska-Schnass and Lim [14] reported the opportunities, risks, and challenges cultural institutions face in adopting AI. One challenge they identified is the uneven level of digitalisation of collections and the resulting varying metadata quality. They emphasised the value of human resources as essential in producing high-quality metadata used to train the models. In a similar context, Neudecker [15] discussed the mutual opportunities of combining cultural heritage data and AI tools, while highlighting criticisms of data practices in the field. The author described the workflows that leverage AI in two projects run by the Berlin State Library.

On a more practical level, an expanding body of research has been utilising transformer-based models for a variety of applications, that have the potential to be leveraged for CH digitalisation workflows. Tang et al. [16] developed Universal Document Processing (UDOP), a vision-text-layout Transformer for document understanding and generation. UDOP integrates pretraining and multi-domain downstream tasks within a prompt-based sequence generation framework. It establishes state-of-the-art performance on eight Document AI tasks, including document understanding and QA, spanning various data domains such as finance reports, academic papers, and websites. However, its application to CH data is yet to be tested. For text recognition of multilingual historical printed and handwritten documents from libraries and archives, Ströbel et al. [9] tested the Transformer-based Optical Character Recognition (TrOCR) model [17] and showed the robustness of the model in different settings. Cao et al. [18] illustrated GPT-4's emergent ability to understand a word despite being subjected to extreme character-level permutations. They showed that the LLM can still answer questions despite being provided with a heavily scrambled context. This capability can be leveraged in correcting bad quality OCR results. In another recent work, Borenstein et al. [19] adapted PIXEL [20], a text-free pixel-based language model that renders text as images, for downstream language understanding tasks on historical documents without relying on OCR.

Domain-specific Information Extraction (IE) pipelines previously required expert-curated training datasets to achieve passable performance. De Toni et al. [21] showed that LLMs are able to extract some entities from historical texts with out-of-distribution languages in a zero-shot setting. Taking it a step further, Petroni et al. [22] showed that LLMs have the potential to fill missing information in a Knowledge Graph (KG). As per their definition, a fact is synonymous to a triple having a subject, a relation, and an object (e.g. <"Dante", :birthplace, "Florence">. An

LLM "knows" a fact, if it can correctly identify the masked relation in a cloze statement (e.g. Dante [MASK] Florence.)

Just as important is the added value of LLMs to the search and retrieval of CH resources. In particular, Retrieval-Augmented Generation (RAG) [23] models combine different kinds of knowledge from LLMs (vast, probabilistic, general) and KGs (expandable, precise, explicit, interpretable) to access and present relevant, factual information without the need for structured queries.

These developments pave the way for more sophisticated information extraction pipelines and enhanced capabilities in handling historical texts and documents.

3. Use Cases

This section presents the intersection of transformer-based technologies and the digitalisation workflows within two concrete CH digitalisation use cases. Through the utilisation of transformer models and experimenting with them, we analyse the ways in which these models can adapt the workflows in such settings.

3.1. *Wiedergutmachung*: Information Extraction from Archival Documents

The project *"Themenportal Wiedergutmachung"* [3] aims to create an information system for contextualisation of historical knowledge derived from archival documents. At its core, this project revolves around collections of documents, records, and materials directly linked to the compensation process for the atrocities committed by the National Socialist regime in Germany. This digitalisation initiative starts with conversion of document images to machine-readable formats. Subsequently, the workflow entails information extraction, ontology design [24, 25] and population, and linking with external sources, to construct the *Wiedergutmachung* Knowledge Graph. Due to the historical nature of the documents, each of these stages present unique challenges that can be tackled using traditional methods, while transformer models offer a more direct approach to addressing them.

3.1.1. Text Recognition for a Variety of Text Types
Traditionally, workflows dealing with a variety of text types on scanned documents employed distinct text recognition models. In [26] and [27] Vafaie at al. propose a pipeline for separation of machine-printed text and handwritten text on historical archival documents that contain both text types. This OCR pre-processing step helps improve the quality of the transcripts, by breaking down each document image into two layers, each containing a particular text type, namely, handwritten text, or machine-printed text, and consequently, feeding the layers into the appropriate OCR or Handwritten Text Recognition (HTR) engines.

The new Transformer-based OCR (TrOCR) models have demonstrated the capability to adapt to variations in fonts, text types, styles, and languages [17, 9]. Utilisation of these models eliminates the need for the pre-requisite steps of dataset synthesis and model training for text type separation.

---

[3]https://www.archivportal-d.de/themenportale/wiedergutmachung

3.1.2. Information Extraction with Rules vs. LLMs

One of the constituent datasets of the *Wiedergutmachung* records is the *"Bundeszentralkartei"*, a card index with more than 2.3 million card files, registering all applications for compensation for National Socialist Injustice. Collected from Compensation Offices across Germany, the cards encompass 32 distinct document types. Traditional rule-based methods require different scripts, customised for each document type, to extract information from these cards. However, with the adoption of LLMs, a paradigm shift is to be expected. Instead of laboriously crafting multiple scripts, LLMs present the opportunity to streamline this process by providing document templates as context. In this LLM-based approach, a unified prompt, coupled with the appropriate context, can facilitate the extraction of information from all cards spanning across the 32 document types.

3.2. *Deutsche Digitale Bibliothek*: Metadata Enhancement with Book Titles

The bibliographical metadata collection of the *DDB* suffers from incompleteness and inaccuracies [28]. Since the age of the objects span several centuries, it leads to uncertainties with respect to authorship and date attributions. In addition, the involvement of numerous libraries poses difficulties in maintaining metadata quality. These challenges are detrimental to user experience.

3.2.1. Information Extraction without Expert Labeled Data

When millions of objects in a collection require further editorial work, a fully manual intervention is laborious and impractical. One solution is to utilise the lengthy titles of pre-modern objects that potentially encode additional pertinent information as in Example 1[4]. What is considered to be a disadvantage by modern bibliographic cataloging standards can be leveraged as inputs to a transformer-based IE pipeline. LLMs have been shown to retrieve some Named Entities on historical documents in non-modern languages in zero-shot setting [21]. By being able to identify which objects have neither Named Entities in the title nor other metadata properties, domain experts can already identify the amount of possible objects that need complete revision.

> **TITLE:** *Die Letzte Predigt,* ***Doctoris Martini Lutheri***<PERSON>, heiliger
> <...> zu **Wittemberg**<GPE> <...> **den 17. Januarij, im 1546. Jar**<DATE>

Example 1: Lengthy title with PERSON, GeoPolitical Entity, and DATE Named Entities.

3.2.2. Efficient Search and Retrieval

Cultural heritage portals like the *DDB* rely heavily on keyword-based filters called facets. Users sift through the search results by using these facets to gauge relevance. This means that the values assigned to the facets have to be precise and curated, as these are normally implemented as drop-down menus. A single-letter difference between the search string and the target object could lead to either elation or frustration. Using LLMs' contextual word embeddings to represent both the search string and the document content for comparison, or concatenating both representations as input to a binary document relevance classifier [29], or estimating the likelihood of the search string given the metadata properties, are just some of the notable examples of the utility of LLMs in efficient search and retrieval.

---

[4]https://ddb.de/item/6563H62JUWEVSVTH3T7TJWCPK2NOMLK7

4. Discussion

Drawing upon our firsthand experiences with transformer-based technologies within the context of *"Themenportal Wiedergutmachung"* and *"Deutsche Digitale Bibliothek"*, in this section, we explore the implications and intricacies associated with the application of these technologies in digitalisation workflows for CH data.

4.1. Opportunities

These technological advancements offer a range of benefits to digitalisation workflows, from increased efficiency and adaptability to improved accuracy and metadata quality.

Increased Efficiency with a unified approach. Transformer-based OCR models have proven to perform well on a variety of text types, without any pre-processing steps [9]. By integrating multiple tasks within a single model, TrOCR unifies the OCR process. The result is a pipeline that eliminates the need for identification of text types and use of distinct engines, reducing resource overhead and thus, an increase in efficiency of the workflow and resource allocation, especially when dealing with large-scale digitalisation projects.

Adaptability and Accuracy. Whether deciphering historical handwritten manuscripts or transcribing printed documents with historical fonts, TrOCR's adaptability ensures more accurate results across the board. Moreover, TrOCR excels in recognising handwritten text – something that often poses challenges for traditional OCR systems [9].

Opting for LLMs over a hybrid approach involving both rule-based and deep-learning-based NLP methods for IE offers an advantage by minimising complexity. This benefit becomes particularly pronounced when dealing with diverse and inconsistent data patterns, necessitating the creation of multiple rule sets due to data heterogeneity.

Improved metadata quality. LLMs can be used for zero-shot IE on historical texts[21]. This can partially aid in enhancing metadata records lacking in context descriptions, such as authorship or date attribution, but it can compensate with more textual content encoded in other properties such as the title.

Search and Retrieval Enhancement. Open-Source LLMs perform relatively well in a zero-shot setting [30] with the Query Likelihood Model used in ranking document relevance according to the probability of generating a query given the content of a document. This lessens the need for overly precise user queries.

4.2. Challenges

The integration of transformer models in the research for and development of digitalisation pipelines is not without its associated costs and challenges. Some of these challenges are outlined below.

Data Privacy. When utilising transformer-based solutions such as OCR services and LLMs, data privacy is a critical concern, given that many of these services are hosted on external servers. Transmission of data to the external servers for processing can potentially expose sensitive information, especially if the input contains proprietary data, or confidential material. On-premise solutions are designed to overcome this drawback. Moreover, it is necessary to secure the infrastructure where the model runs, to ensure data privacy. This might involve using firewalls, and disallowing access to third parties.

Resource Allocation. The computational power required for training and utilising transformer models, and the expenses associated with data storage and "per-token" processing, can strain the budgets of research and GLAM institutions. Hence, a thorough understanding of the associated financial implications is a prerequisite for informed decision making and resource allocation.

Reliability. The performance of transformer models hinges on the quality and diversity of the training data. In the context of CH, the interpretability of transformer-generated outputs becomes crucial for maintaining the integrity and accuracy of processed data. As reliability is paramount in preserving the authenticity of CH, it is essential to establish evaluation mechanisms to increase confidence in digitalisation projects that make use of these models.

Reproducibility. A generative LLM, when asked to respond to the same question repeatedly, will not be able to consistently provide identical answers due to its stochastic nature. For use cases that require precise results, such as metadata enhancement, this can prove detrimental.

Training Data Biases. Inadvertent biases resulting from unfiltered content scraped from the web in training data, can perpetuate societal harm if unaddressed. Various efforts assess biases in transformer models across dimensions such as gender, race, religion, and profession. Additionally, it is crucial to pay attention to social, cultural, and geographic biases in training data [31]. Moreover, application of off-the-shelf LLMs on CH documents with limited accessibility and usually unseen by LLMs, could deteriorate the performance due to popularity bias. Addressing these biases becomes imperative for ethical digitalisation efforts, especially in CH collections.

Overreliance on LLMs. In a short period of time, LLMs experienced improvement by leaps and bounds. However, they still tend to hallucinate or generate convincing answers that are not based on facts. Human oversight is necessary in order to prevent such mistakes. It is important to educate domain experts on the mechanisms behind transformer models in order to drive the point that the purpose of LLMs is not to replace human expertise, but rather, to provide relief from tedious and repetitive tasks.

5. Conclusion

The development of transformer-based technologies has opened new avenues for research in Digital Cultural Heritage. As demonstrated in the two use cases of *"Themenportal Wiedergutmachung"* and *"DDB"*, utilising these technologies streamlines digitalisation processes and offers opportunities to increase accuracy and efficiency. However, moving forward, it is essential to address the challenges associated with integrating these technologies, including data privacy concerns, reliability, and the risk of overreliance on transformer models. Navigating these challenges with a heritage-focused perspective allows for responsible utilisation of transformer-based technologies to advance cultural heritage digitalisation initiatives.

Acknowledgments

References

[1] D. B. Marcum, Digitizing for access and preservation strategies of the Library of Congress, First Monday (2007).

[2] K. J. Borowiecki, T. Navarrete, Digitization of heritage collections as indicator of innovation, Economics of Innovation and New Technology 26 (2017) 227–246.

[3] L. M. Hughes, Digitizing collections: strategic issues for the information manager, volume 2, Facet Publishing, 2004.

[4] J. Purday, Think culture: Europeana.eu from concept to construction, Bibliothek Forschung und Praxis 33 (2009) 170–180. doi:10.1515/bfup.2009.018.

[5] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018).

[6] Open AI, GPT-4 technical report, 2023. URL: https://cdn.openai.com/papers/gpt-4.pdf, [Online; accessed 21-June-2023].

[7] D. Xu, W. Chen, et al., Large language models for generative information extraction: A survey, arXiv preprint arXiv:2312.17617 (2023).

[8] Y. Li, T. Yao, Y. Pan, T. Mei, Contextual transformer networks for visual recognition, IEEE Transactions on Pattern Analysis and Machine Intelligence 45 (2022) 1489–1500.

[9] P. B. Ströbel, T. Hodel, W. Boente, M. Volk, The Adaptability of a Transformer-Based OCR Model for Historical Documents, in: Intl. Conf. on Document Analysis and Recognition, Springer, 2023, pp. 34–48.

[10] A. Vaswani, N. Shazeer, et al., Attention is all you need, Advances in neural information processing systems 30 (2017).

[11] P. Hille, AI: Chatbots replace journalists, 2023. URL: https://www.dw.com/en/ai-chatbots-replace-journalists-in-news-writing/a-65988172, [Online; posted 21-June-2023].

[12] B. D. Lund, T. Wang, et al., ChatGPT and a new academic reality: Artificial Intelligence-written research papers and the ethics of the large language models in scholarly publishing, Journal of the Association for Information Science and Technology 74 (2023) 570–581.

[13] J. G. Meyer, R. J. Urbanowicz, et al., ChatGPT and large language models in academia: opportunities and challenges, BioData Mining 16 (2023) 20.

[14] M. Pasikowska-Schnass, Y.-S. Lim, Artificial intelligence in the context of cultural heritage and museums: Complex challenges and new opportunities, Technical Report PE 747.120, European Parliamentary Research Service, Brussels, 2023.

[15] C. Neudecker, Digital Curation and AI: Opportunities and Risks for Cultural Heritage Institutions, in: S. Thiel, J. C. Bernhardt (Eds.), AI in Museums: Reflections, Perspectives and Applications, transcript Verlag, Bielefeld, 2023, pp. 149–162. doi:10.1515/9783839467107-013.

[16] Z. Tang, Z. Yang, et al., Unifying vision, text, and layout for universal document processing, in: Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition, 2023, pp. 19254–19264.

[17] M. Li, T. Lv, et al., Trocr: Transformer-based optical character recognition with pre-trained models, in: Proc. of the AAAI Conf. on Artificial Intelligence, volume 37, 2023, pp. 13094–13102.

[18] Q. Cao, T. Kojima, Y. Matsuo, Y. Iwasawa, Unnatural error correction: GPT-4 can almost perfectly handle unnatural scrambled text, in: Proc. of the 2023 Conf. on Empirical Methods in Natural Language Processing, 2023, pp. 8898–8913.

[19] N. Borenstein, P. Rust, D. Elliott, I. Augenstein, PHD: Pixel-based language modeling

of historical documents, in: H. Bouamor, J. Pino, K. Bali (Eds.), Proc. of the 2023 Conf. on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Singapore, 2023, pp. 87–107. doi:`10.18653/v1/2023.emnlp-main.7`.

[20] P. Rust, J. F. Lotz, et al., Language modelling with pixels, in: The Eleventh International Conference on Learning Representations, 2022.

[21] F. De Toni, C. Akiki, et al., Entities, Dates, and Languages: Zero-Shot on Historical Texts with T0, in: A. Fan, et al. (Eds.), Proc. of BigScience Episode #5 – Workshop on Challenges & Perspectives in Creating Large Language Models, Association for Computational Linguistics, virtual+Dublin, 2022, pp. 75–83. URL: https://aclanthology.org/2022.bigscience-1.7. doi:`10.18653/v1/2022.bigscience-1.7`.

[22] F. Petroni, T. Rocktäschel, et al., Language Models as Knowledge Bases?, in: K. Inui, J. Jiang, V. Ng, X. Wan (Eds.), Proc. of the 2019 Conf. on Empirical Methods in Natural Language Processing and the 9th Intl. Joint Conf. on Natural Language Processing (EMNLP-IJCNLP), Association for Computational Linguistics, Hong Kong, China, 2019, pp. 2463–2473. doi:`10.18653/v1/D19-1250`.

[23] P. Lewis, E. Perez, et al., Retrieval-augmented generation for knowledge-intensive nlp tasks, in: H. Larochelle, et al. (Eds.), Advances in Neural Information Processing Systems, volume 33, Curran Associates, Inc., 2020, pp. 9459–9474.

[24] M. Vafaie, O. Bruns, N. Pilz, J. Waitelonis, H. Sack, CourtDocs Ontology: Towards a Data Model for Representation of Historical Court Proceedings, in: Proc. of the 12th Knowledge Capture Conference 2023, 2023, pp. 175–179.

[25] M. Vafaie, O. Bruns, N. Pilz, D. Dessí, H. Sack, Modelling Archival Hierarchies in Practice: Key Aspects and Lessons Learned, in: 6th Intl. Workshop on Computational History (HistoInformatics 2021), Online event, September 30-October 1, 2021, volume 2981, Aachen, Germany: RWTH Aachen, 2021, p. 6.

[26] M. Vafaie, O. Bruns, N. Pilz, J. Waitelonis, H. Sack, Handwritten and printed text identification in historical archival documents, in: Archiving Conference, volume 19, Society for Imaging Science and Technology, 2022, pp. 15–20.

[27] M. Vafaie, J. Waitelonis, H. Sack, Improvements in Handwritten and Printed Text Separation in Historical Archival Documents, in: Archiving Conference, volume 20, Society for Imaging Science and Technology, 2023, pp. 36–41.

[28] Tan, Mary Ann and Jiang, Shufan and Sack, Harald, How to Turn Card Catalogs into LLM Fodder, in: Deep Learning and Linguistic Linked Data (DLnLD) Workshop at LREC-COLING, 2024.

[29] R. Nogueira, W. Yang, K. Cho, J. Lin, Multi-Stage Document Ranking with BERT, 2019. `arXiv:1910.14424`.

[30] S. Zhuang, B. Liu, B. Koopman, G. Zuccon, Open-source Large Language Models are Strong Zero-shot Query Likelihood Models for Document Ranking, in: H. Bouamor, J. Pino, K. Bali (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2023, Association for Computational Linguistics, Singapore, 2023, pp. 8807–8817. doi:`10.18653/v1/2023.findings-emnlp.590`.

[31] I. O. Gallegos, R. A. Rossi, et al., Bias and fairness in large language models: A survey, arXiv preprint arXiv:2309.00770 (2023).