KI-generierte Objektbeschreibungen für schriftliches Kulturerbe

Eine Pilotstudie

1. Einleitung

Das Ziel der Pilotstudie[1] war es, den Einsatz von KI-gestützten Verfahren bei der Vermittlung von schriftlichem Kulturerbe für ein diverses Zielpublikum zu erproben.[2]

Gedächtniseinrichtungen wie Bibliotheken, Sammlungen, Archive und Museen erfassen seit vielen Jahren in großem Umfang ihre wertvollen Bestände digital und präsentieren sie in Online-Portalen oder Repositorien. Die Datensätze bestehen aus digitalen Abbildungen und Metadaten (z.B. Angaben zur Entstehungszeit, zum Entstehungsort, zu enthaltenen Autor*innen und Texten oder objektbiografische Informationen wie Vorbesitzer*innen und Aufbewahrungsorte). Diese Metadaten enthalten oft wissenschaftliche Terminologie, Abkürzungen, werden tabellarisch oder in anderer Form strukturiert dargestellt und verlinken auf weitere Metadaten in vergleichbarer Form (z.B. Katalogbeschreibungen). Eine derartige Darstellung und Aufbereitung scheint im Regelfall wenig geeignet, um ein breites Publikum anzusprechen. Um die Informationen und das Wissen über die Kulturobjekte mit möglichst vielen interessierten Menschen zu teilen und so die Zugänglichkeit des kulturellen Erbes zu verbessern, ist daher eine alternative Präsentation erforderlich. Die nachträgliche manuelle Aufbereitung von Objektinformationen, insbesondere das Erstellen von ansprechenden Beschreibungstexten, ist im großen Stil aufgrund der Mengenverhältnisse jedoch nicht bzw. nur mit unverhältnismäßig hohem Aufwand möglich.

Im Rahmen dieser Studie wurde deshalb eine automatisierte Erzeugung von Beschreibungstexten kuratorischer Art auf der Basis vorhandener Metadaten untersucht. Der Ansatz, generative KI in Form sogenannter großer Sprachmodelle (Large Language Models, LLMs) zur Vermittlung von Kulturgut einzusetzen, versprach dabei eine hochgradig skalierbare Lösung, um existierende Kulturdaten allgemeinverständlich aufzubereiten und die so entstandenen Beschreibungstexte für unterschiedliche Zielgruppen bereitstellen zu können (z.B. Leichte Sprache, Übersetzungen).

1 Die Pilotstudie wurde vom Hessischen Ministerium für Wissenschaft und Forschung, Kunst und Kultur (HMWK) in dem Projekt der Universitätsbibliothek Marburg „MATE – Maschinell erstellte Begleittexte für Kulturobjekte mittels generativer künstlicher Intelligenz“ (2023-2024) für die Dauer von sechs Monaten gefördert und vom Marburg Center for Digital Culture and Infrastructure (MCDCI) der Philipps-Universität Marburg wissenschaftlich begleitet; vgl. https://www.uni-marburg.de/de/ub/forschen/digitalisierung/projekte/mate, Stand: 21.07.2024. Das Projekt wurde im Wintersemester 2023/24 flankiert durch eine Lehrveranstaltung im M.A.-Studiengang „Cultural Data Studies“, Projektmodul: „Kulturobjekte und Künstliche Intelligenz“ sowie durch die Veranstaltung „KI meets Kulturerbe“ am 9. Februar 2024 an der Philipps-Universität Marburg.

2 Dieser Beitrag basiert auf dem Vortrag der Mitautorin Diana Müller „Kulturerbe mittels KI zielgruppenspezifisch vermitteln“ am 05.06.2024 auf der 112. BiblioCon2024 in Hamburg.

2. Objektart, Datenquelle und Testcorpus

Für die Pilotstudie wurde bewusst nicht das gesamte Spektrum schriftlichen Kulturerbes in den Blick genommen, sondern exemplarisch eine Objektart betrachtet. Der Fokus wurde dabei auf mittelalterliche Buchhandschriften gelegt. Bücher und Fragmente aus der Zeit vor Erfindung und Verbreitung des Buchdrucks sind unikale Objekte und stellen bereits aufgrund ihrer Einmaligkeit einen besonders bedeutenden Teil der schriftlichen kulturellen Überlieferung dar. Dass diese Schriftzeugnisse über Jahrhunderte hinweg überliefert und erhalten geblieben sind und so Einblicke in vormoderne Zeiten, Kulturen und Gesellschaften ermöglichen, macht sie für viele Menschen zu faszinierenden Objekten. Das vergleichsweise große Publikumsinteresse an mittelalterlichen Büchern sorgt für einen erhöhten Vermittlungsbedarf, so dass sie als Objektart für diese Pilotstudie besonders geeignet erschienen.

Als Datenquelle für die Erprobung des Verfahrens diente das Repositorium „Corvey digital“[3], das von der Universitätsbibliothek Marburg betrieben wird und eine Online-Datenbank zum Nachweis mittelalterlicher Handschriften mit Provenienz der ehemaligen Reichsabtei Corvey darstellt. Das Repositorium basiert technisch auf der Software DSpace 6.3 und nutzt im Wesentlichen Dublin Core als Metadatenschema.[4]

Um die verschiedenen Besonderheiten der Objektart „mittelalterliche Buchhandschrift“ abzudecken und zugleich konsistente und vergleichbare Ergebnisse zu gewährleisten, wurde nach entsprechenden Vortests ein aussagekräftiges, festes Set von insgesamt drei Testobjekten aus dem Repositorium definiert:

- Marburg, Universitätsbibliothek, Ms. 49[5]
- Kassel, Universitäts-, Landes- und Murhardsche Bibliothek, 2o Ms. theol. 60[6]
- Bad Homburg v. d. Höhe, Stadtarchiv, S 08 Hss.-Fragmente, Nr. 16[7]

Diese Auswahl ermöglichte es, die Vielfalt der mittelalterlichen Buchüberlieferung exemplarisch abzubilden und die Leistungsfähigkeit der Sprach-KI bei der Generierung von Beschreibungstexten für sehr unterschiedliche Typen von Handschriften (gewöhnliche Sammelhandschrift, liturgisches Buch mit Prachteinband, Fragment) zu evaluieren.

3 Corvey digital, https://corvey.ub.uni-marburg.de/, Stand: 21.07.2024.

4 Die Entwicklung von „Corvey digital“ erfolgte im Rahmen des DFG-Projekts „Die mittelalterlichen Buchhandschriften der Klosterbibliothek Corvey digital“ (Laufzeit: 2020–2021) an der Universitätsbibliothek Marburg. Siehe auch: Maul, Alexander; Müller, Diana: Das Portal „Corvey digital“ – Ein Digitalisierungsprojekt an der Universitätsbibliothek Marburg, in: AKMB-news 29, 2023, S. 49–53. Maul, Alexander; Müller, Diana: „Corvey digital“. Werkstattbericht zum Webportal für mittelalterliche Handschriften mit Provenienz der ehemaligen Reichsabtei Corvey, in: Maniculae 3, 2022, S. 42–47. https://doi.org/10.21248/maniculae.35.

5 Marburg, Universitätsbibliothek, Ms. 49, http://dx.doi.org/10.48643/b4tm-32.

6 Kassel, Universitätsbibliothek, Landesbibliothek und Murhardsche Bibliothek, 2o Ms. theol. 60, http://dx.doi.org/10.48643/b4tm-80.

7 Bad Homburg v. d. Höhe, Stadtarchiv, S 08 Hss.-Fragmente, Nr. 16, http://dx.doi.org/10.48643/b4tm-59.

3. Vorarbeiten

Neben den zu beschreibenden Kulturerbeobjekten mussten auch die KI-Werkzeuge, d. h. die zu verwendenden Sprachmodelle, für die Pilotstudie bestimmt werden. Die Testreihen wurden zunächst unter Nutzung des Modells GPT-4 der Firma OpenAI durchgeführt, das zum Zeitpunkt der Technologiewahl als leistungsfähigstes am Markt verfügbares Sprachmodell galt. Da es sich bei GPT-4 jedoch um ein proprietäres, kommerzielles Modell handelt, das ausschließlich auf den Servern des Anbieters betrieben werden kann, wurden ergänzend dazu freie Modelle ausgewählt, deren Lizenz insbesondere auch eine lokale Nutzung in der eigenen Infrastruktur der Universitätsbibliothek bzw. des Hochschulrechenzentrums der Universität Marburg ermöglichte: Llama 2 (Meta/Microsoft), Falcon (Technology Innovation Institute) und Mixtral (Mistral AI).

Die Verwendung von GPT-4 erfolgte über eine lizenzpflichtige Programmierschnittstelle (API). Unter Verwendung des Schlüssels für den Zugang zu den API-Diensten wurden Programmroutinen für die Nutzung der Programmierschnittstelle (API) der Firma OpenAI eingerichtet, um darüber eine automatisierte Nutzung des GPT-4-Modells für die Erzeugung von Texten zu ermöglichen. Parallel wurden Programmroutinen für den automatisierten Abruf der benötigten Testdatensätze aus dem „Corvey digital"-Repositorium erstellt, die ihrerseits auf der entsprechenden Programmierschnittstelle (REST-API[8]) der Repositoriumssoftware DSpace (Version 6 und 7) basierten. Ergänzend wurden zudem im weiteren Verlauf des Projektes Programmroutinen für den Re-Import generierter Texte (bzw. für Metadaten allgemein) in Repositorien auf DSpace-Basis erstellt, die es perspektivisch erlauben, die generierten Beschreibungstexte wiederum automatisiert in den Objektmetadaten der Datenquelle zu verankern. Das entsprechende Python-Paket wurde unter MIT-Lizenz auf GitHub und pypi.org zur freien Nutzung veröffentlicht.[9]

Für die systematische Dokumentation aller Abfragen und die Evaluierung der generierten Texte wurde ein internes Dokumentationsformular entwickelt. Dieses Formular erfasst die Abfragen, verwendeten Parameter, Anweisungen an das Modell (Prompts) und die generierten Texte. Dadurch wurde eine systematische Erfassung und Analyse aller Interaktionen mit den LLMs gewährleistet. Die Evaluierung der Texte basierte auf Kriterien wie Relevanz, Neutralität, Wertungsfreiheit, sachlicher Korrektheit und angemessener Terminologie sowie Stil und Tonalität. Nach dem Ampel-Modell wurden Sätze in grün (brauchbar), gelb (noch brauchbar) oder rot (unbrauchbar) klassifiziert.

Während bei der Nutzung der GPT-4-API (und vergleichbarer kommerzieller Modelle) alle aufwendigen Rechenoperationen des LLMs ausschließlich in der Cloud-Infrastruktur des Anbieters ausgeführt werden, erfordert die Verwendung von lokal ausführbaren großen Sprachmodellen besondere Ressourcenausstattung vor Ort. Die Ausführung (Inferenz), insbesondere aber das Training entsprechender Modelle benötigen überaus viel Arbeitsspeicher und Rechenkapazität, wobei letztere aus Effizienzgründen in der Regel nicht von regulären Hauptprozessoren (CPUs), sondern

8 „REST" steht für „Representational State Transfer" und stellt ein äußerst weit verbreitetes Paradigma für die Architektur der Programmierschnittstellen von webbasierten Anwendungen dar.
9 Vgl. https://github.com/ub-unimr/dspyce, Stand: 21.07.2024.

von optimierten Prozessortypen (GPUs bzw. spezialisierter Hardware) in entsprechenden Cluster-Setups stammt. Da die Pilotstudie allerdings bewusst kein eigenes Training von Modellen umfasste und aufgrund des Testcorpus die Anzahl der zu erstellenden Texte insgesamt überschaubar blieb, konnten die ausgewählten Modelle auf entsprechend gut ausgestatteten PCs in der Infrastruktur der Universitätsbibliothek ausgeführt werden.

Auf der Grundlage der durchgeführten Vortests wurden die umfangreichen Metadatensätze der Testobjekte intellektuell gesichtet und eine Auswahl solcher Felder identifiziert, die durch ihren Informationsgehalt für die Beschreibungstexte besonders sinnvoll erschienen und zugleich von der KI potenziell gut zu verarbeiten waren. Zum Abruf dieser Daten wurde die o. g. Programmroutine für den Datenabruf über die REST-API des „Corvey digital"-Repositoriums genutzt, die Daten dabei entsprechend gefiltert und anschließend im JSON-Format[10] gespeichert. JSON wurde als Verarbeitungsformat ausgewählt, da es aufgrund seiner hohen Verbreitung in den Trainingsdaten von LLMs gut repräsentiert und mithin für die Modelle „leicht verständlich" ist. Bei der Überführung in das JSON-Format mussten jedoch insbesondere die originären Feldbezeichnungen angepasst werden, um den Zweck des jeweiligen Feldes für die Modelle mehr oder weniger unmittelbar verständlich zu machen und um eine korrekte Verwendung der Terminologie im Deutschen zu unterstützen. So wurden beispielsweise die Feldbezeichnungen „repository" in „aufbewahrende Einrichtung", „medium" in „Material" (in der Regel Papier oder Pergament), „created" in „Herstellungsdatum" der Handschrift, „signature" in „Signatur", „provenance" in „Provenienz", „title_alternative" in „alternative oder nichtkanonische Titel", „language" in „Sprachen", „format" in „Gattung", „origin" in „Herstellungsort" und „authors" in „Autoren" geändert.

4. Durchführung

Der erprobte Workflow zur Texterzeugung umfasste mehrere Schritte (vgl. Abb. 1): (1) Vom Abruf der Metadaten aus dem ursprünglichen Repositorium via API, (2) der Transformation der entsprechenden Daten im JSON-Format, (3) dem Kombinieren der Metadaten mit den entsprechenden Anweisungen („Prompts") bis hin zur (4) Übergabe an die API für die Generierung von Beschreibungstexten durch die Large Language Models (LLMs). Die auf diese Weise erzeugten und ausgegebenen (5) Beschreibungstexte wurden (6) systematisch gesichert.

10 „JSON" steht für „JavaScript Object Notation" und stellt ein weit verbreitetes und leichtgewichtiges Format für die maschinenlesbare Codierung von strukturierten Daten, z.B. zur Verwendung in APIs, dar.

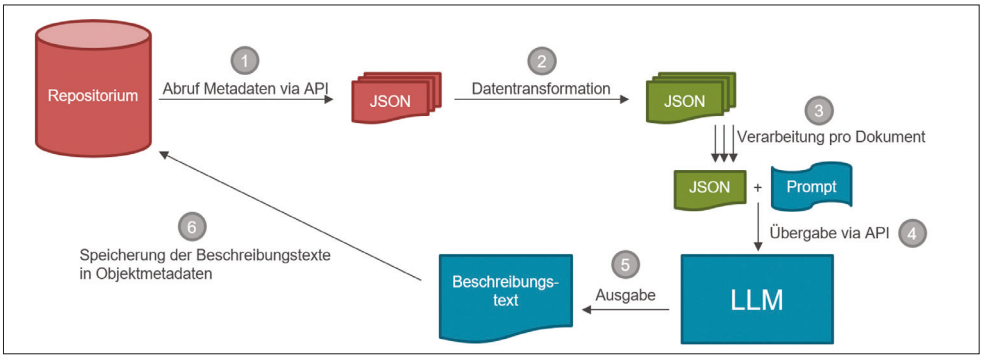

Abb. 1: Workflow der Texterzeugung

Ein Beispiel für abgerufene Metadaten vor ihrer Filterung bzw. weiteren Verarbeitung ist der folgenden Abbildung zu entnehmen:


```
…
"name": "Franciscus Galuani de Ianua - Johannes
Herolt - Theodulus u. a.",
"authors": "Franciscus, Galvanus (um 1348/52),
Herolt, Johannes (-1468),
Nicolaus, de Graetz (-1444), Gerson, Jean (1363-
1429), Johannes, de Francfordia (1380-1440), Paul
II., Papst (1417-1471), Theodulus, Italus (None)",
"location": "Universitätsbibliothek Marburg",
"medium": "paper",
"created": "1444-1470",
"signature": "Ms. 49",
"provenance": "Kloster Corvey and Campill,
Severus",
"title_alternative": [],
"language": ["Latein","Deutsch"],
"format": "codex",
"origin":[{"name":"Deutschland","gndId": "4011882-
4"}]
…
```


Abb. 2: Metadaten (Auszug)

Einen wesentlichen Aspekt bei der Generierung von Ausgabetexten durch Sprachmodelle stellt das Prompten dar, also das Formulieren einer Handlungsanweisung an das jeweilige Modell. Von diesen Anweisungstexten hängt die Qualität der Ausgabe des LLMs, im vorliegenden Fall also der Objektbeschreibungen, maßgeblich ab. Durch die verbreitete Nutzung von LLMs haben sich eine Reihe von typischen Techniken des Promptings entwickelt, die sich gezielt die Funktionsweise der Modelle zunutze machen bzw. experimentell als besonders tauglich für bestimmte Aufgabenstellungen erwiesen haben. Die Herausforderung im Rahmen der Pilotstudie war, dass interaktive Prompting-Methoden (z.B. das sog. Chain-of-Thought-Prompting) nicht infrage kamen, da sie sich nicht für ein automatisiertes Verfahren eignen. Die Prompts wurden aber testweise auf Deutsch und auf Englisch formuliert. Dabei zeigten sich grundsätzlich bessere Ergebnisse, wenn die Eingabeaufforderung in englischer Sprache formuliert wurde, wobei es dennoch möglich war, den generierten Text selbst in Deutsch auszugeben. Beim Erstellen einer solchen Eingabeaufforderung erwies es sich außerdem als

besonders zielführend, dem Modell gewissermaßen eine Rolle zuzuweisen (z.B. „as historian“) und die Textsorte (z.B. „description“) sowie weitere gewünschte Eigenschaften des Textes (z.B. „easy to understand“) möglichst präzise und explizit zu beschreiben.

Einen der frühen verwendeten Beispielprompts stellt die folgende Abbildung dar:

Prompt instruction:
As an Historian write an interesting description in German language which is easy to understand for the medieval manuscript by using the metadata […]

Parameter Temperature:
0.7

Model:
GPT 4

Abb. 3: Prompting-Beispiel

Das zugehörige Ausgabe-Beispiel für die Metadaten (vgl. Abb. 2) und den Prompt (vgl. Abb. 3) ist schließlich der folgenden Darstellung zu entnehmen:

Das Manuskript "Franciscus Galvani de Ianua - Johannes Herolt - Theodulus u. a." ist ein bemerkenswertes historisches Dokument aus der Mitte des 15. Jahrhunderts (1444-1470). Es wurde von einer Reihe von Autoren verfasst, darunter Franciscus Galvanus, Johannes Herolt, Nicolaus de Graetz, Jean Gerson, Johannes de Francfordia, Papst Paul II. und Theodulus Italus. Dieses Manuskript, das auf Papier geschrieben und im Codex-Format präsentiert wurde, befindet sich derzeit in der Universitätsbibliothek Marburg unter der Signatur "Ms. 49".

Das beeindruckende Dokument hat seinen Ursprung in Deutschland, wie aus den Metadaten hervorgeht. Es wurde sowohl auf Latein als auch auf Deutsch verfasst, was auf die weit verbreitete Verwendung dieser Sprachen in der Wissenschaft und Theologie dieser Zeit hinweist. Die Provenienz des Manuskripts weist auf das Kloster Corvey und Campill, Severus hin, was seine historische Bedeutung und seinen geistlichen Kontext unterstreicht.

Das Manuskript ist ein Zeugnis der mittelalterlichen Gelehrsamkeit und Theologie und bietet einen tiefen Einblick in die Gedanken und Diskurse jener Zeit. Es repräsentiert eine Epoche, in der das Schreiben und Kopieren von Manuskripten eine wichtige kulturelle Praxis war, die Wissen und Ideen über Generationen hinweg bewahrte und verbreitete. Es ist ein unverzichtbares Zeugnis für die Geschichte der Schriftlichkeit und Gelehrsamkeit im mittelalterlichen Europa.

*Abb. 4: Ausgabe-Beispiel (Farbige Hervorhebungen von Problemstellen wurden durch die Autor*innen vorgenommen und entstammen nicht der Textausgabe des Modells)*

Erste systematische Testreihen wurden auf der Grundlage des Modells GPT-4 durchgeführt, da dieses zum Zeitpunkt der Durchführung gemeinhin als leistungsfähigstes allgemein verfügbares LLM galt. Die gewonnenen Beschreibungstexte und Beobachtungen aus der Textgenerierung mit diesem Modell wurden ausführlich in einem speziellen Berichtsformat dokumentiert und ausgewertet. Dabei konnten Fehlermuster identifiziert werden, z.B. beziehen sich die Metadaten zu den ausgewählten mittelalterlichen Büchern auf zwei separate Beschreibungsebenen, namentlich die Objektebene (das Buch als materielles Objekt) und die Inhaltsebene (enthaltene Texte), die es strikt zu trennen gilt,

was den Sprachmodellen in ersten Iterationen der Tests unzureichend gelang. Ein typisches Fehlermuster zeigte sich in der falschen Anwendung moderner Formulierungen auf vormoderne Kontexte. Beispielsweise ist die Aussage „Autor schreibt Buch" korrekt in Formulierungen wie „Rainald Goetz verfasste das Buch *Abfall für alle*" oder „Benjamin von Stuckrad-Barre schrieb das Buch *Noch wach?*". Diese Formulierungen bzw. die zugrundeliegenden Konzepte sind in den Trainingsdaten der KI offensichtlich weit verbreitet und bekannt. Übernimmt die KI jedoch diese modernen Formulierungen für mittelalterliche Bücher, führt dies zu Fehlern. Ein Beispiel hierfür wäre die Aussage „Franciscus Galvanus schrieb die Handschrift Ms. 49". Diese Aussage ist nicht nur deshalb falsch, weil Franciscus Galvanus bereits nicht mehr lebte, als die Handschrift Ms. 49 im 15. Jahrhundert entstand. Das vorliegende Problem besteht vielmehr darin, dass „Autor" und „Schreiber" zwei unterschiedliche Rollen sind, die es strikt zu trennen gilt. Der Autor ist der ursprüngliche Verfasser eines überlieferten Textes, während der Schreiber derjenige ist, der den Text kopiert oder niederschreibt. Eine korrekte Formulierung wäre: „Die Handschrift Ms. 49 entstand im 15. Jahrhundert und enthält einen Text von Franciscus Galvanus."

Unbrauchbar war darüber hinaus in ersten Tests oftmals die Relevanzgewichtung von bezifferten Informationen. So wurden nebensächliche Informationen wie z.B. die Abmessung des Objekts nach Länge und Breite in Zentimetern in den ausgegebenen Beschreibungstexten unangemessen prominent platziert und überdeutlich ausformuliert. Außerdem zeigten die Ausgabetexte terminologische Fehler: Z.B. wurde die für die Objektgattung „mittelalterliche Buchhandschriften" gängige Blattzählung als Seitenzählung ausgegeben, was im Ergebnis zu Falschaussagen führt (z.B. 188 Seiten statt 188 Blatt [entsprechend 276 Seiten nach moderner Zählweise]). Nicht zufriedenstellend war zudem die übermäßige Verwendung wertender Adjektive (z.B. „*bemerkenswertes* historisches Dokument", „*tiefe* Einblicke", „*unverzichtbares* Zeugnis").

Diese Probleme wurden im Rahmen der Tests nicht nur identifiziert, sondern auch gründlich analysiert, um iterativ neue Ansätze für das weitere Prompting bzw. die Vorverarbeitung von Metadaten zu entwickeln sowie die Konfiguration von Parametern zu optimieren.[11] So wurde beim weiteren Prompten z.B. auf die explizite Aufforderung verzichtet, eine „interessante" Beschreibung zu erstellen. Im Ergebnis führte dies zu den gewünschten sachlicheren Beschreibungstexten. Überraschenderweise führte die explizite Anweisung, eine „neutrale" oder „sachliche" Beschreibung zu generieren, nicht zu diesem Effekt. Gleiches galt für negative Aufforderungen („Vermeide wertende Formulierungen!"), was die Schwierigkeit verdeutlicht, entsprechende Prompts zu formulieren.

5. Erprobung freier Modelle

Wie eingangs beschrieben, sollten neben GPT-4 als kommerziellem Modell bewusst auch freie und lokal ausführbare LLMs in die Studie einbezogen werden, um ein breiteres Bild bezüglich potenziell einsetzbarer Modelle und ihrer Leistungsfähigkeit, aber auch bezüglich möglicher Stärken und

11 Das Absenken des sog. *temperature*-Wertes, der sich bei gleichen Prompts auf die Varianz der Ausgabe auswirkt, hatte beispielsweise kaum positive Effekte, insbesondere was die häufige Verwendung wertender Adjektive betrifft. Zu den Parametern von GPT-4 s. https://platform.openai.com/docs/api-reference/chat/create, Stand: 21.07.2024.

Schwächen zu erhalten. Aus einem sich sehr dynamisch entwickelnden Markt für diese Modelle[12] wurden Llama 2, Falcon und Mixtral für die Pilotstudie ausgewählt.

Grundsätzlich lässt sich festhalten, dass die freien Modelle hinsichtlich wertender Adjektive und ihrer Anpassbarkeit durch Parameter deutlich besser auf Konfigurationsänderungen reagierten als GPT-4. Die niedrig angesetzten *temperature*-Werte bewirkten – wie zu erwarten – durch den Verzicht auf Varianz die Beschränkung auf die Ausformulierung der übergebenen Metadaten-Inhalte. Dadurch waren die Ergebnisse zugleich naturgemäß deutlich stärker von der vorangehenden Auswahl und Aufbereitung der Metadaten abhängig.

Das Trainingskorpus der freien Modelle wird im Vergleich zu den GPT-Modellen wohl von englischsprachigen Texten dominiert. Die freien Modelle wiesen jedenfalls Schwierigkeiten bei der Generierung deutscher Texte auf. Häufig traten Grammatik-, aber auch Ausdrucksfehler auf. Dem konnte durch einen zweistufigen Ansatz begegnet werden: Die Modelle wurden angewiesen, die Beschreibungstexte in englischer Sprache zu generieren; anschließend wurden diese Texte in einem zweiten Schritt wiederum maschinell ins Deutsche übersetzt.

Abgesehen von den bereits genannten Aspekten unterschieden sich die Modelle auch durchaus in ihren Ausgaben: So fiel beispielsweise die Textlänge – also die Ausführlichkeit der Antworten – der mit Mixtral generierten Beschreibungen um durchschnittlich ein Drittel höher aus als bei ihren Pendants von Falcon und Llama 2, wobei insbesondere die von Falcon generierten Texte oft deutlich kürzer gerieten. Dies korrespondierte zugleich mit der Beobachtung, dass auch bei Mixtral hin und wieder wertende Aussagen zu beobachten waren, wenn auch deutlich seltener als bei GPT-4. Dieses Verhalten wurde bei den anderen beiden freien Modellen im Rahmen der Testläufe nicht beobachtet.

Die Ergebnisse der getesteten freien Modelle erscheinen trotz der aufgezeigten Mängel erfolgversprechend. Sowohl Llama 2 als auch Falcon sind in der Lage, allein auf Basis der übergebenen Informationen sinnvolle Texte zu generieren. Auch Mixtral ließ sich durch Anpassungen am Prompting sowie an der Konfiguration in dieser Hinsicht im Zuge der Tests deutlich verbessern. Die freien Modelle schafften es jedoch ebenso wenig wie GPT-4, die erforderliche Differenzierung der Beschreibungsebenen vorzunehmen. Überdies ist bei einem Einsatz von freien Modellen grundsätzlich zu bedenken, dass für eine lokale Ausführung in Massenverfahren große Rechenkapazitäten zur Verfügung stehen müssen.

6. Ergebnisse und Perspektiven

Die Pilotstudie zielte darauf ab, die Erzeugung von Beschreibungstexten aus Objekt-Metadaten zu testen und dabei Herausforderungen und Probleme zu identifizieren. Die Resultate zeigen, dass dieser Ansatz grundsätzlich vielversprechend ist, um zusätzliche Präsentationsformen für große Mengen digitalisierter Kulturgutobjekte zu entwickeln und verschiedene Zielgruppen effektiv anzusprechen.

12 Vgl. etwa die Auswahl von Textgenerierungs-Modellen auf dem einschlägigen Portal Hugging Face, https://huggingface.co/models?pipeline_tag=text-generation, Stand: 21.07.2024.

Allerdings wurden auch signifikante Herausforderungen aufgedeckt. Die mittelalterliche Buchkultur, eine sehr spezifische Wissensdomäne, ist in den Trainingsdaten großer Sprachmodelle (LLM) nur unzureichend repräsentiert – analog wird dies auch für andere Domänen der kulturellen Überlieferung gelten. Das in den Modellen kodierte Wissen über vergangene Epochen reicht im Ergebnis nicht aus, um präzise und terminologisch korrekte Texte über historische Objekte zu generieren, die als Grundlage für eine automatisierte Erzeugung zielgruppenspezifischer Derivate taugen, wie Texte in Leichter Sprache oder Übersetzungen. Zudem stellt das Genre der Objektbeschreibung[13] eine sehr spezielle Textsorte dar, die in den Trainingsdaten der LLMs kaum vorkommt.

Um diese Herausforderungen zu bewältigen, ist eine doppelte Optimierung notwendig. Erstens kann durch gezieltes nachträgliches Training der vorhandenen Modelle eine bessere Anpassung an die Anforderungen der Beschreibungstexte erreicht werden. Hierfür müssen geeignete Trainingsdaten, also Mustertexte, zusammengestellt werden. Zweitens kann durch die Anbindung geeigneter externer Datenquellen das domänenspezifische Kontextwissen der Modelle verbessert werden. Ein vielversprechender Ansatz hierfür ist das sogenannte „Retrieval Augmented Generation“ (RAG), bei dem Vektordatenbanken oder Wissensgraphen im Textgenerierungsverfahren gezielt abgefragt und eingebunden werden. Dies ermöglicht die dynamische Einbindung von domänenspezifischem Basiswissen und faktualistischen Informationen, wie Lebensdaten von Verfasser*innen oder anderen beteiligten Personen.

Zusammenfassend lässt sich festhalten, dass die Pilotstudie innovative Ansätze und Herausforderungen an der Schnittstelle von Kultur und Technologie beleuchtete und das Bewusstsein für die Potenziale der KI bei der Vermittlung von Kulturerbe stärkte. Mit gezielten Optimierungen und einem klaren Fokus auf Verständlichkeit und Zielgruppenorientierung kann die Integration von KI die künftige Präsentation von Kulturgut prägen und eine breite und vielfältige Rezeption durch ein diverses Publikum ermöglichen.

Corinna Berg, *Marburg Center for Digital Culture and Infrastructure, Philipps-Universität Marburg*, https://orcid.org/0009-0003-4021-0475

Eike Löhden, *Universitätsbibliothek Marburg*, https://orcid.org/0000-0001-9315-3660

Diana Müller, *Universitätsbibliothek Marburg*, https://orcid.org/0000-0003-0092-5217

Tobias Müllerleile, *Universitätsbibliothek Marburg*, https://orcid.org/0000-0002-1609-4836

Zitierfähiger Link (DOI): https://doi.org/10.5282/o-bib/6076

13 Gemeint sind kuratorische Texte, wie sie zum Beispiel in Ausstellungskatalogen zu finden sind.