Erfahrungen und Überlegungen zur Verwendung von LLMs zur automatisierten Datenextraktion von biografischen Informationen

Autor:innen
Klaus Schmidt, Anke Silomon, Dirk Goldhahn, Franziska Naether, Peter Mühleder
(Sächsische Akademie der Wissenschaften zu Leipzig)

Der Knowledge Graph

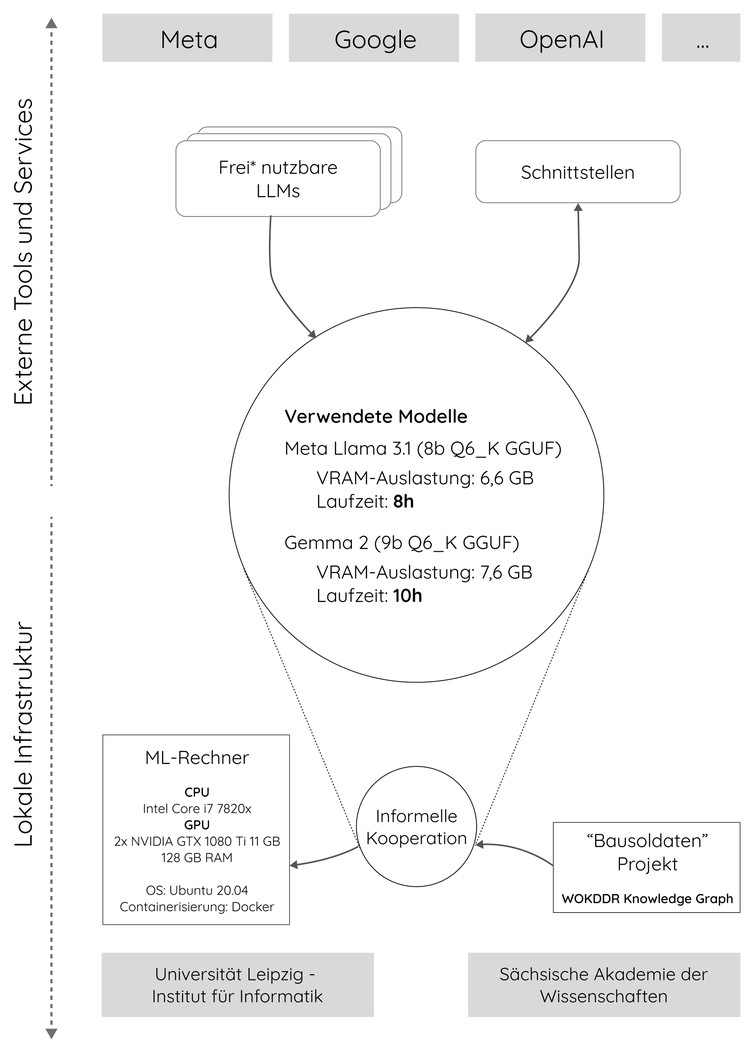
WOKDDR ist eine digital verfügbare Forschungsplattform, die grundlegende Informationen über kirchliche und nichtkirchliche Akteur:innen, Gruppen sowie Forschungsliteratur zur Geschichte der Kirchen in der DDR sowie der Demokratie- und Oppositionsbewegung enthält. Sie ist Teil der Projekte "Kirchliche Praxis in der DDR" (2023-2024) und "Umgang mit Andersdenkenden und die Konsequenzen" speziell zu Bausoldaten (2024-2025), beide gefördert mit Landesmitteln des SMWK.

Problemstellung

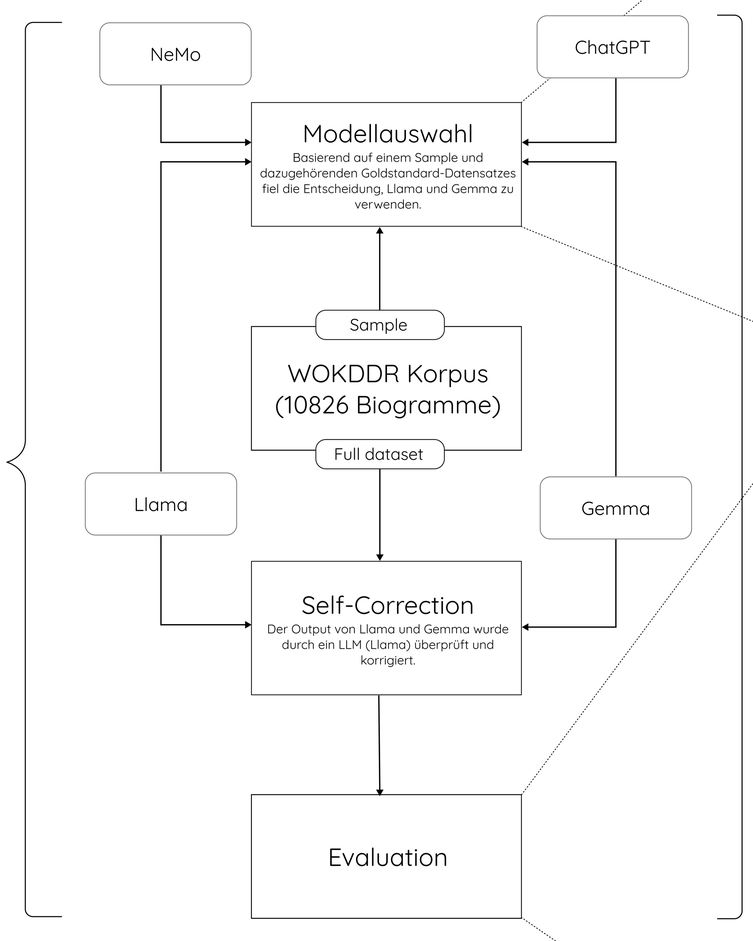
Die Biogramme im Knowledge Graph enthalten stichpunktartige Informationen zu den beruflichen Lebensläufen. In diesen Angaben sollen nun automatisiert Zeitangaben, Berufsangaben, Orte und Organisationen identifiziert und mit den entsprechenden Entitäten im Knowledge Graph verknüpft werden.

Datenextraktion mit LLMs

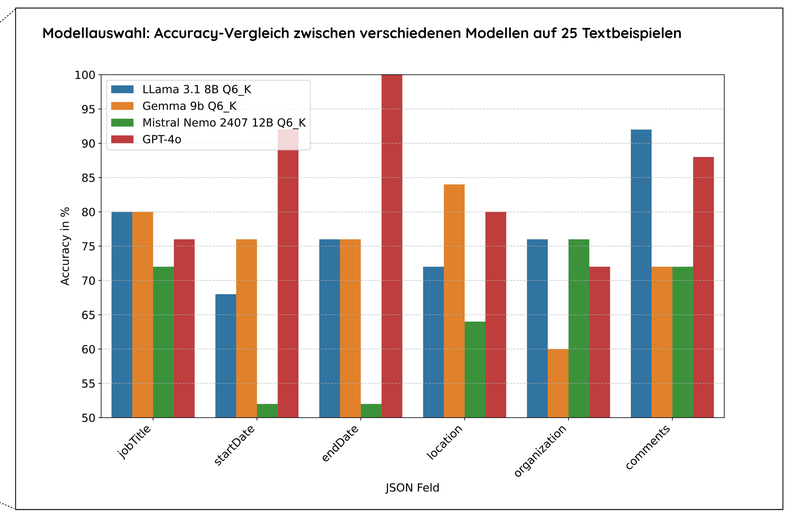
Infrastruktur

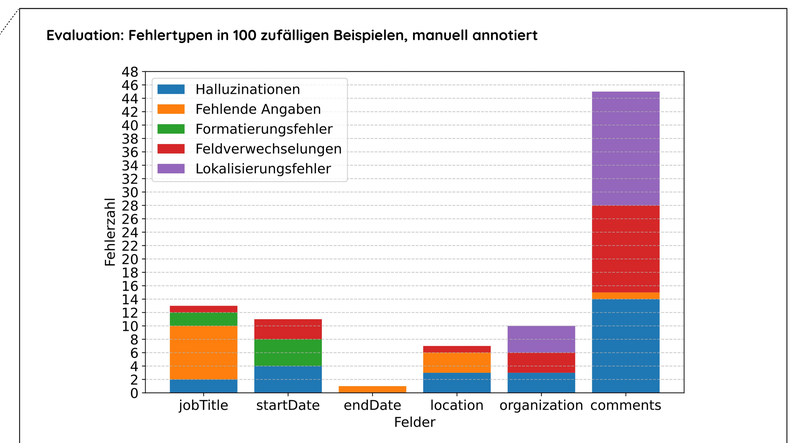


Workflow



Ergebnisse





Halluzinationen
Hinzufügen von Informationen, die nicht in der Quelle enthalten sind

Fehlende Angaben
Weglassen wichtiger Informationen

Formatierungsfehler
Format- oder Orthographie-Fehler

Feldverwechslungen
Informationen in die falschen Kategorien eingeordnet

Lokalisierungsfehler
Informationen in nichtdeutscher Lokalisierung

"1990-1992: Vikariat in Kieve"
Output: {"jobTitle": "Vikariat", "startDate": "1990", "endDate": "1992", "location": "Kiew"}

"1916 Erzieherin und Lehrerin in einem Kinderheim in Bad Kreuznach"
Output: {'jobTitle': 'Erzieherin', 'startDate': '1916', 'location': 'Bad Kreuznach', 'organization': 'Kinderheim', 'comments': "Position in a children's home"}

"1975 Dozent für neutestamentliche Exegese"
Output: {'jobTitle': 'Dozent', 'startDate': '1975', 'organization': 'neutestamentliche Exegese', 'comments': 'neutestamentliche Exegese is related to an evangelical church position'}

Überlegungen zur Methodenkritik

Ethische Fragen

LLMs verursachen massive ökologische und soziale Kosten. Sie sind **„Extraction Technologies"**, die natürliche Ressourcen wie seltene Erden sowie große Mengen an Energie und Wasser beanspruchen. Zudem nutzen sie Daten und kreative Arbeit (z. B. von Künstler*innen und Autor*innen) ohne Zustimmung und erfordern die Ausbeutung schlecht bezahlter Arbeitskräfte, die für das Labeln von Daten und das Prüfen von Inhalten zuständig sind. Dadurch verschärfen generative KI-Modelle zentrale Probleme wie die Klimakrise und ökonomische Ungleichheit.

siehe Kate Crawford. 2021. Atlas of AI - Power, Politics, and the Planetary Costs of Artificial Intelligence. Yale Univserity Press

* **"Open Washing"** beschreibt die Praxis, Offenheit vorzutäuschen, um Vorteile wie Vertrauen oder positive Wahrnehmung zu erlangen, ohne tatsächlich umfassend transparent zu sein. Im Bereich der generativen KI ist diese Problematik besonders relevant, da die Komplexität von Modellen, die enormen Kosten für Trainingsressourcen und rechtliche Risiken häufig eine vollständige Offenlegung verhindern. Oft werden nur bestimmte Teile eines Systems, etwa Modellgewichte, unter einer Open-Source-Lizenz veröffentlicht, während andere wesentliche Komponenten, wie Trainingsdaten oder Algorithmen, verschlossen bleiben.

> "There is ample evidence that as a communication strategy, open-washing is highly effective. The launches of Llama 2 and Llama 3 were greeted with considerable excitement in mainstream media outlets, almost without exception uncritically echoing the open source claim as a major selling point."

siehe Andreas Liesenfeld and Mark Dingemanse. 2024. "Rethinking open source generative AI: open-washing and the EU AI Act". In: Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24). Association for Computing Machinery, New York, NY, USA, 1774–1787. https://doi.org/10.1145/3630106.3659005

Verwendung aus geschichtswissenschafltlicher Sicht

LLMs werden seit einiger Zeit in den Geschichtswissenschaften effektiv eingesetzt, bisher v.a. für die Übersetzung von Texten, die automatisierte Erkennung von handgeschriebenen Quellen oder Mustern in großen Text- und Bildmengen. Für Historiker:innen ist die Methode der automatisierten Datenextraktion attraktiv, weil enorm zeitsparend.

"ChatGPT und Co. in den Geschichtswissenschaften – Grundlagen, Prompts und Praxisbeispiele". Von Mareike König; veröffentlicht 19/08/2024, aktualisiert 21/08/2024
https://dhdhi.hypotheses.org/9197

Der Pilot-Versuch, mit Hilfe von Llama und Gemma Berufsbezeichnungen aus der Forschungsdatenbank WOKDDR auszulesen und das Ergebnis einer nochmaligen Selbstkorrektur zu unterziehen, ergab Folgendes: Die bei der Evaluation manifestierten Fehler wie z.B. Halluzinationen, fehlende Angaben und Formatierungsfehler, müssen manuell korrigiert werden. Da die Qualität des Outputs eines LLM von der Qualität des Inputs abhängt, kann das weniger zeitintensive händische Nachbessern gleichzeitig sinnvoll genutzt werden, um das Vokabular / die Ontologie zu präzisieren und weitere Schwachstellen von WOKDDR zu korrigieren.

Sächsische Akademie der Wissenschaften zu Leipzig
SACHSEN
Diese Maßnahme wird mitfinanziert durch Steuermittel auf der Grundlage des vom Sächsischen Landtag beschlossenen Haushaltes.