Prompt-Engineering und Hermeneutik – Best Practices für die historische und qualitative Forschung

Möbus, Dennis
dennis.moebus@fernuni-hagen.de
FernUniversität in Hagen, Deutschland
ORCID: 0009-0008-9064-7460

Vu, Binh
binh.vu@srh.de
SRH Hochschule Heidelberg, Deutschland

Bayerschmidt, Philipp
philipp.bayerschmidt@fernuni-hagen.de
FernUniversität in Hagen, Deutschland

Einleitung und Fragestellung

Mit dem Einsatz von Large Language Models (LLMs) und Chat-Interfaces ist auch das Design von Suchanfragen bzw. Aufgabenstellungen an die Künstliche Intelligenz zu einem vieldiskutierten Thema geworden. Längst haben ChatGPT und Co. Einzug in die traditionell hermeneutisch arbeitenden Wissenschaften gehalten – und das auf allen Ebenen: Recherche, Zusammenfassung, Analyse und Textproduktion. Der Workshop richtet sich an Anfänger*innen und Nutzer*innen mit ersten Erfahrungen im Prompt Design, die LLMs in der historischen und qualitativen Forschung einsetzen (möchten). Nach der Adaption klassischer Verfahren aus dem maschinellen Lernen und Natural Language Processing (NLP) wird auch das generative Potential von LLMs ausgeschöpft und schließlich der Versuch einer maschinellen Interpretationen unternommen, um Transkripte lebensgeschichtlicher Interviews aus einem Oral-History-Projekt der 1980er Jahre und Briefserien deutscher Amerikaauswanderer zu analysieren.

Hintergrund

In den letzten Jahren hat sich im Bereich des Natural Language Processing (NLP) ein tiefgreifender Wandel vollzogen, der durch das Aufkommen von Deep Learning und der Entwicklung großer Sprachmodelle gekennzeichnet ist. Diese Fortschritte haben sich grundlegend auf verschiedene Anwendungen ausgewirkt – von der maschinellen Übersetzung bis zum *Question Answering* (QA) – und haben sowohl eine neue Ära maschinellen Sprachverstehens als auch maschineller Spracherzeugung (Stichwort generative KI) eingeläutet (Palo et al., 2023; Kasneci et al., 2023). Das Herzstück dieser Entwicklung sind *Transformer-Modelle*, eine Deep-Learning-Architektur, die die Art und Weise der Sprachverarbeitung revolutioniert hat (Wolf et al., 2020). Transformer haben herkömmliche neuronale Modelle wie Faltungsnetze und rekurrente Netze in ihrer Fähigkeit, Texte zu verarbeiten, übertroffen.

Spätestens seit ChatGPT ist KI (in der aktuellen Form von Deep Learning) im Alltag angekommen – und damit auch im Forschungsalltag. Zahlreiche Diskussionsrunden (Rygiel, 2024; Mlynarczyk, 2023), Workshops (Althage et al., 2024; Schreiber, 2023), Blogbeiträge (Eckenstaler, 2023; Mähr, 2024) und Artikel (Hiltmann, 2024; Lieder/Schäffer, 2024) zeugen von den Veränderungen, die KI für Forschung und Lehre bedeuten. Das macht vor den hermeneutisch arbeitenden Disziplinen keinen Halt. Neben den bekannten Gefahren, die von Deep Fakes und KI-generierten Texten ausgehen, bietet die LLM-basierte KI allerdings große Chancen für die inhaltliche Erschließung von Quellen.

Prompting zwischen Algorithmus und Hermeneutik

Voraussetzung für die Nutzung großer Sprachmodelle ist das Formulieren präziser Suchanfragen (*Prompting*), was mittlerweile unter den Begriffen *Prompt Design* und *Prompt Engineering* zu einer Subdisziplin der Data Science geworden ist. Da die Formulierung eines Erkenntnisinteresses zu den Grundlagen wissenschaftlichen Arbeitens gehört, scheint Prompting hermeneutisch orientierten Disziplinen näherzuliegen als klassisch algorithmenbasierte Machine-Learning-Anwendungen. Für komplexere Fragestellungen sind jedoch auch komplexere Anweisungen vonnöten, die dann mitunter ähnlich formuliert und formalisiert werden müssen wie ein Algorithmus – allerdings in natürlicher Sprache, denn die Barriere der Programmiersprache entfällt beim Prompting (Hiltmann, 2024).

Durch die Nutzung der vortrainierten Fähigkeiten großer Sprachmodelle und ihre Feinabstimmung auf bestimmte Aufgaben durch sorgfältig ausgearbeitete Prompts, können bemerkenswerte Ergebnisse erzielt werden. Der Prompting-Ansatz ermöglicht eine effektivere Nutzung von Sprachmodellen, da sie ohne ein Neutraining schnell an neue Aufgaben angepasst werden können (Mogavi et al., 2023). Mittlerweile haben sich verschiedene Methoden und Strategien im Prompt Design/Engineering entwickelt.

Während im *Zero-Shot-Learning* nur eine einfache Anfrage an das LLM gesendet wird, werden beim *Few-Shot-Learning* Beispiele formuliert, an denen sich die Ausgabe (die Completion) des LLMs orientieren soll, um das Ergebnis zu optimieren. Das ist etwa bei inhaltsextrahierenden Verfahren wie der Named Entity Recognition hilfreich, um

die Kategorien (wie Person, Ort, Organisation) zu definieren oder durch die Übergabe von Goldstandards (die man als große Sammlung von Beispielen begreifen kann) das Modell zu verfeinern (Finetuning).

Geht es um Schlussfolgerungen, die das LLM ziehen soll (wie etwa bei dem Auftrag, eine bestimmte Textstelle zu interpretieren), ist das *Chain-of-Thought-Prompting* sinnvoll. Bei diesem Ansatz wird das LLM aufgefordert, jeden Schritt, den es bei der Lösung der Aufgabe unternimmt, zu erklären. Diese Art der Reflexionsleistung befähigt das LLM einerseits zu komplexeren „Denkweisen", andererseits kann die Arbeitsweise für Außenstehende besser nachvollzogen werden (Liu et al., 2021).

Eine ganz andere Qualität haben *System Prompts* oder *Role Prompts*. Über eine möglichst detaillierte Personenbeschreibung soll das LLM für die Bearbeitung der Aufgabe eine ganz bestimmte Rolle einnehmen – etwa die einer Neuzeithistorikerin oder die eines Sozialforschers (Kong et al. 2023). Für welche Herangehensweise man sich letztlich entscheidet, hängt vom jeweiligen Einsatzgebiet ab. Der Workshop adressiert zwei Disziplinen, es ist explizit erwünscht, dass Teilnehmer*innen eigene Fragestellungen mit in die hands-on-Phasen einbringen.

Ziele und Ablauf des Workshops

Der Workshop möchte sich diese Strategien des Prompting aneignen, um exemplarisch lebensgeschichtliche Interviews und Briefserien zu interpretieren. Dazu werden zunächst bekannte Verfahren der Informationsextraktion und des NLP adaptiert. Mit einer Named Entity Recognition (NER) können inhaltliche Marker extrahiert werden, um einen Überblick erwähnter Personen, Orte, Unternehmen oder Ereignisse zu erhalten – dazu werden die Erkenntnisse erster Ansätze von Prompting-basierter NER herangezogen (Ashok, 2023; González-Gallardo, 2023, Hiltmann et al., 2024). Die Ergebnisse können mit einem derzeit im Rahmen von Oral-History.Digital entwickelten NER-Modell verglichen werden.

Ein Novum des Prompting ist die Mensch-Maschine-Interaktion mit natürlicher Sprache. Damit ist das Formulieren aller Arten von Prompts von unendlich vielen Variablen abhängig (Mishra et al., 2023). Allerdings wird erst dadurch das volle Potential der Sprachmodelle „abrufbar". In einem nächsten Schritt lösen wir uns entsprechend von etablierten NLP-Methoden und „befragen" die Texte, um weitere Inhalte zu erschließen. Ein erster Ansatz wird eine Kurzzusammenfassung sein, anschließend können Prompts zum Schreiben einer Biographie der Schreibenden/Interviewten entwickelt werden. Dabei ist den Teilnehmenden die Form (tabellarisch/Fließtext) freigestellt. Eine Aufgabe dieses Schritts wird in allen Fällen die systematische Veränderung von Details im Prompt und der Vergleich der KI-generierten Kondensate sein. Zur Dokumentation werden Padlets oder ähnliche kollaborative Werkzeuge für die Teilnehmenden vorbereitet. Durch das Vergleichen zweier LLMs, der Auseinandersetzung mit unterschiedlichen Parameterisierungen und dem Vergleich der Ergebnisse – auch unter den Vorzeichen unterschiedlicher disziplinärer Fragestellungen – wird nicht nur die hermeneutische Arbeit mit den Quellen optimiert, sondern gleichzeitig eine Algorithmenkritik betrieben. Über die sich verändernden Outputs können im Optimalfall auch Rückschlüsse auf die Interpretations- und „Verstehens"-prozesse der KI gewonnen und somit einer digitalen Hermeneutik auf einer weiteren Ebene Vorschub geleistet werden.

Als abschließender Schritt soll am zweiten Tag ein Template zur hermeneutischen Textinterpretation erarbeitet werden, um die KI analytisch arbeiten zu lassen (Henrickson/Meroño -Peñuela, 2023). Da sich der Workshop an interdisziplinäre Teilnehmer*innen richtet, können auf Grundlage des aktuellen State-of-the-Art im Prompt-Engineering jeweils individuelle Interpretationstemplates entwickelt werden. Vorarbeiten und teils beeindruckende Beispiele gibt es aus der Geschichtswissenschaft und der qualitativen Forschung (Hiltmann, 2024, Lieder/Schäffer, 2024).

Technischer Rahmen

Die Prompts und Templates werden in den LLMs Meta Llama und Mixtral entwickelt. Llama 3.1 war bis September 2024 die neueste Familie großer Sprachmodelle von Meta und umfasst Modelle mit 8B, 70B und zum ersten Mal 405B Parametern (Fuhrmann, 2014). Mit 141 Milliarden Gesamtparametern und explizit multilingualer Ausrichtung ist Mixtral 8x22B eine ernstzunehmende europäische Alternative zu Llama 3.1. (Bastian, 2024).

Hintergrund der Verwendung dieser beiden LLMs ist, dass Daten der qualitativen Sozialforschung in der Regel sensibel sind und perspektivisch nur mit Hilfe lokaler, DSGVO-konform gehosteter LLMs analysiert werden dürfen – sowohl Llama als auch Mistral laufen bereits auf KI-Workstations des Instituts für Geschichte und Biographie der FernUni Hagen. Für den Workshop werden Daten vorbereitet, die auch auf Plattformen genutzt oder über APIs geteilt werden dürfen, um die gemeinsame Arbeit zu erleichtern.

Technisch werden Google Colabs bereitgestellt, die wiederum auf together.ai, einer End-to-End-Plattform für generative KI, zugreifen. Together.ai unterstützt den gesamten Zyklus von der Bereitstellung vortrainierter Modelle bis hin zur Feinabstimmung und Erstellung benutzerdefinierter Modelle. Für die Teilnehmenden bedeutet das, dass sie sowohl einen Google-Account als auch ein together.ai-Konto benötigen (bei Registrierung wird ein ausreichender Vorrat an Credits zur Verfügung gestellt). In den Arbeitsphasen ist es den Teilnehmer*innen freigestellt, mit welchem der beiden Modelle sie arbeiten – möglich ist, sich in Einzel- oder Gruppenarbeit auf ein Modell zu konzentrieren oder beide zu vergleichen.

Die Beschränkung auf zwei Modelle ermöglicht einerseits einen systematischen Vergleich untereinander, andererseits ist so noch genügend Freiraum, um die Modelle

auch in der Tiefe zu erschließen – beispielsweise durch das Ändern der Hyperparameter. Diese umfassen *Output Length*, also die maximale Anzahl von Tokens (hier Wörter), die das Modell für seine Antwort benutzen darf. Die *Temperature* beeinflusst den Zufallsfaktor der Ausgabe durch die Auswahl mehr oder weniger wahrscheinlicher Wörter. Während *Top-k* die Wortauswahl auf die *k* wahrscheinlichsten Wörter in jedem Schritt der Textgenerierung begrenzt, engt *Top-p* (oder *Nucleus Sampling*) die kumulative Wahrscheinlichkeit aller bereits verwendeten Wörter der Completion ein (De la Vega, 2023).

Abb. 1 Überblick der Parameter, die im Workshop genutzt werden können

Parameter	Purpose	Recommended Settings
Output Length	Sets response size	Adjust based on task complexity
Temperature	Controls randomness	0.2–0.4 for precision, 0.6–0.8 for creativity
Top-p	Balances diversity and coherence	0.8–0.95 for general-purpose use
Top-k	Limits word selection	40–100 for flexible tasks

Format und Zielgruppe

Der zweitägige Workshop richtet sich an Historiker*innen und Vertreter*innen qualitativ forschender Disziplinen, die mit autobiographischen Textquellen arbeiten. Um eine gute Betreuung der hands-on-Phasen zu gewährleisten, wird die Anzahl der Teilnehmenden auf 30 Personen beschränkt. Erste Erfahrungen im Prompt Design sind hilfreich, aber nicht zwingend erforderlich – im Workshop wird ohnehin zur Gruppenarbeit aufgerufen, wodurch sich Erfahrungswerte ergänzen und ausgleichen können. Zur aktiven Teilnahme ist ein digitales Endgerät und die (kostenlose) Registrierung bei Google und together.ai notwendig.

Beitragende zum Workshop

Dennis Möbus ist Historiker und promovierte zu Freiheitsbegriffen und -erfahrungen deutscher Amerikaauswanderer. Er ist Mitarbeiter im DFG-Projekt Oral-History.Digital sowie am Institut für Geschichte und Biographie (IGB) der FernUniversität in Hagen und koordiniert die Forschungsgruppe digital humanities – Forschen im digitalen Raum, wo er sich mit Verfahren des Text Mining, Algorithmenkritik und digitaler Hermeneutik auseinandersetzt.

Binh Vu ist Professor für Data Science an der SRH Hochschule Heidelberg. Er ist außerdem University Ambassador und Ausbilder des NVIDIA Deep Learning Institute. Seine Forschungsschwerpunkte sind maschinelles Lernen, Deep Learning und die Entwicklung von Wissensmanagementsystemen, die in der Spieleindustrie und der medizinischen Forschung zum Einsatz kommen, sowie datenwissenschaftlichen Innovationen.

Philipp Bayerschmidt ist Historiker. In seiner Masterarbeit verglich er die Aussagen von ehemaligen Auschwitz-Häftlingen im 1. Frankfurter Auschwitzprozesses mit ihren lebensgeschichtlichen Interviews vierzig Jahre später. Seit 2021 ist er wissenschaftlicher Mitarbeiter im DFG-Projekt Oral-History.Digital am IGB. Er promoviert zu den Themen Migration und Heimat in lebensgeschichtlichen Interviews sowie zur Anwendung von Topic Modeling in der qualitativen und historischen Forschung.

Bibliographie

Althage, Melanie, Martin Dröge, Anna Faust und Mareike König. 2024. ChatGPT und Co. in der Geschichtswissenschaft: eine Praxis-Einführung für Einsteiger:innen. https://digigw.hypotheses.org/5523.

Ashok, Dhananjay und Zachary C. Lipton. 2023. PromptNER: Prompting For Named Entity Recognition. https://doi.org/10.48550/arXiv.2305.15444.

Bastian, Matthias. 2024. Mistrals Mixtral 8x22B setzt neue Bestwerte bei Open-Source-LLMs. The Decoder. https://the-decoder.de/mistrals-mixtral-8x22b-setzt-neue-bestwerte-bei-open-source-llms/.

Bayerschmidt, Philipp und Dennis Möbus: Quantität und Qualität. Im Erscheinen. Inhaltverzeichnisse für Korpora lebensgeschichtlicher Interviews computergestützt erstellen. In: BIOS – Zeitschrift für Biographieforschung, Oral History und Lebensverlaufsanalysen.

Eckenstaler, Sophie. 2023. ChatGPT in den Geschichtswissenschaften? Ein Praxisversuch. https://hochschulforumdigitalisierung.de/chatgpt-in-den-geschichtswissenschaften-ein-praxisversuch-2/.

Fuhrmann, Marvin. 2024. Llama 3.1: Was die neue KI kann und warum Meta sie verschenkt. T3N. https://t3n.de/news/llama-3-1-warum-meta-die-ki-verschenkt-1637247/.

González-Gallardo, Carlos-Emiliano, Emanuela Boros, Nancy Girdhar, Ahmed Hamdi et al. 2023. Yes but… Can ChatGPT Identify Entities in Historical Documents? In: 2023 ACM/IEEE Joint Conference, S. 184-185. https://doi.org/10.1109/JCDL57899.2023.00034.

Henrickson, Leah und Albert Meroño -Peñuela. 2023. Prompting meaning: a hermeneutic approach to optimising prompt engineering with ChatGPT. In: AI & Society. Online: https://doi.org/10.1007/s00146-023-01752-8.

Hiltmann, Torsten. 2024. Hermeneutik in Zeiten der KI. Large Language Models als hermeneutische Instrumente in den Geschichtswissenschaften. In: Gerhard Schreiber, Lukas Ohly (Hg.): KI:TEXT. Berlin/Boston, S. 201-232.

Hiltmann, Torsten, Martin Dröge, Nicole Dresselhaus, Sophie Eckenstaler et al. 2024. NER, aber prompto! Named Entity Recognition mit Large Language Models für historische Texte. https://dhistory.hypotheses.org/7870.

Ibrahim, Adam, Benjamin Thérien, Kshitij Gupta, Mats L Richter et al. 2024. Simple and scalable strategies to continually pre-train large language models. https://arxiv.org/abs/2403.08763.

Kasneci, Enkelejda, Kathrin Seßler, Stefan Küchemann, Maria Bannert et al. 2023. ChatGPT for good? On opportunities and challenges of large

language models for education. https://doi.org/10.1016/j.lindif.2023.102274.

Kong, Aobo, Shiwan Zhao, Hao Chen, Qicheng Li et al. 2023. Better Zero-Shot Reasoning with Role-Play Prompting. https://arxiv.org/abs/2308.07702

Lieder, Fabio Roman und Burkhard Schäffer. 2024. Reconstructive Social Research Prompting (RSRP). Distributed Interpretation between AI and Researchers in Qualitative Research. https://osf.io/preprints/socarxiv/d6e9m.

Liu, Pengfei, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang et al. 2021. Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing. https://doi.org/10.48550/arxiv.2107.13586.

Mähr, Moritz. 2024. Mit ChatGPT Texte schreiben: Prompting-Methoden für Historiker:innen. https://doi.org/10.5281/zenodo.10823269 .

Mlynarczyk, Olga. 2023. Learning to think like the past. Applicability of retrieval augmented LLMs in historical research. https://dhistory.hypotheses.org/6334.

Mogavi, Reza Hadi, Chao Deng, Justin Juho Kim, Pengyuan Zhou et al. 2023. Exploring User Perspectives on ChatGPT: Applications, Perceptions, and Implications for AI-Integrated Education. https://doi.org/10.48550/arxiv.2305.13114.

Palo, Norman Di, Arunkumar Byravan, Leonard Hasenclever, Markus Wulfmeier, et al. 2023. Towards A Unified Agent with Foundation Models. https://doi.org/10.48550/arxiv.2307.09668.

Pham, Chau Minh, Alexander Hoyle, Simeng Sun, Philip Resnik et al. 2023. TopicGPT: A Prompt-based Topic Modeling Framework. https://doi.org/10.48550/arXiv.2311.01449.

Rygiel, Philippe. 2024. Are AI and LLM important for historians? https://www.c2dh.uni.lu/de/events/are-ai-and-llm-important-historians.

Schreiber, Gerhard. 2023. KI – Text und Geltung. Wie verändern KI-Textgeneratoren wissenschaftliche Diskurse? https://www.hsozkult.de/event/id/event-137654.

Vega, Miguel de la. 2023. Understanding OpenAI's "Temperature" and "Top_p" Parameters in Language Models. Medium. https://medium.com/@1511425435311/understanding-openais-temperature-and-top-p-parameters-in-language-models-d2066504684f.

Wang, Han, Nirmalendu Prakash, Nguyen Khoi Hoang, Ming Shan Hee et al. 2023. Prompting Large Language Models for Topic Modeling. https://doi.org/10.48550/arXiv.2312.09693.

Wolf, Thomas, Lysandre Debut, Victor Sanh, Julien Chaumond et al. 2020. Transformers: State-of-the-Art Natural Language Processing. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, online, S. 38-45. https://aclanthology.org/2020.emnlp-demos.6/.