LLMs unter Kontrolle: Offene Modelle in Forschung und Praxis.

Hermes, Jürgen
hermesj@uni-koeln.de
Universität zu Köln, Deutschland
ORCID: 0000-0002-8367-8073

Niebes, Kai
kai.niebes@tib.eu
TIB Hannover, Deutschland
ORCID: 0009-0005-3141-4809

Sarah, Oberbichler
oberbichler@ieg-mainz.de
DH Lab Leibniz Institut für Europäische Geschichte Mainz (IEG), Deutschland
ORCID: 0000-0002-1031-2759

Wagner, Andreas
wagner@lhlt.mpg.de
Max-Planck-Institut für Rechtsgeschichte und Rechtstheorie, Frankfurt/M., Deutschland
ORCID: 0000-0003-1835-1653

1. Ziele des Workshops

Auch im vierten Jahr nach der Veröffentlichung von ChatGPT sind die Bedingungen für den Einsatz großer, vortrainierter KI-Modelle in den Geisteswissenschaften allenfalls vage ausgelotet (Simons, Zichert und Wüthrich 2025). Mit jedem neuen Modell wächst die Faszination für die beeindruckenden Fähigkeiten dieser Systeme – sei es in der Textproduktion, beim Sprachverstehen oder der Informationsstrukturierung. Zugleich mehren sich die Stimmen, die auf grundlegende Probleme und Risiken verweisen: Epistemologische Verschiebungen (Mollick 2025; Livingstone und Stricker 2025), methodische Intransparenz (Dobson 2023) und ethische Fallstricke (Reynoldson 2025; Attard-Frost und Widder 2025). Nicht zuletzt in seiner Keynote zur DHd2025 hat Marc Dingemanse dafür plädiert, den wissenschaftlichen Umgang mit großen Sprachmodellen kritisch zu beleuchten. Dingemanse hob insbesondere die mangelnde Transparenz vieler Modelle, ihre begrenzte Aussagekraft über Bedeutung sowie die Risiken eines ungenügend reflektierten Einsatzes im akademischen Kontext hervor. Die Nutzung proprietärer LLMs kann weder dauerhaft garantiert noch im Detail nachvollzogen werden. Viele dieser Modelle sind durch den Einsatz von RLHF auf gefällige Antworten (z. B. übermäßige Zustimmung) optimiert (Malmqvist 2024), filtern oder verzerren Inhalte, lügen gar bisweilen (Mitchell 2025; Liang et al. 2025) und ermöglichen keine Kontrolle über Training, Fine-Tuning oder sogar die genaue Modellversion (Manchanda et al. 2025).

Das Spannungsfeld zwischen omnipräsenten und (noch) einfach zugänglichen Tools auf der einen und der notwendigen, aber komplexen Abwägung zu den Bedingungen und Auswirkungen ihres Einsatzes auf der anderen Seite verlangt eine grundlegende und differenzierte Auseinandersetzung, die technisches Wissen mit (geistes)wissenschaftlicher Reflexion verbindet. Mit unserem Workshop wollen wir zumindest einen Teilbereich dieses Themenkomplexes abdecken (Reproduzierbarkeit, Interpretierbarkeit, methodische Offenheit) und uns auf die Nutzung offener Sprachmodelle (OpenLLMs) fokussieren. Dabei wollen wir ausloten, inwiefern sie eine transparente, nachvollziehbare und kontrollierbare, aber auch handhabbare Alternative zu kommerziellen Closed-Source-Modellen darstellen können (Kukreja et al. 2024). Der Workshop soll vor allem dazu dienen, in die technische Handhabung einzuführen, mit den Teilnehmenden sollen aber auch kritische Perspektiven auf Potenziale und Grenzen von OpenLLM-Systemen in der (digitalen) Geisteswissenschaft diskutiert werden.

Ziel des Workshops (ein ganzer Tag oder zwei Halbtage) ist es, Wissenschaftler:innen aus den Digital Humanities mit OpenLLMs vertraut zu machen – sowohl auf theoretischer als auch auf praktischer Ebene. Der Workshop richtet sich an Forschende mit Interesse an der Nutzung generativer KI-Modelle, die sich kritisch mit der Frage auseinandersetzen wollen, wie sprachmodellgestützte Verfahren in geisteswissenschaftlichen Kontexten transparent, reproduzierbar und nachhaltig eingesetzt werden können.

2. Gliederung des Workshops

Die Veranstaltung gliedert sich in drei Theorie- und zwei Praxisblöcke:

A. Einführender Theorieblock: Warum offene LLMs? Begriffsklärung und Modelle

Nach einer kurzen Einführung in Aufbau und Funktionsweise großer Sprachmodelle klärt der erste Theorieblock, was Offenheit bei LLMs konkret bedeutet (Liesenfeld und Dingemanse 2024) und welche rechtlichen und technischen Einschränkungen sich hinter scheinbar freien Modellen verbergen (Gibney 2024). Die Teilnehmenden erhalten einen Überblick zu zentralen Modellfamilien (aktuell wären das BERT, OLMo, Qwen und Gemma, kann sich durch aktuellere Entwicklungen bis Februar 2026 noch ändern), zu deren Lizenzmodellen, Performanz, Hardwareanforderungen und verfügbare Schnittstellen.

B. Erster Praxisblock: Ein kleines OpenLLM auf dem eigenen Rechner

Ein zentrales Anliegen des Workshops ist es, Hemmschwellen abzubauen und erste praktische Erfahrungen mit OpenLLMs zu ermöglichen. Im Vorfeld der Veranstaltung

erhalten alle Teilnehmenden ein Installations-Tutorial, mit dem sie ein kompaktes LLM (z.B. OLMo-2 7B) mithilfe der Plattform Ollama (alternativ: Huggingface) auf ihrem eigenen Rechner installieren können. Während des Workshops wird ein Zeitfenster von ca. 30 Minuten eingeräumt, in dem verbliebene Installationsprobleme gemeinsam gelöst werden können. Wer die Installation bereits erfolgreich abgeschlossen hat, kann in dieser Zeit eine Kaffeepause machen, erste Prompts testen oder sich informell austauschen. Ziel des Blocks ist ein erstes Aha-Erlebnis: Sprachmodelle können lokal laufen, ohne Cloud-Anbindung, auf dem eigenen Gerät – wir halten dies für einen bedeutsamen Perspektivwechsel im Vergleich zur gewohnten Nutzung von kommerziellen Plattformen.

C. Mittlerer Theorieblock: Mögliche Anwendungen für OpenLLMs in den DH

Dieser Abschnitt des Workshops zeigt konkrete Nutzungsmöglichkeiten offener Sprachmodelle in geisteswissenschaftlichen Arbeitsprozessen auf. Anhand kleinerer Fallbeispiele (u.a. Audio-Transkription, Textnormalisierung, Informationsextraktion) wird diskutiert, wo OpenLLMs eine Alternative zu geschlossenen Systemen darstellen können – und wo ihre Grenzen liegen (z.B. bei geringen Kontextfenstern, schwächerer multilingualer Abdeckung, beschränkter Modellkapazität).

D. Zweiter Praxisblock: Ein leistungsfähiges OpenLLM über Serverzugang

Für eine intensivere Erprobung der Möglichkeiten erhalten die Teilnehmenden Zugang zu einem dedizierten Server (bereitgestellt von den Workshop-Veranstaltenden), auf dem mehrere leistungsfähige OpenLLMs (z.B. Gemma-27B) bereitgestellt werden. Der Zugriff erfolgt über eine JupyterHub-Umgebung mit vorbereiteten Notebooks. Diese bieten drei verschiedene Zugänge, darunter ein vordefinierter Klassifikations-Task zum schnellen Austesten und ein offenes Forschungsnotebook zur Entwicklung eigener Prompts oder Experimente. Je nach Interesse und Vorkenntnissen können Teilnehmende unterschiedliche Pfade wählen. Die Workshopleitung unterstützt bei der Interpretation der Ergebnisse und gibt Hinweise auf technische Hintergründe (Promptstruktur, Tokenisierung, Temperature etc.).

E. Abschließender Theorieblock: Vergleich von lokalen, serverbasierten und kommerziellen LLMs, Ausblick

Der Abschluss des Workshops dient der gemeinsamen Reflexion. Diskutiert werden soll, welche Erfahrungen mit lokalen vs. serverbasierten vs. kommerziellen LLMs gemacht wurden, ob sich bereits abzeichnet, welche Modelle für welche Aufgaben genutzt werden können und wo die tatsächlichen Unterschiede in Bezug auf Transparenz, Kontrolle, Reproduzierbarkeit, Interpretierbarkeit liegen. Dabei spielen auch Tools und Workflows für systematische Evaluationen (DeepEval, trackio, codecarbon, EcoLogits u.ä.), sowie Erfahrungen mit diesen eine Rolle (Husain 2024; Schmid 2024; Rudd, Andrews und Tully 2025; Luccioni, Trevelin und Mitchell 2024). Abschließend soll noch diskutiert werden, welche Infrastrukturen aus Sicht der DH-Community wünschenswert wären. Der Workshop versteht sich auch als Beitrag zur nachhaltigen Kompetenzentwicklung innerhalb der deutschsprachigen DH-Community.

3. Informationen zum Workshop

Format und Zeitstruktur

Der Workshop ist als zwei halbtägige oder eine ganztägige Veranstaltung konzipiert (8 Zeitstunden inkl. Pausen).

Zielpublikum

Der Workshop richtet sich an Wissenschaftler:innen aus den Digital Humanities, die bereits erste Erfahrungen mit generativer KI (z.B. via ChatGPT) gemacht haben, aber tiefer in das Thema einsteigen und insbesondere Alternativen zu kommerziellen Lösungen kennenlernen wollen. Programmierkenntnisse sind nicht erforderlich, wohl aber die Bereitschaft, mit eigenen Daten oder Fragestellungen zu experimentieren. Hilfreich sind Grundkenntnisse im Umgang mit Kommandozeile oder Webtools (z.B. Jupyter Notebooks). Teilnehmende werden gebeten, im Vorfeld eine Installationsanleitung durchzuarbeiten.

Maximale Teilnehmerzahl

Aufgrund der technischen Infrastruktur (Serverkapazität) ist die Zahl der Teilnehmenden auf **30 Personen** beschränkt.

Technische Ausstattung

- Eigener Laptop mit Internetzugang, mindestens 8 GB RAM (empfohlen: 16 GB)
- Vorab-Installation von Ollama oder vergleichbarer Plattform nach Anleitung
- WLAN und Stromversorgung vor Ort
- Zugang zu bereitgestelltem Server (wird durch die Organisatoren eingerichtet)
- Beamer für Präsentationen und Systemdemos

Kein gesonderter Call for Papers

Kontakt und Beitragende

Jürgen Hermes – Institut für Digital Humanities, Universität zu Köln – hermesj@uni-koeln.de – Forschungsinteressen: Grenzen, Nutzen und Grundlagen generativer KI, auch im Bezug zu den DH; Tools für die Public Humanities; Prozessierung von Texten.

Sarah Oberbichler – Luxembourg Centre for Contemporary and Digital History (C²DH) – sarah.oberbichler@uni.lu **–** Forschungsinteressen: Digitale Medien; Generative KI und NLP; Umweltgeschichte; Migrationsgeschichte.

Andreas Wagner – Max-Planck-Institut für Rechtsgeschichte und Rechtstheorie, Frankfurt/M. – wagner@lhlt.mpg.de – Forschungsinteressen: NLP; digitale Editorik; Funktionen und Leistungsfähigkeit von Encoding Modellen; KI Sprachverständnis und (rechtl.) Argumentationen; Information Extraction

Kai Niebes – TIB Hannover – kai.niebes@tib.eu – Forschungsinteressen: Verarbeitung von Textdaten; Verarbeitung von 3D-Daten (Punktwolken, Meshes, Gaussian

Splatting); Prozessautomatisierung; Tools für die Digital Humanities

Bibliographie

Attard-Frost, Blair und David Gray Widder. 2025. The ethics of AI value chains. *Big data & society* 12, Nr. 2 (6. Mai). https://doi.org/10.1177/20539517251340603.

Dobson, James E. 2023. On reading and interpreting black box deep neural networks. *International journal of digital humanities* 5 (20. November). https://doi.org/10.1007/s42803-023-00075-w.

Gibney, Elizabeth. 2024. Not all 'open source' AI models are actually open: here's a ranking. *Nature*. Juni. https://doi.org/10.1038/d41586-024-02012-5.

Husain, Hamel. 2024. Your AI product needs evals. 29. März. https://hamel.dev/blog/posts/evals/index.html (zugegriffen: 15. November 2025).

Kukreja, Sanjay, Tarun Kumar, Amit Purohit, Abhijit Dasgupta und Debashis Guha. 2024. A Literature Survey on Open Source Large Language Models. In: *Proceedings of the 2024 7th International Conference on Computers in Management and Business*, 133–143. ICCMB '24. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3647782.3647803.

Liang, Kaiqu, Haimin Hu, Xuandong Zhao, Dawn Song, Thomas L. Griffiths und Jaime Fernández Fisac. 2025. Machine bullshit: characterizing the emergent disregard for truth in large language models. *ArXiv preprint*. 10. Juli. https://doi.org/10.48550/arXiv.2507.07484.

Liesenfeld, Andreas und 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*, 1774–1787. FAccT '24. New York, NY, USA: Association for Computing Machinery, Juni. https://doi.org/10.1145/3630106.3659005.

Livingstone, Victoria und Jeppe Klitgaard Stricker. 2025. The disappearance of the unclear question. 24. Juli. https://www.unesco.org/en/articles/disappearance-unclear-question (zugegriffen: 1. August 2025).

Luccioni, Sasha, Bruna Trevelin und Margaret Mitchell. 2024. The environmental impacts of AI -- primer. 3. September. https://huggingface.co/blog/sasha/ai-environment-primer (zugegriffen: 15. November 2025).

Malmqvist, Lars. 2024. Sycophancy in Large Language Models: Causes and Mitigations. November. https://doi.org/10.48550/arXiv.2411.15287.

Manchanda, Jiya, Laura Boettcher, Matheus Westphalen und Jasser Jasser. 2025. The Open Source Advantage in Large Language Models (LLMs). Februar. https://doi.org/10.48550/arXiv.2412.12004.

Mitchell, Melanie. 2025. Why AI chatbots lie to us. *Science*. 6758. Juli. https://doi.org/10.1126/science.aea3922.

Mollick, Ethan. 2025. Against „Brain Damage". AI can help, or hurt, our thinking. 7. Juli. https://www.oneusefulthing.org/p/against-brain-damage (zugegriffen: 1. August 2025).

Reynoldson, Miriam. 2025. Tracing the AI value chain: the ethical costs of generative AI. 27. Mai. https://themindfile.substack.com/p/tracing-the-ai-value-chain (zugegriffen: 1. August 2025).

Rudd, Ethan M., Christopher Andrews und Philip Tully. 2025. A practical guide for evaluating llms and LLM-reliant systems. *ArXiv preprint*. 16. Juni. https://doi.org/10.48550/arXiv.2506.13023.

Schmid, Phil. 2024. LLM evaluation doesn't need to be complicated. 11. Juli. https://www.philschmid.de/llm-evaluation (zugegriffen: 15. November 2025).

Simons, Arno, Michael Zichert und Adrian Wüthrich. 2025. Large Language Models for History, Philosophy, and Sociology of Science: Interpretive Uses, Methodological Challenges, and Critical Perspectives. Juni. https://doi.org/10.48550/arXiv.2506.12242.