EIN NICHT-ENGLISCHES KI-SPRACHMODELL

ZUSAMMENFASSUNG

Mit der Veröffentlichung von ChatGPT und anderen Sprachmodellen (Large Language Modells – LLMs) entstanden ein Hype und große Erwartungen auf Disruptionen in Wirtschaft und Gesellschaft. Basis für die Modelle sind vor allem Trainingsdaten auf US-Englisch. Das hat technisch bedingt Auswirkungen auf die Qualität in anderen Sprachen aber auch kulturelle Einflüsse, weil die Sprachenlogiken z. B. im Deutschen anders sind. Dazu gesellen sich rechtliche Probleme in der Anwendung (z. B. Datenschutzkonformität). Weiters ergeben sich wirtschaftliche Fragen hinsichtlich der Geheimhaltung von Geschäftsgeheimnissen. Als politische Dimension gilt eine noch größere Abhängigkeit von den globalen Konzernen und eine weiter verringerte digitale Souveränität. In den letzten Monaten gab es Fortschritte, (offene) Sprachmodelle auf nicht-englischer Sprache zu entwickeln und Infrastruktur in der EU aufzubauen.

USA und China beherrschen den Markt

Wo bleibt Europa?

ÜBERBLICK ZUM THEMA

LLM mit disruptivem Potential

Seit der Veröffentlichung von ChatGPT im Herbst 2022 besteht kein Zweifel, dass die großen Sprachmodelle (LLM) einen großen Sprung in der Entwicklung von Systemen der sog. Künstlichen Intelligenz darstellen. Inwieweit der Hype um große Sprachmodelle (wie ChatGPT, Gemini, Mistral, Claude, Grok etc.) berechtigt ist und diese tatsächlich die in sie gesteckten Erwartungen erfüllen, muss sich erst zeigen. Bereits jetzt zeichnen sich aber Verwerfungen in Wirtschaft und Gesellschaft ab, etwa durch Pläne für umfangreiche Entlassungen unter Verweis auf KI.[1]

AI-Innovation-Package der Europäischen Kommission

Es zeigt sich, dass durch die hauptsächliche Nutzung von US-englischen Trainingsdaten für die bestehenden Modelle technisch bedingte Beschränkungen in der Qualität der Nutzung und des Outputs in anderen Sprachen entstehen. Dazu kommen kulturelle Einflüsse über Ausdrucksformen und Sprachlogiken sowie rechtliche und wirtschaftspolitische Aspekte. Absehbar sind mittlerweile zwei unterschiedliche Entwicklungsstrategien, die entweder darauf abzielen, das größte und beste Modell zu haben, oder sehr speziell auf Anforderungen aus der wirtschaftlichen Realität (z. B. Prozesse bei Kunden) abgestimmt zu sein (Bomke 2024). Die zweite Strategie könnte sein, Eigenheiten bestehender Modelle, z. B. in Fragen von Datenschutz, Urheberrecht etc. entgegenzuwirken. Diese Eigenschaften wurden unter europäischen Gesichtspunkten als Problem erkannt. Dreh- und Angelpunkt europäischer Anstrengungen ist das AI-Innovation-Package der Europäischen Kommission,[2] im Rahmen dessen in der EU ***KI-Rechenzentren*** aufgebaut werden sollen, fünf davon besonders große sogenannte Gigafactories.[3]

US-Englisch dominiert die Trainingsdaten – mit Auswirkungen auf die Effektivität des Modells

Alle derzeit diskutierten großen Sprachmodelle bauen größtenteils auf US-englischen Trainingsdaten auf. Diese Struktur in den Trainingsdaten hat technische Auswirkungen auf die Qualität der Ergebnisse in anderen Sprachen (Nicholas/Bhatia 2023). Texte müssen für die Erstellung der Modelle in maschinell verarbeitbare Einheiten, genannt Token, zerlegt werden. Ein Token kann als ein Stück Text betrachtet werden, das ein Modell auf einmal verarbeitet. In der deutschen Sprache könnte ein Token ein Wort, ein Satzzeichen, ein Teil eines Wortes oder sogar nur ein Buchstabe sein. Ein Token kann in verschiedenen Sprachen unterschiedliche Längen haben, z. B. kann es in Englisch oft ein Wort sein, während es in Sprachen wie Chinesisch oft ein Zeichen ist.[4] Diese Arbeit übernehmen Software-Programme, sogenannte „Tokenizer“. Sind diese auf Englisch trainiert, verwenden sie die Strukturmerkmale der englischen Sprache. Da andere Sprachen andere Strukturmerkmale aufweisen, führt das dazu, dass Worte zerschnitten und auf mehrere Tokens verteilt werden. Eine bewährte Faustregel für das Verhältnis ist etwa im Englischen: 1 Wort ≈ 1,3 Token, im Deutschen: 1 Wort ≈ 1,8 Token,

---

1 Derzeit trifft dies z.B. die IT-Branche, wo Berufseinsteiger:innen zunehmend schlechte Aussichten auf Jobs haben: *derstandard.at/story/3000000290507/accenture-intel-microsoft-tech-konzerne-streichen-zehntausende-jobs-was-ist-los*.

2 *digital-strategy.ec.europa.eu/en/factpages/ai-innovation-package*.

3 *digital-strategy.ec.europa.eu/en/policies/ai-factories*.

4 *online-marketing-leipzig.de/das-tokenlimit-ein-einstieg-fuer-anfaenger-mit-tipps-zum-prompting/*.

im Spanischen: 1 Wort ≈ 2 Token und im Französischen: 1 Wort ≈ 2 Token.[5] Da alle LLMs ein (technisches) Token-Limit[6] haben (und die Leistungen der Anbieter in der Regel nach Token abgerechnet werden) hat die Anwendung auf Englisch trainierter Tokenizer auf anderssprachige Trainingsdaten zur Folge, dass weniger Texte (je Prompt) produziert werden können und vor allem sprachliche Sinneinheiten nicht gut abgebildet werden.

Auch 2025 noch große Qualitätsunterschiede zwischen Sprachen

Evaluierungen verschiedener Sprachmodelle haben gezeigt, dass deutliche Abweichungen mehrsprachiger LLMs zwischen verschiedenen Sprachen existieren: Beim gleichen Modell schnitt die am schlechtesten unterstützte Sprache zwischen 8 % und beinahe 50 % besser als die beste ab – und Englisch immer am besten (Hupkes & Bogoychev, 2025). Dabei zeigte sich auch, dass viele der üblichen Benchmarks von Sprachmodellen in nicht-englischer Sprache auf maschinell übersetzten Fragen aus dem Englischen aufbauen – und damit lokale Feinheiten nicht abgebildet werden. Schlussendlich sind damit viele Benchmarks nicht besonders aussagekräftig für den Einsatz in nicht-angelsächsischen Kontexten.

Kulturelle, ethische und rechtliche Probleme

Neben den technischen Auswirkungen sind es somit auch kulturelle Eigenheiten, die sich in den Trainingsdaten und damit den Antworten der Modelle abbilden (siehe Ramesh et al. 2024, Fenech-Borg et al., 2025). Mit der Vorherrschaft US-englischer Trainingssätze werden auch US-amerikanische Formulierungen, Denkweisen, Ethik-Richtlinien etc. mittransportiert, was zu einer tendenziellen Angleichung von Kommunikationsformen im Sinne einer globalisierten Sprech- und Denkweise führen kann. Neben den möglichen kulturellen Auswirkungen werden aber vor allem rechtliche Probleme in der Anwendung offenbar. Viele der Modelle wurden mit Daten aus dem Internet, teilweise aus Social Media Plattformen trainiert. Es ist bei vielen der derzeitigen Modelle nicht klar, woher die Daten stammen, ob personenbezogene Daten darin verarbeitet wurden, ob die davon betroffenen Datensubjekte um ihr Einverständnis gefragt wurden, ob sie also im Einklang mit der DSGVO und damit wesentlichen europäischen Grundwerten genutzt werden können. Bei vielen Daten stehen auch Urheberrechtsverletzungen im Raum, wie etwa im prominenten Fall der New York Times, die OpenAI verklagt hat (Chen, 2025).

Offenes, mehrsprachiges Modell Apertus aus der Schweiz

Im Sommer 2025 haben die ETH Zürich, EPFL und das Schweizer Nationale Supercomputing Center das Modell Apertus veröffentlicht: Ein großes Sprachmodell unter Berücksichtigung Schweizer Werte: Mehrsprachigkeit, Transparenz und dem Gemeinwohl verpflichtet (ETH Zürich, 2025). Im Sinne der Transparenz sind sowohl die gesamte Trainingsarchitektur als auch alle Trainingsdaten, Quellcode, Anleitungen und die trainierten Modelle offen verfügbar. Gleichzeitig wurden die Daten im Einklang mit schweizerischem und europäischem Urheberrecht ausgewählt und persönliche Daten entfernt. Es ist auch vollständig mit dem EU-AI-Act kompatibel. Anders als viele andere Sprachmodelle wurde es von Grund auf mehrsprachig entwickelt – mit besonderer Rücksicht auch auf Schweizerdeutsch und Rätoromanisch.

---

5 *deinkikompass.de/blog/openai-gpt-token-guide*.

6 Z. B. GPT 32.765, Llama2 2.048, Claude_2 100.000 und PaLM 8.000 siehe *empolis.com/blog/entfesselte-llms-trotz-token-beschraenkung/*.

Wirtschaftliche Überlegungen

Neben der eingeschränkten Eignung für nicht-englische Sprachen stellen sich noch wirtschaftliche Überlegungen Fragen, insbesondere zur Offenlegung von z. B. Geschäftsgeheimnissen durch den Einsatz großer Sprachmodelle. Dabei steigt die Gefahr für gezielte Angriffe, je tiefer KI-Systeme in andere digitale Infrastrukturen integriert werden, um etwa gezielt sensible Daten zu stehlen (Constantin, 2025). Nichtdestotrotz wird das wirtschaftliche Potenzial großer KI-Sprachmodelle von vielen als vielversprechend angesehen. Ihre Anpassungsfähigkeit an unterschiedlichste branchen- und unternehmensspezifische Anforderungen und ihre hohe Wiederverwendbarkeit eröffnen unzählige Anwendungsfälle. Anhand von Praxisbeispielen verdeutlicht eine Publikation der deutschen Akademie der Technikwissenschaften (acatech) die Chancen sowie Herausforderungen der Sprachmodelle. Die Autor:innen empfehlen den Aufbau eines offenen, kommerziell nutzbaren Datensatzes in deutscher Sprache, der europäischen Werten und Regeln entspricht und die Entwicklung von Sprachmodellen in Deutschland unterstützt (Löser 2023).

Geopolitik und digitale Souveränität

Zu den o. a. technischen, rechtlichen, kulturellen und wirtschaftlichen Fragestellungen kommt noch die politische Dimension hinzu. Die Nutzung vornehmlich in den USA und China produzierter Modelle führt zu noch größeren Abhängigkeiten von den großen amerikanischen sowie chinesischen Technologiekonzernen, was im Sinne anzustrebender ***Digitaler Souveränität*** kontraproduktiv erscheint. Ziel ist deshalb, ein offenes Sprachmodell auf Basis europäischer, nicht-englischer oder deutscher bzw. österreichischer Sprache zu entwickeln[7] und so sicherzustellen, dass sowohl sprachliche als auch kulturelle und rechtliche Eigenheiten des deutschen und europäischen Sprachraumes Berücksichtigung finden. Eine wichtige Initiative in diese Richtung stellt die von der EU geförderte EuroLLM Initiative dar: Wie Apertus (s. o.) soll EuroLLM dezidiert mehrsprachig (alle offizielle EU-Sprachen abbildend), transparent und mit EU-Recht vereinbar sein (Martins et al., 2025). Erste kleine Modelle dieser Initiative zeigen bereits vielversprechende Ergebnisse im Vergleich zu kommerziellen LLMs kommerzieller Anbieter.

EuroLLM: ein offenes Sprachmodell für die EU

Höhere Effizienz bei sprachlicher Fokussierung

Vorteile eines dezidiert deutschsprachigen Modells könnten sein, dass die Modelle schlanker sein können. Das wirkt sich auch positiv auf die verringerten Hardware-Anforderungen (z. B. bei den schnellen, für KI-Modelle notwendigen Grafikkarten) und die erhöhte Betriebsgeschwindigkeit aus. Ein gezielt vorselektiertes Datenmaterial kann die Qualität erhöhen, den Fokus auf bestimmte Sachgebiete ermöglichen und eine optimale Nutzer:innenführung mit Sensibilisierung für die Ergebnisse beinhalten. Außerdem entstünden geringere Kosten während des Betriebs (vgl. Meffert 2023).

[7] D. h. Trainingsdaten aus den fünf größten europäischen Sprachen oder auch konkret dem deutschen Sprachraum zu verwenden und geeignete Tokenizer für die europäischen Sprachen bzw. die deutsche Sprache zu entwickeln und zu verwenden.

RELEVANZ DES THEMAS FÜR DAS PARLAMENT UND FÜR ÖSTERREICH

Förderung der österreichischen KI-Landschaft sowie der sprachlichen und wirtschaftlichen Eigenständigkeit

Aus dem oben Dargelegten geht hervor, dass die Entwicklung eines Sprachmodells auf Basis nicht-englischer Trainingsdaten Vorteile technischer, kultureller und ökonomischer Natur bringen kann. Um die Unterschiede zwischen den Sprachen, auch zwischen den Varietäten Deutsch und Österreichisch innerhalb des Modells hochzuhalten und den Standort Österreich in diese Entwicklung einzubringen, erscheint es notwendig, über mögliche Kooperationen nachzudenken (z. B. mit Apertus oder EuroLLM), sodass auch genuin österreichische Inhalte in die Trainingsdaten integriert werden. Die österreichische KI-Landschaft[8] kann dazu sicher wichtige Beiträge liefern. Nicht zuletzt sei auf die Einrichtung des Exzellenz-Clusters zu Bilaterale KI verwiesen.[9]

AI Factory Austria AI:AT

Grundlage dafür könnte auch die oben genannte EU-Initiative zum Aufbau von KI-Rechenzentren bilden (siehe auch ***Nachhaltige KI-Rechenzentren***), wovon eines in Österreich im Entstehen ist, die AI Factory Austria AI:AT.[10] Diese Recheninfrastruktur soll dabei unter anderem auch der österreichischen Verwaltung zur Verfügung stehen, um für die Verwaltung maßgeschneiderte KI-Lösungen zu entwickeln.

VORSCHLAG WEITERES VORGEHEN

Wie gezeigt werden konnte, macht es einen Unterschied, welche Sprache für das Training von großen Sprachmodellen verwendet wird. Durch die Sprache wird auch ein kultureller und Werte-Bias durch das Modell mitgelernt.

TA-Studie zum Sprach- und Kultur-Bias von LLM

Das Parlament könnte eine TA-Studie beauftragen, die auf Basis bestehender Erkenntnisse zu den Einflüssen unterschiedlicher Sprachen auf die Leistungsfähigkeit von Sprachmodellen ermittelt, inwieweit es aus Sicht der österreichischen Wirtschaft, der Konsument:innen und Bürger:innen angezeigt wäre, ein eigenes Modell zu entwickeln. Besonderes Augenmerk könnte auch auf österreichische sprachliche Eigenheiten gelegt werden, um diese auch in Zukunft zu erhalten.

---

8 *aiaustria.com*, *ai-landscape.at*, *asai.ac.at*, *brutkasten.com/artikel/nxai-ai-experte-sepp-hochreiter-gruendet-neues-ki-startup*.

9 *fwf.ac.at/aktuelles/detail/oesterreichs-naechste-exzellenzcluster-starten*.

10 *ai-at.eu*.

ZITIERTE LITERATUR

Bomke, L., 2024, Intel, Allianz und Unilever setzen auf Silo AI, *Handelsblatt, Wochenende 5./6./7. April 2024.*

Chen, N. (2025, 9. Juni). *Stolen Stories or Fair Use? The New York Times v. OpenAI and the Limits of Machine Learning*. Columbia Undergraduate Law Review. ***culawreview.org/ddc-x-culr-1/nyt-v-openai-and-microsoft***.

Constantin, L. (2025, 8. August). Black Hat: Researchers demonstrate zero-click prompt injection attacks in popular AI agents. *CSO Online*. ***csoonline.com/article/4036868/black-hat-researchers-demonstrate-zero-click-prompt-injection-attacks-in-popular-ai-agents.html***.

ETH Zürich. (2025, 2. September). Apertus: Ein vollständig offenes, transparentes und mehrsprachiges Sprachmodell. *ETH Zürich*. ***ethz.ch/de/news-und-veranstaltungen/eth-news/news/2025/09/medienmitteilung-apertus-ein-vollstaendig-offenes-transparentes-und-mehrsprachiges-sprachmodell.html***.

Fenech-Borg, E. Z., et al. (2025). *The Cultural Gene of Large Language Models: A Study on the Impact of Cross-Corpus Training on Model Values and Biases* (No. arXiv:2508.12411). arXiv. ***doi.org/10.48550/arXiv.2508.12411***.

Hupkes, D., & Bogoychev, N. (2025). *MultiLoKo: A multilingual local knowledge benchmark for LLMs spanning 31 languages* (No. arXiv:2504.10356). arXiv. ***doi.org/10.48550/arXiv.2504.10356***.

Löser, A., Tresp, V. et al., 2023, *Große Sprachmodelle entwickeln und anwenden – Ansätze fuŕ ein souveränes Vorgehen*, München: Whitepaper aus der Plattform Lernende Systeme, acatech, ***doi.org/10.48669/pls_2023-6***.

Martins, P. H., et al. (2025). EuroLLM: Multilingual Language Models for Europe. *Procedia Computer Science*, *255*, 53–62. ***doi.org/10.1016/j.procs.2025.02.260***.

Meffert, K., 2023, Künstliche Intelligenz: Deutsche Texte in KI-Sprachmodellen, *Dr. DSGVO Blog*, ***dr-dsgvo.de/kuenstliche-intelligenz-deutsche-texte-in-ki-sprachmodellen***.

Nicholas, G. und Bhatia, A., 2023, *Lost in Translation: Large Language Models in Non-English Content Analysis*, im Auftrag von: Center for Democracy & Technology: cdt Research, ***cdt.org/insights/lost-in-translation-large-language-models-in-non-english-content-analysis/***; auch veröffentlicht als: ***arxiv.org/abs/2306.07377***.

Ramesh, K., Sitaram, S. und Choudhury, M., 2024, *Fairness in Language Models Beyond English: Gaps and Challenges*; ***arxiv.org/pdf/2302.12578.pdf***.