ECONSTOR
Make Your Publications Visible.

A Service of
ZBW
Leibniz-Informationszentrum Wirtschaft
Leibniz Information Centre for Economics

Bienert, Joerg; Klös, Hans-Peter

Research Report

Große KI-Modelle als Basis für Forschung und wirtschaftliche Entwicklung

IW-Kurzbericht, No. 58/2022

Provided in Cooperation with:
German Economic Institute (IW), Cologne

Suggested Citation: Bienert, Joerg; Klös, Hans-Peter (2022) : Große KI-Modelle als Basis für Forschung und wirtschaftliche Entwicklung, IW-Kurzbericht, No. 58/2022, Institut der deutschen Wirtschaft (IW), Köln

This Version is available at:
https://hdl.handle.net/10419/261457

Standard-Nutzungsbedingungen:

Die Dokumente auf EconStor dürfen zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden.

Sie dürfen die Dokumente nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, öffentlich zugänglich machen, vertreiben oder anderweitig nutzen.

Sofern die Verfasser die Dokumente unter Open-Content-Lizenzen (insbesondere CC-Lizenzen) zur Verfügung gestellt haben sollten, gelten abweichend von diesen Nutzungsbedingungen die in der dort genannten Lizenz gewährten Nutzungsrechte.

Terms of use:

Documents in EconStor may be saved and copied for your personal and scholarly purposes.

You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.

If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.

WWW.ECONSTOR.EU

Mitglied der Leibniz-Gemeinschaft

IW-Kurzbericht 58/2022

Große KI-Modelle als Basis für Forschung und wirtschaftliche Entwicklung

Joerg Bienert / Hans-Peter Klös, 05.07.2022

Unternehmen und Forschungseinrichtungen in den USA und China haben seit 2020 bahnbrechende Ergebnisse mit sogenannten großen KI-Modellen erzielt. Große Sprachmodelle wie z.B. GPT-3 erlauben bisher nicht realisierbare Anwendungen, vor allem in den Bereichen der Text-Erstellung und Dokumenten-Verarbeitung. Solchen großen KI-Modelle revolutionieren den Markt für Künstliche Intelligenz und werden sich disruptiv auf die gesamte Wirtschaft auswirken. Deutschland und Europa dürfen hier nicht den Anschluss verlieren.

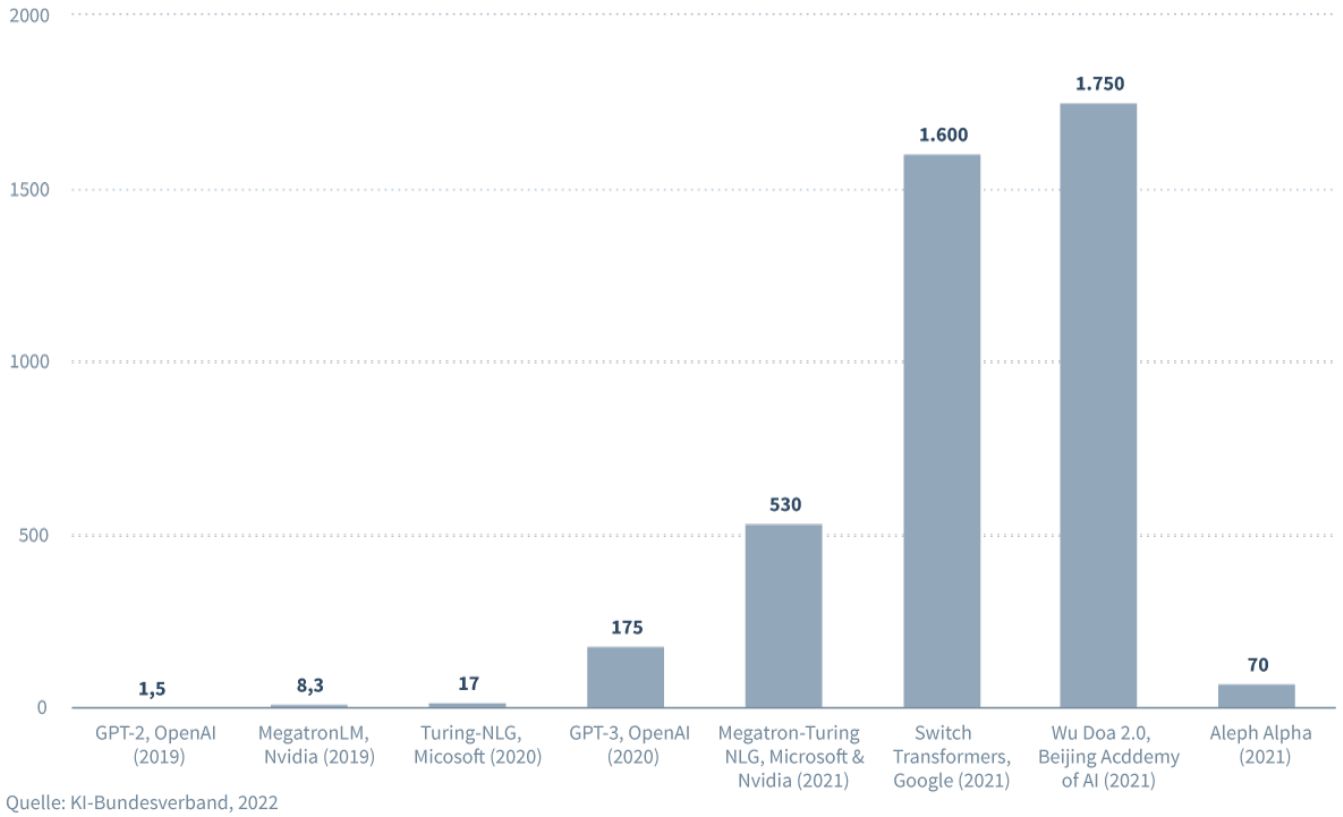
Aktuelle Fortschritte in Deutschland und Europa, aber noch deutlicher Rückstand

Das European High Performance Computing Joint Undertaking (EuroHPC JU) hat vor wenigen Tagen das Forschungszentrum Jülich als Partner im deutschen Gauss Centre for Supercomputing ausgewählt, um den ersten europäischen Supercomputer der nächsten Generation zu betreiben (BMBF, 2022). Erst wenige Wochen zuvor hat das Leibniz-Rechenzentrum der Bayerischen Akademie der Wissenschaften als erstes wissenschaftliches Rechenzentrum in Europa mit der Nutzung der Wafer Scale Engine 2 mit 2,6 Billionen Transistoren und 850.000 Rechenkernen eine weitere Grundlage speziell für Verfahren künstlicher Intelligenz (KI) entwickelt (StMWK, 2022).

Trotzdem hinken die aktuellen Entwicklungen in Deutschland und Europa im Bereich KI hinter dem schon erreichten Stand in den USA und China hinterher: Seit der Veröffentlichung von GPT-2 (Generative Pretrainierte Transformer-2) - eine Open-Source-künstliche Intelligenz, die von OpenAI im Februar 2019 erstellt wurde - hat sich die Anzahl der Parameter, mit denen große KI-Modelle trainiert werden, mehr als vertausendfacht. Das größte bis heute bekannte europäische Modell Aleph Alpha mit ca. 70 Milliarden Parametern ist im Vergleich mit den neuesten Modellen amerikanischer und chinesischer Entwickler mit mehr als 1,6 Billionen Parametern noch sehr klein (Grafik)

Ein Grund für diesen Rückstand: Europäischen KI-Entwicklern fehlt es an verfügbarer und speziell für die KI-Entwicklung abgestimmter Infrastruktur. Deshalb sind erheblich Investitionen in die Entwicklung und Bereitstellung großer KI-Modelle sowie KI-Infrastrukturen erforderlich. Dies hat auch damit zu tun, dass es bisher noch kein großes privatwirtschaftliches Investment in Deutschland und Europa in große KI-Modelle und die erforderliche Infrastruktur gibt. Die bereits von der EU-Kommission erkannte Gefahr einer strategischen Abhängigkeit im Bereich Cloud und Edge Computing (Europäische Kommission, 2021, S. 91 ff). erstreckt sich nun auch auf den Bereich der KI.

Anzahl Parameter großer KI-Modelle, in Milliarden



Quelle: KI-Bundesverband, 2022

Große KI-Modelle revolutionieren den Markt für Künstliche Intelligenz

Die Entwicklung einer Künstlichen Intelligenz beruhte in der Vergangenheit vor allem auf dem Trainieren individueller und für eine einzige Aufgabe spezialisierter KI-Modelle. Vor einigen Jahren gelang dann Forschern amerikanischer Industrielabore der entscheidende Durchbruch. Neue generelle KI-Modelle, die mit riesigen Text- und Datenmengen trainiert wurden, zeigten enorme Performanzen

Heute ermöglichen große KI-Sprachmodelle nicht nur das automatisierte Schreiben von Texten in nahezu menschenähnlicher Schreibqualität, sondern sie können auch selbstständig programmieren. KI-Sprachmodelle sind also nicht nur wichtige Grundlage für die Mensch-Maschine-Kommunikation, sondern in Zukunft auch für die Automatisierung weitreichender Wirtschaftsprozesse sowie als praktisch unendlich skalierbare Software-Entwicklungsressourcen einsetzbar.

Multimodale Modelle, die Text- und Bilddaten miteinander kombinieren, erstellen bereits Fotos aus Texten oder erkennen den Inhalt eines Videos - Fähigkeiten, die traditionelle KI-Modelle nur schwer entwickeln konnten. Der Vorteil der großen KI-Modelle ist, dass sie nicht hochspezialisiert sind und damit für jede Anwendung neu entwickelt werden müssen, sondern mit entsprechenden Anpassungen vielseitig einsetzbar sind.

Namentlich für das Geschäftsmodell Deutschlands (Industrie 4.0) ist die Verarbeitung großer Datenmengen absolut erfolgskritisch. Die zunehmende Bedeutung digitaler Zwillinge, die Bemühungen um einen gemeinsamen europäischen „Industrial Dataspace“ (z.B. Gaia-X und Catena-X) machen die Verfügbarkeit großer KI-Rechenkapazitäten als neuer kritischer Infrastruktur deutlich. Aber auch für neue Technologiefelder (z.B autonomes Fahren; Nanorobotik; Human Augmentation; grüne Technologie) ist eine ausreichende Rechnerkapazität für KI-Anwendungsfelder absolut essenziell.

Eine Abhängigkeit von nicht-europäischen KI-Lösungen muss daher dringend vermieden werden, um nicht wertvolle Unternehmensdaten preisgeben zu müssen, nicht in vollem Umfang an den Vorteilen der Künstlichen Intelligenz partizipieren zu können oder um nicht KI-Monopole in den Händen großer US-amerikanischer und chinesischer Technologiekonzerne entstehen zu

lassen. Alles in allem geht es darum, die digitale Souveränität Deutschlands und Europas zu sichern und eine Spitzenposition in dieser zentralen Schlüsseltechnologie einzunehmen.

Was getan werden kann

Die beiden jüngsten Förderentscheide und auch weitere Initiativen, wie das vom BMWK geförderte Projekt Open-GPT-X oder die Förderbekanntmachung zu KI-Rechenzentren, zielen aktiv darauf ab, den technologischen Rückstand zu verkleinern. Zwar verfügt Europa zur Erstellung großer KI-Modelle in der Forschung über die notwendigen Grundlagen bzgl. Systemen und Algorithmen und auch über die Top-Talente in Wissenschaft und Forschung im Bereich der Sprachmodelle. Dennoch fehlt es derzeit noch an einer ausreichenden und für die Entwicklung großer KI-Modelle dedizierten Infrastruktur.

Deshalb scheint eine breite europäische Initiative für große KI-Modelle (Large European AI-Models, LEAM) erforderlich zu sein (KI-Bundesverband, 2022). Dazu gehört in erster Linie der Aufbau einer gemeinsamen europäischen Infrastruktur, auf der große KI-Modelle aufgebaut werden können, die Zusammenstellung großer Korpora von Trainingsdaten nach europäischen Bedürfnissen und Werten sowie das Training und die Bereitstellung großer KI-Modelle. Folgende Eckpfeiler sind für eine solche Initiative relevant:

Etablierung eines KI-Hochleistungs-Rechenzentrums: Der Aufbau und der Betrieb eines dezidierten KI-Rechenzentrums könnte Europa selbst zum Motor für Innovationen in diesem Bereich machen. Dazu muss die technische Infrastruktur von Anfang an in Hardwareauswahl und – konfiguration sowie in Rechen- und Speicherkapazität konkurrenzfähig mit den Laboren sein, in denen die Durchbrüche der letzten Jahre erzielt wurden.

Aufbau von Kompetenzen und Personalkapazitäten: Parallel muss personelle Kompetenz in hinreichender Kapazität aufgebaut werden, um die ersten großen europäischen KI-Modelle zu schaffen. Dabei müssen auch Forschung und Technologieentwicklung konzentriert gefördert werden, um in wichtigen Anwendungsgebieten die Führung zu übernehmen.

Sammlung von Daten und Entwicklung von Algorithmen: Bei der Auswahl der Daten und Algorithmen müssen die besonderen Bedürfnisse der europäischen Gesellschaft berücksichtigt werden (z.B. die Einhaltung ethischer Normen und Verlässlichkeit in kritischen Wissensbereichen). Zudem müssen die Planung und der Einsatz der Infrastruktur, die Auswahl der Algorithmen, die Gestaltung der Trainingsprozesse und auch die Organisation und technische Ausführung so angelegt sein, dass eine maximal mögliche Energieeffizienz erreicht wird.

Etablierung einer eigenständigen Organisationseinheit: Es bedarf einer Organisation, welche die entstandenen Modelle für die europäische Industrie und Forschung zur Verfügung stell, mit hinreichend mächtigen Cloudstrukturen für Nutzer aus Wirtschaft, Wissenschaft und Verwaltung, die das Training der Modelle für spezielle Anwendungen (Fine-Tuning), den Einsatz der Modelle in fertigen Anwendungen (Inferencing) und das Experimentieren für die Verbesserung der Modelle und die Erforschung neuer Anwendungen ermöglichen.

Die Vision ist mithin ein Fokussierungspunkt, um den herum sich ein leistungsfähiges, vielfältiges Ökosystem aus Wirtschaft, Forschung, neuen Geschäftsmodellen und Startups bildet. In enger Abstimmung und Zusammenarbeit mit bestehenden Forschungseinrichtungen, Unternehmen, Verbänden und Initiativen könnte so ein Kristallisationspunkt für die zielgerichtete Sammlung von Daten (OpenData), breit gefächerte Forschungsvorhaben für die Entwicklung von wertschöpfenden Anwendungen für alle Branchen und wirtschaftliches Wachstum entstehen.

Literatur

BMBF, 2022, Stark-Watzinger: Supercomputer kommt nach Deutschland, Pressemitteilung 45/22 vom 15.06.2022

Europäische Kommission, 2021, Strategic dependencies and capacities, Commission Staff Working Document, SWD (2021) 352 final, Brüssel

KI-Bundesverband, 2022, Large European AI Models (LEAM) als Leuchtturmprojekt für Europa, Konzeptpapier, o.O.

StMWK Bayern, 2022, Weltweit größter Superchip zur Verarbeitung von Big Data jetzt in Bayern am Leibniz-Rechenzentrum, Pressemitteilung 090/22, 25. Mai 2022