Deutsches Forschungszentrum für Künstliche Intelligenz
German Research Center for Artificial Intelligence

dfki ainext

Vertrauenswürdige KI und Sprachtechnologie

Herausforderungen, Methoden und Infrastruktur für die Qualitätssicherung von KI-Systemen

KI für den Menschen –
Intelligente Lösungen für die Wissensgesellschaft

Das DFKI forscht seit über 35 Jahren an KI für den Menschen und orientiert sich an gesellschaftlicher Relevanz und wissenschaftlicher Exzellenz in den entscheidenden zukunftsorientierten Forschungs- und Anwendungsgebieten der Künstlichen Intelligenz.

Die Deutsche Forschungszentrum für Künstliche Intelligenz GmbH (DFKI) wurde 1988 als gemeinnützige Public-Private Partnership gegründet. Das DFKI unterhält Standorte in Kaiserslautern, Saarbrücken, Bremen, Niedersachsen (Osnabrück und Oldenburg), Labore in Berlin, Darmstadt und Lübeck sowie eine Außenstelle in Trier.

In 27 Forschungsbereichen, zehn Kompetenzzentren und acht Living Labs werden ausgehend von anwendungsorientierter Grundlagenforschung Produktfunktionen, Prototypen und patentfähige Lösungen im Bereich der Informations- und Kommunikationstechnologie entwickelt. Die Finanzierung erfolgt über Zuwendungen öffentlicher Fördermittelgeber sowie durch Entwicklungsaufträge aus der Industrie.

Projektergebnisse und Meilensteine werden periodisch institutionell und durch ein international besetztes Expertengremium (Wissenschaftlicher Beirat) begutachtet. Neben dem Bundesministerium für Bildung und Forschung und den Bundesländern Rheinland-Pfalz, Saarland, Bremen und Niedersachsen sind im DFKI-Aufsichtsrat zahlreiche namhafte deutsche und internationale Hochtechnologie-Unternehmen aus einem breiten Branchenspektrum vertreten.

www.dfki.de

Vertrauenswürdige KI für die Wirtschaft

Die EU-Verordnung zur Regulierung Künstlicher Intelligenz ist am 1. August 2024 in Kraft getreten, wobei es für betroffene Unternehmen und Organisationen unterschiedliche Übergangsfristen gibt. Beim AI Act verfolgt die Europäische Union einen risikobasierten Ansatz: Anwendungen, die erhebliche Auswirkungen auf Gesundheit, Sicherheit oder Grundrechte haben, werden als Hochrisikosysteme eingestuft. Um in kritischen Infrastrukturen, in der Finanzbranche oder im Gesundheitswesen eingesetzt werden zu können, müssen solche Systeme vertrauenswürdig sein und strenge Anforderungen erfüllen an Transparenz, Korrektheit, Erklärbarkeit, technisch-inhaltliche Robustheit, menschliche Kontrolle und Datenschutz.

Für Anbieter und kommerzielle Nutzer bedeutet dies, dass jede Hochrisiko-KI-Anwendung vor ihrem Einsatz einer Prüfung unterzogen wird. Es muss nachvollziehbar sein, wie ein System funktioniert und wie es Entscheidungen trifft (Transparenz und Erklärbarkeit). Im operativen Einsatz müssen KI-Systeme technisch stabil, fehlertolerant und vor Manipulation geschützt sein (Robustheit und Sicherheit). In sensiblen Bereichen dürfen keine vollständig autonomen Entscheidungen ohne menschliche Kontrolle getroffen werden.

Der aktuelle Stand der Technologie erfüllt diese Anforderungen nur begrenzt. Insbesondere große Sprachmodelle sind intransparent und deshalb schwer nachvollziehbar. An dieser Stelle kommt der Forschung eine zentrale Bedeutung zu, da sie vertrauenswürdige KI entscheidend vorantreibt. Forschungsinitiativen konzentrieren sich auf die Entwicklung von Garantien für Algorithmen und Verfahren, sodass KI-Anwendungen auch in unsicheren oder unvorhersehbaren Umgebungen sicher funktionieren und fair entscheiden.

Forschung ist der Schlüssel für vertrauenswürdige KI. Sie kann das Wissen und die Instrumente bereitstellen, um KI-Systeme robust, fair und transparent zu gestalten. Sie ermöglicht es der Industrie, KI sicher und verantwortungsvoll einzusetzen, Innovationen voranzutreiben und gleichzeitig regulatorischen, ethischen und gesellschaftlichen Anforderungen gerecht zu werden. Dies eröffnet neue Geschäftschancen, mindert Risiken und schafft Vertrauen.

Prof. Dr. Antonio Krüger
CEO

Potenziale vertrauenswürdiger Sprachtechnologie für die europäische Wirtschaft

Unternehmen aller Branchen stehen vor der Herausforderung, KI-Anwendungen zu implementieren, ohne die genauen Folgen der politischen Regulierung durch den EU AI Act zu kennen. Nur im gemeinsamen Dialog können Wissenschaft, Wirtschaft und Politik verlässliche Standards für vertrauenswürdige KI definieren, die für Anwender Sicherheit und für Unternehmen Innovationsanreize schaffen.

Autor: Prof. Dr. Antonio Krüger

Innovative KI-Verfahren sind von hoher Relevanz für die Wertschöpfung in verschiedenen Branchen und für Unternehmen unterschiedlicher Größe. Ganz gleich, ob wir neue KI-Technologien im Gesundheitswesen, in der Energieversorgung oder im Verkehrswesen einsetzen: Mit der zunehmenden Verbreitung von KI wächst auch das Bewusstsein, dass diese Systeme vertrauenswürdig sein müssen, damit sie auch in missionskritischen Entscheidungen eingesetzt werden können. Doch die Schlussfolgerungen der Systeme sind teilweise schwer nachvollziehbar, die Datenbasis unbekannt und die Algorithmen intransparent. Für marktfähige Lösungen und vor dem Hintergrund der politischen Regulierung durch den EU AI Act müssen belastbare Kriterien und Anforderungen für vertrauenswürdige KI entwickelt werden.
Unternehmen brauchen praktische Werkzeuge und Standards sowie Prüf- und Zertifizierungsverfahren, mit denen sie den technologischen Anforderungen an vertrauenswürdige KI-Anwendungen gerecht werden können.

Anforderungen an vertrauenswürdige KI

Vertrauen ist vor allem dann erforderlich, wenn die Risiken zunehmen: Der europäische Ansatz für vertrauenswürdige KI unterscheidet verschiedene Stufen der Kritikalität von KI-Anwendungen. So müssen KI-Systeme, die innerhalb der Europäischen Union entwickelt oder eingesetzt werden, je nach Risikoeinstufung unterschiedliche Anforderungen erfüllen.
Um die Vertrauenswürdigkeit von KI-Systemen zu verbessern, sollten Aspekte wie Robustheit, algorithmische Fairness, Erklärbarkeit und Transparenz berücksichtigt werden. Darüber hinaus muss die Vertrauenswürdigkeit während des gesamten Lebenszyklus eines KI-Systems systematisch ermittelt und bewertet werden.

DFKI forscht zu Vertrauenswürdigkeit großer Sprachmodelle

Die verschiedenen Anforderungen an vertrauenswürdige KI lassen sich gut im Bereich der KI-basierten

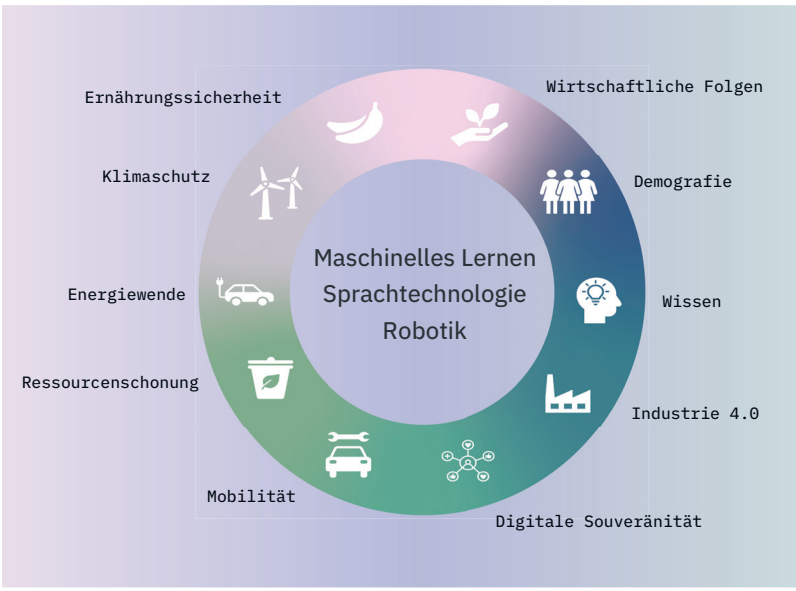
Sprachtechnologie verdeutlichen. Large Language Models (LLMs) haben ein disruptives Potenzial und sprachbasierte KI-Werkzeuge wie GPT-4 sind mittlerweile im privaten und öffentlichen Bereich allgegenwärtig. Oftmals erfüllen sie die Anforderungen an die Vertrauenswürdigkeit nur in geringem Maße, was ihre Einsetzbarkeit in Anwendungsszenarien mit hohem Sicherheitsbedarf (wie Medizin oder Rechtsprechung) stark einschränkt.

Das liegt daran, dass Parameter, Trainingsdaten, Trainingsmethoden und Inferenzeinstellungen oft nicht zugänglich sind. Diese und weitere offene Fragen im Bereich der Grundlagenforschung zur Vertrauenswürdigkeit großer Sprachmodelle werden derzeit am DFKI erforscht: Das DFKI Speech and Language Technology Lab in Berlin erarbeitet verlässliche Evaluationskriterien für vertrauenswürdige KI-basierte Sprachtechnologie. Diese beziehen sich unter anderem auf Robustheit und Zuverlässigkeit und gehen beispielsweise der Frage nach, wie die Performanz von Sprachmodellen in spezifischen Domänen, Anwendungen oder auch über Sprachgrenzen hinweg gemessen werden kann. Auch im Hinblick auf Transparenz und Erklärbarkeit muss u. a. weiter erforscht werden, wie Wissen in großen Sprachmodellen lokalisiert und wie bestimmtes Modellverhalten einzelnen Modellkomponenten zugeordnet werden kann. Im Rahmen des Centre for European Research on Trusted AI (CERTAIN) untersuchen Forscherinnen und Forscher am DFKI derzeit mit Methoden der mechanistischen Interpretierbarkeit, ob LLMs komplexe Regeln lernen oder nur Fakten abrufen können.

Im Projekt OpenGPT-X entwickelt das DFKI verschiedene Ansätze, darunter auch Open-Science-Konzepte. Der offene Ansatz ist ein möglicher Weg zur Entwicklung einer sichereren und vertrauenswürdigeren KI und fördert zudem die Beteiligung aller an der Gestaltung der Zukunft dieser Technologie. In diesem Sinne verfolgt auch die vom DFKI ins Leben gerufene Forschungsinitiative Occiglot das Ziel einer digitalen Sprachgerechtigkeit in Europa. Hochwertige Grundlagenforschung und wirtschaftlich verwertbare technologische Anwendungen erfordern den direkten Zugang zu Sprachmodellen und zu den Daten, mit denen sie trainiert wurden. Zu diesem Zweck engagiert sich das akademische, gemeinnützige Forschungskollektiv Occiglot für eine offene, transparente und vollständig dokumentierte Wissenschaft und damit für die Entwicklung von Open Source LLMs.

Innovationshemmnisse durch Regulierung abbauen

Die Europäische Union arbeitet bereits an Standards, um den AI Act als europäisches Recht umzusetzen und die Anwendung in der Praxis zu konkretisieren. Damit Unternehmen in Europa die Potenziale innovativer KI-basierter Sprachtechnologien voll ausschöpfen können, müssen aus den Konzepten für die technischen und ethischen Anforderungen an vertrauenswürdige Systeme überprüfbare Standards und belastbare Testverfahren abgeleitet werden. Langfristig könnten höhere Sicherheitsanforderungen zur breiteren Akzeptanz bei den Anwendern und mehr Klarheit bei Investitionsentscheidungen für Unternehmen führen. Im besten Fall sollten innovative Entwicklung, rechtliche Ausgestaltung und technische Umsetzung reibungslos synchronisiert werden. Dazu bedarf es eines lebendigen Dialogs zwischen Wirtschaft, Wissenschaft und Politik.



Trusted AI – Vertrauenswürdigkeit und große Sprachmodelle

Interview mit Dr. André Meyer-Vitali und Dr. Simon Ostermann

ChatGPT, LLaMA oder Mistral liefern erstaunliches Fachwissen und ganz offensichtlich falsche Antworten. Was ist der Grund für diese Unzuverlässigkeit?

Dr. André Meyer-Vitali forscht zu hybrider und verteilter Künstlicher Intelligenz. Er ist Principal Investigator am CERTAIN-Zentrum für Europäische Forschung zu Vertrauenswürdiger KI.

Meyer-Vitali: Generative Pretrained Transformer sind große, hochgradig vernetzte Modelle, dadurch opak und nur schwer kontrollierbar. Die zugrunde liegende Technologie – Deep Learning – baut kein wirkliches Verständnis des Problems auf, sondern bildet lediglich komplexe statistische Zusammenhänge ab.

Ostermann: Das liegt nicht nur daran, dass die Technologie eine Black Box ist. Viele Anbieter stellen den Quellcode ihrer Systeme nicht zur Verfügung. Ohne Einblick in Parameter, Trainingsdaten, Trainingsmethoden und Inferenzeinstellungen ist es schwierig nachzuvollziehen, wie diese Modelle zu ihren Ergebnissen gelangen.

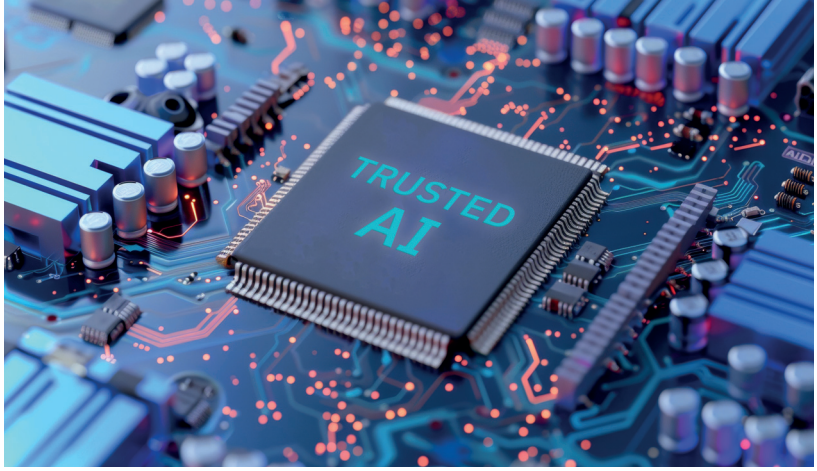
Wie kann erreicht werden, dass Generative KI vertrauenswürdiger wird?

Meyer-Vitali: Unter dem Begriff „Trusted AI“ wird ein neuer Gesamtansatz für zuverlässige Systeme vorangetrieben. Das Ziel ist eine neue KI-Generation, die Garantien über ihre Funktionalität bereitstellt und so eine Verwendung insbesondere auch in kritischen Anwendungen ermöglicht.

Trusted AI ist gekennzeichnet durch ein hohes Maß an Zuverlässigkeit, Sicherheit, Transparenz, Robustheit, Fairness und Verifizierbarkeit, wobei die Funktionalität bestehender Systeme keineswegs beeinträchtigt, sondern sogar verbessert werden soll. Entwickler, Benutzer und Regulatoren sollen auf die Leistung und Zuverlässigkeit auch in komplexen sozio-technischen Systemen vertrauen können.

Brauchen wir einen technologischen Neustart?

Meyer-Vitali: Nicht ganz, aber die Grundlage für eine neue Generation von KI-Systemen werden eher hybride Systeme sein, die sich nicht nur auf datengetriebene Ansätze stützen, sondern das gesamte Spektrum von KI-Techniken nutzen, einschließlich symbolischer KI-Methoden, Suche, Argumentation und Planung.



Was ist von hybriden KI-Systemen zu erwarten?

Meyer-Vitali: Der Einsatz neuro-symbolischer Modelle könnte die Validierung erleichtern, mehr Transparenz schaffen und eine stärkere Rechenschaftspflicht fördern, während kausale Modelle verständliche Erklärungen und Einblicke in die Entscheidungsprozesse der KI bieten. Durch die Kombination von Machine Learning mit symbolischem Schlussfolgern und der expliziten Repräsentation von Wissen in hybriden KI-Systemen wird „Trust by Design“ ermöglicht. Semantische oder andere explizite Modelle können Wissen darstellen, das dann nicht mehr maschinell gelernt werden muss. Außerdem können sie den Lernprozess in die richtige Richtung steuern, wodurch das Generalisieren, die Robustheit und die Interpretierbarkeit verbessert werden. Dieser hybride Ansatz wird gerne auch als dritte Welle der KI bezeichnet („3rd Wave of AI“).

Welchen Forschungsfragen für zuverlässige LLMs geht die Sprachtechnologie zurzeit nach?

Ostermann: Für uns steht die Frage im Fokus, wie Sprachmodelle insbesondere in sicherheitskritischen Bereichen wie Medizin oder Rechtsprechung deterministische und angemessene Entscheidungen treffen können. Am CERTAIN Centre for European Research on Trusted AI untersuchen wir derzeit, ob LLMs in der Lage sind, komplexe Regeln zu lernen, oder ob sie nur Fakten wiedergeben können. Dabei verwenden wir die mechanistische Interpretierbarkeit, einen neuen Ansatz zum Verständnis neuronaler Netze. Ziel ist es, die Fähigkeiten dieser Modelle durch „Reverse Engineering“ besser zu verstehen und das menschliche Verständnis ihrer Funktionsweise zu vertiefen.

Hilft das Konzept der mechanistischen Interpretierbarkeit auch bei der Beherrschung des Bias-Problems bei KI-Systemen?

Ostermann: Wissenschaftliche Studien konnten zeigen, dass sexistisches Verhalten auf spezifische Modellkomponenten zurückzuführen ist, die deaktiviert werden können, um solche Reaktionen zu unterdrücken. Wir wollen herausfinden, wie sich Wissen in großen Sprachmodellen lokalisieren lässt und wie bestimmtes Verhalten auf einzelne Modellkomponenten zurückgeführt werden kann.

Lässt sich damit auch verhindern, dass Sprachmodelle manipulierte Antworten geben?

Ostermann: Ein besonderes Augenmerk unserer Arbeit liegt auf dem Schutz vor Manipulation durch Prompt-Injektionen: Wie kann verhindert werden, dass Dritte die Ausgabe eines Sprachmodells durch „Zwischenprompts“ verfälschen? Die Forschenden setzen hier auf den Einsatz konkreter Beispiele und untersuchen, wie solche sorgfältig ausgewählten Beispiele dabei helfen können, die Grenzen eines Modells aufzuzeigen.

Dr. Simon Ostermann ist Computerlinguist und forscht zu transparenten und robusten Sprachmodellen. Am DFKI ist er Senior Researcher im Forschungsbereich Sprachtechnologie und Multilingualität sowie Co-Leiter des Kompetenzzentrums Generative KI.

MISSION KI Nationale Initiative für Künstliche Intelligenz und Datenökonomie

Qualitäts- und Prüftechniken für Vertrauenswürdigkeit in medizinische KI-Systeme

Autoren: Ludger van Elst und Adriano Lucieri

« Die Ansiedlung des Innovations- und Qualitätszentrums am DFKI ermöglicht es uns, vorhandene Strukturen zu nutzen und über die breite Kompetenz unserer Forschungsbereiche sowie die weitreichenden Netzwerke im Bereich vertrauenswürdiger KI wertvolle Synergien zu schaffen. Als Forschungspartner von Unternehmen verschiedenster Branchen wissen wir um die besonderen Herausforderungen hinsichtlich Sicherheit, Zuverlässigkeit und ethische Vertretbarkeit, die mit der Entwicklung von marktfähigen KI-Anwendungen einhergehen. Mit dem Innovations- und Qualitätszentrum möchten wir deshalb eine Anlaufstelle für alle Unternehmen schaffen, um diesen Zugang zu KI-Spitzenforschung zu bieten. »

Prof. Dr. Andreas Dengel, Geschäftsführender Direktor DFKI Kaiserslautern

Mangelnde Vertrauenswürdigkeit ist nicht nur ein großer Risikofaktor bei der Anwendung von KI-Systemen, sondern stellt auch ein mögliches Hemmnis für deren Verbreitung und Erfolg dar. Für Unternehmen ist die Investition in KI-basierte Anwendungstechnologien aufgrund unsicherer Marktchancen und der technischen Komplexität mit einem hohen wirtschaftlichen Risiko verbunden. Nutzerinnen und Nutzer können die Qualität der Ergebnisse oft kaum überprüfen, sodass sie die neuen Technologien unter Umständen nur zögerlich akzeptieren oder aber schwer einschätzbare Sicherheitsrisiken eingehen. Um diese Situation zu verbessern, sind zwei Fragen grundlegend: Wie können Anbieter bereits bei der Entwicklung eine hohe Qualität und Vertrauenswürdigkeit ihrer KI-basierten Systeme sicherstellen? Und umgekehrt: Wie können Nutzerinnen und Nutzer möglichst einfach erkennen, welchen Systemen auf dem Markt sie tatsächlich in dem für sie notwendigen Maße vertrauen können?

Ein wichtiger Baustein für eine fundierte Einschätzung der Vertrauenswürdigkeit von KI-Komponenten und -Systemen kann die Entwicklung transparenter und einheitlicher freiwilliger Qualitäts- und Prüfstandards sein, die im Einklang mit europäischen Werten stehen und den Anforderungen des EU AI Act entsprechen.

Im Projekt TrustifAI entwickelt das DFKI am Standort Kaiserslautern spezifische Qualitätskriterien sowie Prüfverfahren und -methoden, um die Einhaltung eines festgelegten Mindeststandards sicherzustellen. Dieser Standard zielt darauf

ab, die Vertrauenswürdigkeit der KI zu gewährleisten und umfasst unter anderem die Bereiche Transparenz, Fairness, Datenschutz, Robustheit und Cybersicherheit. Dabei wird das Prinzip „Trust by Design“ verfolgt: Qualitätssicherungsverfahren werden frühzeitig in den Entwicklungsprozess von Software integriert. Zur Sicherstellung der Praxistauglichkeit werden Prüfmethoden zur Messung der Vertrauenswürdigkeit erarbeitet. Die Validierung erfolgt anhand von fünf ausgewählten Anwendungsfällen aus den Bereichen Medizin und Gesundheitswesen und wird in speziell entwickelten Testumgebungen durchgeführt.

Dass sich die Arbeiten des DFKI zunächst auf den Einsatz von KI im Gesundheitswesen konzentrieren, liegt auch an der starken Präsenz von Unternehmen und Forschungseinrichtungen aus Medizin, Pharma und Biotechnologie in Rheinland-Pfalz. Zu den Anwendungsfällen, die im Rahmen von TrustifAI untersucht werden, gehören die Erkennung von Hautkrebs anhand dermatoskopischer Bilder, die Entscheidungsunterstützung bei der Behandlung von Tumoren, die Unterstützung bei der Bedarfsprognose und -planung im Rettungsdienst, die Verbesserung von Behandlungsplänen für psychiatrische Patienten sowie die Erkennung potenzieller Probleme bei der präoperativen Intubation anhand medizinischer Videodaten.

Die Forschung im Projekt TrustifAI ist Teil von MISSION KI, einer Initiative zur Förderung künstlicher Intelligenz in Deutschland. Die Initiative wird vom Bundesministerium für Digitales und Verkehr mit einem Gesamtbudget von 32 Millionen Euro gefördert und durch acatech – Deutsche Akademie der Technikwissenschaften – koordiniert. Sie schafft eine Plattform für die Zusammenarbeit führender Organisationen aus den Bereichen Testen, Standardisierung und Wissenschaft. Das IQZ Kaiserslautern ist das erste von zwei durch MISSION KI geplanten Zentren, mit denen die Bundesregierung den Einsatz von vertrauenswürdiger KI an der Schnittstelle zwischen Spitzenforschung und praktischer Anwendung vorantreibt. Das Zentrum in Kaiserslautern dient hauptsächlich kleinen und mittleren Unternehmen sowie potenziellen Gründern als zentrale Anlaufstelle für alle Fragen rund um vertrauenswürdige KI. Es bietet ein umfassendes Programm sowie Beratungsleistungen, die den Wissenstransfer im Bereich vertrauenswürdiger KI und KI-Prüfung gezielt fördern. Durch regelmäßige Vorträge, praxisnahe Workshops und branchenübergreifende Netzwerkveranstaltungen wird zugleich der Aufbau einer starken, regionalen KI-Community vorangetrieben, die den Innovationsstandort nachhaltig stärkt. Das Zentrum ist dabei nicht auf den medizinischen Bereich beschränkt, sondern steht Innovatoren aus allen Branchen offen. Ein zweites Zentrum soll durch die MISSION KI-Initiative des Bundesdigitalministeriums und acatech bis Jahresende in Berlin entstehen.

« Um Vertrauen in neuartige Anwendungen zu schaffen, muss KI-Qualität von Beginn der Entwicklung an mitgedacht werden. Daher freuen wir uns darauf, gemeinsam systematisch an Methoden zur Qualitätsverbesserung von KI-Anwendungen zu arbeiten und entsprechende Prüfkriterien zu etablieren, die perspektivisch auch Sektoren außerhalb des Gesundheitsbereichs zugutekommen sollen. »

Prof. Dr. Sebastian Vollmer
Leiter des DFKI-Forschungsbereichs Data Science und ihre Anwendungen

Occiglot: Open-Source-Sprachmodelle *aus* Europa *für* Europa

Autoren: Prof. Dr. Georg Rehm, Dr. Patrick Schramowski

Die jüngsten Fortschritte bei großen Sprachmodellen (LLMs) zeigen das disruptive Potenzial dieser Technologie. Hohe Entwicklungs- und Trainingskosten sowie Fachkenntnisse führen jedoch dazu, dass das Feld von wenigen großen Tech-Unternehmen und Deep-Tech-Startups dominiert wird. Diese stellen zentrale europäische Werte wie sprachliche Vielfalt und kulturellen Reichtum zugunsten wirtschaftlich motivierter Entscheidungen oft in den Hintergrund.

Die maßgeblich vom DFKI vorangetriebene Forschungsinitiative Occiglot betont die Bedeutung dedizierter Sprachmodellierungslösungen, um Europas digitale Souveränität und Wettbewerbsfähigkeit im Bereich der KI zu sichern. Dies ist entscheidend für das Ziel der digitalen Sprachgerechtigkeit in Europa, das von Projekten wie European Language Equality (ELE) und dem Europäischen Parlament angestrebt wird. Hochwertige Forschung und wirtschaftlich relevante Technologien erfordern dabei offenen Zugang zu Sprachmodellen und den zugrundeliegenden Trainingsdaten. Als gemeinnütziges Forschungskollektiv setzt sich Occiglot für Open Science und die Entwicklung von Open-Source-LLMs ein.

Open Data – Occiglot Fineweb

Zwei große Herausforderungen bei der Entwicklung leistungsfähiger LLMs sind der Bedarf an Hochleistungsrechenzentren und großen Mengen hochwertiger Sprachdaten. Obwohl in Europa durch das EuroHPC Joint Undertaking eine Infrastruktur für Hochleistungsrechner geschaffen wird, sind Zugang und Nutzung oft mit hohem Aufwand und technischen Hürden verbunden. Deshalb nutzt Occiglot vortrainierte, offen verfügbare Modelle wie die von META (LLaMA) oder MistralAI, obwohl die verwendeten Daten nicht dokumentiert und somit schwer nachvollziehbar sind. Die mangelnde Transparenz erschwert es jedoch, Arbeiten zu reproduzieren oder darauf aufzubauen.

Occiglot schließt diese Lücke, indem es robuste, mehrsprachige Datensätze kuratiert und transparent bereitstellt. Die Datensammlung umfasst bereits etwa 230 Millionen bereinigte Dokumente in zehn Sprachen und basiert auf bestehenden Arbeiten sowie kuratierten Datensätzen. Alle Dokumente wurden mit speziellen Filtermethoden bearbeitet, was zu einem deutschsprachigen Datensatz von über 60 Milliarden Tokens geführt hat. Diese Daten werden öffentlich zugänglich gemacht, ebenso wie die daraus resultierenden Sprachmodelle, die mit Partnern wie HessianAI vortrainiert werden.

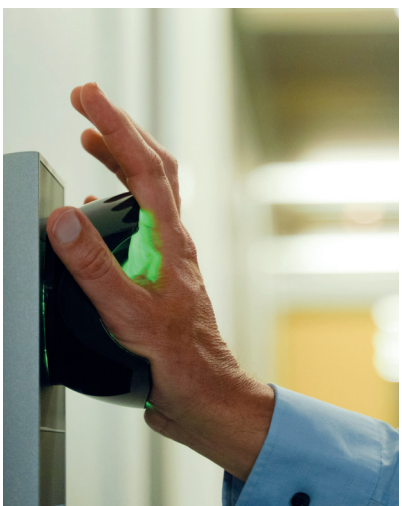
Aktuelle Forschung

Das DFKI arbeitet derzeit in mehreren Forschungsbereichen an großen Sprachmodellen, darunter das vom Bundesministerium für Wirtschaft und Klimaschutz geförderte Projekt OpenGPT-X. Zum Occiglot-Forschungskollektiv gehören neben dem DFKI auch Organisationen wie HessianAI, die TU Darmstadt, Common Crawl, Ontocord.AI, das Barcelona Supercomputing Center sowie die Netzwerke European Language Grid und European Language Equality.

Vertrauenswürdige Forschungsumgebung für sensible personenbezogene Daten

Vertrauenswürdigkeit von KI bezieht sich nicht nur auf die Funktionsweise und Algorithmik maschineller Lernverfahren und großer Sprachmodellen, sondern auf den gesamten Entstehungsprozess von KI-Systemen. Dazu gehört auch die Verarbeitung von Daten für das Training neuronaler Netze. In der Medizin handelt es sich dabei in der Regel um besonders geschützte personenbezogene Daten wie Ultraschall-, Röntgen-, MRT- oder CT-Bilder. Aber auch Foto-, Video- und Audioaufnahmen von Patientinnen und Patienten, etwa aus der sozialen Interaktion mit dem medizinischen Personal, können zu Trainingsdaten für KI-Systeme werden. Diese werden zur Unterstützung von Diagnose und Therapie bei psychiatrischen und psychosomatischen Erkrankungen eingesetzt.

Damit diese hochsensiblen Daten in KI-Modelle einfließen können, bedarf es vertrauenswürdiger Forschungsumgebungen, die die Einhaltung der DSGVO und anderer Datenschutzbestimmungen sowie der Grundsätze guter wissenschaftlicher Praxis gewährleisten. Eine solche Forschungsinfrastruktur – Secure Machine Learning Architecture (SEMLA) – hat das DFKI an den Standorten Saarbrücken und Kaiserslautern aufgebaut. Kernstück von SEMLA ist die Umsetzung sogenannter technischer und organisatorischer Maßnahmen (TOMs) für Datenschutz und Datensicherheit, deren Berücksichtigung in datensensiblen Forschungsprojekten gefordert wird.



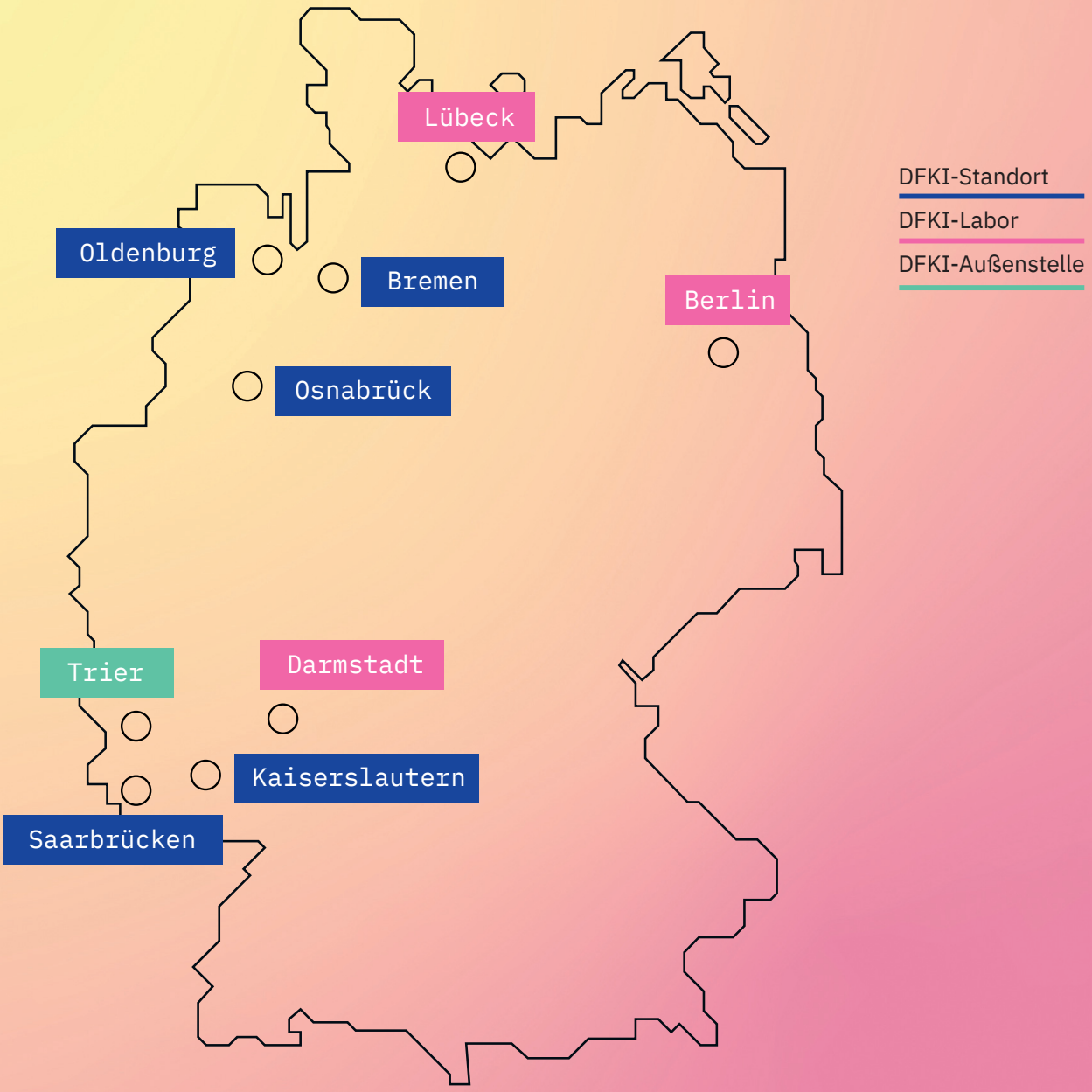
Der Zugang zum SEMLab ist mit einem Handvenen-Scanner biometrisch gesichert.

„SEMLA versetzt uns überhaupt erst in die Lage, mit höchstsensiblen Daten zu arbeiten. Wir werten so zunächst Daten aus Medizinprojekten aus, an denen das DFKI beteiligt ist, und trainieren dann auf diesen Daten neuronale Netze. In Zukunft soll SEMLA quelloffen zur Verfügung gestellt werden“, sagt Projektleiter Dr. Jan Alexandersson.

Im Gegensatz zu Cloud-Lösungen speichert und verarbeitet SEMLA die Daten ausschließlich „on premises“, also am DFKI. Die Forschungsumgebung besteht aus einer Recheninfrastruktur (CPU, GPU, Speicher), die in Kaiserslautern betrieben und geschützt wird, sowie einem biometrisch gesicherten Annotations- und Experimentierlabor in Saarbrücken, dem SEMLab.

Die Forschungsinfrastruktur ist so ausgelegt, dass mit Daten der höchsten Sensitivitätsklasse 4 nach dem Klassifikationsschema des Alan Turing Institute und darüber hinaus geforscht werden kann.

Zukünftig soll es auch Dritten möglich sein, auf den von SEMLA gehosteten Datensätzen über das Internet Modelle zu berechnen. Dazu wird auch eine Zertifizierung nach ISO 2700X und TISAX sowie nach EuroPriSe – dem Europäischen Datenschutzgütesiegel (EuroPriSe, 2022) – angestrebt.



Impressum

dfki ai next; Ausgabe September | 2024; **Herausgeber:** Deutsches Forschungszentrum für Künstliche Intelligenz GmbH (DFKI); Trippstadter Str. 122, 67663 Kaiserslautern; **Tel.:** +49 631 20575 0; **E-Mail:** news@dfki.de; **Redaktion:** Heike Leonhard (verantwortlich); **Lektorat:** Sandra Antakli, Armindo Ribeiro; **Layout:** Lando Lehmann; **Satz:** One Vision Design

Credits

Titelseite: KI-generiertes Bild, Stable Diffusion, Adobe Firefly; **Seite 3:** DFKI, Jürgen Mai; **Seite 4:** metamorworks - stock.adobe.com; **Seite 5:** DFKI; **Seite 6:** DFKI, Armindo Ribeiro; **Seite 7, oben:** KI-generiertes Bild, Adobe Stock, **unten:** privat; **Seite 8:** DFKI; **Seite 9, oben:** acatech, **unten:** DFKI, Jürgen Mai; **Seite 10:** https://occiglot.eu; **Seite 11:** DFKI, Jaron Hollax; **Seite 12:** DFKI.