SLMs könnten theoretisch besser auf individuelle Nutzer angepasst werden, da jeder Nutzer sein eigenes Modell verwenden könnte.

Sebastian Schuster
Professor für Sprachtechnologie an der Fakultät für Informatik der Universität Wien

was die Effizienz steigert. Sie spielen auch eine wichtige Rolle bei der Erstellung und Analyse von Wartungsberichten sowie bei der Echtzeitanalyse von Energieverbrauchsdaten, um Einsparungen zu identifizieren.
In der Produktion können SLMs Anweisungen in natürlicher Sprache geben und Statusmeldungen von Maschinen interpretieren, wodurch die Effizienz der Mitarbeiter erhöht wird.
Im Bildungsbereich ermöglichen SLMs personalisierte Lernmodule, die auf individuelle Schülerfragen eingehen, ohne dass umfangreiche Cloud-Ressourcen benötigt werden. Sie können auch Lehrmaterialien nach Schwierigkeitsgrad oder Relevanz kategorisieren.
Für die Telekommunikationsbranche bieten SLMs Möglichkeiten zur Verbesserung der Kundeninteraktion, indem sie in Kommunikationsplattformen integriert werden, um personalisierte Nachrichten oder Antworten auf häufige Fragen zu liefern. Darüber hinaus können sie technische Support-Tickets und Netzwerkdokumentationen analysieren

DEMOKRATISIERUNG DER KI. „Die Modelle generieren immer noch deutlich schneller Texte auf Systemen mit spezieller Hardware, aber es ist möglich, relativ kleine KI-Systeme auch lokal auf einem normalen Laptop zu nutzen", sagt Schuster. „Manche Hersteller bauen vermehrt auch Chips in Laptops ein, die KI-Anwendungen schneller machen, und auf solchen Laptops könnte man auch schneller Texte und Codes generieren."
Demnach könnte man tatsächlich von einer Demokratisierung der KI sprechen. „SLMs sind definitiv einfacher auf eigenen Rechnern zu nutzen als LLMs und tragen damit zu einem gewissen Grad zur Demokratisierung bei, da man weniger abhängig von kommerziellen Betreibern wie OpenAI oder Google ist." Das Training von SLMs benötigt allerdings noch sehr viel Rechenleistung und ist daher kostspielig, betont der Computerlinguistiker: „Somit ist weiterhin nicht der Fall, dass ohne spezielle Hardware und die damit anfallenden Kosten KI-Systeme entwickelt werden können." Einen weiteren Engpass sieht der Experte bei den Daten: „Kommerzielle Betreiber wie OpenAI und Google haben sehr viel in die Kuration von Daten investiert und gute Trainingsdaten zu haben ist wichtig, damit die Modelle gut funktionieren." Schuster verweist darauf, dass es Bestrebungen gibt, hochwertige Datensätze öffentlich verfügbar zu machen. Hier ist etwa „Dolma" zu nennen. Es umfasst den offenen Datensatz mit drei Billionen Tokens aus vielfältigen Quellen sowie das Dolma Toolkit zur Kuratierung von Sprachmodelldatensätzen. „Aber diese können qualitätsmäßig immer noch nicht jenen Daten, die kommerzielle Betreiber zur Verfügung haben, das Wasser reichen", stellt Schuster klar. „Zudem verbessern kommerzielle Betreiber ihre Systeme basierend auf den Anfragen, die Nutzer absetzen und das sind wiederum Daten, auf die man keinen Zugriff hat, wenn man allein sein Modell verwendet."

CHANCE & RISIKEN. „Durch die Senkung der Einstiegshürden werden SLMs auch kleineren Unternehmen und Schwellenländern die Möglichkeit geben, KI zu nutzen, was zu einer breiteren Innovation und Akzeptanz führt", sagt Urbantschitsch.
Zukünftiges Potenzial von SLMs liegt sicher in der weiteren Erforschung ihrer Energieeffizienz und der Integration in Edge-Computing-Geräte wie IoT-Geräte, um Entscheidungen in Echtzeit zu ermöglichen. „Da sich SLMs immer leichter feinabstimmen lassen, werden wir mehr branchenspezifische Modelle sehen, die generische LLMs in Bezug auf Genauigkeit und Relevanz für spezielle Aufgaben übertreffen", sieht Urbantschitsch Chancen bei SLMs für branchenspezifische Innovation. Generell wachse das Potenzial für SLMs sehr schnell. Etwa bei Edge- und On-Device-KI: „SLMs werden intelligentere, datenschutzfreundliche Anwendungen auf Smartphones, Wearables und IoT-Geräten ermöglichen, die eine Echtzeitverarbeitung ohne Abhängigkeit von der Cloud erlauben", so der Red-Hat-Vizepräsident. „Aber auch bei personalisierter und adaptiver KI. Ihre Anpassungsfähigkeit ermöglicht hochgradig personalisierte Benutzererfahrungen im Bildungs- und Gesundheitswesen sowie bei Produktivitätstools für Unternehmen."
Der Vergleich des Energieverbrauchs zwischen großen und kleinen Sprachmodellen wird ebenfalls ein zentrales Thema sein, insbesondere im Hinblick auf nachhaltige und ressourcenschonende KI-Anwendungen.
Man darf aber nie außer Acht lassen, dass SLMs eben klein sind und damit auch rasch an Grenzen stoßen. „Zum Beispiel bei der Vielsprachigkeit", sagt Sebastian Schuster. „SLMs haben deutlich weniger Parameter, die Informationen abbilden können." Daher erlauben sie nicht, Informationen in vielen Sprachen abzubilden. „Dies limitiert auch das Wissen, das in diesen Modellen abgebildet werden kann", fährt Schuster fort. „Außerdem sind kleinere Modelle im Vergleich zu LLMs im Allgemeinen schlechter darin, komplexe Aufgabenstellungen zu lösen, und das wird vermutlich auch in naher Zukunft so bleiben."
Ein Problem generativer KI ist u. a. die Förderung von Falschinformationen. Zwar betrifft das LLMs auch, aber Schuster vermutet, dass SLMs hier einen Nachteil habe: „Je kleiner das Modell ist und je weniger Daten verwendet wurden, um das Modell zu trainieren, desto größer ist die Wahrscheinlichkeit, dass keine Informationen zu einem Thema im Modell abgebildet sind." Das birgt die Gefahr, dass die Modelle zu Falschaussagen tendieren, die eventuell plausibel klingen, jedoch keinen Bezug zur Realität haben.