Whitepaper
Forschungsprojekt ADLER – effiziente und souveräne Sprachmodelle für die öffentliche Verwaltung

Die Restriktionen der Verwaltung als Innovationsmotor: kleine Sprachmodelle als Schlüssel zur digitalen Souveränität

Gerade die Restriktionen der öffentlichen Verwaltung können zum Innovationstreiber werden: Begrenzte Infrastruktur, strenge Datenschutzanforderungen und eine fehlende US-Cloud-Anbindung schaffen den Druck, der ressourceneffiziente und souveräne KI-Lösungen hervorbringt.
In immer kürzeren Abständen bringen die großen KI-Labore neue, leistungsstärkere Modelle heraus, deren Betrieb immer größere Rechenkapazitäten erfordert. Werden Daten außerhalb der eigenen Einflusssphäre in externen Rechenzentren verarbeitet, besteht das Risiko, dass sie von Dritten ausspioniert oder von den Modellanbietern für das eigene Training verwendet werden. Vor dem französischen Senat konnte Microsoft nicht ausschließen, dass Daten französischer Bürger und Bürgerinnen in die Vereinigten Staaten transferiert werden (Woollacott 2025). Der Aufbau eigener Rechenzentren ist aufwendig und langwierig.
Kleine Sprachmodelle, die auf handelsüblicher und bestehender Hardware laufen, bieten einen Ausweg aus der Abhängigkeit von Drittanbietern und eröffnen neue Möglichkeiten für sichere, souveräne KI-Anwendungen. Limitierte Kapazitäten bei verfügbarer Infrastruktur und hohe Anforderungen an sichere Datenverarbeitung fördern die Nutzung kleiner Modelle.
Wie kann eine KI-Innovation mit kleinen Sprachmodellen aussehen? In diesem Whitepaper zeigen wir konkrete Innovationspfade auf.

Was sind kleine Sprachmodelle?

Sprachmodelle sind KI-Systeme, die menschliche Sprache verstehen und erzeugen können. Man kann sie sich als Maschinen mit Millionen fein justierbarer Stellschrauben vorstellen: Je mehr Schrauben, desto leistungsfähiger, aber auch ressourcenhungriger. Die Größe eines Modells wird in Parametern gemessen, also der Anzahl dieser Schrauben.

Was heißt klein?

Eine allgemein anerkannte Unterscheidung zwischen kleinen und großen Sprachmodellen hat sich bisher noch nicht etabliert (Subramanian et al. 2025). Ein Forschungsteam von NVIDIA etablierte folgende Definition: Ein Small Language Model (SLM) ist ein Modell, das auf gängiger Consumer-Hardware ausgeführt werden kann und Inferenz mit einer Latenz durchführt, die niedrig genug ist, um die Anfragen von einzelnen Nutzenden zu bedienen (Belcak et al. 2025).

Was unterscheidet kleine von großen Modellen in ihren Fähigkeiten?

Kleine Sprachmodelle entfalten ihr Potenzial durch gezielte Anpassung. Trainiert auf spezifischen Datensätzen, können sie bei definierten Aufgaben genauso gut oder besser abschneiden als deutlich größere Modelle (Subramanian et al. 2025; Wang et al. 2025), während große Sprachmodelle als Generalisten ein breites Spektrum ohne zusätzliches Training bewältigen. Die Stärken und Limitationen von SLMs zeigen die nächsten Abschnitte.

Klein, lokal, souverän: Was kleine Sprachmodelle der Verwaltung konkret bringen

Datensouveränität by Design

Eine zentrale Herausforderung beim Einsatz von Künstlicher Intelligenz ist die bisher nicht flächendeckend verfügbare Infrastruktur für die sichere Verarbeitung von Daten mit höheren Schutzniveaus. Kleine und lokale Sprachmodelle lösen dieses Problem durch die Art der Verarbeitung, indem die Daten bleiben, wo sie sind:

- Datensouveränität „by Design": Mit lokalen Sprachmodellen verbleiben die Daten auf der jeweiligen Hardware-Infrastruktur und sind damit vor unberechtigtem Zugriff geschützt. Da keine Datenströme über das offene Internet oder an Drittanbieter (Cloud-Provider) fließen, entfallen komplexe rechtliche Hürden wie der Abschluss von Auftragsverarbeitungsverträgen für außereuropäische Cloud-Lösungen.
- Schutz sensibler Informationen: Besonders bei der Verarbeitung von Sozialdaten, Gesundheitsdaten oder Verschlusssachen bietet die lokale Trennung die höchstmögliche Sicherheit. Die KI agiert in einem geschlossenen, kontrollierbaren Raum.
- Anonymisierung durch kleine Modelle: Kleine und lokale Sprachmodelle können sensible Informationen vorverarbeiten, beispielsweise anonymisieren oder aggregieren, sodass diese anschließend auch vom größeren Modell in der Cloud ohne Datenschutzbedenken verarbeitet werden können.

Ressourceneffizienz

Sprachmodelle haben durch Training und laufenden Betrieb einen hohen Ressourcenverbrauch. Prognosen zeigen, dass dieser rapide zunehmen wird, die Modelle damit einen Engpass im Stromnetz erzeugen und das Wachstum bremsen können (Jessner 2026). Der Einsatz kleiner Sprachmodelle leistet einen Beitrag zu einer ressourceneffizienten Wirtschaft. Ökologische Nachhaltigkeit geht einher mit wirtschaftlicher Effizienz. Durch gezielte Einsatzplanung lassen sich Kosten senken:

- Der Einsatz kleiner, lokal betriebener Modelle kann die Kosten im Vergleich zu OpenAI-Modellen um das 5- bis 29-Fache senken (Irugalbandara et al. 2024) – bei gleicher Leistung für spezifische Aufgaben.
- Die Kapazitäten vorhandener Hardware werden effizient genutzt, statt neue Spezialhardware teuer zu beschaffen.
- Kleine Modelle übernehmen einfachere Aufgaben wie Extraktion oder Klassifikation. Dies sind Aufgaben, für die ein großes Cloud-Modell ein Vielfaches kosten würde.
- Die Verarbeitung einer Anfrage auf lokaler Hardware verursacht einen kleineren ökologischen Fußabdruck als die Verarbeitung derselben Anfrage durch ein großes Modell im Rechenzentrum.

Spezialist statt Alleskönner

Kleine Modelle, die auf einem handelsüblichen Laptop laufen, können ChatGPT, Gemini oder Claude bei spezifischen Aufgaben übertreffen. Trainiert auf fokussierten Datensätzen, lernen diese Modelle genau die Fähigkeiten, die für die zu erledigende Aufgabe notwendig sind. Für repetitive Verwaltungsaufgaben braucht es nicht das Wissen des ganzen Internets (Heaven 2025).
Kleine Modelle lassen sich einfach an die Bedarfe von Organisationseinheiten anpassen. Trainiert auf das notwendige Domänenwissen und spezifische Aufgaben, können sie direkt in Workflows als Komponenten eingebunden werden. Diese Modularität erlaubt es, komplexe Verwaltungsprozesse in kleine, automatisierbare Schritte zu zerlegen, die nacheinander von spezialisierten Modellen abgearbeitet werden.

Resilienz und Offline-Fähigkeit

Lokal betriebene Modelle funktionieren unabhängig von der Internetverfügbarkeit. Dies ist besonders in Bereichen kritisch, in denen Konnektivität nicht garantiert werden kann oder die Nutzung externer Netze ein Sicherheitsrisiko darstellt.

- Echtzeitverarbeitung vor Ort: Mitarbeitende im Außendienst können Informationen sofort und zuverlässig auswerten, beispielsweise bei der Erfassung von Umweltschäden oder baulichen Begehungen.
- Resilienz in Krisensituationen: Bei Netzausfall oder in Notfallszenarien sind lokale KI-Anwendungen voll einsatzfähig. Ihr Einsatz erstreckt sich von der digitalen Assistenz für Einsatzkräfte bis hin zur Information der Bevölkerung.

Kontrolle und Auditierbarkeit

Der lokale Betrieb ermöglicht Verantwortlichen, Hoheit über die Gestaltung und den Betrieb der KI-Anwendung auszuüben, und schafft Transparenz und Kontrolle.

- Für Verantwortliche ist die KI-Modellauswahl nachvollziehbar, ebenso die Kriterien, die dazu herangezogen wurden, da Beschaffung und Betrieb in ihrer Hand liegt. Es gibt keine unangekündigten Updates durch externe Provider, die das Verhalten der KI plötzlich verändern könnten.
- Lokale und kleine Modelle lassen sich kostengünstig so anpassen, dass sie die Werte der Organisation widerspiegeln, beispielsweise ethische Leitlinien.
- Lokale Systeme sind jederzeit überprüfbar. IT-Sicherheitsbeauftragte und Datenschutzverantwortliche haben den gesamten Software-Stack im eigenen Zugriff und können die Anwendung bis auf die Code-Ebene prüfen.

Grenzen für den Einsatz kleiner Modelle

Kleine Modelle sind Spezialisten und es gibt Aufgaben, die nicht zu ihrem Profil passen. Für die Bearbeitung öffentlicher Dokumente ohne schützenswerte Daten bieten große Sprachmodelle oft die flexiblere und aufwandsärmere Lösung. Der Mehrwert kleiner Modelle entfaltet sich dort, wo ihre spezifischen Stärken (Datensouveränität, Offline-Fähigkeit oder Kosteneffizienz bei hohen Fallzahlen) den Aufwand für Anpassung und Betrieb rechtfertigen.

Bei komplexen Schlussfolgerungen über mehrere Denkschritte oder bei Aufgaben, die umfassendes Weltwissen erfordern, stoßen kleine Modelle an ihre Grenzen. Aktuelle Ansätze

zeigen, dass diese Limitationen durch Aufgabenteilung durch agentische Architekturen aus kleinen Sprachmodellen kompensiert werden können. Wie eine Analyse und praktische Umsetzung aussieht, zeigt das Projekt ADLER.

Das Forschungsprojekt ADLER: souveräne KI aus der Praxis

Wie lassen sich die Vorteile von SLMs in die praktische Umsetzung bringen? Das erprobt das Team des Forschungsprojekts ADLER im KI-Kompetenzcenter (KI-KC) der Bundesdruckerei anhand unterschiedlicher Anwendungsfälle. Anlass sind wiederkehrende Erkenntnisse aus der Projektarbeit des KI-KC. Fehlende eingestufte Infrastruktur verhindert die Verarbeitung von eingestuften Inhalten. Daraus entstanden die Kernfragen des Forschungsprojekts:

- Was ist auf Basis der bestehenden IT-Infrastruktur der Verwaltungseinheiten möglich?
- Wie können kleine Modelle so angepasst werden, dass sie für Aufgaben der öffentlichen Verwaltung einsetzbar sind?

Das Projekt ADLER liefert erste praktische Antworten auf diese Fragen und zeigt, welche Leistungspotenziale SLMs mit Optimierungen haben.

Ziele des Projekts sind:

- **Technische Spezialisierung:** Kleine Sprachmodelle werden durch automatisierte Prompt-Optimierung (DSPy, Khattab et al. 2024) und Finetuning auf künstlichen Daten gezielt auf Verwaltungsaufgaben zugeschnitten.
- **Pragmatische Umsetzung:** Die vorhandene IT-Infrastruktur bildet den Rahmen. Modelle werden nicht isoliert entwickelt, sondern von Anfang an in bestehende Workflows integriert.
- **Nutzenorientierung:** Verwaltungsmitarbeitende erhalten schnell individuelle KI-Anwendungen, die ihren konkreten Arbeitsalltag unterstützen.

Unser Vorgehen

Das Forschungsfeld ist neu und viele Publikationen stammen aus den letzten drei Jahren. Diese sichtet das Projektteam kontinuierlich und bezieht sie in die Projektarbeit ein. Experimente zeigen, wie gut sich die Ansätze in der Praxis bewähren. Diese sind so gewählt, dass kleine Sprachmodelle eng an die Anforderungen der Verwaltungspraxis angelehnt sind: Datenschutz, Nachvollziehbarkeit und Zuverlässigkeit. Das folgende Beispiel zeigt, wie dieses Vorgehen konkret aussieht.

Ein Praxisbeispiel

Prüfung und Validierung von Anträgen gehört zu den Kernaufgaben der Verwaltung. Mit kleinen Sprachmodellen erprobte das Projektteam, wie gut diese Informationen extrahieren und die Vollständigkeit des Antrags zuverlässig bewerten können. Das Ergebnis ist eindeutig: Kleine Sprachmodelle können Anträge zuverlässig prüfen – wenn die richtige Optimierungsstrategie gewählt wird.

Zunächst wurde ein Hardware-Setting festgelegt, das den üblichen Endgeräten in der öffentlichen Verwaltung entspricht, beispielsweise einem Laptop mit 8 GB RAM. Daraus ergaben sich Grenzen für die Modellgröße.

Anhand eines mehrstufigen Evaluationsdatensatzes wurden Modelle auf ihre Eignung bewertet:

- Wie viele richtige Antworten gibt das Modell?
- Inwieweit werden korrekte Inhalte extrahiert?
- Kann es Anweisungen befolgen, etwa nur mit „Ja" oder „Nein" zu antworten?

Die Experimente zeigten: Nicht jedes Modell eignet sich gleich gut und die Wahl der richtigen Optimierungsstrategie macht einen erheblichen Unterschied.

- Prompting: Durch gezielte Anweisungen an das Modell (Prompting) verbesserten sich die Ergebnisse deutlich. Diese Anweisungen können manuell formuliert und mit Beispielen hinterlegt sein oder automatisch von einem größeren Modell erstellt werden. Prompting ist aufwandsarm und erfordert kein zusätzliches Training.
- Finetuning: Noch größere Leistungssprünge ermöglicht das Finetuning. Das Modell wird mit eigenen Beispielen weitertrainiert, wie eine neue Mitarbeiterin, die gezielt auf eine bestimmte Aufgabe eingearbeitet wird. Dabei kommt es auf die Qualität der Trainingsdaten an.
- Synthetische Daten: Stehen keine oder zu wenige reale Daten zur Verfügung, können synthetische Daten eingesetzt werden. Diese werden von einem größeren Sprachmodell generiert und spiegeln die Eigenschaften der Originaldaten wider.

Eine zentrale Erkenntnis: Finetuning mit synthetischen Daten erzielt die höchste Genauigkeit. Entscheidend ist dabei das Trainingsformat – ob auf einzelnen Abschnitten oder vollständigen Anträgen trainiert wird, beeinflusst die Leistung erheblich. Diese Erkenntnisse geben erste Hinweise darauf, wo kleine Sprachmodelle in der Verwaltungspraxis sinnvoll eingesetzt werden können.

Praxiseinsatz von kleinen Sprachmodellen

Die Forschungsergebnisse und der Austausch mit Verwaltungsmitarbeitenden zeigen, dass sich kleine Sprachmodelle für spezifische Anwendungsfälle besonders eignen:

1. Verarbeitung sensibler Daten

Daten und Dokumente sind eingestuft und können mit kleinen und lokal betriebenen Modellen direkt verarbeitet werden. Denkbar ist auch, dass kleine Sprachmodelle für Anonymisierung und Aggregation von Daten eingesetzt werden, sodass diese anschließend ohne sensible Information auf größeren Sprachmodellen verarbeitet werden können.

2. Zivil- und Katastrophenschutz

Ereignisse wie Stromausfälle zeigen, dass Kommunikationskanäle ausfallen können. In zivile Notfall-Apps eingebunden, können Informationen auf die individuelle Situation der Menschen abgestimmt werden – von der pflegebedürftigen Seniorin bis zur Familie mit Kleinkindern –, und das in der jeweiligen Muttersprache, völlig autark vom Mobilfunknetz.

3. Kosteneffiziente Antragsbearbeitung

Kleine Sprachmodelle sind ideal als Assistenten für das Bearbeiten von Anträgen. Sie übernehmen repetitive Aufgaben wie Extraktion oder Klassifizierung. Zudem können kleine Sprachmodelle die Vollständigkeit von Anträgen validieren. Bei diesen Aufgaben erbringen sie

die gleiche Leistung wie große Modelle zu einem Bruchteil der Kosten. Insbesondere wenn Aufgaben oft anfallen, bringen kleine Modelle erhebliche Kostenvorteile mit sich.

4. Wissensmanagement für vertrauliche Daten

Wissen ist oft unstrukturiert über verschiedene Dateien und Ordner verteilt. Kleine Sprachmodelle können Mitarbeitenden in Behörden mit gewachsenen Dokumentationsbeständen helfen, geeignete vertrauliche Dokumente für eine Anfrage schnell zu finden oder Informationen zu extrahieren.

5. Anomalie-Erkennung im Offline-Einsatz

In einem Projekt konnten im Rahmen von Bilderkennung kleine Modelle für das Erkennen von Anomalien bei 3D-Objekten genutzt werden. Die Objekterkennung funktionierte ohne Internetverbindung, die Analyse fand lokal auf Smartphones statt.

Kleine Sprachmodelle sind dort einsetzbar, wo es darauf ankommt: sicher, effizient und unter voller Kontrolle der Verwaltung

Quellen

Belcak, Peter; Heinrich, Greg; Diao, Shizhe; Fu, Yonggan; Dong, Xin; Muralidharan, Saurav; Lin, Yingyan C.; Molchanov, Pavlo: Small Language Models are the Future of Agentic AI. In: *arXiv preprint arXiv:2506.02153v2* (2025). https://arxiv.org/pdf/2506.02153

Heaven, Will D.: Small language models: 10 Breakthrough Technologies 2025. In: *MIT Technology Review* (2024). https://www.technologyreview.com/2025/01/03/1108800/small-language-models-ai-breakthrough-technologies-2025

Irugalbandara, Chandra; Mahendra, Ashish; Daynauth, Roland; Kasthuri Arachchige, Tharuka; Dantanarayana, Jayanaka; Flautner, Krisztian; Tang, Lingjia; Kang, Yiping; Mars, Jason: Scaling Down to Scale Up: A Cost-Benefit Analysis of Replacing OpenAI's LLM with Open Source SLMs in Production. In: *arXiv preprint arXiv:2312.14972v3* (2024). https://arxiv.org/pdf/2312.14972v3

Jessner, Oliver: KI-Rechenzentren vs. Stromnetz: Wer zahlt die Zeche für den KI-Boom? In: Golem.de (2026). https://www.golem.de/news/ki-rechenzentren-vs-stromnetz-wer-zahlt-die-zeche-fuer-den-ki-boom-2602-205079.html

Khattab, Omar; Singhvi, Arnav; Maheshwari, Paridhi; Zhang, Zhiyuan; Santhanam, Keshav; Vardhamanan, Sri; Haq, Saiful; Sharma, Ashutosh; Joshi, Thomas T.; Moazam, Hanna; Miller, Heather; Zaharia, Matei; Potts, Christopher: DSPy: Compiling Declarative Language Model Calls into State-of-the-Art Pipelines. In: International Conference on Learning Representations (ICLR) (2024). https://openreview.net/forum?id=sY5N0zY5Od

Subramanian, Shreyas; Elango, Vikram; Gungor, Mecit: Small Language Models (SLMs) Can Still Pack a Punch: A survey. In: *arXiv preprint arXiv:2501.05465* (2025). https://arxiv.org/abs/2501.05465

Wang, Feng; Shi, Zesheng; Wang, Bo; Wang, Nan; Xiao, Han: ReaderLM-v2: Small Language Model for HTML to Markdown and JSON. In: arXiv preprint arXiv:2503.01151v1 (2025). https://arxiv.org/pdf/2503.01151.pdf

Woollacott, Emma: Microsoft Can't Keep EU Data Safe From US Authorities. In: *Forbes* (2025). https://www.forbes.com/sites/emmawoollacott/2025/07/22/microsoft-cant-keep-eu-data-safe-from-us-authorities