

KI-Modellvergleich für die öffentliche Verwaltung

[15.07.2026] Die Bundesdruckerei veröffentlicht das fortlaufende Forschungsprojekt MÖVE. Auf einer interaktiven Website können Behörden KI-Sprachmodelle systematisch vergleichen. Der Bewertungsrahmen ist speziell auf den Public Sector abgestellt und prüft hinsichtlich Leistung, Sicherheit und Einhaltung europäischer Vorgaben.

Im [Forschungs- und Innovationsprojekt](#) „Modelle für die öffentliche Verwaltung evaluieren“ – kurz: MÖVE – hat die [Bundesdruckerei](#) einen Bewertungsrahmen für KI-Sprachmodelle entwickelt, der das Augenmerk speziell auf den Einsatz im Public Sector legt. Nun liegen die ersten Ergebnisse vor: Behörden und öffentliche Einrichtungen können ab sofort einsehen, wie über 50 Sprachmodelle (Large Language Models, LLMs) im Hinblick auf Leistungsfähigkeit, Sicherheit, Nachhaltigkeit, Transparenz und Vereinbarkeit mit demokratischen Werten abschneiden.

Internationale Rankings reichen nicht

Das seit Januar 2025 laufende Forschungs- und Innovationsprojekt MÖVE ist nach Angaben der Bundesdruckerei der erste ganzheitliche KI-Modellvergleich, der speziell für die Anforderungen der öffentlichen Verwaltung entwickelt wurde. Das ist notwendig, weil gängige Modellvergleiche vor allem englisches Textverständnis, mathematische Fähigkeiten oder allgemeines Weltwissen messen. Im Behördenalltag entscheidende Kriterien wie der Zugang zu nachvollziehbaren Modellinformationen sowie Sicherheits- und Nachhaltigkeitsaspekte werden bei solchen KI-Benchmarks nicht berücksichtigt. „Ob Open Source oder proprietär – entscheidend ist, wie ein Modell im konkreten Einsatzfall reagiert: im deutschen Rechts- und Sprachkontext, verlässlich, sicher und wertekonform. MÖVE macht genau das messbar“, sagt Frederik Blachetta, Chief Technology, Data & Innovation Officer (CTO) und Geschäftsführer der Bundesdruckerei-Gruppe.

Auch Grundwerte auf dem Prüfstand

MÖVE verbindet erstmals technische Leistungsmessung und Governance-Anforderungen in einem einheitlichen Framework. Grundlage sind neun eigens entwickelte, deutschsprachige Testdatensätze aus dem realen Verwaltungskontext. Für die Bewertung der Sprachmodelle wird neben Leistungsaspekten auch die Einbeziehung von Zielen und Werten der Europäischen Union untersucht. Dafür analysiert MÖVE die Anzahl von Halluzinationen in KI-generierten Texten, die Vereinbarkeit der Antworten mit demokratischen Grundwerten, die Effizienz der Arbeitsweise und die Informations-Transparenz. Auf Basis dieser Kriterien durchläuft jedes Modell eine automatisierte Bewertung, die dann in einem interaktiven Online-Vergleich einzusehen ist. Dabei können einzelne Kriterien gezielt ein- oder ausgeblendet werden, sodass jede Behörde erkennen kann, welches Modell zu den eigenen Anforderungen passt.

Sicherheit zuverlässig messen

Weitere Bewertungskriterien, die MÖVE ergänzen sollen, sind bereits in der Entwicklung. Gemeinsam mit dem Bundesamt für Sicherheit in der Informationstechnik (BSI) entwickelt die Bundesdruckerei Verfahren

zur Beurteilung in den Bereichen Cyber-Sicherheit, Robustheit und Faktentreue von Sprachmodellen weiter; das Fraunhofer-Institut für Angewandte und Integrierte Sicherheit AISEC bringt seine wissenschaftliche Expertise in die kontinuierliche Weiterentwicklung der Sicherheitskriterien ein.

Das ist insofern relevant, weil LLMs zwar Maßnahmen zur Cyber-Sicherheit unterstützen können, aber auch das Risiko bergen, für Cyber-Angriffe genutzt zu werden. Dies betrifft insbesondere die Entwicklungsfortschritte von Frontier-KI-Modellen, deren weit vorangeschrittene Fähigkeiten neue Sicherheitsfragen aufwerfen können. Hier sind ein kontinuierliches Benchmarking und eine frühzeitige Einordnung essenziell.

Damit die Vorgehensweise nachvollzogen werden kann, ist die Bewertungsmethodik hinter MÖVE als Open-Source-Repository frei verfügbar, zudem wird sie in einem Forschungspaper erörtert.

(sib)

Stichwörter: Künstliche Intelligenz, BSI, Bundesdruckerei, Fraunhofer AISEC, MÖVE