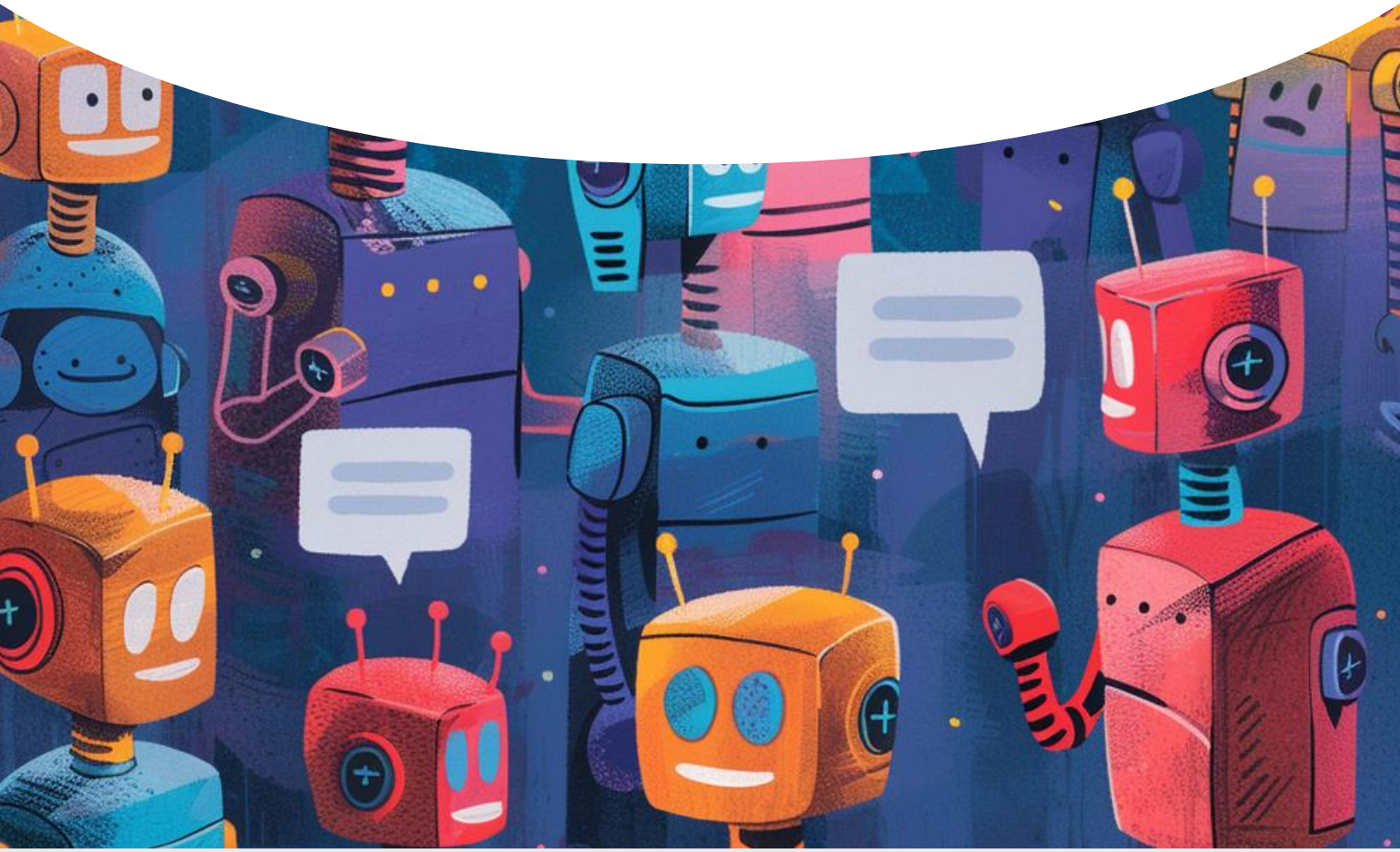
BERICHT

Gestaltung und Anwendung von Demonstratoren



Gefördert durch:
Bundesministerium für Wirtschaft und Klimaschutz
aufgrund eines Beschlusses des Deutschen Bundestages

Bericht: Gestaltung und Anwendung von Demonstratoren

DEMONSTRATOR FÜR EIN NUTZERZENTRIERTES EVALUATIONSFRAMEWORK VON GROSSEN SPRACHMODELLEN

Die Evaluation von großen Sprachmodellen (Large Language Models, LLMs) ist eine zentrale Aufgabe, um deren Anwendbarkeit und Genauigkeit in einem konkreten Kontext sicherzustellen. Durch die aktuellen Fortschritte im Bereich der Künstlichen Intelligenz ist eine Vielzahl von Sprachmodellen in unterschiedlichen Sprachen und auf Basis verschiedener Trainingskorpora entstanden. Faktoren wie Sprache, Wissensbasis und ein präzises Nutzerverständnis sind dabei entscheidend dafür, ob ein Sprachmodell für einen spezifischen Anwendungskontext geeignet ist.

Gängige automatisierte Evaluationsmethoden, etwa Multiple-Choice-Tests oder die Bewertung von Sprachmodellen durch andere Sprachmodelle, eignen sich zwar, um Grundfähigkeiten vergleichbar zu machen. Sie sind jedoch zu allgemein, um individuelle Anwendungskontexte wie medizinische Beratung, Bildung, Recht oder den privaten Haushalt zuverlässig abzubilden. Auch bestehende nutzerbasierte Tools wie die „Chatbot Arena“ erzeugen lediglich allgemeine Ranglisten und berücksichtigen die konkreten Aufgaben einzelner Nutzerinnen und Nutzer nicht. Zudem sind die verbreiteten Metriken überwiegend auf die englische Sprache ausgerichtet, was die belastbare Bewertung von Modellen in anderen Sprachen erschwert.

Das Mittelstand-Digital Zentrum Fokus Mensch entwickelt einen Demonstrator für eine nutzerzentrierte Evaluation von Sprachmodellen. Das Werkzeug ermöglicht es, eigene Prompts gleichzeitig an mehrere Sprachmodelle zu senden, deren Antworten direkt gegenüberzustellen und mithilfe einer fünfstufigen Likert-Skala sowie ergänzender Kommentare zu bewerten. Um Verzerrungen durch Markenbekanntheit oder eine feste Reihenfolge zu vermeiden, werden die Modellnamen während der Bewertung verborgen und die Position der Modelle randomisiert.

ZIELGRUPPE

Die Zielgruppe für diesen Demonstrator sind insbesondere KMU, die Sprachmodelle in ihre Geschäftsprozesse integrieren wollen und vor der Aufgabe stehen, ein geeignetes Modell für ihren konkreten Anwendungsfall auszuwählen. Ebenso relevant ist der Demonstrator für Entwicklerinnen und Entwickler, Beratungen sowie Forschende, die Sprachmodelle systematisch und nutzerzentriert miteinander vergleichen möchten, insbesondere dann, wenn nicht-englische Sprachen oder fachspezifische Anforderungen eine Rolle spielen.

STAND DER ENTWICKLUNG

Ein erster funktionaler Prototyp des Evaluationswerkzeugs wurde auf Basis von Feldbeobachtungen mit Nutzenden entwickelt und bereits in mehreren Workshops erfolgreich eingesetzt. Die Benutzeroberfläche besteht aus vier zentralen Komponenten: einer Eingabemaske für Prompts, mehreren parallelen Chat-Modulen für die Antworten der Sprachmodelle (aktuell drei), einem Feld

Gefördert durch:
Bundesministerium für Wirtschaft und Klimaschutz
aufgrund eines Beschlusses des Deutschen Bundestages

für Aufgabentitel und -beschreibung sowie einem Bewertungsbereich mit Likert-Skala und Kommentarfunktion.

Auf Grundlage der gesammelten Bewertungen generiert das System einen Evaluationsbericht. Dieser stellt alle bewerteten Aufgaben in einem Liniendiagramm dar, sodass auf einen Blick erkennbar ist, welches Modell im allgemeinen Trend am besten abschneidet und in welchen Aufgabenbereichen es stärker oder schwächer ist. Ergänzend enthält der Bericht für jede bewertete Aufgabe eine detaillierte Tabelle mit dem ursprünglichen Prompt, der numerischen Bewertung, dem Kommentar der Nutzenden und der jeweiligen Modellantwort.

Gefördert durch:
Bundesministerium für Wirtschaft und Klimaschutz
aufgrund eines Beschlusses des Deutschen Bundestages