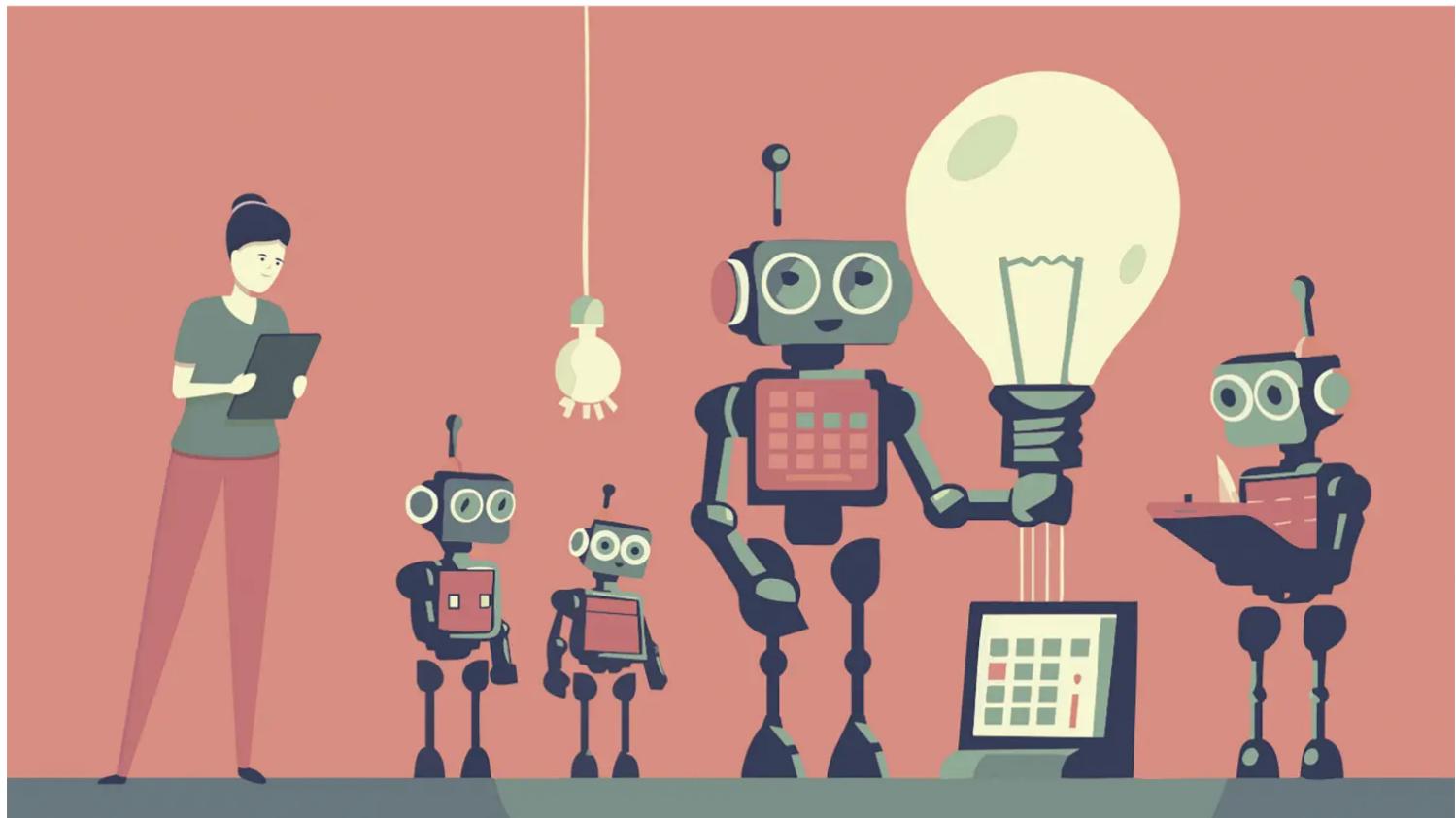


Künstliche Intelligenz: Benchmarks für generative Sprachmodelle im Überblick

21.05.2024 14:30 Uhr Andreas Christian, Kush Varshney



Zahlreiche Metriken und Benchmarks wollen dabei helfen, das "beste" LLM auszuwählen. Ganz so einfach ist es aber nicht.

Zwar verfügen große Sprachmodelle (Large Language Models, LLMs) über nützliche Fähigkeiten, das Einbinden in die alltägliche Arbeit gestaltet sich jedoch schwierig: Oft fehlen verlässliche Informationen zur Qualität der Modelle und der für das Training verwendeten Daten, es gibt eine stetig wachsende Zahl von Benchmarks und Metriken zur Beurteilung der Modelle und schließlich entstehen weltweit umfangreiche gesetzliche Regelwerke für den Einsatz von KI, die verstanden und eingehalten werden müssen. Für Anwender wird es deshalb immer wichtiger, generative KI-Modelle sinnvoll zu bewerten.

Große Foundation-Modelle, zu denen die LLMs gehören, können zwar immer öfter auch mit multimedialen Daten wie Text, Bild oder Audio umgehen. Dieser Artikel konzentriert sich jedoch

auf aktuelle Bewertungsmethoden für große generative Sprachmodelle, also Metriken und Benchmarks für die Verarbeitung von Sprachdaten (Natural Language Processing).

MEHR ZUM THEMA KÜNSTLICHE INTELLIGENZ (KI)

Warum sich Datenschutzbehörden mit ChatGPT und Co. schwertun

Künstliche Intelligenz: Benchmarks für generative Sprachmodelle im Überblick [1]

Marktübersicht: KI-Server mit GPUs im Überblick [2]

Künstliche Intelligenz: teuer, US-amerikanisch, Big-Tech-dominiert [3]

PyTorch: Eigene Bildgenerierungs-KI mit Python bauen [4]

Website per KI hacken: Browser-Skripte mit ChatGPT und Co. generieren [5]

Trend-Beruf: Mit diesen Fähigkeiten wird man KI-Experte [6]

Transkriptionsdienste: Whisper V3 im Vergleich mit Online-Diensten [7]

Projekt noFake trainiert Datenmodelle für Faktenchecks [8]

Lokale KI verschlagwortet Fotosammlung auf NAS [9]

Multi-Agenten-Systeme: Automatisierte Leistungsanpassung für bessere KI [10]

Fremdsprachen lernen: Wie man ChatGPT zum Sprechtrainer aufrüstet [11]

ANDREAS CHRISTIAN



Andreas Christian ist Senior Information Architecture und Technical Sales Specialist bei IBM DACH.

KUSH R. VARSHNEY

Kush R. Varshney ist IBM Fellow für AI Governance bei IBM Research.



Trotz des weitreichenden Einsatzes großer Sprachmodelle fehlt es derzeit an einem klaren Verständnis dafür, wie sie funktionieren, **wozu sie grundsätzlich fähig sind und wann sie versagen könnten [12]**. Weiterhin decken die aktuell verfügbaren Bewertungsansätze für LLMs nicht alle Risikobereiche ab. Es gibt also viele gute Gründe, sich vor dem Einsatz von LLMs mit deren Risiken vertraut zu machen.

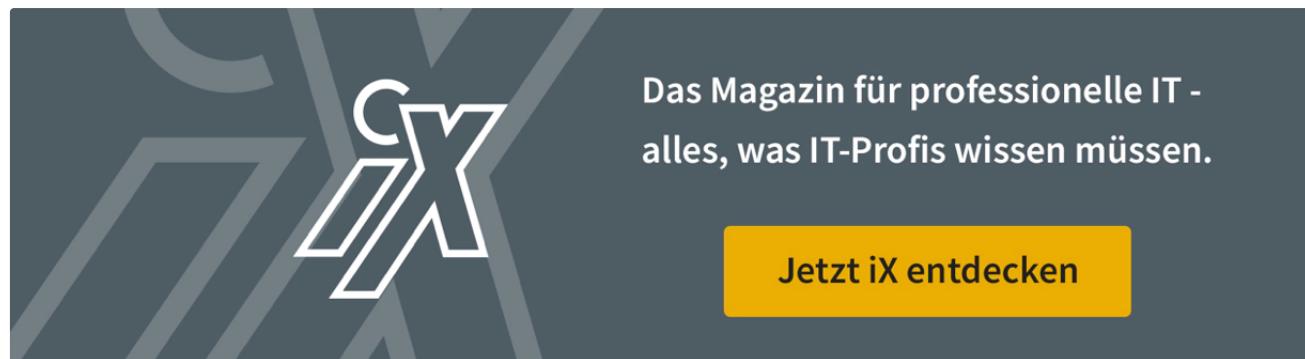
IX-TRACT

- Aufgrund der Vielzahl an offenen und geschlossenen Modellen ist die Bewertung generativer KI zu einer wesentlichen Anforderung für Nutzer und Unternehmen geworden.
- Metriken sind Kennzahlen, die die Leistung eines Modells hinsichtlich einer spezifischen Aufgabe beurteilen. Benchmarks bewerten die aufgabenübergreifende Leistung eines Modells.
- Benchmarks haben häufig eine begrenzte Lebensdauer. Weiterhin können deren Bewertungsansätze zu Testergebnissen führen, die eine höhere Modellqualität vortäuschen, als das tatsächlich der Fall ist.
- Metriken und Benchmarks helfen bei der initialen Auswahl eines Modells, das Unternehmen dann für ihren speziellen Use Case testen und bewerten müssen.

Neben Problemen wie Bias und Drift, die bereits aus dem klassischen KI-Umfeld bekannt sind, gibt es eine ganze Reihe neuer Risiken. Zu einigen Gefahren, die etwa aus den verwendeten Trainingsdaten resultieren können, gehören unbeabsichtigte Copyright-Verletzungen, das versehentliche Offenlegen personenbezogener Daten und das ungewollte Verbreiten von Desinformation. Es gibt jedoch automatisierte Ansätze, solche Risiken zu analysieren.

Beispielsweise kann man Text Classifier benutzen, um die Wahrscheinlichkeit zu bestimmen, dass ein Text in eine Kategorie wie Hassrede oder Belästigung fällt – bei OpenAI gibt es hierfür eine Moderation API. Solche Verfahren basieren oft auf proprietärem Code und es fehlen standardisierte Ansätze und allgemein anerkannte Metriken. Das Red Teaming zum Aufdecken von Schwachstellen in LLMs ist ein weiterer Ansatz, um Risiken zu analysieren, den man allerdings manuell durchführen muss.

Einige drastische Beispiele möglicher Risiken von LLMs sind im Anhang zum GPT-4 Technical Report gelistet. **Sie verdeutlichen die damit verbundenen Herausforderungen für die Entwickler der Modelle [13].**



[14]

Übersicht zu Bewertungsansätzen für LLMs

Eine Besonderheit von LLMs im Vergleich zu klassischen KI-Modellen ist deren Vielseitigkeit, also die Fähigkeit der Modelle, sehr unterschiedliche Aufgaben zu erledigen. In der Praxis aktuell besonders häufig vorkommende Aufgaben sind in der Tabelle gelistet.

Häufige Anwendungsfälle von LLMs

| Task | Anwendungsbeispiel |
|--------------------------|--|
| Q&A | Beantwortung von Kundenfragen |
| Summarization | Zusammenfassung von Besprechungsprotokollen |
| Content Generation | Erstellung von E-Mails |
| Named Entity Recognition | Extraktion wichtiger Fakten aus Anträgen |
| Insight Extraction | medizinische Diagnosen, basierend auf Fallbeschreibungen |
| Classification | Klassifikation von Kundenbeschwerden |

Für die Bewertung von LLMs gibt es verschiedene Ansätze, die hier kurz vorgestellt und weiter

unten genauer erläutert werden: Metriken sind numerische Kennzahlen, die bestimmte Eigenschaften eines Modells beurteilen und seine Leistung hinsichtlich einer spezifischen Aufgabe bewerten. Hier kommen referenzbasierte Metriken wie ROUGE oder BLEU zum Einsatz. Benchmarks dienen zur Analyse der aufgabenübergreifenden Leistung eines Modells.

Menschliches Feedback kommt da zum Einsatz, wo Benutzer etwa ihre (subjektive) Bewertung des Modells abgeben. Das Verfahren LLM as a Judge ist ein rein maschinerller Bewertungsansatz, bei dem LLMs andere LLMs bewerten. **Dieser Ansatz eignet sich beispielsweise für ein automatisiertes Ranking von LLMs [15].**

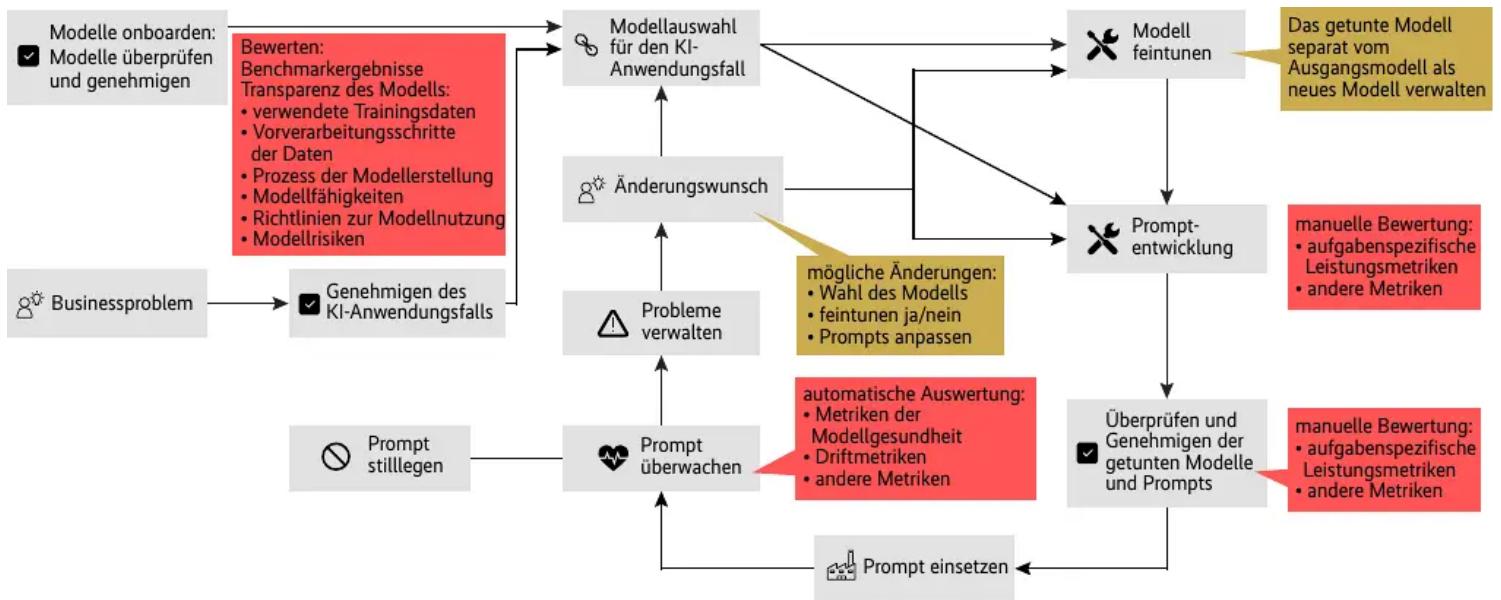
Die unterschiedlichen Aufgaben und zugehörigen Bewertungsansätze lassen sich entsprechend der Tabelle unterteilen. Bei den Predictive Tasks wird der Eingabetext hinsichtlich bestimmter Eigenschaften analysiert, beispielsweise klassifiziert. Für die Bewertung gibt es hier typischerweise gelabelte Beispieldaten mit den objektiv richtigen Ergebnissen. Bei der zielgerichteten Texterzeugung, also etwa Zusammenfassungen oder Übersetzungen in eine andere Sprache, sind in der Regel Referenzbeispiele verfügbar, die als Grundlage für die Bewertung dienen. Schließlich gibt es kreative Aufgabenstellungen wie das Schreiben einer Geschichte zu einem bestimmten Thema. Hierzu gibt es kein objektiv "richtiges" Ergebnis.

Bewertungsansätze für große Sprachmodelle (LLMs)

| Task-Typ | Predictive Tasks | zielgerichtete Texterzeugung | creative Texterzeugung |
|-------------------|---|--|---|
| Beispiele | Textklassifikation, Stimmungsanalyse | Textübersetzung, Textzusammenfassung | Geschichten schreiben, E-Mails generieren |
| Ergebnistyp | Es liegt ein objektiv richtiges Ergebnis vor. | Es gibt Referenzbeispiele (Texteingaben und erwartete Textausgaben). | Es gibt kein objektiv richtiges Ergebnis. |
| Bewertungsansätze | traditionelle ML-Metriken wie Accuracy, Precision | <ul style="list-style-type: none"> • referenzbasierte Metriken wie BLEU oder ROUGE • menschliches Feedback • LLM as a Judge • Benchmarks | <ul style="list-style-type: none"> • menschliches Feedback • LLM as a Judge • Benchmarks |

Bewertung von LLMs im Modelllebenszyklus

Die folgende Abbildung zeigt einige Phasen des Lebenszyklus von LLMs und der zugehörigen Prompts. Die Phasen, in denen man die Modelle typischerweise bewertet, sind mit roten Kommentaren versehen. Neben dem hier dargestellten Prompting und Feintuning gibt es weitere Ansätze zur Leistungsverbesserung von LLMs, etwa Retrieval Augmented Generation.



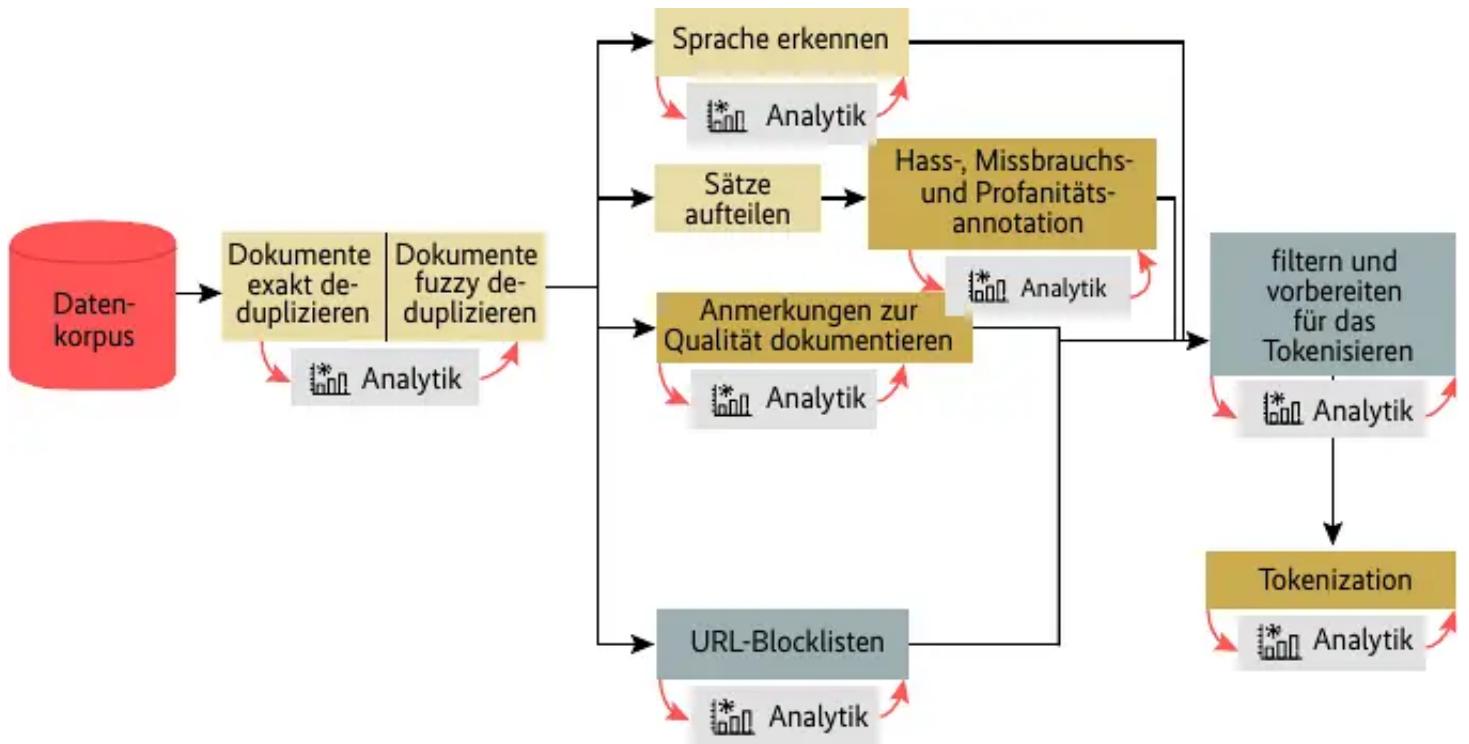
Im Lebenszyklus eines KI-Modells kann es an vielen Stellen nötig sein, das Modell und seine Ausgabe zu bewerten. Was geprüft wird, zeigen die roten Sprechblasen.

(Bild: IBM)

LLMs und Prompts evaluiert man insbesondere in vier Phasen ihres Lebenszyklus. Beim Onboarding von Modellen externer Anbieter sind Benchmarks bei der initialen Auswahl von Modellen hilfreich, da sie die allgemeine Leistungsfähigkeit generativer Sprachmodelle vergleichen. Auch externe Anbieter führen selbst Evaluierungen und Benchmarking durch, während sie LLMs vorab trainieren.

Benchmarks sind allerdings kein Ersatz für die spezifische Leistungsbewertung der Modelle in Bezug auf einen konkreten Anwendungsfall. Weiterhin müssen Unternehmen während des Onboardings weitere Eigenschaften der Modelle beurteilen, die Benchmarks oft nicht berücksichtigen. Ein besonders wichtiger Bereich ist die Beurteilung der Transparenz der Modelle, also die Verfügbarkeit von Informationen darüber, wie sie entwickelt wurden, welche Trainingsdaten verwendet und wie diese aufbereitet wurden.

Die nächste Abbildung zeigt eine beispielhafte Pre-Processing-Pipeline für Trainingsdaten. Das CRFM (Center for Research on Foundation Models) hat kürzlich ein viel beachtetes Paper herausgebracht, das einen Foundation Model Transparency Index (FMTI) definiert, **der 100 verschiedene Aspekte der Modelltransparenz beinhaltet und LLMs entsprechend diesen Kriterien bewertet [16]**. Das Ergebnis zeigt, dass bezüglich der Modelltransparenz noch erheblicher Verbesserungsbedarf bei den Anbietern besteht.



Ein Aspekt der Modelltransparenz sind Informationen zur Aufbereitung der verwendeten Trainingsdaten.
(Bild: IBM)

Feintunen, Prompt Engineering und Monitoring

Um die Qualität von Modellausgaben zu verbessern, kann man einen Prompt mit Beispieldatensätzen (Eingaben und gewünschten Ausgaben) versehen, die das Modell beim Scoring erhält. Das bezeichnet man im ML-Umfeld als Few-Shot Prompting. Eine andere Möglichkeit zur aufgabenspezifischen Leistungsverbesserung und **zum Sicherstellen eines kontrollierten Verhaltens von Modellen ist das Feintunen [17]** – wobei ein vollständiges Feintunen von Modellen zumeist aufwendig ist. **Parameter-efficient Fine-Tuning (PEFT) passt durch zusätzliches Training nur bestimmte Gewichte des Modells an [18].** Nach dem Feintunen sollte man das modifizierte Modell hinsichtlich der Aufgaben und des gewünschten Verhaltens erneut evaluieren.

Für den Einsatz von LLMs in konkreten Anwendungsfällen muss man spezifische Prompts entwickeln. Hierbei gilt es, eine möglichst optimale Anweisung an ein Modell zu schreiben, beispielsweise: "Erstelle eine Liste der nächsten sinnvollen Schritte zur Bearbeitung der vorliegenden Kfz-Schadensmeldung." Das Prompting ist ein iterativer Prozess, bei dem man unterschiedliche Modelle, Modellparameter und Anweisungen testet. Anschließend ist es sinnvoll, den Prompt mit geeigneten Performancemetriken zu evaluieren. Aufgrund der wachsenden Zahl an Metriken kann die Auswahl des passenden Verfahrens hierbei eine Herausforderung sein.

In Produktion sollte man die optimierten Modelle und zugehörigen Prompts fortlaufend überwachen. Beim Monitoring bewertet man daher operative KPIs wie Performance, Durchsatz, aufgelaufene Kosten oder auch Aspekte wie Modelldrift. Drift bedeutet, dass es Veränderungen

in der aktuellen Datenverteilung gegenüber den verwendeten Trainingsdaten gibt, was zu einer Leistungsabnahme eines Modells führen kann. Im Falle von LLMs kann man beispielsweise überwachen, ob sich die Art der Anfragen ändert oder das Modell Ausgaben generiert, die außerhalb der vorgesehenen Domäne liegen.

Metriken für die Bewertung von LLMs

Die Liste von Metriken für die Bewertung von LLMs ist lang. Dieser Artikel konzentriert sich auf Metriken zur aufgabenspezifischen Leistungsbewertung. Die Grundidee von referenzbasierten Metriken wie BLEU, ROUGE und METEOR besteht darin, den durch ein Modell oder einen Prompt generierten Text mit einem qualitativ hochwertigen, von Menschen erstellten Referenztext zu vergleichen. Je näher der generierte Text am Referenztext liegt, desto höher ist der Metrikwert. Der Hauptunterschied zwischen den Metriken besteht darin, wie sie Nähe messen.

Metriken für die aufgabenspezifische Bewertung von LLMs

| Metrik | Entwicklungsjahr | Anwendungsbereich | referenzbasiert |
|-----------|------------------|----------------------|-----------------|
| BLEU | 2002 | Übersetzen | ja |
| METEOR | 2005 | Übersetzen | ja |
| ROUGE | 2004 | Zusammenfassen | ja |
| BERTScore | 2019 | aufgabenübergreifend | ja |

BLEU (Bilingual Evaluation Understudy), ein Urgestein unter den Metriken, wurde bereits vor über 20 Jahren von IBM für die Evaluierung maschinell erzeugter Textübersetzungen entwickelt. Es misst die Übereinstimmung zwischen kleinen Segmenten (N-Grammen) des generierten Textes und des Referenztextes. Der Fokus liegt hierbei auf der Genauigkeit (Precision) der Übereinstimmung: Generierte Texte, die ausschließlich Textfragmente enthalten, die im Referenztext vorgegeben sind, bewertet BLEU tendenziell hoch.

ROUGE (Recall-oriented Understudy for Gisting Evaluation) dient zur Evaluierung maschinell generierter Textzusammenfassungen. Es bewertet, in welchem Maße die Textfragmente aus dem Referenztext wirklich auch *alle* im generierten Text enthalten sind. Die Messgröße fokussiert also eher darauf, dass das Modell nichts auslässt (Recall).

METEOR (Metric for Evaluation of Translation with Explicit Ordering) ist eine weitere Metrik zur Bewertung von Textübersetzungen. METEOR misst das harmonische Mittel aus Precision und Recall und korreliert stark mit der menschlichen Beurteilung von Satzhähnlichkeit. Es bezieht weiterhin die semantische Übereinstimmung in die Beurteilung ein, indem METEOR etwa

Synonyme und Wortstämme mitberücksichtigt.

BERTScore: Schwächen älterer Metriken ausgleichen

BERTScore ist eine jüngere Metrik, die sich für die Bewertung unterschiedlicher Aufgaben eignet. Traditionelle Bewertungsmaßnahmen, die auf N-Grammen basieren, haben diverse Einschränkungen. Sie neigen beispielsweise dazu, semantisch korrekte Umschreibungen nicht ausreichend zu berücksichtigen, Abhängigkeiten zwischen Textabschnitten nicht zu erfassen oder Texte niedriger zu bewerten, die zwar semantisch korrekt sind, bei denen die Wortreihenfolge jedoch stark vom Referenztext abweicht.

Diese Schwachpunkte versucht BERTScore auszugleichen. Analog zu anderen gängigen Maßnahmen berechnet BERTScore einen Ähnlichkeitswert für jedes Token des generierten Textes mit jedem Token im Referenztext. Anstatt jedoch exakte Übereinstimmungen zu berechnen, ermittelt BERTScore die Tokenähnlichkeit unter Verwendung kontextueller Embeddings.

Neben Maßnahmen zur Bewertung der Qualität generierter Texte gibt es Maßnahmen zum Messen von Risiken und schädlichen Effekten von LLMs. Diese Maßnahmen sind in der Regel referenzfrei. Beispielsweise kann man Halluzination (das LLM liefert Ausgaben, die nicht dem Kontext entsprechen) erkennen oder den Grad an hasserfüllten Inhalten messen [19]. Hierfür kommen Textklassifikatoren zum Einsatz, die mit Beispielen für Halluzination und Hassrede trainiert sind.

Benchmarks und Bewertungsframeworks für LLMs

Benchmarks sind Sammlungen von Aufgaben und Referenzdaten, um Modelle zu bewerten und zu vergleichen. Die Tabelle zeigt eine Liste bekannter Benchmarks seit dem Jahr 2021. Die zugehörigen Benchmarkergebnisse sind in Form von Ranglisten öffentlich zugreifbar. Ein Problem vieler Benchmarks ist, dass ihre Bewertungsansätze zu Testergebnissen führen, die eine höhere Qualität der Modelle vortäuschen, als es in der Realität der Fall ist. **Daneben haben Benchmarks oft eine begrenzte Lebensdauer und einen eingeschränkten Anwendungsbereich [20]:** Sie bewerten nur einen kleinen Ausschnitt der Fähigkeiten der Modelle.

Populäre Benchmarks und Frameworks für die Bewertung von LLMs

| Name | Aufgabenbereich | Beschreibung |
|----------------|-----------------|--|
| MMLU (2021) | allgemein | MMLU besteht aus 57 Aufgaben, darunter elementare Mathematik, US-Geschichte, Informatik, Recht und mehr. |
| BIG-bench | allgemein | BIG-bench besteht derzeit aus 204 Aufgaben, die 450 Experten aus unterschiedlichsten Bereichen |

(2022)

beigesteuert haben.

HELM
(2022)

allgemein

HELM umfasst 42 Aufgaben und berücksichtigt nicht nur die Genauigkeit der Modelle, sondern auch Fairness, Toxizität, Effizienz und Robustheit.

EleutherAI
LM Eval
Harness
(2020)

allgemein

Language Model Evaluation Harness ist ein Open-Source-Framework für den Test von LLMs. Es unterstützt über 60 Standardbenchmarks für LLMs.

FM-eval
(2023)

allgemein

Framework für die Evaluierung von Prompts und durch Feintuning optimierten Modellen. Beinhaltet Benchmarks für akademische und geschäftliche Anwendungsbereiche.

MT-Bench
(2023)

Chatbots

MT-Bench besteht aus 80 aufeinander aufbauenden Fragen zur Bewertung von Chatbot-Fähigkeiten wie logischem Schlussfolgern und Wissen.

Chatbot
Arena
(2023)

Chatbots

Crowd-sourced Plattform für das Ranking von Chatbots und den zugehörigen LLMs durch menschliche Teilnehmer.

Unter anderem die UC Berkeley hat MMLU (Measuring Massive Multitask Language Understanding) mit dem Ziel entwickelt, Schwachpunkte in früheren Benchmarks wie GLUE (Generalized Likelihood Uncertainty Estimation) und SuperGLUE zu beheben, die eher sprachliche Fähigkeiten als das Gesamtverständnis von Sprache bewerten. Um gute Ergebnisse bei diesem Test zu erreichen, müssen Modelle über umfangreiches Weltwissen und Problemlösungsfähigkeiten verfügen.

BIG-bench (Beyond the Imitation Game) umfasst ähnlich wie MMLU eine breite Palette an Aufgaben, die hier von Experten aus unterschiedlichen Fachbereichen zusammengetragen wurden. Die Musterlösungen als Baseline für die Bewertung der Aufgaben hat ebenfalls ein Expertenteam erarbeitet. Etwa 80 Prozent der Benchmarkaufgaben sind einstufige Interaktionen. Hierbei misst man die Übereinstimmung der vom Modell generierten Ausgaben mit den Referenzergebnissen, wobei Standardmetriken wie ROUGE zum Einsatz kommen.

Etwa 20 Prozent der Aufgaben werten Programme aus, wobei pro Aufgabe jeweils über mehrere Runden mit dem Modell interagiert wird und die Leistung auch mithilfe benutzerdefinierter Metriken messbar ist. Neben den Fähigkeiten von Modellen untersucht BIG-bench auch die Auswirkung der Modellgröße auf die Leistung.

HELM, Evaluation Harness und FM-eval

Das CRFM (Center for Research on Foundation Models) in Stanford hat **HELM (Holistic Evaluation of Language Models)** zur Bewertung von 30 populären Modellen entwickelt [21], darunter etwa Llama 2 von Meta. Der Benchmark nutzt eine neue Taxonomie für die verwendeten Szenarien und Metriken und soll dadurch die Einschränkungen im Bewertungsprozess transparenter machen.

Szenarien sind Kombinationen von Aufgaben (Q&A, Summarization), Domänen (Nachrichten, Bücher) und der zugrunde liegenden Sprache. Neben der Genauigkeit der Modelle verwendet HELM viele weitere Metriken wie Accuracy, Calibration, Robustness, Fairness, Bias, Toxicity, Efficiency. HELM erfordert für eine Bewertung eine vergleichsweise große Menge an Rechenkapazität. Der Benchmark soll kontinuierlich weiterentwickelt und mit neuen Szenarien, Metriken und Modellen aktualisiert werden.

Der Evaluation Harness von EleutherAI ist ein Open-Source-Framework, das ein standardisiertes Vorgehen für den Test von Sprachmodellen ermöglicht. Das beschleunigt das Durchführen der Tests deutlich, Anwender können ihre Modelle mit einer Kombination von über 60 unterstützten Benchmarks testen. Eine der Zielsetzungen dieses Tools ist es, dass Anwender öffentlich zugreifbare Benchmarkergebnisse selbst möglichst einfach nachprüfen können. Das Framework kommt als Backend für das öffentlich zugängliche LLM Leaderboard von Hugging Face zum Einsatz.

Das Framework FM-eval von IBM verfolgt einen modularen Bewertungsansatz, beginnend mit einer einfachen Bewertung während des Modelltrainings bis hin zu einer umfassenden Bewertung, die Faktoren wie Unbedenklichkeit, Robustheit und Privatsphäre miteinbezieht [22]. Das Framework ist auf einen flexiblen Workflow hin ausgelegt und ermöglicht das einfache Hinzufügen von Aufgaben, Datensätzen und Metriken in den Evaluierungsprozess. Es beinhaltet die Open-Source-Komponente unitxt. Sie bietet eine Vorgehensweise und ein Interface zur Definition von Datensätzen und entsprechende Metriken für die Evaluierung. Das ermöglicht auch die Konvertierung von Rohdatensätzen in die von LLMs benötigten Eingaben.

Vollautomatisierte Bewertung – LLM as a Judge

Menschliche Bewertungen sind nach wie vor der Goldstandard für die Leistungsbewertung von LLMs. Leider ist dieser manuelle Ansatz teuer und vor allem langsam. 2023 wurden daher diverse Frameworks für eine automatisierte Bewertung von LLMs veröffentlicht, darunter GPTScore, G-Eval und LLM-Eval. Der interessante Aspekt dabei ist, dass diese Frameworks LLMs zur Bewertung von LLMs verwenden. Eine kürzlich veröffentlichte Studie untersucht, ob LLMs bei der Bewertung von Chatbots zu ähnlichen Ergebnissen kommen wie menschliche Bewerter [23] . Dafür haben die Forscher zwei Benchmarks entwickelt, die auf menschlichen Bewertungen von Chatbots beruhen: MT-Bench und Chatbot Arena. MT-Bench besteht aus 80

mehrstufigen Expertenfragen. Jeweils zwei zufällig ausgewählte Chatbots stellen sich dabei den Fragen und Experten bestimmen dann die bessere Antwort. Dieser Benchmark konzentriert sich also auf fest umrissene Aufgabenstellungen und Themengebiete. Bei Chatbot Arena hingegen können Benutzer beliebige Chats mit jeweils zwei zufällig ausgewählten Chatbots führen und anschließend für den besseren Chatbot abstimmen.

Für die automatische Bewertung durch LLM-Judges haben die Forscher verschiedene Varianten getestet: Einem LLM-Judge kann man eine Frage und zwei von Chatbots erzeugte Antworten vorlegen – zusammen mit der Anweisung, die bessere Antwort zu bestimmen. Alternativ können Nutzer einen LLM-Judge anweisen, einer Chatbot-Antwort direkt eine Punktzahl zuzuweisen.

Die in MT-Bench und Chatbot Arena von Menschen durchgeführten Bewertungen hat man anschließend mit den von LLMs durchgeführten Bewertungen verglichen. Das Ergebnis der Untersuchung zeigt, dass die besten aktuell verfügbaren LLM-Judges zu ähnlichen Bewertungsergebnissen kommen wie menschliche Tester.

Diese Ansätze zum Einsatz von LLMs als Schiedsrichter basieren allerdings auf Modellen, die man mittels Referenzdaten auf diese Aufgabe hin angepasst hat. Eine Alternative für das Ranking von LLMs ganz ohne Referenzdaten ist ein Ansatz, bei dem sich LLMs gegenseitig evaluieren. Inspiriert durch Menschen, bei denen sowohl ein Experte als auch eine sachkundige Person einen Neuling identifizieren können, besteht die Idee darin, Dreiergruppen von Modellen zu betrachten. Jedes Modell bewertet die beiden anderen und identifiziert mit hoher Wahrscheinlichkeit korrekt das schlechteste Modell im Triplet.

Human Red Teaming

Die bisher vorgestellten Methoden bewerten nur die Leistungsfähigkeit von LLMs. Da LLMs mit einer breiten Palette von Risiken behaftet sind, arbeiten viele Teams aktuell an neuen Methoden, die beim Erkennen dieser Risiken helfen sollen. Red Teaming ist eine Methode, um Schwachstellen von LLMs aufzudecken. Das Red Team fungiert dabei als Angreifer und testet mögliche schädliche Aktionen. Es kann etwa Prompts konstruieren, die das Modell dazu veranlassen, unerwünschte Texte zu generieren. Die Ergebnisse des Red Teamings helfen dabei, das Modell durch weiteres Training und Anpassen robuster zu machen.

Auch hier gibt es Bestrebungen zur Automatisierung. **Es gibt erste Vorschläge für ein Red Teaming Attack Framework [24]**, das zuerst einige manuell konstruierte Attack-Prompts sammelt und dann weitere LLMs nutzt, um diese Angriffsanleitungen nachzuahmen und weitere Anleitungen zu generieren. Die AttaQ-Methode automatisiert das Red Teaming, indem sie Bereiche der Eingabesemantik identifiziert, **für die das Modell wahrscheinlich schädliche Ausgaben erzeugt [25]**.

Fazit

Ein Schwerpunkt bei der Bewertung generativer Sprachmodelle ist die Beurteilung der aufgabenspezifischen Leistung. Hierfür eignen sich von Experten erstellte Referenzdaten in Kombination mit aufgabenspezifischen Metriken, die die Übereinstimmung der generierten Texte mit den Referenzdaten ermitteln. Benchmarks eignen sich nur bedingt für die aufgabenspezifische Leistungsbewertung, da sie die für Unternehmen relevanten Anwendungsfälle oft nur ungenügend berücksichtigen.

Daher sollten Unternehmen Benchmarks eher als unterstützendes Hilfsmittel für die initiale Auswahl von Modellen nutzen. Darüber hinaus erfordert jedes Unternehmen, Land, jede Branche und Anwendung eine Feinabstimmung des Verhaltens der Modelle und eine anschließende Bewertung, anstatt sich auf die Entscheidungen der externen Modellanbieter zu verlassen.

Bisher gibt es noch keine allgemein anerkannten Standards für die Bewertung von LLMs. Sinnvoll wäre ein Industriestandard, der es Anwendern auf einfache Weise ermöglicht, unterschiedliche Bereiche wie Risiken, ethische Aspekte und die Leistungsfähigkeit der Modelle zu beurteilen. Auch fehlen Standards zur Interpretation und effektiven Kommunikation der Bewertungsergebnisse, etwa standardisierte Verfahren zur Aggregation und Gewichtung der verschiedenen Bewertungsdimensionen. Hier möchte die kürzlich gegründete AI Alliance Abhilfe schaffen. Mitglieder sind Anbieter wie AMD, IBM, Intel und Meta, Forschungseinrichtungen wie das CERN, die Harvard University oder die UC Berkeley und Unternehmen, die KI einsetzen.

Der aktuelle Trend, Trainingsdaten und Modellparameter immer weiter zu vergrößern, ist kein Garant für bessere Leistung. Der Einsatz kleinerer Modelle, die man mit qualitativ höherwertigen Daten trainiert hat, ist deutlich nachhaltiger. Insofern wären leistungsfähigere und standardisierte Data-Governance-Prozesse für die qualitative Beurteilung, Nachverfolgbarkeit und die automatisierte Bereinigung der Trainingsdaten wünschenswert.

(pst [26])

URL dieses Artikels:

<https://www.heise.de/-9724539>

Links in diesem Artikel:

- [1] <https://www.heise.de/hintergrund/Kuenstliche-Intelligenz-Benchmarks-fuer-generative-Sprachmodelle-im-Ueberblick-9724539.html>
- [2] <https://www.heise.de/hintergrund/Marktuebersicht-KI-Server-mit-GPUs-im-Ueberblick-9724539.html>

9720404.html

- [3] <https://www.heise.de/hintergrund/Trends-bei-KI-teuer-US-amerikanisch-Big-Tech-dominiert-9718996.html>
- [4] <https://www.heise.de/ratgeber/PyTorch-Eigene-Bildgenerierungs-KI-mit-Python-bauen-9710438.html>
- [5] <https://www.heise.de/ratgeber/Website-per-KI-hacken-Browser-Skripte-mit-ChatGPT-und-Co-generieren-9706903.html>
- [6] <https://www.heise.de/ratgeber/Trend-Beruf-Mit-diesen-Faehigkeiten-wird-man-KI-Experte-9699530.html>
- [7] <https://www.heise.de/tests/Transkriptionsdienste-Whisper-V3-im-Vergleich-mit-Online-Diensten-9675869.html>
- [8] <https://www.heise.de/hintergrund/Faktencheck-Wie-KI-Assistenztools-helfen-koennen-Fake-News-zu-entlarven-9686264.html>
- [9] <https://www.heise.de/ratgeber/Wie-eine-lokale-KI-die-Fotosammlung-auf-dem-NAS-verschlagworten-kann-9685509.html>
- [10] <https://www.heise.de/hintergrund/Multi-Agenten-Systeme-Automatisierte-Leistungsanpassung-fuer-bessere-KI-9677800.html>
- [11] <https://www.heise.de/ratgeber/Fremdsprachen-lernen-Wie-man-ChatGPT-zum-Sprechtrainer-aufraestet-9675269.html>
- [12] <https://arxiv.org/abs/2108.07258>
- [13] <https://arxiv.org/abs/2303.08774>
- [14] <https://www.heise.de/ix/>
- [15] <https://arxiv.org/abs/2402.14860>
- [16] <https://arxiv.org/abs/2310.12941>
- [17] <https://www.heise.de/hintergrund/Sprachmodelle-verbessern-So-geht-s-mit-OpenLLaMA-und-der-Transformer-Bibliothek-9300588.html>
- [18] <https://deci.ai/blog/fine-tuning-peft-prompt-engineering-and-rag-which-one-is-right-for-you>
- [19] <https://arxiv.org/abs/2402.05624>
- [20] <https://hai.stanford.edu/news/ai-benchmarks-hit-saturation>
- [21] <https://arxiv.org/abs/2211.09110v2>
- [22] <https://arxiv.org/abs/2206.04615>
- [23] <https://arxiv.org/abs/2306.05685>
- [24] <https://arxiv.org/abs/2310.12505>
- [25] <https://arxiv.org/abs/2311.04124>
- [26] <mailto:pst@heise.de>