M17a - Modellauswahl

Stand: 05.2025

1 KI-Modelllandschaft: Ein Überblick

Die moderne KI-Landschaft bietet verschiedene spezialisierte Modelltypen für unterschiedliche Anwendungsfälle:

- **Reasoning-Modelle**: Spezialisiert auf logisches Denken und systematische Problemlösung (z.B. o3-mini) - diese Modelle lösen komplexe Aufgaben durch schrittweises, strukturiertes Denken.
- **Sprachmodelle**: Konzipiert für natürlichsprachliche Aufgaben wie Textgenerierung, Zusammenfassungen und Konversationen (z.B. GPT-4) - sie verstehen und erzeugen menschenähnliche Texte.
- **Codex-Modelle**: Optimiert für Codegenerierung und Programmieraufgaben - diese Modelle können Code schreiben, analysieren und debuggen.
- **Bildgenerierungsmodelle**: Erzeugen Bilder aus textlichen Beschreibungen (z.B. DALL-E) - sie wandeln Textanweisungen in visuelle Ergebnisse um.
- **Sprachverarbeitungsmodelle**: Spezialisiert auf Spracherkennung und -transkription (z.B. Whisper) - sie wandeln gesprochene Sprache in Text um.

2 Vergleich wichtiger Modelle

Die Wahl des richtigen Modells ist entscheidend für optimale Ergebnisse, Ressourcenschonung und maximale Effizienz. Hier ein Überblick der wichtigsten Modelle:

Modell	Hauptmerkmale	Empfohlene Anwendungsfälle
GPT-4o	Multimodales Allround-Modell: Versteht Text, Bilder und Audio, kann Bilder generieren. Sehr schnell und vielseitig.	Alltägliche Aufgaben, Brainstorming, Texterstellung, Content-Ideen, Bildanalysen, E-Mails, Konzepte. Gut für schnelle Dialoge und allgemeine Fragen.
GPT-4o Mini	Leichtere Version von GPT-4o: Verarbeitet Text und Bilder, ressourcenschonend und günstiger. Deutlich intelligenter als GPT-3.5-turbo.	Einfachere Aufgaben, Bildverarbeitung, schnelle und unkomplizierte Anwendungen, kostengünstige Chatbots.

Modell	Hauptmerkmale	Empfohlene Anwendungsfälle
o3-mini	Reasoning-Modell: Hohe Intelligenz bei niedrigen Kosten und geringer Latenz. Konzipiert für strukturiertes Denken.	Wissenschaftliche, mathematische und Programmieraufgaben, technische und logische Probleme, faktenbasierte Recherchen.
o4-mini	Kompaktes Reasoning-Modell: Optimiert für Geschwindigkeit und Kosteneffizienz. Stark in mathematischen, Programmier- und visuellen Aufgaben.	Komplexe Argumentationsstrukturen, technische Aufgaben, Programmierprojekte, visuelles Denken, wissenschaftliche Fragestellungen.
o3	Leistungsstärkster "Denker": Herausragend in Programmierung, Mathematik, Wissenschaft und visueller Analyse. Arbeitet mit verknüpften Einzelschritten ("Chain-of-Thought").	Komplexe Recherchen, anspruchsvolle Programmieraufgaben, Datenanalyse, strategische Planung, Code-Review und Debugging. Beste Wahl für höchste Präzision.

3 Modellauswahlprozess: Schritt für Schritt

Die Auswahl des optimalen KI-Modells erfordert einen strukturierten Prozess:

3.1 Anforderungsanalyse

- **Definition der Aufgaben**: Legen Sie fest, welche spezifischen Funktionen das Modell erfüllen soll (z.B. Textgenerierung, Fragebeantwortung).
- **Qualitätskriterien**: Bestimmen Sie, welche Qualitätsstandards (Kohärenz, Genauigkeit) erfüllt werden müssen.
- **Domänenkenntnisse**: Identifizieren Sie, welches Fachwissen für Ihre Aufgabe notwendig ist.
- **Antwortgeschwindigkeit**: Definieren Sie die akzeptable Reaktionszeit des Modells.
- **Budget**: Setzen Sie einen finanziellen Rahmen für Ihre KI-Lösung.

3.2 Bewertungskriterien

- **Verständlichkeit**: Wie klar und nachvollziehbar sind die Modellausgaben?
- **Effizienz**: Wie schnell verarbeitet das Modell Eingaben und liefert Ausgaben?
- **Skalierbarkeit**: Kann das Modell mit steigenden Anforderungen mitwachsen?
- **Kosten**: Wie hoch sind die Betriebs- und Nutzungskosten des Modells?

3.3 Recherche und Vorauswahl

- Analysieren Sie verfügbare Modelle anhand Ihrer festgelegten Kriterien und erstellen Sie eine Vorauswahl geeigneter Kandidaten.

3.4 Praktische Modellbewertung

- **Quantitative Methoden**: Verwenden Sie Benchmarks und Metriken, um die Leistung objektiv zu messen.
- **Qualitative Verfahren**: Sammeln Sie Nutzerfeedback zur praktischen Verwendbarkeit.
- **Testphase**: Erproben Sie die Modelle in einer realistischen Umgebung.

3.5 Finale Auswahl und Implementierung

- Treffen Sie eine fundierte Entscheidung für das am besten geeignete Modell und integrieren Sie es in Ihre Systeme.

Beta-WebApp für Modellauswahl. 😊

4 Modellkaskade: Mehrere Modelle klug kombinieren

Die Modellkaskade kombiniert mehrere KI-Modelle, um ihre jeweiligen Stärken zu nutzen und Schwächen auszugleichen:

4.1 Beispiel für eine Modellkaskade

1. **Datenanalyse mit pandas**: Analysiert große Datensätze und erstellt statistische Zusammenfassungen
2. **Logische Strukturierung mit o3-mini**: Strukturiert die Ergebnisse und erstellt eine logische Gliederung
3. **Kreative Textgenerierung mit GPT-4o**: Verfasst ansprechende Texte basierend auf der Struktur
4. **Multimodale Präsentation**: Ergänzt den Text mit visuellen Elementen

4.2 Vorteile einer Modellkaskade

1. **Effizienzsteigerung**: Jedes Modell wird für seine Stärken optimal eingesetzt
2. **Kostenoptimierung**: Ressourcenschonende Modelle für einfache Aufgaben, teurere nur wo nötig
3. **Flexibilität**: Bearbeitung unterschiedlichster Anforderungen durch spezialisierte Modelle

5 Bewertungsmethoden für KI-Modelle

5.1 Wichtige Benchmarks

- **MMLU (Massive Multitask Language Understanding)**: Standard-Benchmark über 57 Fachgebiete, der die Allgemeinbildung und Fachkenntnisse von Modellen misst.

Modell	MMLU-Score
GPT-4o	88,7%
Gemini 2.0 Ultra	90,0%
Claude 3 Opus	88,2%
Llama 3.1 405B	87,3%
gpt-4o-mini	70,0%

5.2 Bewertungsdimensionen

Die Bewertung von KI-Modellen umfasst verschiedene Aspekte:

1. **Wissens- und Fähigkeitsbewertung**:
   - Wie gut beantwortet das Modell Fragen verschiedener Schwierigkeitsgrade?
   - Wie zuverlässig ergänzt es fehlendes Wissen?
   - Wie gut löst es logische und mathematische Probleme?
   - Wie effektiv nutzt es externe Werkzeuge?
2. **Alignment-Bewertung**:
   - Inwieweit stimmt das Modellverhalten mit menschlichen Werten überein?
   - Wie ethisch und moralisch sind die Antworten?
   - Wie fair und unvoreingenommen ist das Modell?
   - Wie wahrhaftig sind die gelieferten Informationen?
3. **Sicherheitsbewertung**:
   - Wie robust ist das Modell gegenüber Störungen und Angriffen?
   - Welche potenziellen Risiken birgt die Nutzung des Modells?

5.3 Konkrete Bewertungsmethoden

5.4 Automatisierte Metriken

- **BLEU**: Misst die Übereinstimmung zwischen generiertem und Referenztext durch Vergleich von Wortgruppen.
- **ROUGE**: Bewertet die Qualität von Zusammenfassungen durch Analyse übereinstimmender Wortsequenzen.

5.5 Menschliche Bewertung

- Bewertung nach Kriterien wie Grammatik, Zusammenhang, Lesbarkeit und Relevanz
- Elo-System für den direkten Vergleich verschiedener Modelle (ähnlich wie bei Schach-Ratings)

5.6 KI-basierte Bewertung

- Einsatz leistungsfähiger Modelle zur Bewertung anderer Modelle
- Automatische Erkennung von Fehlinformationen in KI-Antworten

6 Praktische Anwendungsbereiche

Die Modellevaluierung und -auswahl findet in verschiedenen Szenarien Anwendung:

6.1 Kundenservice-Chatbots

- Auswahl eines schnellen Modells mit guter Verständlichkeit und Mehrsprachigkeit
- Bewertung nach Kundenzufriedenheit und Lösungsrate

6.2 Content-Erstellung

- Nutzung kreativer Modelle für Marketing, Social Media und Blogbeiträge
- Bewertung nach Originalität, Engagement und Konversionsraten

6.3 Technische Assistenz

- Einsatz von Reasoning-Modellen für Programmierung und Fehlerbehebung
- Bewertung nach Codequalität und Lösungsgeschwindigkeit

7 Fazit

Fazit

Zusammenfassend lässt sich sagen, dass die **Evaluierung von Large Language Models (LLMs) ein wichtiges Forschungsgebiet** ist, um ihre Fähigkeiten und Grenzen zu verstehen. Die Evaluierung umfasst verschiedene **Attribute wie Grammatikalität, Kohäsion, Gefallen, Relevanz, Flüssigkeit und Bedeutungserhalt**. Sowohl **menschliche Evaluatoren als auch LLMs selbst werden zur Bewertung eingesetzt**. Es gibt **spezifische Benchmarks und Datensätze** zur Bewertung von LLMs in verschiedenen Bereichen wie **Textgenerierung, Fragebeantwortung und Zusammenfassung**.

Ein wichtiger Aspekt der LLM-Evaluierung ist die **Sicherheitsbewertung**, die **Robustheit gegenüber adversarialen Angriffen** (manipulierte Eingaben, um LLM in die Irre zu führen) und die Identifizierung von **Risiken wie Bias und Toxizität** umfasst. Die Evaluierung kann auch auf **spezialisierte LLMs** in Bereichen wie Medizin, Recht und Finanzen zugeschnitten sein.

Verschiedene **Metriken, darunter Likert-Skalen und der BLEU-Score**, werden zur Quantifizierung der LLM-Leistung verwendet. Es gibt auch **Tools und Frameworks wie DeepEval**, die die Evaluierung erleichtern. Es ist wichtig zu beachten, dass **Evaluierungsbias existieren können**, beispielsweise eine Präferenz für längere Texte. Die **ethischen Aspekte** spielen ebenfalls eine Rolle bei der Entwicklung und Nutzung von LLMs.

8 A | Aufgabe

Die Aufgabestellungen unten bieten Anregungen, Sie können aber auch gerne eine andere Herausforderung angehen.

Anforderungsanalyse für ein KI-Projekt

Entwickeln Sie eine strukturierte Anforderungsanalyse für ein fiktives oder reales KI-Projekt.

Aufgabenstellung:

1. Wählen Sie einen konkreten Anwendungsfall (z.B. Kundenservice-Chatbot für eine Bank, Content-Generator für Social Media, oder Übersetzungstool für technische Dokumentation).
2. Definieren Sie:
   - Die primären Funktionen, die das KI-Modell erfüllen soll
   - Die spezifischen Anforderungen an das Sprachverständnis
   - Notwendige Fachkenntnisse in relevanten Domänen
   - Anforderungen an die Antwortgeschwindigkeit
   - Budget-Rahmenbedingungen
3. Erstellen Sie eine Prioritätenliste dieser Anforderungen (unbedingt erforderlich, wichtig, wünschenswert).
4. Beschreiben Sie, welche Kompromisse Sie bei konkurrierenden Anforderungen eingehen würden.

Abgabeformat:
Erstellen Sie ein Dokument mit Ihrer Anforderungsanalyse (1-2 Seiten).

Vergleichsanalyse bekannter KI-Modelle

Führen Sie eine vergleichende Analyse von mindestens drei verschiedenen KI-Modellen anhand vorgegebener Bewertungskriterien durch.

Aufgabenstellung:

1. Wählen Sie drei KI-Modelle aus der folgenden Liste aus:
   - GPT-4o
   - Claude 3 Opus
   - Gemini 2.0 Ultra
   - Llama 3.1
   - Mistral 7B
   - Ein anderes aktuelles KI-Modell Ihrer Wahl
2. Recherchieren Sie die Leistungsmerkmale dieser Modelle anhand der folgenden Kriterien:
   - MMLU-Score oder vergleichbare Benchmark-Ergebnisse
   - Kontextfenstergröße
   - Antwortlatenz
   - Kosten (pro Token oder alternativer Maßstab)
   - Verfügbarkeit (API, Open-Source, etc.)
   - Unterstützte Sprachen
   - Multimodale Fähigkeiten (falls vorhanden)
3. Erstellen Sie eine Bewertungstabelle mit den recherchierten Informationen.
4. Verfassen Sie eine begründete Empfehlung, welches dieser Modelle sich für folgende Szenarien am besten eignen würde:
   - Entwicklung eines kostengünstigen Chatbots für ein kleines Unternehmen
   - Erstellung von KI-generierten Inhalten für ein internationales Nachrichtenportal
   - Unterstützung bei der Software-Entwicklung

Abgabeformat:
Vergleichstabelle mit Bewertungen und einer Seite mit Ihren Empfehlungen.

Konzept für die qualitative Evaluation eines Sprachmodells

Entwickeln Sie ein strukturiertes Testverfahren zur qualitativen Bewertung eines Sprachmodells.

Aufgabenstellung:

1. Entwerfen Sie ein Bewertungsschema mit 5-7 qualitativen Kategorien, die für Ihre gewählte Anwendung relevant sind (z.B. Genauigkeit, Kreativität, Nützlichkeit der Antworten, Verständnis komplexer Anweisungen, Kulturelle Sensibilität).
2. Erstellen Sie für jede Kategorie:
   - Eine klare Definition, was in dieser Kategorie bewertet wird
   - Eine Bewertungsskala (z.B. 1-5 oder 1-10)
   - 2-3 konkrete Testfragen oder -aufgaben, die diese Kategorie prüfen
   - Bewertungskriterien: Was wäre eine ausgezeichnete (5/5) vs. eine unzureichende (1/5) Antwort?

3. Beschreiben Sie den Evaluationsprozess:
    - Wie viele Bewerter sollten eingesetzt werden?
    - Wie würden Sie die Bewertungen zusammenfassen?
    - Welche Maßnahmen würden Sie ergreifen, um Bewertungsverzerrungen zu vermeiden?
4. Erläutern Sie, wie Sie die Ergebnisse dieser qualitativen Bewertung mit quantitativen Metriken (wie MMLU) kombinieren würden, um ein Gesamtbild der Modellleistung zu erhalten.

Abgabeformat:
Ein 2-3 seitiges Konzeptpapier mit Ihrem Evaluationsschema, den Testfragen und dem geplanten Prozess.