Bundesministerium der Finanzen

Methodische Evaluierung von Sprachmodellen

Modelle für die Öffentliche Verwaltung Evaluieren

René Walter, Dr. Thilo Michael

Agenda

1. Übersicht KI-Kompetenzcenter
2. Was ist MÖVE?
3. Was wird evaluiert?
4. Ergebnisse
5. Ausblick

1. KI-Kompetenzcenter

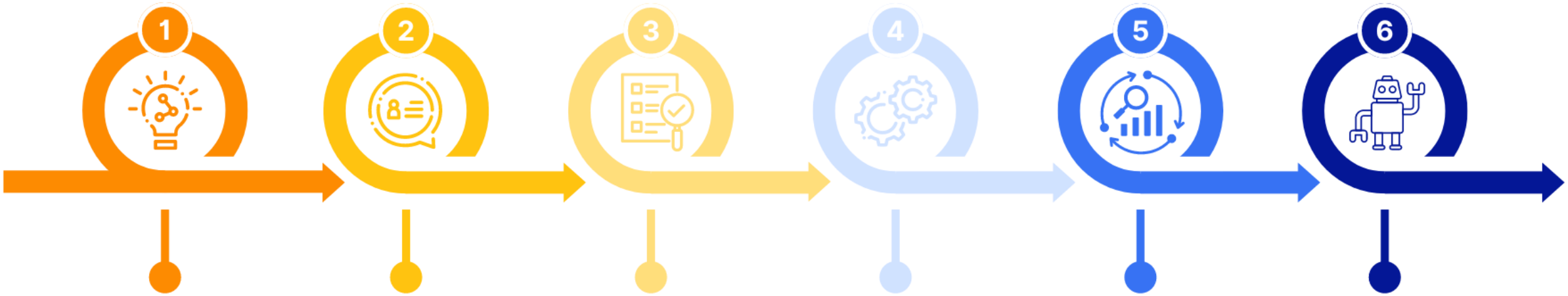
Übersicht und Vorgehen

Befähigung der Bundesverwaltung zum souveränen Einsatz von KI

Mehrwerte

- Nutzerzentrierte KI-Lösungen
- Schnelle und einfache Verprobung innovativer KI-Anwendungsideen
- Direkter Zugang zu KI-Experten

Iterativ von der Idee zum KI-Prototyp



1 Use Cases identifizieren & Machbarkeit bewerten

- Definition der Aufgabe, die mit KI gelöst werden soll
- Bewertung der Datenlage

2 Mitarbeitende der Verwaltung einbinden

- Analyse von Arbeitskontext und Rahmen-bedingungen für einen KI-Einsatz
- Verstehen der Anforderungen der Nutzenden

3 Analyse der verfügbaren Daten

- Bereinigen der Daten
- Explorative Datenanalyse

4 Auswahl und/oder Entwicklung eines KI-Modells

- Auswahl geeigneter Modelle
- Trainieren von Modellen

5 Evaluation von Modellen und Mehrwerten

- Bewerten und auswählen des KI-Modells mit Zielnutzenden
- Iterative Entwicklung einer prototypischen Anwendung

6 Projekt- und Prototyp-Übergabe

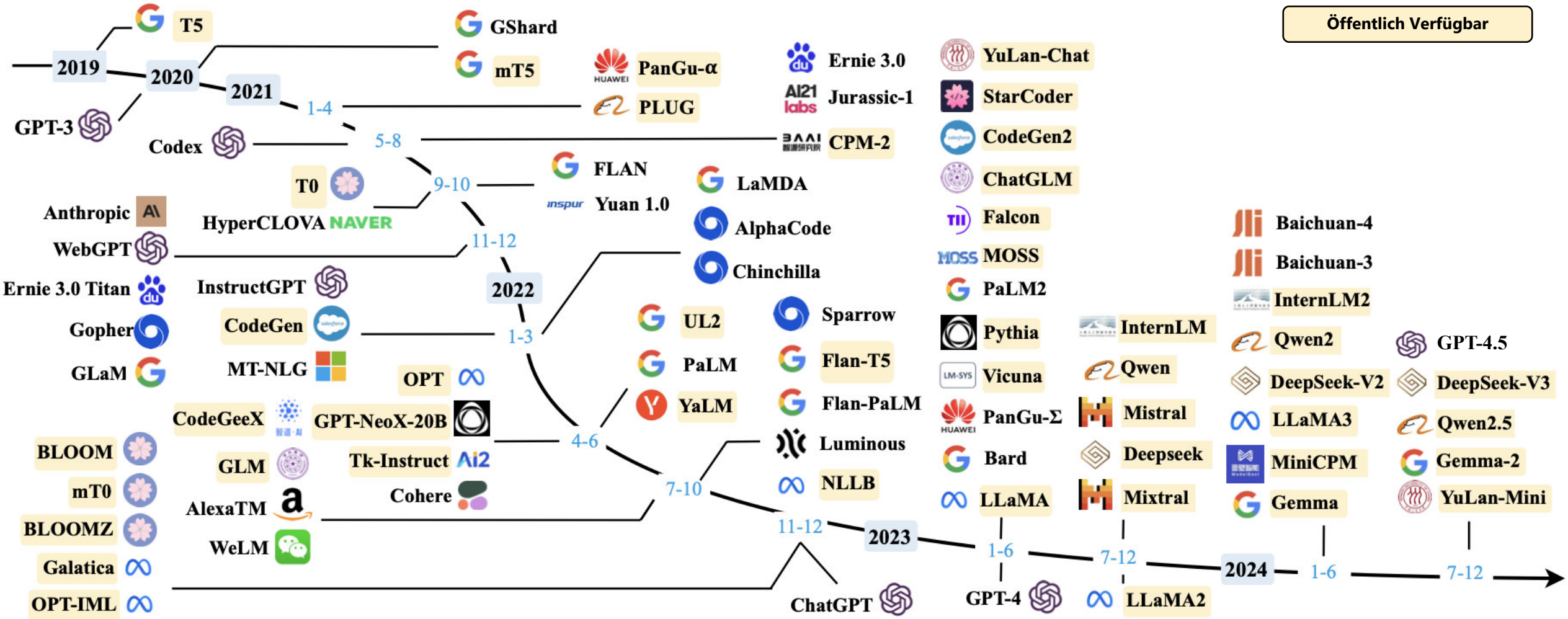
- KI-Modell für Nutzenden-gruppen implementieren
- Prototyp und fachliche Erkenntnisse bereitstellen

2. Was ist MÖVE?

Übersicht des Forschungsvorhabens

Sprachmodelle: *Nicht nur ChatGPT*

Öffentlich Verfügbar



Quelle: https://arxiv.org/pdf/2303.18223

Was ist MÖVE

- Ein Sprachmodell-Benchmark für die öffentliche Verwaltung
- Forschungsprojekt im Rahmen des KI-Kompetenzcenters
- Aktive Entwicklung seit ca. 6 Monaten

Ziel

Der Benchmark soll ein frei verfügbares, nützliches Werkzeug für den Einsatz von Sprachmodellen in der öffentlichen Verwaltung sein

Es gibt schon viele LLM-Benchmarks! Wieso noch einer?

- **Fokus auf relevanten Aufgaben für die Bundesverwaltung**

  Andere: Puzzle, Generelles Wissen, "AGI"

  MÖVE: Relevante Aufgaben, die in der öffentlichen Verwaltung Anwendung finden

- **Performance und Governance Kriterien**

  Andere: Fokus auf Performance in einer spezifischen Aufgabe

  MÖVE: beinhaltet "Governance Kriterien" wie Bias, Sicherheit, Halluzinationen

- **Deutsche Sprache**

  Andere: Häufig in englische oder automatisch übersetzte Datensätze

  MÖVE: speziell ausgewählte, teils händisch erstellte Datensätze, meistens direkt auf Deutsch

- **Validierte Datensätze**

  Andere: Sehr große Datensätze mit teils schlechter Qualität

  MÖVE: Goldstandard Datensätze mit Domänenwissen direkt aus der Verwaltung

Übersicht der Performance und Governance Kriterien

PERFORMANCE Evaluierung

GOVERNANCE Evaluierung

Übersicht der Performance und Governance Kriterien

Eine Performance oder Governance Evaluierung besteht aus mehreren **"Szenarien"**:

Ein Szenario wird definiert durch:

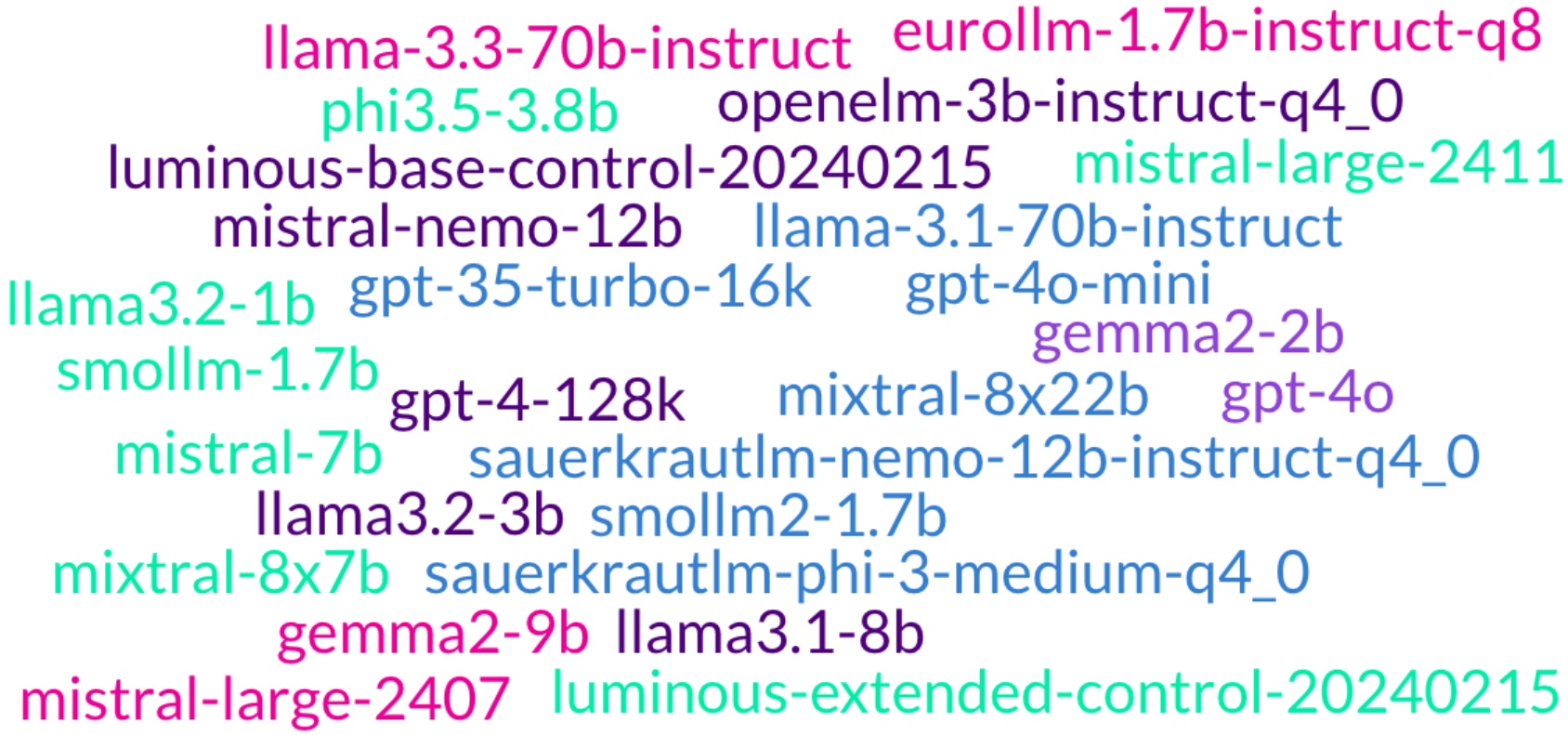
- Ein LLM (z.B. GPT-4o, mistral-large-2411, ...)
- Eine Aufgabe (z.B. Zusammenfassen, Question Answering, ...)
- Einen Datensatz
- Einen Prompt (z.B. "*Fasse das folgende Gesetz zusammen: {Gesetzestext}*")
- Mehrere Metriken ("Klassische" Metriken, ML-basierte Metriken, LLM-as-a-judge Metriken)

3. Was wird evaluiert?

Modelle, Datensätze, Metriken

LLMs: Welche Modelle werden evaluiert

Wir evaluieren über 32 Modelle (hauptsächlich **Open-Weights Modelle** und OpenAI's **GPT** Modelle)



Performance-Evaluierung

Aufgaben und Datensätze

Aktuell evaluieren wir 2 Aufgaben mit insgesamt 6 Datensätzen:

ZUSAMMENFASSEN

- Eur Lex Sum
  Zusammenfassen von europäischen Gesetzen
- Swiss Leading Decisions
  Zusammenfassen von Schweizer Gerichtsentscheidungen
- **KIKC-Summary**
  Händisch erstellte Zusammenfassungen von Dokumenten aus verschiedenen Ministerien

BEANTWORTUNG VON FRAGEN

- German QuAD
  Fragen aus der deutschen Wikipedia
- **KIKC-QA**
  Händisch erstellte Fragen und Antworten aus Dokumenten aus verschiedenen Ministerien
- **FAQ Law**
  FAQ zu mehreren deutschen Gesetzen

Woran wir gerade arbeiten...

KIKC-Keywords (händisch erstellt), **SUM_Q1** (Veröffentlichungen des Umweltbundesamt), **SUM_Q4** (DIP Veröffentlichungen), **SUM_Q10** (social science open access repository), Wahlomat (Datensatz zur deutschen Politik)

***fett-markierte** Datensätze wurden für MÖVE erstellt und kuratiert

Unsere Metriken

- Die Aufgabe bestimmt, welche Metriken angewendet werden
- **7** "Metriken"
- **17** *Scores* (z.B. "f1")
- **2** Aggregationstypen (Median und Durchschnitt)

Klassische Metriken: Exact match, ROUGE, BLEU

ML-basierte Metriken: BERTScore, SemScore

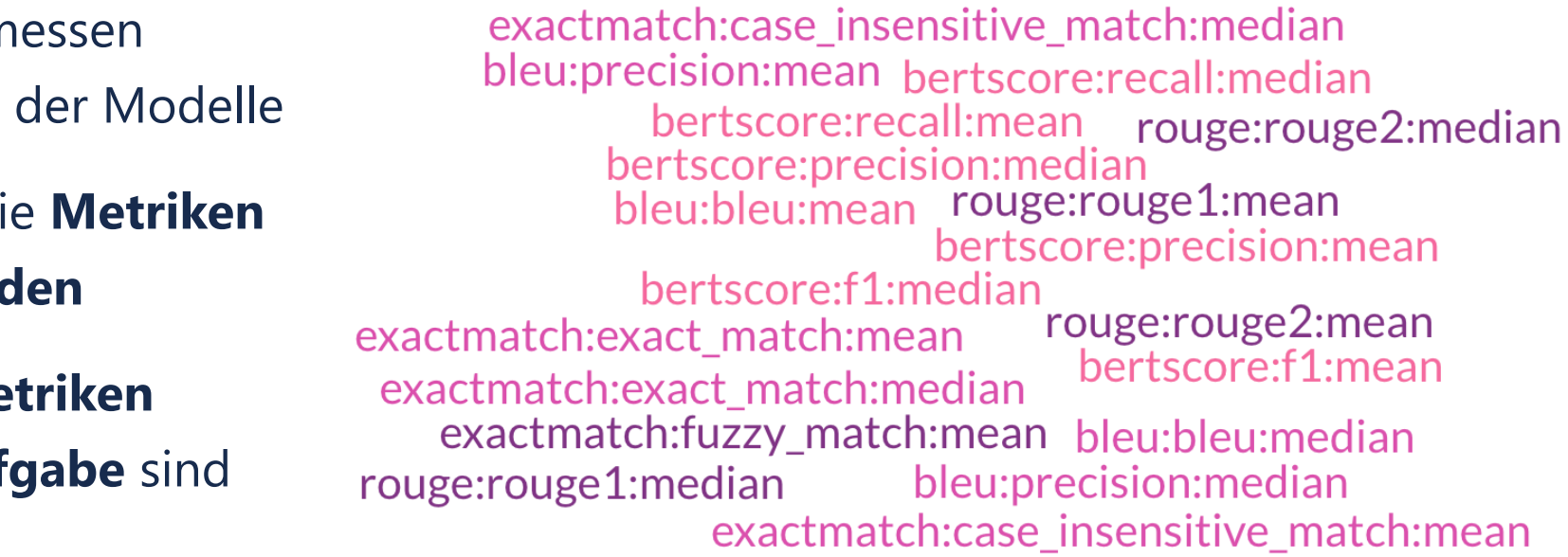
LLM-as-a-judge Metriken: RAGAS

4. Ergebnisse

Wie schneiden die Modelle ab

Unsere Metriken

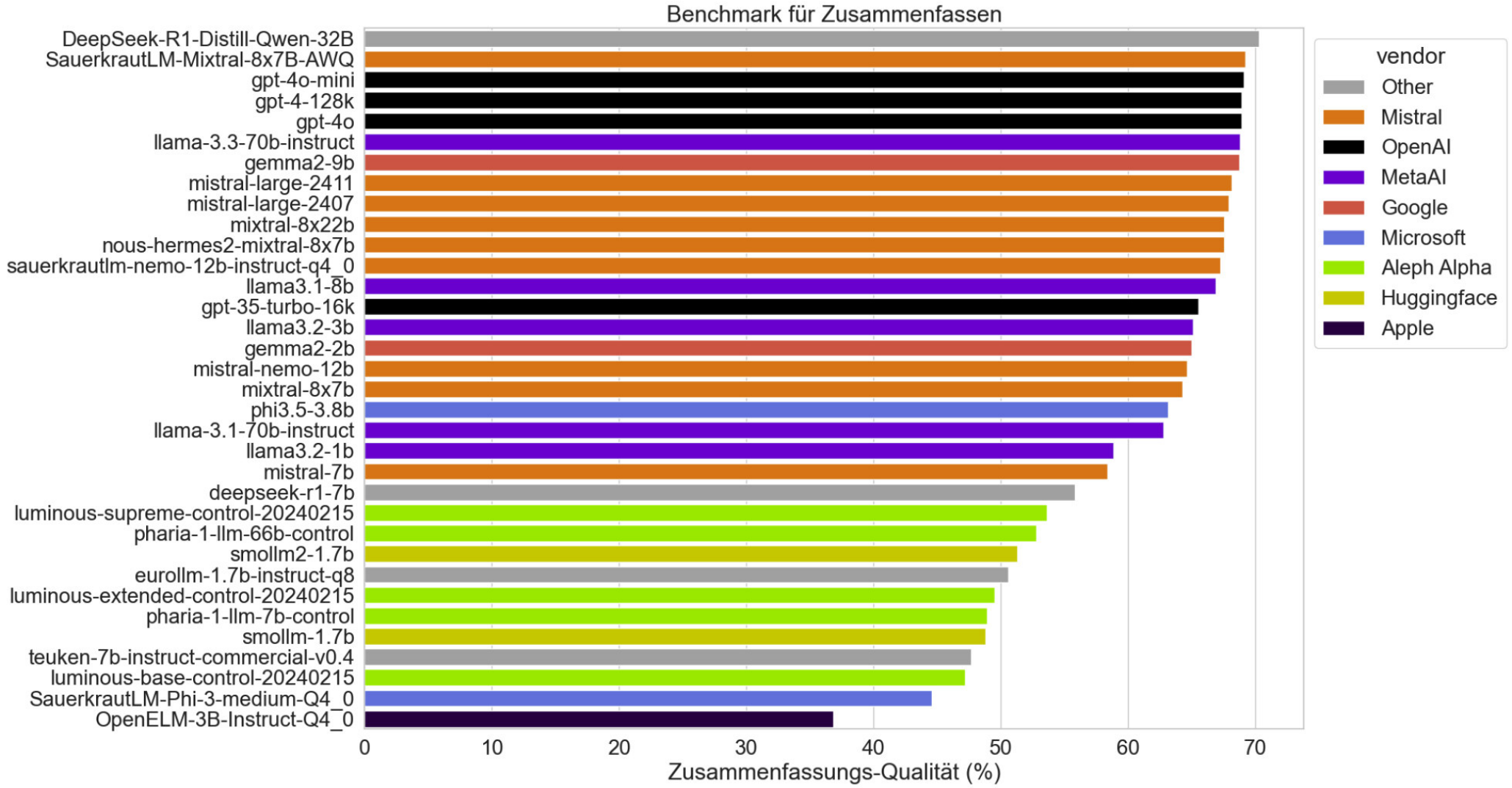
- Die **gewählten Metriken** messen **unterschiedliche Facetten** der Modelle
- Wir untersuchen wie sich die **Metriken untereinander unterscheiden**
- Wir wählen aus, **welche Metriken repräsentativ für eine Aufgabe** sind



Aufgaben-basierter Score

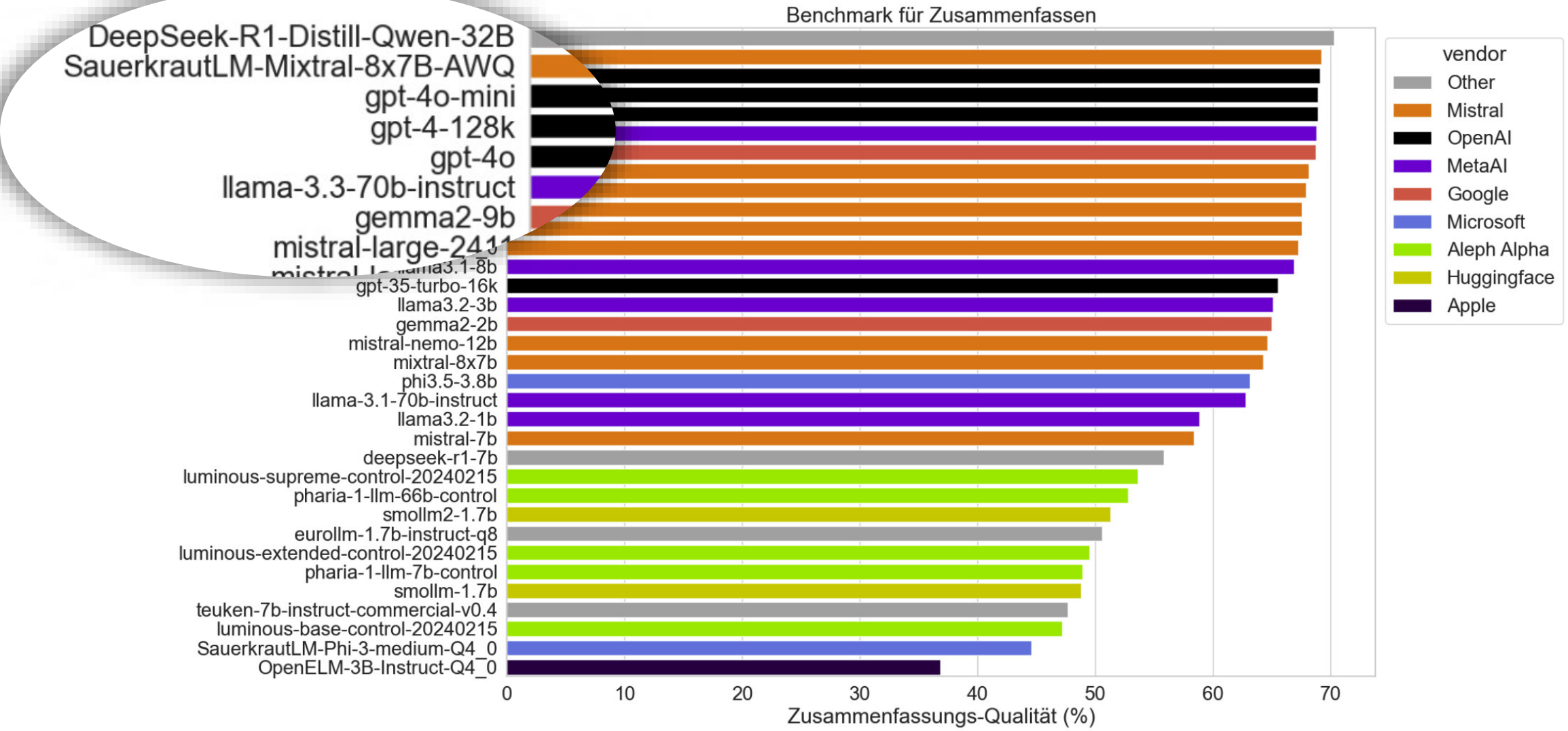
- Wir berechnen einen "Overall Score" für jede Aufgabe (**Summarization, Question Answering**)
- Dieser Score ist eine Kombination aus repräsentativen Metriken der verschiedenen Metrikentypen: **klassische, ML-basiert, LLM-as-a-judge**
- Wir arbeiten an einem wissenschaftlich fundierteren Ansatz um aufgabenbasierte Metriken zu berechnen

Ergebnisse: Zusammenfassen



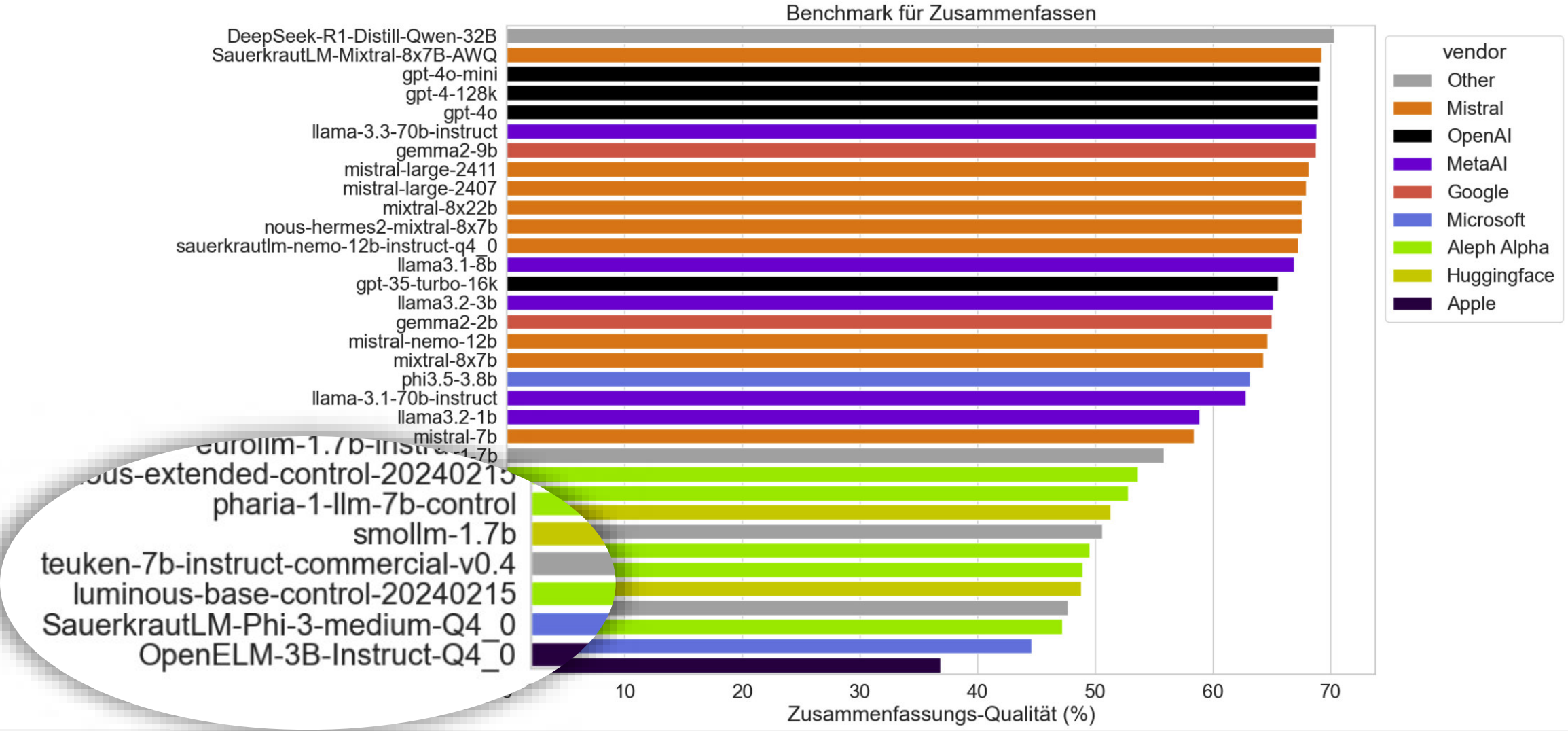
VORLÄUFIGE ERGEBNISSE

Ergebnisse: Zusammenfassen



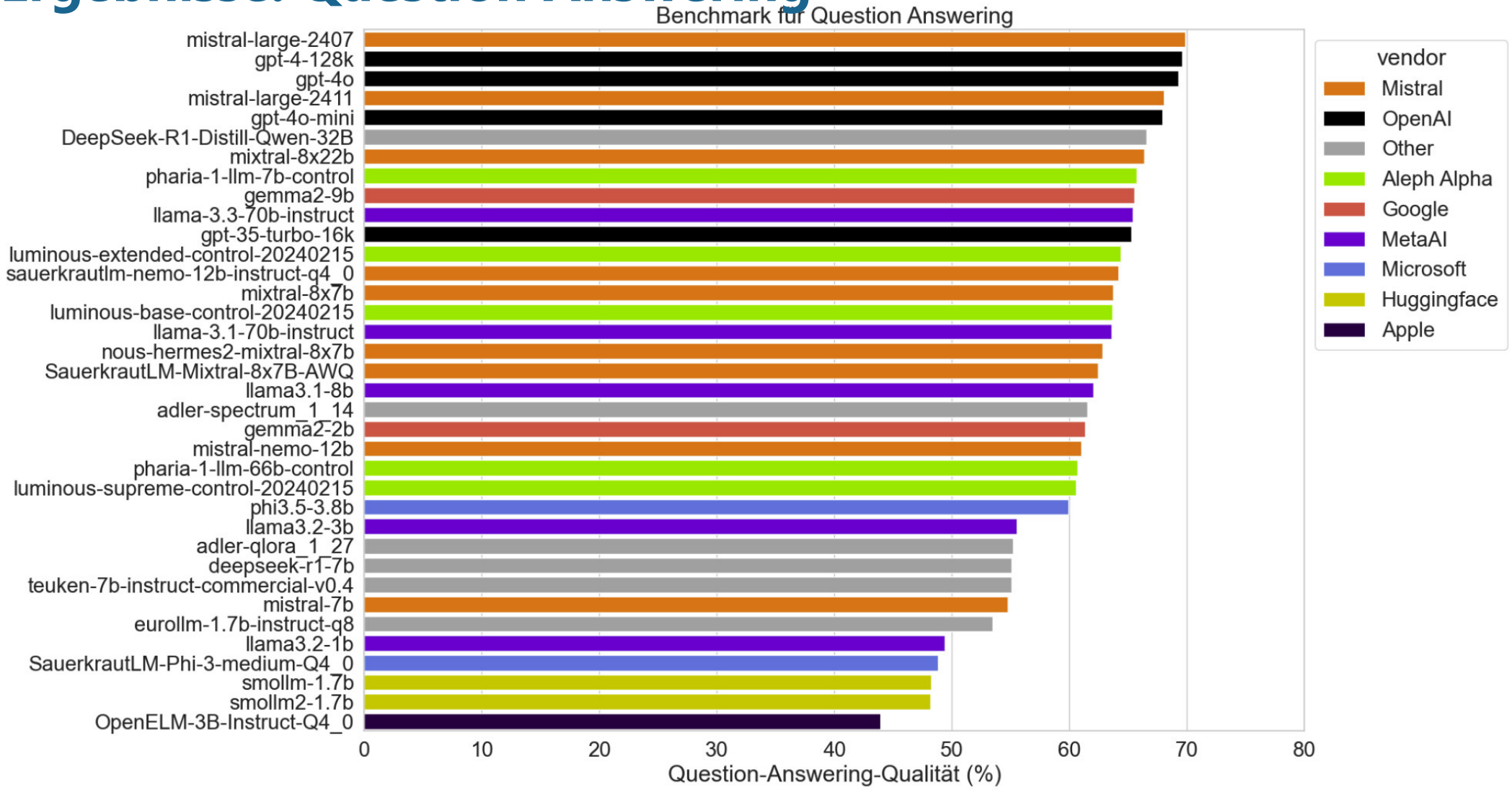
VORLÄUFIGE ERGEBNISSE

Ergebnisse: Zusammenfassen



VORLÄUFIGE ERGEBNISSE

Ergebnisse: Question Answering



VORLÄUFIGE ERGEBNISSE

5. Ausblick

Woran wir gerade arbeiten

Governance-Kriterien

Soziale Fairness
- Biases (z.B. Misogynie)
- Daten
  - En: Fairness (DecodingTrust), Microagressions (Social Bias Inference Corpus)
  - De: German NLG Bias

Politischer Bias
- Extrahieren von politischen Ansichten aus den Modellen
- Sehr relevant u.a. wegen der politischen Entwicklungen in den USA
- Daten
  - GermanPartiesQA, Political Compass Test

Andere Governance Kriterien
- Potentielle Kooperationen (TÜV AI.Lab, TU/DFKI)

Performance-Kriterien

Keyword extraction
- Händisch erstellter Goldstandard-Datensatz verfügbar
- Große Nachfrage bei Ministerien

Übersetzen
- Finden von öffentlich verfügbaren, ministeriumsnahen Datensätzen
- Fokus: Deutsch <-> Englisch
- Austausch mit dem BAMF zu weiteren Sprachen

Reasoning Modelle
- Evaluierung von Reasoning Modellen
  - OpenAI (o1, o1-mini, o3, o3-mini)
  - Deepseek (R1 & distills)
  - Alibaba Qwen (QwQ)

Ergebnisse & Präsentation

Parametrisches Modell zur Vorhersage von "Subjektiver Qualität"

- Geplante Kooperation mit dem DFKI SLT und der TU Berlin
- Kombination von Metriken in ein Modell um die Qualität eines generierten Textes vorherzusagen

Öffentliche Webseite

- Interaktives Leaderboard
- Selektion von Aufgaben und/oder Modellen und automatisches kombinieren von Qualitäts-Scores

Forschungspaper

- Umfangreiche Dokumentation über den Benchmark
- Einsichten in die Details der Datensätze

Mehr Interesse an KI?

KI-Report

Bedarfsanalyse

Anwendungsfälle identifizieren

Lernen

Aufbauseminar KI-Kompetenz

3 Tage

Praxisnahes Seminar für Bundesbehörden zur Stärkung von KI-Kompetenz

Kontakt

Dr. Thilo Michael
Senior Innovations Developer
thilo.michael@bdr.de

Camilla Dalerci
Senior Innovations Developer
camilla.dalerci@bdr.de

Wir freuen uns auf regen Austausch!
kontakt-kikc@bdr.de

für Ihre Aufmerksamkeit