

Leistung großer KI-Sprachmodelle in der Kinder- und Jugendpsychiatrie

Johanna Charlotte Neubauer¹, Anna Kaiser^{1,2,3}, Leon Lettermann⁴, Tobias Volkert⁵, Alexander Häge¹

¹Klinik für Psychiatrie und Psychotherapie des Kindes- und Jugendalters, Zentralinstitut für Seelische Gesundheit, Medizinische Fakultät Mannheim, Universität Heidelberg, Deutschland

²Klinik für Kinder- und Jugendpsychiatrie, Psychosomatik und Psychotherapie, Universitätsklinikum Ulm, Ulm, Deutschland

³Klinische Psychologie und Psychotherapie des Kindes- und Jugendalters, Universität Freiburg, Freiburg, Deutschland

⁴BioQuant, Universität Heidelberg; Institut für Theoretische Physik, Universität Heidelberg, Heidelberg, Deutschland

⁵Unabhängiger Wissenschaftler

Ziele

- Evaluation der Genauigkeit großer künstlicher Intelligenz (KI)-Sprachmodelle bei Multiple-Choice-Fragen in der Kinder- und Jugendpsychiatrie
- Vergleich verschiedener Modelle (ChatGPT vs. Gemini)

Hintergrund

- Große Sprachmodelle gewinnen zunehmend an Bedeutung in Medizin und Ausbildung
- Einsatzmöglichkeiten in der Kinder- und Jugendpsychiatrie bislang unzureichend untersucht
- Unklar: Wie gut schneiden gängige große Sprachmodelle in für Ärzt:Innen gängigen Wissensprüfungen ab?

Methoden

- 150 Multiple-Choice-Fragen (Fachärzt:innen-Niveau) [1]
- 5 Antwortoptionen, 1 korrekt
- Reduktion von Zufallseffekten (jede Frage 10× gestellt)
- Randomisierte Antwortreihenfolge
- Befragte Modelle:
 - ChatGPT-4o, ChatGPT o1-mini
 - Gemini 2.0 Flash, Gemini 1.5 Flash
- Analyse: Genauigkeit (%), t-Tests, Wilcoxon-Tests, Korrelationsanalyse

Überlegene Performance von Gemini 2.0 Flash gegenüber Gemini 1.5 Flash

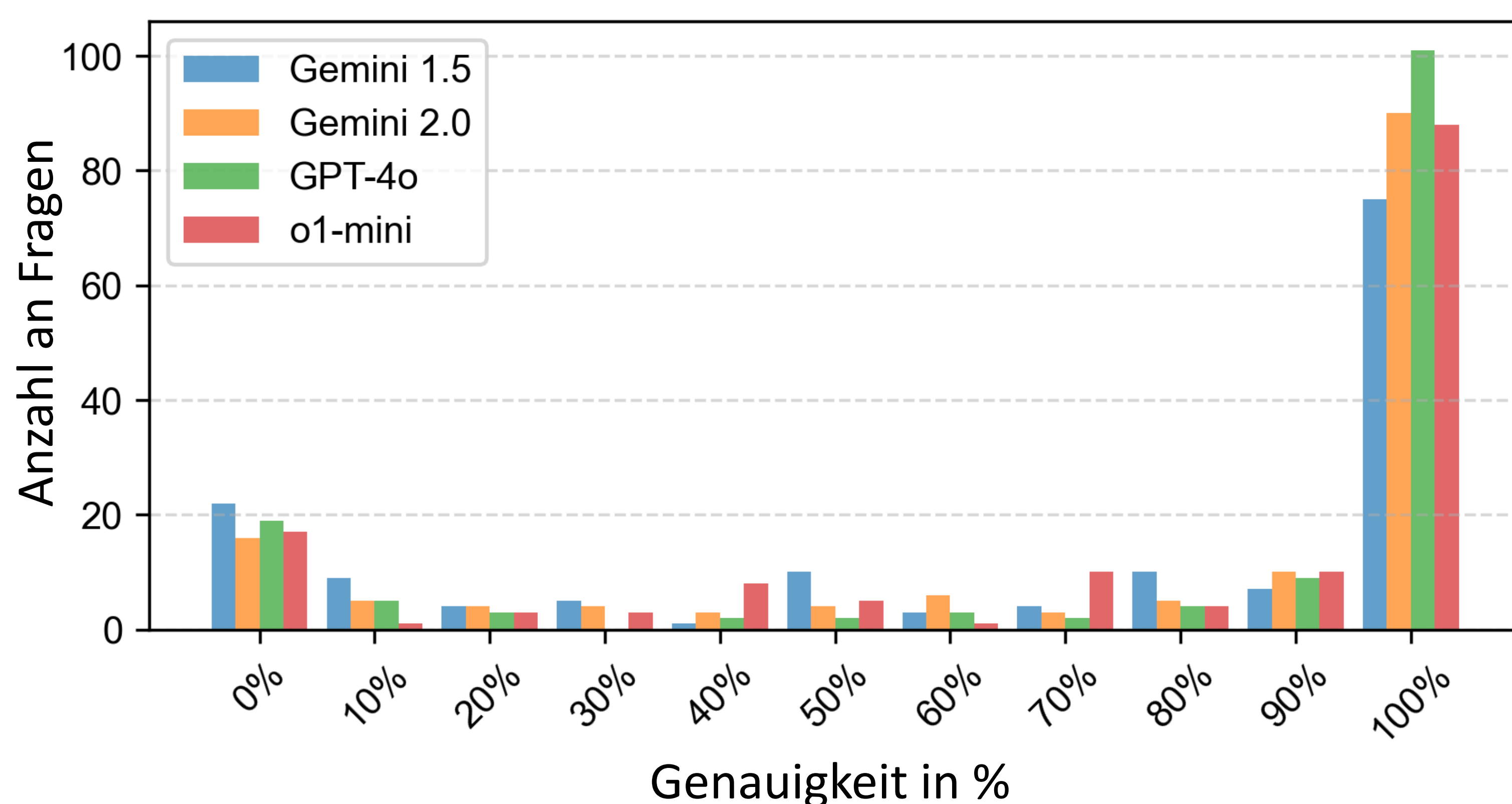


Abb. 1: Verteilung der Fragen nach erreichter Genauigkeit für vier große Sprachmodelle

- Gemini 2.0 Flash zeigte eine signifikant höhere Genauigkeit als Gemini 1.5 Flash ($p = .003$)
- Insgesamt hohe Gesamtgenauigkeit aller Modelle
- Keine signifikanten Unterschiede zwischen ChatGPT-4o, ChatGPT o1-mini, Gemini 2.0 Flash
- Genauigkeit der Modelle in %:
 - 78,9 % ChatGPT-4o
 - 76,7 % o1-mini
 - 76,3 % Gemini 2.0 Flash
 - 68,3 % Gemini 1.5 Flash

Referenz

[1] Costello C, Schumacher L. Child and adolescent psychiatry for the specialty board review. 5th ed. Taylor & Francis; 2023.

Konsistente Genauigkeit über verschiedene Themenbereiche

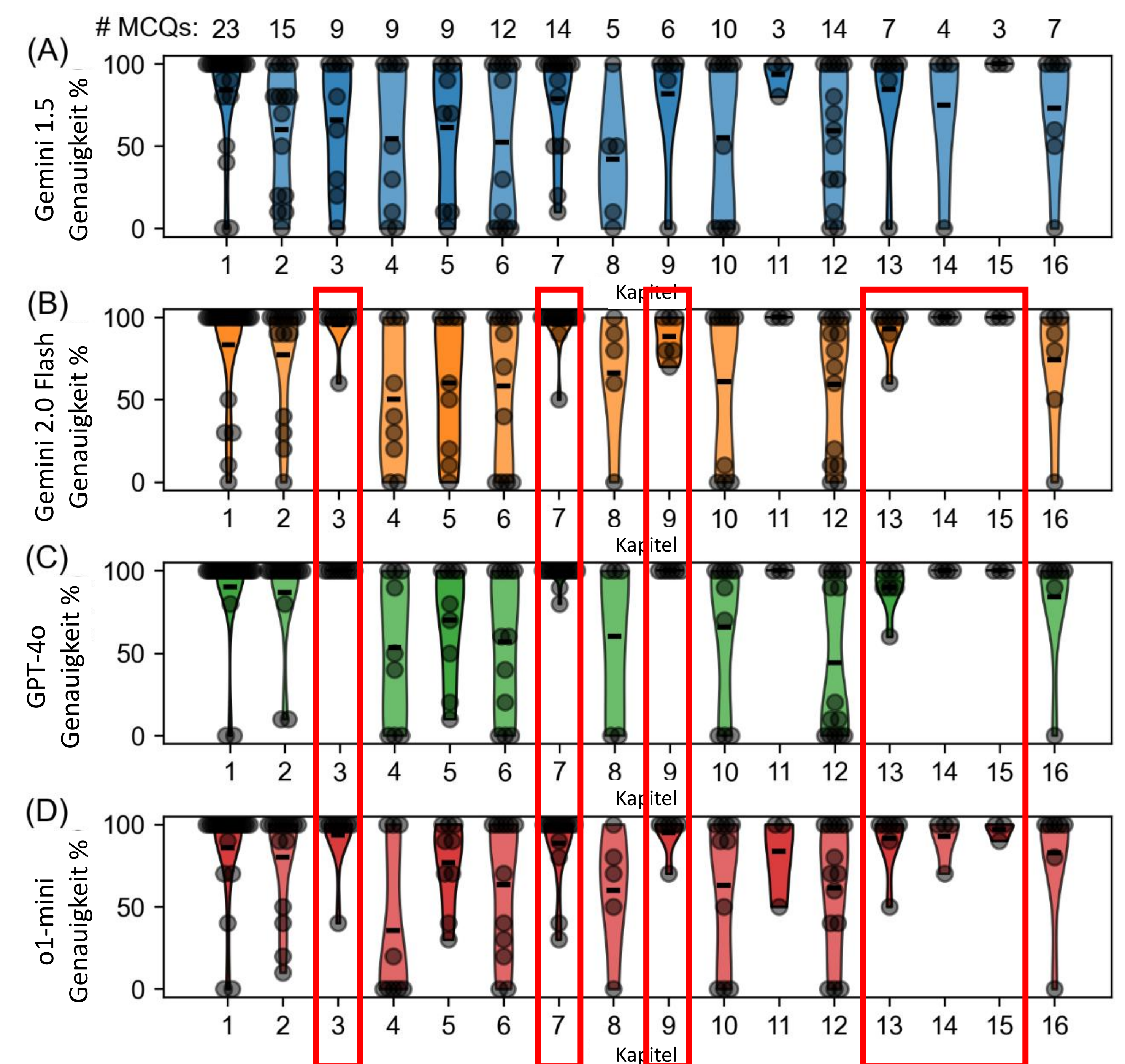


Abb. 2: Genauigkeit der getesteten Sprachmodelle nach Themenkapiteln im Buch [1]. x-Achse: Buchkapitel (z. B. Kapitel 3: psychotische Erkrankungen, Kapitel 7: Essstörungen, Kapitel 9: Abhängigkeitserkrankungen, Kapitel 13: Psychotherapie, Kapitel 14: Behandlungssettings, Kapitel 15: spezielle Themen wie Forensik und Public Health). y-Achse: Genauigkeit der Antworten. Jeder Punkt entspricht dem Anteil korrekt beantworteter Antworten pro Frage (10 Durchläufe). Rot markiert sind Kapitel mit vergleichbar hoher Genauigkeit. #MCQs: Anzahl der Fragen pro Kapitel.

- Konsistent hohe Genauigkeit der Modelle Gemini 2.0 Flash, ChatGPT-4o, ChatGPT o1-mini in unterschiedlichen Themenkapiteln
- 12 Fragen mit sehr niedriger Genauigkeit (<10 %) in allen vier getesteten Sprachmodellen
- Hinweise auf systematische Wissenslücken in komplexen Themenbereichen

Diskussion

- Große Sprachmodelle erreichen Leistungsniveau von Facharzt:innen in Wissenserhebung durch Multiple-Choice-Fragen
- Gute Performance bei klar definierten Störungsbildern
- Schwächen bei komplexen Themen (z. B. Psychopharmakologie)
- Schwächen der Untersuchung:
 - Nur Multiple-Choice-Format (keine klinischen Entscheidungen)
 - Keine realen Patientenszenarien
- möglicher Einsatz großer Sprachmodelle in Lehre und Wissensvermittlung erscheint denkbar
- Einsatz in der klinischen Praxis nur mit Vorsicht
- Weitere Forschung zu klinischer Anwendbarkeit notwendig

Interessenkonflikte

Alexander Häge hat Vergütungen für Tätigkeiten als Berater oder Referent für Shire/Takeda und Medice erhalten, die nicht im Zusammenhang mit dieser Arbeit stehen.

Kontakt Johanna.Neubauer@zi-mannheim.de