Benchmarking von Large Language Modellen in der Pflege: Ein Proof of Concept zur Entwicklung pflegespezifischer Evaluationsmethoden in Deutschland

Inhaltsverzeichnis

Einleitung

Die Leistungsbewertung von Large Language Models (LLM) ist in vielen Disziplinen zu einem zentralen Forschungsfeld geworden. In der Medizin existiert bereits eine Vielzahl spezialisierter Benchmarks, die unterschiedliche Aufgabentypen abdecken, von klassischen Prüfungsfragen bis hin zu kliniknahen Szenarien. Diese Entwicklungen zeigen, wie entscheidend standardisierte Testverfahren für die Bewertung von LLM in sicherheitskritischen Bereichen sind.

Für die Pflegewissenschaft hingegen existieren bislang keine etablierten Benchmarks. Insbesondere im deutschsprachigen Raum fehlen sowohl Datensätze als auch offene Plattformen, die eine systematische und reproduzierbare Evaluierung ermöglichen würden. Zwar liegen erste internationale Ansätze wie MedNurse-QA (Dicheva et al., 2025) und NurValues (Yao et al., 2025) vor, diese stammen jedoch aus dem US-amerikanischen und asiatischen Kontext und sind nicht auf die spezifischen Anforderungen der deutschen Pflege übertragbar.

Damit zeigt sich eine grundlegende Forschungslücke: Ohne sprach- und kontextadäquate Benchmarks bleibt unklar, wie zuverlässig LLM pflegerisches Wissen, Entscheidungen und Werteorientierungen abbilden können. Der hier vorgestellte Proof of Concept (PoC) adressiert diese Lücke, indem er eine erste methodische Grundlage für ein pflegespezifisches Benchmarking-Framework in Deutschland schafft.

Hintergrund

Die Entwicklung und Bewertung von Large Language Models (LLM) erfordert spezifische Verfahren, um Leistungsfähigkeit, Grenzen und Risiken systematisch zu erfassen. In der Informatik und den Sprachwissenschaften haben sich dafür in den letzten Jahren zahlreiche Benchmarks etabliert. Beispiele wie MMLU (Massive Multitask Language Understanding) (Hendrycks et al., 2020) oder BIG-bench (Srivastava et al., 2023) prüfen Modelle über eine Vielzahl von Wissensdomänen hinweg und erlauben standardisierte Leistungsvergleiche. Solche Benchmarks sind methodisch bedeutsam, da sie Aufgaben reproduzierbar operationalisieren und so eine Vergleichbarkeit von Modellen ermöglichen.

In der Medizin wurde dieses Prinzip konsequent aufgegriffen und an domänenspezifische Anforderungen angepasst. Klassische Datensätze wie MedQA (Jin et al., 2020), PubMedQA (Jin et al., 2019) oder MedMCQA (Pal et al., 2022) bilden Fachwissen in Prüfungsformaten ab. Mit MultiMedQA (Singhal et al., 2023) entstand eine integrierte Suite, die verschiedene Aufgaben zusammenführt und unter anderem für die Evaluation von Med-PaLM genutzt wurde. Neuere Ansätze wie ClinicalBench (Chen et al., 2024) oder EHRNoteQA (Kweon et al., 2024) gehen über Faktenwissen hinaus und orientieren sich stärker an klinischen Entscheidungssituationen sowie patientenbezogenen Daten. Damit entsteht in der Medizin ein umfassendes Ökosystem, das Wissensbestände, klinische Prozesse und zunehmend auch ethische Aspekte berücksichtigt.

Für die Pflegewissenschaft existieren bislang nur erste internationale Pilotansätze. Mit MedNurse-QA (Dicheva et al., 2025) wurde ein großer, frei zugänglicher Datensatz mit mehr als 21.000 Frage-Antwort-Paaren aus US-amerikanischen Lehrbüchern geschaffen. NurValues (Yao et al., 2025) wiederum rückt pflegeethische Dimensionen wie Altruismus, Würde, Integrität, Gerechtigkeit und Professionalität in den Mittelpunkt. Beide Initiativen markieren wichtige Fortschritte, beruhen jedoch auf englisch- oder chinesischsprachigen Daten.

Damit bleibt für die deutsche Pflegewissenschaft eine zentrale Lücke bestehen: Es gibt weder sprachlich angepasste Datensätze noch domänenspezifische Benchmark-Frameworks, die sich an den Ausbildungsinhalten, Leitlinien und pflegerischen Versorgungspraxen hierzulande orientieren. Die Schaffung solcher Ressourcen ist eine Grundvoraussetzung, um die Leistungsfähigkeit von LLM für die Pflege valide zu bewerten und ihren Einsatz in Forschung, Lehre und Praxis verantwortungsvoll zu gestalten.

Methodik

Ziel des Proof of Concept (PoC) war es, einen prototypischen Ansatz für ein pflegespezifisches Benchmarking-Framework zu entwickeln. Dabei sollte geprüft werden, wie sich etablierte Evaluationsmethoden aus der Informatik und der Medizin auf die Pflegewissenschaft übertragen lassen.

Aufgabenformate

Für die erste Umsetzung wurden zwei Aufgabenformate gewählt, die unterschiedliche Dimensionen pflegerischen Wissens und Handelns abbilden:

1. Multiple-Choice-Fragen (MC)
    - Aufbau analog zu klassischen Prüfungsformaten (z. B. NCLEX oder MedQA).
    - Abbildung von faktischem Wissen zu pflegerischen Grundlagen, Interventionen und klinischen Situationen.
    - Auswertung mittels Accuracy (Trefferquote).
2. Natural Language Inference (NLI)
    - Prüfung von Schlussfolgerungen aus kurzen Fallbeschreibungen.
    - Modell muss entscheiden, ob eine gegebene Hypothese aus einer pflegerischen Prämisse folgt, ihr widerspricht oder neutral ist.
    - Auswertung über Accuracy und Macro-F1-Score, um auch bei unbalancierten Klassen faire Ergebnisse zu erzielen.

Datenstruktur und Annotation

Die Aufgaben werden in einem standardisierten JSONL-Format hinterlegt, sodass jedes Item eine Zeile darstellt. Dieses Format ist für beide Aufgabentypen konsistent und enthält Felder für ID, Sprache, Schwierigkeitsgrad, Quelle, Tags und – falls vorhanden – eine kurze Begründung (Rationale).

Beispiel Multiple Choice (MC)


```
{
  "id":"MC-0001",
  "stem":"Eine Patientin mit Diabetes zeigt Anzeichen einer Hypoglykämie
(Schwitzen, Zittern, Hunger). Was ist die korrekte Erstmaßnahme?",
  "options":[
    "Sofort Insulin geben",
    "15–20 g schnell wirksame Kohlenhydrate oral geben, Blutzucker
kontrollieren",
    "Flüssigkeitsrestriktion",
    "Insulinpumpe ausschalten ohne Messung"
  ],
  "answer_idx":1,
  "rationale":"Hypoglykämie: orale Glukosegabe und Kontrolle; Insulingabe
wäre kontraindiziert.",
  "tags":["diabetes","akutmanagement","patientensicherheit"],
  "difficulty":"leicht",
  "source":"synthetic_poc_v0_0",
  "lang":"de"
}
```


Beispiel Natural Language Inference (NLI)


```
{
  "id":"NLI-0101",
  "premise":"Pflegebericht: Patient, 82 J., bettlägerig, feuchte Haut,
starkes Nachschwitzen, Inkontinenz, Mobilisation nur mit 2 Personen, Braden
13.",
  "hypothesis":"Es besteht ein erhöhtes Dekubitusrisiko.",
  "label":"entails",
  "rationale":"Braden 13 und zusätzliche Faktoren sprechen für erhöhtes
Risiko.",
  "tags":["pflegedoku","risiko","dekubitus"],
  "difficulty":"mittel",
  "source":"synthetic_poc_v0_0",
  "lang":"de"
}
```


Auswertungsansatz

Um die Vergleichbarkeit verschiedener Modelle sicherzustellen, wurde ein gepaartes Design implementiert. Alle Modelle bearbeiten denselben Item-Pool, sodass Unterschiede in den Ergebnissen statistisch überprüfbar sind. Aufgrund der geringen Datensätze dient die statistische Auswertung nur zur Demonstration, da ein signifikantes Ergebnis bei n=10 (MC) und n=5 (NLI) nicht zu erwarten ist.

- Für dichotome Entscheidungen (z. B. richtig/falsch) wird der McNemar-Test eingesetzt, um Unterschiede in den Trefferquoten zwischen Modellen zu prüfen.
- Für größere Stichproben wäre eine Erweiterung auf Chi-Quadrat-Tests angemessen.
- Zur Validierung von Robustheit werden Stichprobenpläne vorgeschlagen: Beginn mit 100 Items pro Task, Erweiterung auf 300–500 Items für stabile Schätzungen, bis hin zu 1.000+ Items für präzise Evaluationen.

Technische Umsetzung

Der PoC wurde in Python umgesetzt und im Repository PflegeBench-PoC (Kolb, 2025) dokumentiert. Kernbestandteile sind:

- Dataset-Definitionen im JSONL-Format,
- Auswertungsskripte zur Berechnung von Accuracy, F1-Score und statistischen Tests,
- eine Pipeline für Modellabfragen, die lokale Modelle oder API-basierte Systeme (z. B. GPT-4) einbindet.

Modellauswahl

Für die Evaluation im Rahmen des PoC wurden bewusst nicht die jeweils leistungsstärksten „Flaggschiffmodelle“ der Anbieter eingesetzt. Stattdessen kamen unterschiedliche Modelltypen zum Einsatz, um auch mit einer kleinen Testmenge potenzielle Unterschiede sichtbar zu machen. Eingesetzt wurden: Claude-3.5-sonnet-latest (Anthropic) als mittelgroßes, generisches Modell, Mistral-medium-2505 als leistungsfähiges, aber ressourcenschonendes Open-Weight-Modell, Gemini-2.5-flash-lite als latenzoptimierte Leichtvariante, GPT-4.1-nano-2025-04-14 als kompaktes Modell aus der GPT-4.1-Familie sowie Grok-code-fast-1, ein auf Programmierung spezialisiertes Modell. Diese Auswahl diente nicht dem Anspruch einer vollständigen Marktübersicht, sondern verfolgte zwei Ziele: Erstens sollte geprüft werden, ob bereits bei kleineren Benchmark-Sets Unterschiede in den Leistungsprofilen erkennbar sind. Zweitens sollte durch die Einbeziehung eines fachfremden Modells (Grok-code-fast-1) gezeigt werden, dass pflegespezifische Aufgaben eine domänensensible Evaluation erfordern und nicht jedes LLM gleichermaßen geeignet ist.

Abgrenzung

Der vorliegende PoC ist als Machbarkeitsstudie zu verstehen. Er bildet noch kein vollwertiges Benchmarking-Framework, zeigt jedoch exemplarisch, wie pflegespezifische Aufgaben operationalisiert, strukturiert und ausgewertet werden können. Zentrale Limitation ist die geringe Größe und Abdeckung der Item-Pools. Diese sollen in zukünftigen Arbeiten durch die Entwicklung offener, deutschsprachiger Datensätze und die Beteiligung der Pflegecommunity erweitert werden.

Governance-Aspekte

Es wurden ausschließlich synthetische Items genutzt, ohne personenbezogene Daten oder geschützte Kataloge (z. B. NANDA, ICNP, ENP). Bias wurde durch exemplarische Variation von Szenarien (Alter, Geschlecht, Setting) reduziert, stereotype Zuschreibungen wurden vermieden.

Ergebnisse / Analyse

Die erste Testrunde des Proof of Concept umfasste 10 Multiple-Choice-Fragen (MC) sowie 5 NLI-Paare. Damit handelt es sich um eine sehr kleine Stichprobe, die in erster Linie die Machbarkeit demonstrieren sollte. Entsprechend sind die Konfidenzintervalle breit und die statistische Aussagekraft eingeschränkt. Die Darstellung dient nur zur Komplettierung des Konzepts.

Multiple Choice (MC)

Alle getesteten Modelle erreichten eine Trefferquote von 100 % (10/10). Dies ist weniger ein Hinweis auf eine vollständige Beherrschung pflegerischen Wissens als vielmehr ein Effekt der geringen Zahl und relativen Einfachheit der Testitems.

- Das 95 %-Konfidenzintervall (Wilson) für 10/10 liegt bei ca. [0,72; 1,00].
- Interpretation: Auch bei perfekter Trefferquote bleibt bei so wenigen Aufgaben ein erheblicher Unsicherheitsbereich. Mit einer größeren Stichprobe würde sich das Intervall deutlich verengen.

Natural Language Inference (NLI)

Hier zeigten sich Unterschiede zwischen den Modellen:

- claude-3-5-sonnet-latest und mistral-medium-2505 erzielten jeweils eine Trefferquote von 100 % und einen Macro-F1-Score von 1,00.
- gemini-2.5-flash-lite und gpt-4.1-nano erreichten 80 % Trefferquote (Macro-F1 0,82).
- grok-code-fast-1 blieb mit 60 % Trefferquote (Macro-F1 0,61) deutlich zurück.

Auch hier sind die 95 %-Konfidenzintervalle breit (z. B. [0,57; 1,00] bei 5/5 Treffern), was die Unsicherheit der Ergebnisse betont.

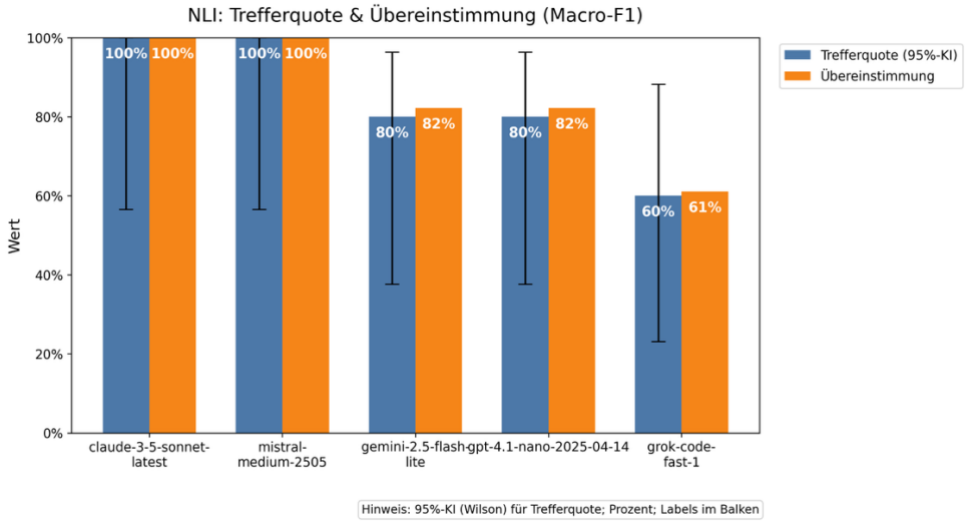

Abbildung 1 NLI: Trefferquote & Übereinstimmung (Macro-F1)

Gesamtwert (PflegeBench-Index)

Zur besseren Vergleichbarkeit wurde ein Gesamtwert gebildet, der den Mittelwert aus MC-Trefferquote, NLI-Trefferquote und NLI-Übereinstimmung darstellt. Hier ergaben sich folgende Werte:

- Claude-3.5-sonnet-latest: 1,00
- Mistral-medium-2505: 1,00
- Gemini-2.5-flash-lite: 0,874
- GPT-4.1-nano: 0,874
- Grok-code-fast-1: 0,737

Damit zeigen sich erste Tendenzen einer Modellhierarchie, die allerdings aufgrund der geringen Fallzahl nur als explorativ einzuordnen ist.

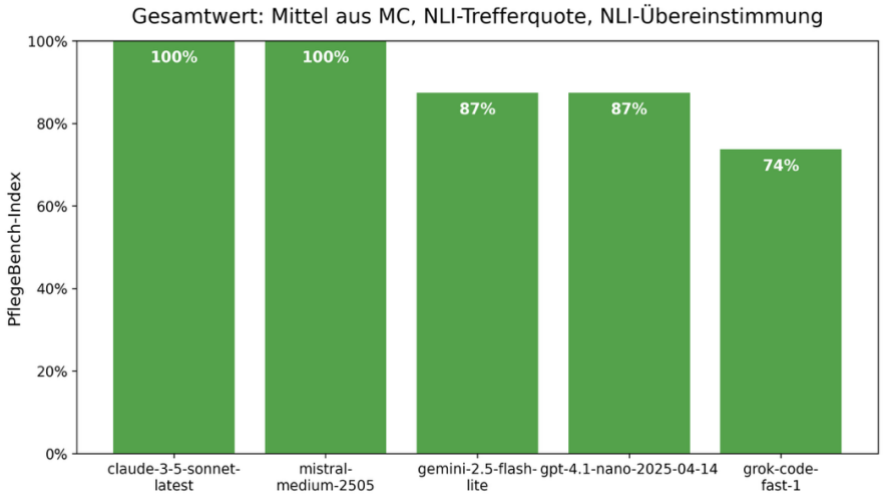

Abbildung 2 Gesamtwert Mittel aus MC, NLI-Trefferquote, NLI-Übereinstimmung

McNemar-Test

Mit dem McNemar-Test wurden paarweise Unterschiede auf identischen Fällen geprüft. Aufgrund der kleinen Fallzahl traten nur wenige Diskordanzen auf, sodass keine signifikanten Unterschiede festgestellt werden konnten. So zeigte sich etwa im Vergleich *mistral vs. grok* eine Tendenz zugunsten von mistral (b10=2), bei einem p-Wert von 0,5, was weit von einer Signifikanz entfernt ist.

Interpretation und Limitationen

- Die Ergebnisse belegen, dass sich die gewählten Aufgabenformate (MC, NLI) grundsätzlich für pflegespezifisches Benchmarking eignen.
- Gleichzeitig verdeutlicht die Analyse die Grenzen kleiner Testmengen: trotz hoher oder perfekter Trefferquoten sind die Konfidenzintervalle breit und Unterschiede zwischen Modellen statistisch unsicher.
- Für robuste Aussagen ist eine deutlich größere Itemzahl erforderlich (empfohlen: 300–500 Items pro Task, perspektivisch >1.000).

Diskussion

Die vorliegenden Ergebnisse sind explorativ und dienen primär dem Demonstrationszweck: Der PoC zeigt, *wie* pflegespezifische Aufgaben (MC, NLI) operationalisiert und ausgewertet werden können. Mit diesem PoC lassen sich keine belastbaren Modellrangfolgen darstellen. Das spiegelt sich in den breiten 95%-Konfidenzintervallen und den nicht signifikanten McNemar-Tests wider (bei MC keine Diskordanzen; bei NLI nur vereinzelte, p≈0,5). Ursache ist die sehr kleine Stichprobe (10 MC, 5 NLI). Für robuste Schlüsse bedarf es deutlich größerer Itempools (300–500+, perspektivisch >1.000) und differenzierter Teilanalysen nach Tags/Themen, wie im PoC vorgesehen.

Ein niedrigschwelliger Einstieg ins Benchmarking ist die Evaluation auf Examensfragen. Eine aktuelle Studie in *Nursing Education Perspectives* testete NCLEX-RN-Praxisfragen mit Gemini, GPT-3.5 und GPT-4 und berichtet die jeweilige Genauigkeit (García-Rudolph et al., 2025), methodisch simpel, reproduzierbar und gut kommunizierbar. Solche Setups zeigen schnell Leistungsunterschiede, bleiben jedoch auf Faktenwissen im MC-Format begrenzt und erfassen weder pflegeprozessuale Entscheidungen noch wertebezogenes Handeln. Damit unterstreichen sie den Bedarf an pflegespezifischen, kontextualisierten Benchmarks, wie sie der PoC skizziert (MC + NLI als Startpunkt, später Falltexte, Langantworten, Dokumentations-NLI).

Im medizinischen Umfeld wächst parallel die Infrastruktur für offene Vergleiche. Das Medical LLM Leaderboard auf Hugging Face aggregiert Ergebnisse über diverse Datensätze und Modelle, erleichtert Transparenz und Reproduzierbarkeit (Liu, 2024). Genau diese öffentliche Plattform-Logik fehlt der deutschen Pflegewissenschaft bislang vollständig: Es gibt weder offene, deutschsprachige Pflegedatensätze noch ein Fach-Leaderboard, das Modelle mit Blick auf deutsche Curricula, Leitlinien und Dokumentationsstandards verlässlich vergleicht. Der PoC adressiert diese Lücke und kann als Keimzelle für ein „PflegeBench-DE“ auf Hugging Face dienen.

Für die deutsche Pflege reicht es nicht, internationale Datensätze 1:1 zu übernehmen. Aussagen werden erst praxisrelevant, wenn Sprache, Ausbildungsinhalte (z. B. Pflegeberufegesetz-konforme Kompetenzen), Leitlinien (z. B. DNQP-Expertenstandards) und Systemlogiken (z. B. Pflegegrade nach SGB XI) abgebildet sind. Zudem variieren LLM-Fähigkeiten sprachabhängig; Unterschiede in Tokenisierung und Pretraining-Daten können Ergebnisse verzerren, unabhängig von der inhaltlichen Güte der Items. Das spricht für explizite de-Sets mit deutschen Items, Glossaren und, wo sinnvoll, de-spezifischen Tokenizer-/Vokab-Konfigurationen, um Fairness und Aussagekraft zu erhöhen. (Der PoC legt dafür mit klaren JSONL-Schemata, Tagging und NLI-Labels die methodische Grundlage.)

Der AI Index 2025 dokumentiert den anhaltenden Schub bei Evaluationspraktiken und die Ausweitung domänenspezifischer Benchmarks (Maslej et al., 2025). Er macht zugleich deutlich, dass die Leistungsfähigkeit moderner Modelle nicht automatisch Zuverlässigkeit im Fachkontext bedeutet. Die systematische, aufgabenspezifische Prüfung bleibt essenziell. Für Deutschland folgt daraus: Wer sichere und passfähige KI-Unterstützung in der Pflege will, braucht eigene Benchmarks, Datenpipelines und Veröffentlichungsorte (z. B. HF-Spaces) mit transparenten Metriken und Protokollen.

Neben Wissens- und Entscheidungsaufgaben sollten werte-/verhaltensorientierte Dimensionen (z. B. Respekt, Würde, Gerechtigkeit) in die Evaluation einfließen. Internationale Vorarbeiten (z. B. NurValues) zeigen die Machbarkeit, stammen aber nicht aus dem deutschen/europäischen Kontext. Eine deutsche Erweiterung, mit Bezug zu Pflegeethik, Berufskodizes und hiesigen Versorgungssituationen, ist erforderlich, um Modellantworten auch professionell-normativ zu bewerten. (Der PoC ist dafür kompatibel, etwa durch Erweiterung der Tags/Taxonomien und Begründungsfelder.)

Die PoC-Ergebnisse sind nicht signifikant und dürfen nicht überinterpretiert werden. Sie illustrieren ein Verfahren (Accuracy, Macro-F1, Wilson-KI, McNemar), kein finales Ranking. Methodisch korrekt wurde ausschließlich mit synthetischen, urheberrechtskonformen Items gearbeitet; personenbezogene Daten oder geschützte Kataloge wurden nicht verwendet; Bias-Risiken wurden über Szenariovielfalt reduziert. Für die nächste Ausbaustufe wären Goldstandard-Kuration, Inter-Rater-Reliabilität, größere Stichproben und eine offene Publikations-/Leaderboard-Infrastruktur wünschenswert.

Schlussfolgerung für die deutsche Pflegewissenschaft

1. Aufbau eines offenen, deutschsprachigen Pflege-Leaderboards (HF-Space), das MC, NLI und später Fall-/Langantwort-Aufgaben integriert. Orientierung an medizinischen Leaderboards, aber mit deutscher Fachlogik.
2. Skalierung der Itempools gemäß PoC-Roadmap (≥300–500/Task), inkl. Tag-Analysen (Setting, Thema, Schwierigkeitsgrad) und gepaarten Tests.
3. Sprachspezifische Qualität: deutschsprachige Items, Glossare, ggf. de-optimierte Tokenizer/Prompts; systematische Tests auf Sprach-Bias.
4. Ethik einbeziehen: werteorientierte Subbenchmarks nach deutschem Kontext (Erweiterung von NurValues-Konzepten).
5. Transparenz & Reproduzierbarkeit: Protokolle, Seeds/Temperatur, Promptversionen, Auswertungsskripte und Grafiken offen dokumentieren (wie im PoC).

Kurz: Der PoC zeigt, wie es geht. Nun braucht es deutsche Daten, Maßstäbe und Infrastruktur, damit Aussagen über KI in der Pflege relevant, fair und belastbar werden. Ergänzende Evidenz aus einfachen MC-Benchmarks und die Praxis etablierter medizinischer Leaderboards stützen diese Richtung, können sie aber nicht ersetzen.

„PflegeBench-PoC“ auf GitHub: https://github.com/goldikolb/PflegeBench-PoC

Literaturverzeichnis

Chen, C., Yu, J., Chen, S., Liu, C., Wan, Z., Bitterman, D.,...Shu, K. (2024). *ClinicalBench: Can LLMs Beat Traditional ML Models in Clinical Prediction?* https://arxiv.org/abs/2411.06469

Dicheva, N. K., Rehman, I. U., Husamaldin, L., & Aleshaiker, S. (2025). *MedNurse-QA (v1.0)*. https://huggingface.co/datasets/NevenaD/MedNurse-QA

García-Rudolph, A., Sanchez-Pinsach, D., Fernandez, M. C., Cunyat, S., Opisso, E., & Hernandez-Pena, E. (2025). How Chatbots Respond to NCLEX-RN Practice Questions: Assessment of Google Gemini, GPT-3.5, and GPT-4. *Nursing Education Perspectives*, *46*(2), E18-E20. https://doi.org/10.1097/01.NEP.0000000000001364

Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., & Steinhardt, J. (2020). Measuring Massive Multitask Language Understanding. *arXiv preprint*.

Jin, D., Pan, E., Oufattole, N., Weng, W.-H., Fang, H., & Szolovits, P. (2020). *What Disease does this Patient Have? A Large-scale Open Domain Question Answering Dataset from Medical Exams*. https://arxiv.org/abs/2009.13081

Jin, Q., Dhingra, B., Liu, Z., Cohen, W. W., & Lu, X. (2019). *PubMedQA: A Dataset for Biomedical Research Question Answering*. https://arxiv.org/abs/1909.06146

Kolb, C. (2025). *PflegeBench-PoC*. https://github.com/goldikolb/PflegeBench-PoC

Kweon, S., Kim, J., Kwak, H., Cha, D., Yoon, H., Kim, K. H.,...Choi, E. (2024). *EHRNoteQA: An LLM Benchmark for Real-World Clinical Practice Using Discharge Summaries (version 1.0.1)*. https://doi.org/10.13026/acga-ht95

Liu, F. (2024). *Medical LLM Leaderboard*. https://huggingface.co/spaces/fenglinliu/medical_llm_leaderboard

Maslej, N., Fattorini, L., Perrault, R., Gil, Y., Parli, V., Kariuki, N.,...et al. (2025). *Artificial Intelligence Index Report 2025*. https://hai.stanford.edu/ai-index/2025-ai-index-report

Pal, A., Umapathi, L. K., & Sankarasubbu, M. (2022). *MedMCQA : A Large-scale Multi-Subject Multi-Choice Dataset for Medical domain Question Answering*. https://arxiv.org/abs/2203.14371

Singhal, K., Azizi, S., Tu, T., Mahdavi, S. S., Wei, J., Chung, H. W.,...Natarajan, V. (2023). Large language models encode clinical knowledge. *Nature*, *620*(7972), 172-180. https://doi.org/10.1038/s41586-023-06291-2

Srivastava, A., Rastogi, A., Rao, A., Shoeb, A. A. M., Abid, A., Fisch, A.,...Garriga-Alonso, A. (2023). *Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models* (2835-8856). https://openreview.net/forum?id=uyTL5Bvosj

Yao, B., Li, Q., Zhang, Y., Yang, S., Zhang, B., Tiwari, P., & Qin, J. (2025). *NurValues: Real-World Nursing Values Evaluation for Large Language Models in Clinical Context*. https://arxiv.org/abs/2505.08734