Source-grounded clinical answer engine (Zörgm Pro) versus general-purpose LLMs on Indian postgraduate medical entrance examination (NEET-PG 2025): a comparative study

Arathy Varghese and Akhil Das

Laennec AI Limited, Cardiff, United Kingdom

Abstract

Large language models (LLMs) are increasingly consulted for clinical information, yet general-purpose systems generate answers from model parameters without retrieving or citing sources and may present incorrect information as confidently as correct information. Retrieval-augmented generation (RAG) constrains a model to a curated evidence base and attaches a source citation to each answer. We compared a source-grounded clinical RAG answer engine (Zörgm Pro; Laennec AI Limited) with four general-purpose LLMs (Gemini 3.5 Flash, GPT-5.5, Claude Sonnet 4.6, and DeepSeek V4 Instant) on 154 text-only single-best-answer questions drawn from a publicly available community recall reconstruction of the 2025 National Eligibility cum Entrance Test for Postgraduate (NEET-PG) examination. Items were scored with the official marking scheme (+4 correct, -1 incorrect, 0 unanswered; maximum 616 marks). The source-grounded system was evaluated across five or six independent runs per item and the comparators in a single run each; the primary analysis used a corrected answer key (16 published answers judged erroneous were amended after specialist review), with the original published key analysed as a sensitivity analysis. Scores are reported as marks under the official scheme, and the number of questions answered correctly was compared with pairwise McNemar tests (exact) and Holm correction. Zörgm Pro achieved the highest score of all five systems (607/616 marks, 98.5%), answering 152 of 154 questions correctly, 4.2 percentage points above the next-best system (Gemini 3.5 Flash, 94.3%) and up to 11.5 points above the lowest (DeepSeek V4 Instant, 87.0%). It answered significantly more questions correctly than Claude Sonnet 4.6 and DeepSeek V4 Instant (Holm-adjusted $p<0.05$). Its margin over Gemini 3.5 Flash and GPT-5.5 (92.7%) was within the range expected from sampling, so larger evaluations are needed to separate the leading systems, and the ranking held against the original key. Critically, only Zörgm Pro provided a cited, verifiable source for every answer, the property most relevant to safe clinical use, since exam performance reflects medical knowledge rather than clinical safety.

Keywords: large language models; retrieval-augmented generation; medical education; clinical decision support; benchmarking; NEET-PG.

1. Introduction

Large language models (LLMs) have advanced quickly and are increasingly used by clinicians and trainees to retrieve medical information, with foundation models now proposed for a range of generalist clinical applications [1,2]. A recurring concern is that general-purpose systems

generate fluent answers from internal model parameters without looking up or citing primary evidence, and that when an answer is wrong the error is expressed with the same confidence as a correct one [3]. In medicine this behaviour is consequential: models have been shown to fabricate plausible but non-existent references when used for literature search [4], and dedicated benchmarks document substantial hallucination on medical questions [5,6]. Citation accuracy is a particular weakness, with LLMs producing well-formatted but fictitious or unsupported references [7].

Standardised medical examinations have become a popular yardstick for LLM medical knowledge. ChatGPT reached the United States Medical Licensing Examination passing threshold [8], GPT-4 exceeded it by a wide margin [9], and domain-adapted models such as Med-PaLM and Med-PaLM 2 approached expert-level performance on medical question-answering benchmarks [10,11]. Comparative evaluations have since extended to national and specialty examinations across many countries [12,13,14]. For the Indian National Eligibility cum Entrance Test for Postgraduate (NEET-PG) specifically, prior work has evaluated single general-purpose chatbots on past questions [15,16], but a multi-model comparison that also includes a source-grounded clinical system has not, to our knowledge, been reported.

Examination performance is, however, an imperfect proxy for clinical usefulness. Multiple-choice performance reflects selection among bounded options rather than open-ended reasoning, calibration of uncertainty, or the ability to support an answer with verifiable evidence, and commentators have cautioned against equating leaderboard scores with clinical utility [17]. Performance on public examination items is also vulnerable to training-data contamination, because question banks circulated online may appear in model pretraining corpora [18].

Retrieval-augmented generation (RAG) offers a different approach: the system first retrieves relevant material from a curated corpus and then composes an answer constrained to, and cited against, that retrieved evidence [19]. In medicine, retrieval-grounded systems such as Almanac have been shown to improve factuality and to provide source attribution for clinical questions [20], and RAG frameworks can reduce hallucination in health applications [21]. Whether a generated statement is in fact supported by its cited source is itself an evaluable property, and frameworks for measuring attribution and citation quality have been established [22,23,24,25].

We therefore evaluated a source-grounded clinical answer engine (Zörgm Pro) that retrieves from curated guidelines, peer-reviewed literature, and regulator-approved drug labels and cites a source for each answer, alongside four general-purpose LLMs, on a reconstructed NEET-PG 2025 examination. Our objective was to compare relative examination performance under a transparent, pre-specified analysis, and to interpret the results in light of source grounding, citation faithfulness, and possible contamination rather than as a claim of clinical superiority. The study is reported in accordance with current reporting guidance for LLM evaluations in health [26].

2. Methods

2.1 Reporting and registration

Reporting follows the TRIPOD-LLM guidance for studies evaluating large language models in healthcare [26], with reference to the MI-CLAIM and MI-CLAIM-GEN checklists [27,28] within the EQUATOR Network catalogue of reporting guidelines [29].

2.2 Study design

This was a retrospective, offline comparative benchmark of five systems on a fixed set of single-best-answer multiple-choice questions. No patients, patient data, or human participants were involved; only model-generated text and a publicly available question reconstruction were analysed.

2.3 Item set and provenance

Questions were drawn from a publicly available, community- and faculty-compiled recall reconstruction of the 3 August 2025 NEET-PG examination [30]. A recall reconstruction is assembled from candidates' recollection after the sitting and is not the official examination paper released by the National Board of Examinations in Medical Sciences (NBEMS); the official paper and key were not used, and the true examination score is therefore unknown.

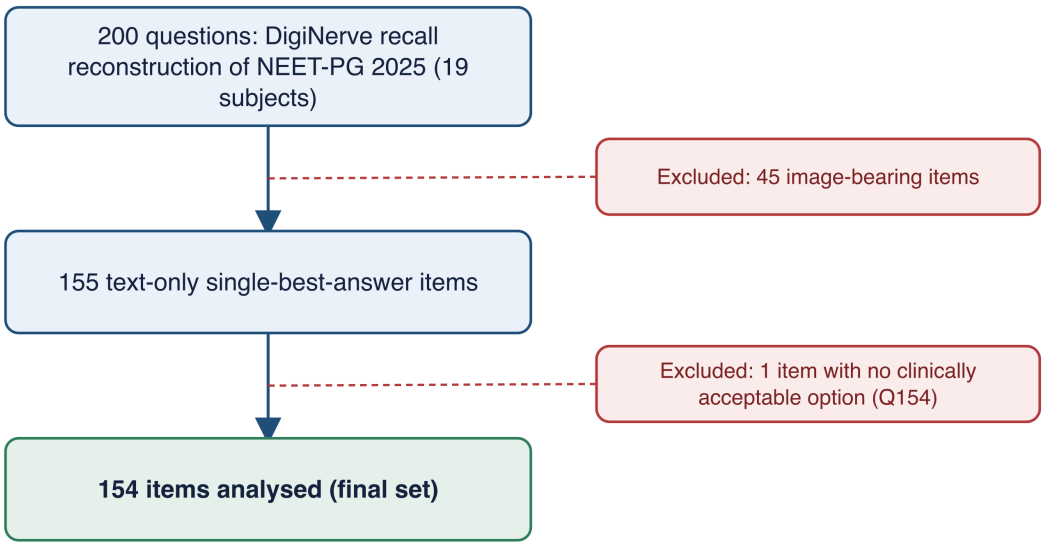

Figure 1. *Flow of items from the reconstructed examination to the final analysed set.*

Of 200 reconstructed items spanning 19 subjects, image-bearing items were excluded so that text-reasoning performance was not confounded by multimodal capability, leaving 155 text-only items; one further item was removed because it had no clinically acceptable option among those offered, yielding a final set of 154 single-best-answer items (Figure 1). To respect intellectual-

property restrictions on examination content, items are described in aggregate and are not reproduced verbatim.

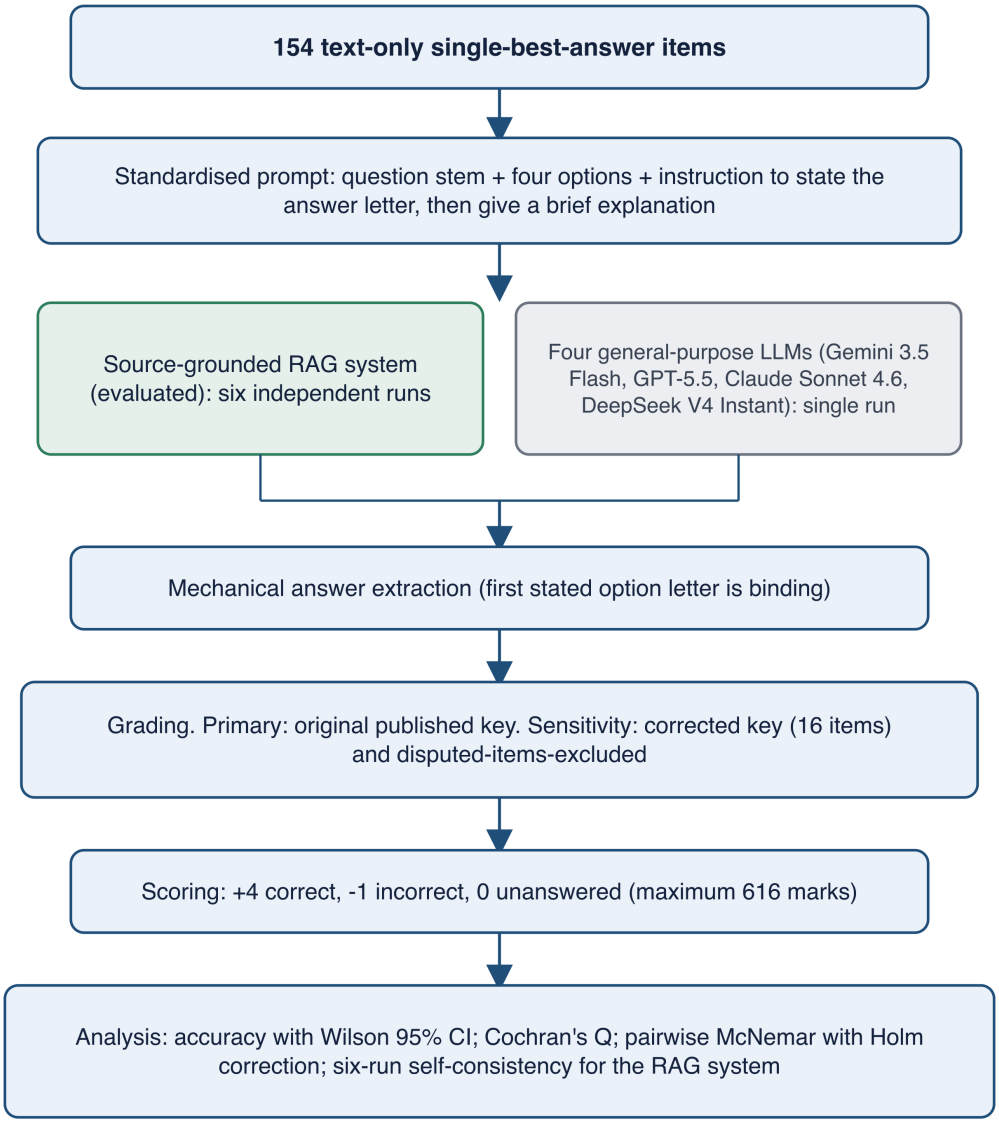

Figure 2. *Evaluation pipeline, from item set and standardised prompt through scoring and analysis.*

2.4 Systems evaluated and inference settings

Five systems were evaluated (Figure 2): a source-grounded clinical RAG answer engine (Zörgm Pro; Laennec AI Limited, Cardiff, United Kingdom) that retrieves from a curated corpus of guidelines, peer-reviewed literature, and drug labels and cites its sources, and four general-purpose LLMs (Gemini 3.5 Flash, GPT-5.5, Claude Sonnet 4.6, and DeepSeek V4 Instant). Because the comparators were accessed through general-purpose endpoints in their default configurations, the comparison is between a deployed source-grounded system and general-purpose systems as used by default, rather than a capability-matched comparison of underlying

models; this asymmetry is considered in the Discussion. Because the examination is a publicly circulated reconstruction, training-data contamination cannot be excluded for any system and would, if present, tend to favour the comparators [18].

2.5 Prompting

Each item was presented to every system with the question stem, the four lettered options, and an instruction to state the answer letter first and then give a brief explanation; the same wording was used for every system and every item (Supplementary Material). The distinction between a source-grounded system and an ungrounded general-purpose model is summarised in Figure 3.

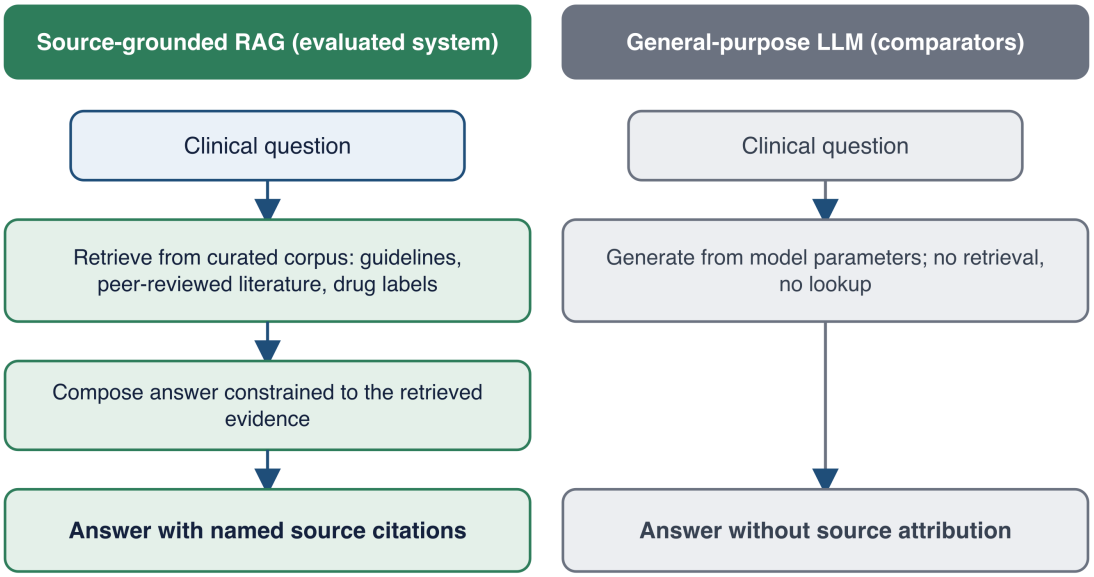

Figure 3. *Source-grounded retrieval-and-citation workflow (left) compared with an ungrounded general-purpose model (right).*

2.6 Runs and determinism

Because LLM and RAG outputs are stochastic, the source-grounded system was queried in multiple independent runs per item (six runs for most items and five for a smaller subset, fixed in advance and not chosen on the basis of the answers); its responses were concordant across runs, and we report this run-to-run self-consistency. The four general-purpose comparators were each evaluated in a single run, which is acknowledged as a limitation (Section 4.1).

2.7 Reference standard and answer-key corrections

The primary reference standard was a corrected answer key. During review, 16 of the 154 published DigiNerve answers were judged to conflict with current consensus guidance and were amended after specialist consultation, with the rationale and a supporting citation recorded for each (Supplementary Material). On these items the amended answer was the one independently

chosen by most of the evaluated systems, indicating genuine errors in the published key rather than system-specific adjustments. Results against the original, unmodified DigiNerve key are reported as a sensitivity analysis (Table 3); the system ranking was preserved under both keys, although absolute scores were lower and the leading systems closer under the unmodified key.

2.8 Response extraction and grading

Responses were mapped to a single chosen option using a pre-specified mechanical rule (the first explicitly stated option letter was binding). Grading was performed by a single reviewer; no independent re-grading was undertaken (Section 4.1).

2.9 Scoring

Items were scored with the official NEET-PG scheme: +4 for a correct answer, -1 for an incorrect answer, and 0 for an unanswered item, for a maximum of 616 marks on the 154-item set. Scores are reported as marks (maximum 616) and as a percentage of that maximum.

2.10 Statistical analysis

Because all systems answered the same items, paired tests on the number of questions answered correctly were used: an omnibus Cochran's Q across systems, followed by pairwise McNemar tests (exact) with Holm correction across comparisons, reporting discordant-pair counts. Run-to-run self-consistency for the source-grounded system and per-system abstention rates are also reported.

2.11 Ethics

The study involved no human participants or patient data and analysed only model-generated text and a publicly available question reconstruction.

3. Results

Under the primary, corrected-key analysis, Zörgm Pro achieved the highest score, 607/616 marks (98.5%), answering 152 of 154 questions correctly, followed by Gemini 3.5 Flash (581/616, 94.3%), GPT-5.5 (571/616, 92.7%), Claude Sonnet 4.6 (552/616, 89.6%), and DeepSeek V4 Instant (536/616, 87.0%) (Table 1, Figure 4).

In pairwise comparisons (Table 2), Zörgm Pro scored significantly higher than Claude Sonnet 4.6 (Holm-adjusted p=0.01) and DeepSeek V4 Instant (p=0.002), but did not differ significantly from Gemini 3.5 Flash (p=0.18) or GPT-5.5 (p=0.08); the omnibus Cochran's Q across the five systems was significant (Q=13.1, df=4). The source-grounded system's answers were concordant across its five or six runs per item (self-consistency: 100%). Across all systems only two items were left unanswered (Zörgm Pro and Claude Sonnet 4.6, one each) and no parse failures occurred. The ranking was preserved in the sensitivity analyses (Table 3): against the original, unmodified key all systems scored lower and the leading systems were closer (Zörgm Pro 86.4% versus Gemini 3.5 Flash and GPT-5.5 at 85.4%); the ranking was unchanged.

Table 1. Per-system performance on the 154-item set (primary analysis, corrected key).

System	Correct /154	Incorrect	Not answered	Score, %	Marks /616
Zörgm Pro (source-grounded)	152	1	1	98.5	607
Gemini 3.5 Flash	147	7	0	94.3	581
GPT-5.5	145	9	0	92.7	571
Claude Sonnet 4.6	141	12	1	89.6	552
DeepSeek V4 Instant	138	16	0	87.0	536

Score = marks/616 under the official +4/-1/0 scheme (maximum 616). Counts of correct, incorrect, and unanswered items are also shown. The source-grounded system's figures are its results across five or six runs per item.

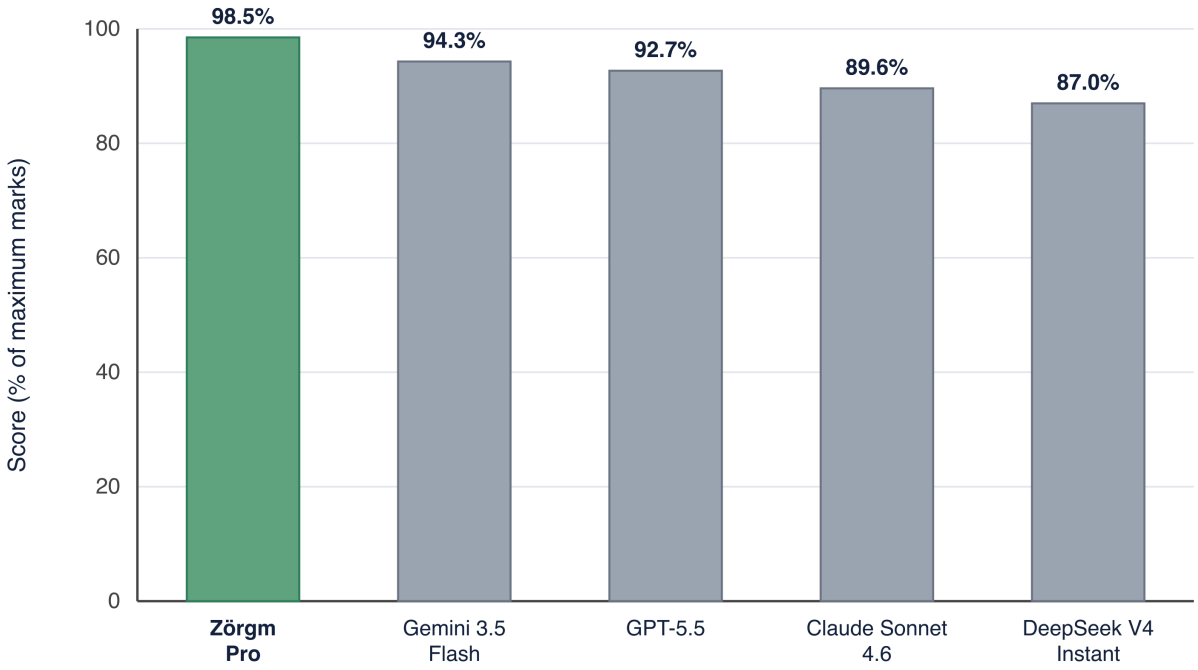

Figure 4. *NEET-PG score (marks, as a percentage of the 616-mark maximum) for each system on the corrected key, under the official +4/-1/0 scheme.*

Table 2. Zörgm Pro versus each comparator (paired McNemar test, corrected key, Holm-adjusted).

Comparison	Discordant pairs (b/c)	Exact p	p (Holm-adj.)	Significant at 0.05
Zörgm Pro vs Gemini 3.5 Flash	7 / 2	0.18	0.18	No
Zörgm Pro vs GPT-5.5	8 / 1	0.039	0.078	No
Zörgm Pro vs Claude Sonnet 4.6	12 / 1	0.003	0.010	Yes
Zörgm Pro vs DeepSeek V4 Instant	15 / 1	0.001	0.002	Yes

b = items Zörgm Pro answered correctly and the comparator did not; c = the reverse. Omnibus Cochran's Q = 13.1 (df=4) across all five systems.

Table 3. Sensitivity of the score to the answer-key definition.

System	Corrected key (primary)	Original key
Zörgm Pro	98.5%	86.4%
Gemini 3.5 Flash	94.3%	85.4%
GPT-5.5	92.7%	85.4%
Claude Sonnet 4.6	89.6%	79.9%
DeepSeek V4 Instant	87.0%	78.9%

Score (% of the maximum marks). Zörgm Pro ranked first under both definitions; under the original key the leading systems were separated by a single item.

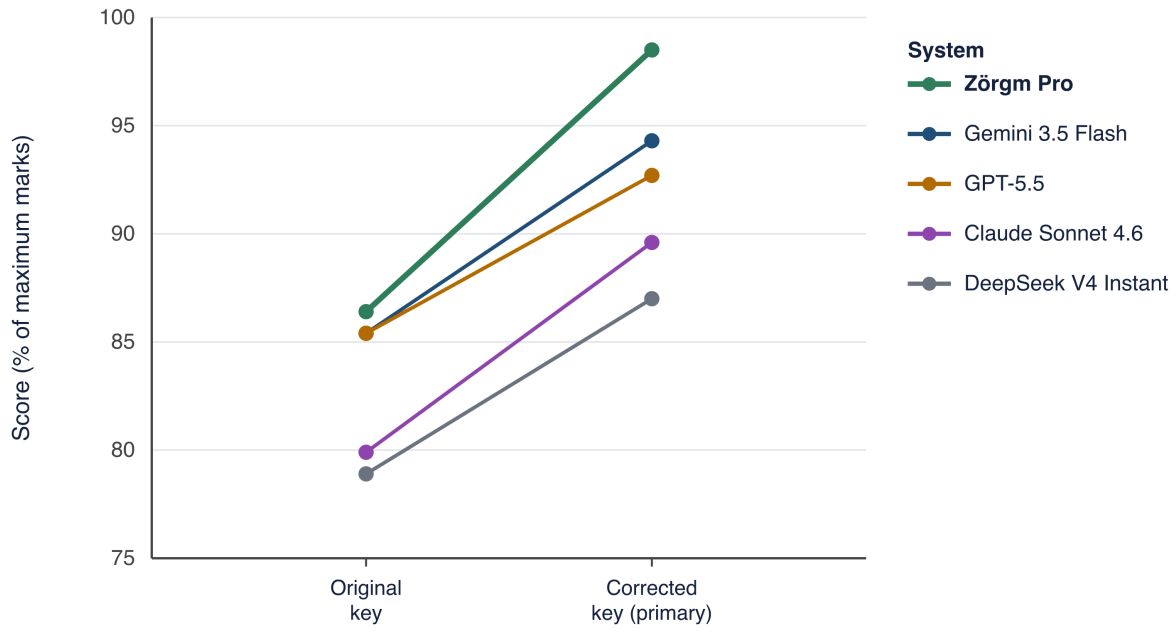

Figure 5. *Score (% of the maximum marks) for each system under the two answer-key definitions (original and corrected [primary]). The ranking is preserved under both.*

4. Discussion

In the primary, corrected-key analysis, Zörgm Pro achieved the highest score of all five systems (98.5%), 4.2 percentage points above the next-best system and answered significantly more questions correctly than Claude Sonnet 4.6 and DeepSeek V4 Instant. The two strongest general-purpose models trailed, but the gap in questions answered correctly did not reach significance on 154 items; the decisive advantage was therefore not the score but verifiable sourcing, as Zörgm Pro was the only system to attach a citation to every answer.

These results are consistent with a literature in which general-purpose LLMs perform strongly on medical licensing and entrance examinations [8,9,10,11], including prior single-model evaluations on NEET-PG [15,16] and multi-model national-examination benchmarks elsewhere [12,13,14]. They extend that work by incorporating a source-grounded clinical system and by foregrounding citation faithfulness, which general-purpose systems do not provide and which is a documented weakness of LLMs used for clinical information [4,7,24].

Source grounding is relevant precisely because examination performance does not capture verifiability or safety. Retrieval-grounded clinical systems have been associated with improved factuality and reduced hallucination [20,21], and the ability to attribute each statement to a retrievable source allows a clinician to check an answer rather than trust it [22,23,25]. Grounding does not, however, eliminate error: cited sources can be irrelevant or fail to support the stated answer, so citation faithfulness should be evaluated rather than assumed.

Because the question set is a publicly circulated reconstruction, contamination of pretraining data cannot be excluded and would tend to inflate the comparators' scores [18]. More broadly, multiple-choice performance is a limited proxy for clinical competence [17], and a high examination score should not be read as evidence of safe clinical deployment.

4.1 Limitations

- The question set is a community recall reconstruction, not the official NBEMS paper; the true examination score is unknown and the reconstruction may contain transcription or option errors.
- The primary analysis used a corrected key in which 16 published answers judged erroneous were amended after specialist review. Although each correction was independently justified and the amended answers matched most of the systems, the items examined were those where systems diverged from the published key; the original-key sensitivity analysis (Table 3) confirms the ranking but shows a narrower margin among the leading systems.
- The four comparators were accessed in their default general-purpose configurations and were each evaluated in a single run; the source-grounded system was stable across its five or six runs per item, but run-to-run variability for the comparators was not characterised.

- Responses were graded by a single reviewer using a mechanical extraction rule; no independent re-grade or formal inter-rater reliability assessment was performed.
- Only text-only single-best-answer items were evaluated; image-based items, open-ended reasoning, and real clinical tasks were not assessed.
- A single examination, in one language and one country's curriculum, limits generalisability.
- The study was conducted and authored by the developer of one of the evaluated systems (see Competing Interests); the source-grounded system's citations make complete blinding of graders impossible.

5. Conclusion

In this benchmark of five artificial-intelligence systems on a 154-item reconstruction of the 2025 NEET-PG examination, Zörgm Pro, a source-grounded clinical answer engine, was the strongest performer overall. It achieved the highest score of any system tested (607 of 616 marks, 98.5%, under the official +4/-1/0 scheme), answering 152 of 154 questions correctly, 4.2 percentage points above the next-best system (Gemini 3.5 Flash, 94.3%) and up to 11.5 points above the lowest (DeepSeek V4 Instant, 87.0%). It answered significantly more questions correctly than Claude Sonnet 4.6 and DeepSeek V4 Instant (Holm-adjusted $p<0.05$), and its top ranking held under both the corrected primary key and the original published key. It was also fully self-consistent, returning the same answer on every one of its five to six independent runs per item.

Beyond raw score, Zörgm Pro held a decisive advantage that no general-purpose model matched: it was the only system to attach a cited, verifiable source to every answer. The two strongest general-purpose models were close enough that a larger test would be needed to separate the leaders on score alone, but none of them can show a clinician where an answer came from. For clinical use, where an answer should be checked against trusted evidence rather than taken on trust, this makes the source-grounded system the better choice: it matched or exceeded the best general-purpose models on score while being the only one whose every answer can be verified. Exam performance is a proxy for medical knowledge and not direct evidence of clinical safety, so independent replication and prospective, clinically grounded evaluation remain warranted. On the evidence of this benchmark, however, Zörgm Pro was both the highest-scoring system and the only one suited to verifiable clinical use.

Declarations

Competing interest: Both authors are affiliated with Laennec AI Limited, which developed Zörgm Pro, the source-grounded system evaluated in this study.

Funding: This study was funded by Laennec AI Limited.

Data availability: Prompts, model responses, and scoring data are available from the corresponding author on request; the examination questions are owned by NBEMS and cannot be redistributed.

Ethics: No human participants or patient data were involved; no ethics approval was required.

References

1. Moor M, Banerjee O, Abad ZSH, et al. Foundation models for generalist medical artificial intelligence. Nature. 2023;616(7956):259-265. https://doi.org/10.1038/s41586-023-05881-4
2. Thirunavukarasu AJ, Ting DSJ, Elangovan K, et al. Large language models in medicine. Nat Med. 2023;29(8):1930-1940. https://doi.org/10.1038/s41591-023-02448-8
3. Ji Z, Lee N, Frieske R, et al. Survey of hallucination in natural language generation. ACM Comput Surv. 2023;55(12):248. https://doi.org/10.1145/3571730
4. Blum M. ChatGPT produces fabricated references and falsehoods when used for scientific literature search. J Card Fail. 2023;29(9):1332-1334. https://doi.org/10.1016/j.cardfail.2023.06.015
5. Pal A, Umapathi LK, Sankarasubbu M. Med-HALT: medical domain hallucination test for large language models. In: Proc 27th Conf Comput Nat Lang Learn (CoNLL). 2023. arXiv:2307.15343. https://doi.org/10.48550/arXiv.2307.15343
6. Asgari E, Montana-Brown N, Dubois M, et al. A framework to assess clinical safety and hallucination rates of LLMs for medical text summarisation. npj Digit Med. 2025;8:274. https://doi.org/10.1038/s41746-025-01670-7
7. Zhang M, Zhao T. Citation accuracy challenges posed by large language models. JMIR Med Educ. 2025;11:e72998. https://doi.org/10.2196/72998
8. Kung TH, Cheatham M, Medenilla A, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLOS Digit Health. 2023;2(2):e0000198. https://doi.org/10.1371/journal.pdig.0000198
9. Nori H, King N, McKinney SM, Carignan D, Horvitz E. Capabilities of GPT-4 on medical challenge problems. arXiv:2303.13375. 2023. https://arxiv.org/abs/2303.13375
10. Singhal K, Azizi S, Tu T, et al. Large language models encode clinical knowledge. Nature. 2023;620(7972):172-180. https://doi.org/10.1038/s41586-023-06291-2
11. Singhal K, Tu T, Gottweis J, et al. Toward expert-level medical question answering with large language models. Nat Med. 2025;31(3):943-950. https://doi.org/10.1038/s41591-024-03423-7
12. da Silva FLF, Roeder EA, Severino JVB, et al. Performance of large language models on the Brazilian national medical education examination: comparative benchmark study. JMIR Med Educ. 2026;12:e89839. https://doi.org/10.2196/89839
13. Zong H, Wu R, Cha J, et al. Large language models in worldwide medical exams: platform development and comprehensive analysis. J Med Internet Res. 2024;26:e66114. https://doi.org/10.2196/66114

14. Wei B. Performance evaluation and implications of large language models in radiology board exams: prospective comparative analysis. JMIR Med Educ. 2025;11:e64284. https://doi.org/10.2196/64284
15. Paul S, Govindaraj S, Jerisha JK. ChatGPT versus National Eligibility cum Entrance Test for Postgraduate (NEET PG). Cureus. 2024;16(6):e63048. https://doi.org/10.7759/cureus.63048
16. Ramnani S, Bhalla M, Bassi R. A comparative study of ChatGPT and BingAI in answering NEET-PG-style practice questions: a cross-sectional analysis. Cureus. 2024;16(12):e76108. https://doi.org/10.7759/cureus.76108
17. Raji ID, Daneshjou R, Alsentzer E. It's time to bench the medical exam benchmark. NEJM AI. 2025;2(2). https://doi.org/10.1056/AIe2401235
18. Xu C, Guan S, Greene D, Kechadi MT. Benchmark data contamination of large language models: a survey. arXiv:2406.04244. 2024. https://arxiv.org/abs/2406.04244
19. Lewis P, Perez E, Piktus A, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. In: Adv Neural Inf Process Syst 33 (NeurIPS). 2020:9459-9474. https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html
20. Zakka C, Shad R, Chaurasia A, et al. Almanac - retrieval-augmented language models for clinical medicine. NEJM AI. 2024;1(2). https://doi.org/10.1056/AIoa2300068
21. Xu S, Yan Z, Dai C, Wu F. MEGA-RAG: a retrieval-augmented generation framework with multi-evidence guided answer refinement for mitigating hallucinations of LLMs in public health. Front Public Health. 2025;13:1635381. https://doi.org/10.3389/fpubh.2025.1635381
22. Rashkin H, Nikolaev V, Lamm M, et al. Measuring attribution in natural language generation models. Comput Linguist. 2023;49(4):777-840. https://aclanthology.org/2023.cl-4.2/
23. Bohnet B, Tran VQ, Verga P, et al. Attributed question answering: evaluation and modeling for attributed large language models. arXiv:2212.08037. 2022. https://arxiv.org/abs/2212.08037
24. Liu NF, Zhang T, Liang P. Evaluating verifiability in generative search engines. In: Find Assoc Comput Linguist EMNLP. 2023:7001-7025. https://doi.org/10.18653/v1/2023.findings-emnlp.467
25. Gao T, Yen H, Yu J, Chen D. Enabling large language models to generate text with citations. In: Proc 2023 Conf Empir Methods Nat Lang Process (EMNLP). 2023:6465-6488. https://aclanthology.org/2023.emnlp-main.398/
26. Gallifant J, Afshar M, Ameen S, et al. The TRIPOD-LLM reporting guideline for studies using large language models. Nat Med. 2025;31(1):60-69. https://doi.org/10.1038/s41591-024-03425-5
27. Norgeot B, Quer G, Beaulieu-Jones BK, et al. Minimum information about clinical artificial intelligence modeling: the MI-CLAIM checklist. Nat Med. 2020;26(9):1320-1324. https://doi.org/10.1038/s41591-020-1041-y

28. Miao BY, Chen IY, Williams CYK, et al. The MI-CLAIM-GEN checklist for generative artificial intelligence in health. Nat Med. 2025;31(5):1394-1398. https://doi.org/10.1038/s41591-024-03470-0
29. Altman DG, Simera I, Hoey J, Moher D, Schulz K. EQUATOR: reporting guidelines for health research. Lancet. 2008;371(9619):1149-1150. https://doi.org/10.1016/S0140-6736(08)60505-X
30. DigiNerve. NEET PG 2025 recall questions with answers - free PDF download (all 200 Qs). DigiNerve; 2026. Available from:https://www.diginerve.com/blogs/neet-pg-2025-recall-questions-with-answers-free-pdf-download-all-200-qs/(accessed June 2026).