

Pattern Recognition or Medical Knowledge? The Problem with Multiple-Choice Questions in Medicine

Maxime Griot^{1,2,3}, Jean Vanderdonckt², Demet Yuksel^{1,3}, Coralie Hemptinne^{1,3}

¹Institute of NeuroScience, Université catholique de Louvain, Brussels, Belgium

²Louvain Research Institute in Management and Organizations, Louvain-la-Neuve, Belgium

³Cliniques universitaires Saint-Luc, Brussels, Belgium

Correspondence: maxime.griot@uclouvain.be

Abstract

Large Language Models (LLMs) such as ChatGPT demonstrate significant potential in the medical domain and are often evaluated using multiple-choice questions (MCQs) modeled on exams like the USMLE. However, such benchmarks may overestimate true clinical understanding by rewarding pattern recognition and test-taking heuristics. To investigate this, we created a fictional medical benchmark centered on an imaginary organ, the **Glianorex**, allowing us to separate memorized knowledge from reasoning ability. We generated textbooks and MCQs in English and French using leading LLMs, then evaluated proprietary, open-source, and domain-specific models in a zero-shot setting. Despite the fictional content, models achieved an average score of 64%, while physicians scored only 27%. Fine-tuned medical models outperformed base models in English but not in French. Ablation and interpretability analyses revealed that models frequently relied on shallow cues, test-taking strategies, and hallucinated reasoning to identify the correct choice. These results suggest that standard MCQ-based evaluations may not effectively measure clinical reasoning and highlight the need for more robust, clinically meaningful assessment methods for LLMs.

1 Introduction

Large Language Models (LLMs), such as ChatGPT, have demonstrated significant potential in the medical field, with studies evaluating their performance on tests originally designed for humans, including the United States Medical Licensing Examination (USMLE) (Jin et al., 2020; Pal et al., 2022; Jin et al., 2019; Nori et al., 2023). Furthermore, the domain-specific research shows that these models perform well on specialized medical exams in areas such as pediatrics, radiology, ophthalmology, plastic surgery, and oncology (Rydzewski et al., 2024; Bhayana et al., 2023; Barile et al., 2024; Mihalache

et al., 2023; Humar et al., 2023). The common reliance on MCQs in these assessments reflects their widespread use as a testing method for medical students around the globe (Al-Wardy, 2010).

However, while MCQs are easy to administer and grade, they have notable limitations, often promoting surface learning and pattern recognition over deep understanding (Veloski et al., 1999). Despite their widespread use, few studies have addressed the potential issues unique to LLMs, such as their reliance on statistical patterns rather than genuine understanding. Notably, when trained on synthetic questions, Meerkat-7b outperformed its base model, Mistral-7b, on medical benchmarks by 18.6%. This performance surpassed Meditron-7b, which improved by only 4% despite being trained on a considerably larger, higher-quality clinical corpus (Kim et al., 2024; Chen et al., 2023). This discrepancy highlights that extensive MCQ-based training can be more effective for benchmark performance than training on comprehensive medical content, raising concerns about the true depth of understanding being evaluated. This is further supported by more complex and realistic evaluations, such as patient interactions (Johri et al., 2025) or free-text questions (Arvidsson et al., 2024), which reveal that LLMs perform poorly compared to medical experts.

These concerns are particularly relevant for LLMs, which are trained on large datasets that likely contain statistical patterns. This reliance can lead models to produce correct answers for incorrect reasons (Jin et al., 2024), such as identifying melanoma based on the presence of a ruler in an image (Narla et al., 2018). While the limitations of MCQ-based medical benchmarks have begun to surface, recent work further underscores their fragility. The MedFuzz experiment (Ness et al., 2024), for example, showed that LLMs could be induced to provide incorrect answers by violating assumptions in the formulation of the questions. Like-

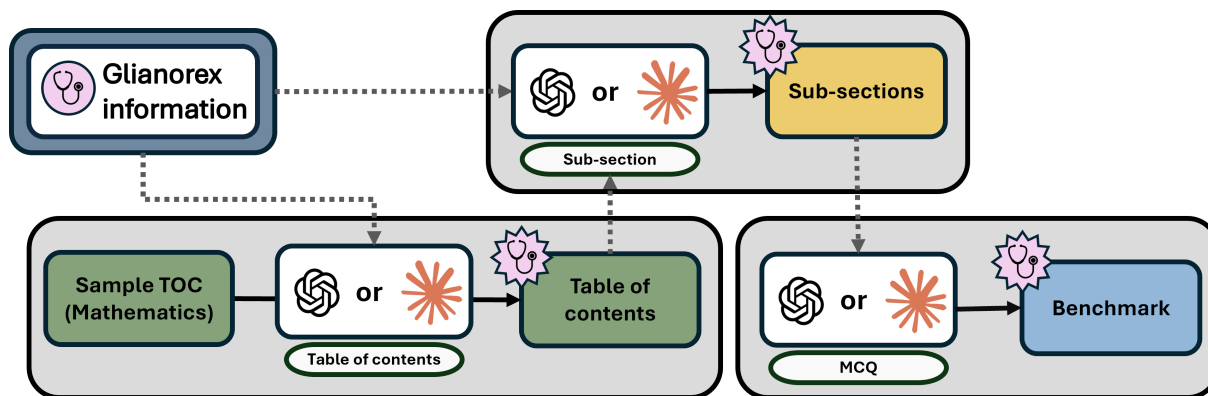


Figure 1: Three-stage pipeline for benchmark generation: (1) Create a structured JSON table of contents using a math textbook template and Glianorex grounding data; (2) Generate detailed subsection content using hierarchical section titles and the same grounding data; (3) Iteratively generate questions for each subsection until at least 200 are created. Medical professionals perform quality assurance at every stage, marked by a stethoscope stamp.

wise, replacing drug names with generic or branded alternatives led to performance drops of up to 10% (Gallifant et al., 2024). Moreover, models struggle to identify when none of the options is correct or when the question cannot be answered due to missing information (Griot et al., 2025). Collectively, these findings suggest that strong performance on MCQs often reflects superficial pattern-matching rather than genuine medical understanding, as evidenced by the models’ sensitivity to minor perturbations in otherwise familiar inputs.

Based on these findings, our work evaluates the performance of LLMs on novel medical concepts absent from the training data, thus investigating their ability to address questions about unfamiliar medical content. This approach investigates whether MCQ-based evaluations are primarily vulnerable to pattern variations or whether models can leverage test-taking strategies when encountering unfamiliar content. To achieve this, we developed a benchmark centered on a fictional organ, the Glianorex, designed to more effectively separate test-taking abilities from training data dependencies than previous studies have achieved.

2 Related work

The evaluation of medical knowledge and clinical skills remains an active research area, with new methods such as oral examinations and competency evaluations being proposed to better assess medical students and residents (Veloski et al., 1999; Prediger et al., 2020; Goins et al., 2023). Globally, medical evaluations are heavily based on MCQs, such as the USMLE in the United States, which sig-

nificantly influences residency placements (Gauer and Jackson, 2017). LLMs are similarly evaluated using MCQs to assess their medical knowledge. Google introduced MultiMedQA with its Med-PaLM model, which combines several existing medical benchmarks and has become a standard for evaluating medical proficiency in AI models (Singhal et al., 2023; Pal et al., 2024). Recently, Google incorporated additional physician-led evaluations into its Med-Gemini model (Saab et al., 2024). These multiple-choice items have raised concerns regarding their relevance for clinical use (Raji et al., 2025). MultiMedQA remains the most commonly reported benchmark to date and is composed of the following benchmarks:

MedQA-USMLE This subset of the MedQA dataset was sourced from the National Board of Medical Examiners (NBME), the organization responsible for the USMLE (NBME, 2024). The dataset is composed of a total of 12723 questions, split into a training set of 10178 samples, a validation set of 1272 questions, and a test set of 1273 questions. The questions have 4 options with only one correct answer (Jin et al., 2020). Most questions present a clinical vignette and require the test-taker to apply clinical or foundational science knowledge to select the best answer.

MedMCQA The Multi-Subject Multi-Choice Dataset for Medical domain Question Answering is composed of 194k multiple choice questions obtained from the All India Institute of Medical Science (AIIMS) and National Eligibility cum Entrance Test Postgraduate (NEET PG) entrance ex-

aminations (AIIMS, 2024; NBEMS, 2024). These questions are split into 3 subsets, one training subset composed of 183k samples, a validation subset of 4.18k samples, and a test subset comprising 6.15k samples, with the distinctive feature that the test subset omits the correct answers to prevent data leakage. The questions have 4 options each and can be either single or multiple choice. Most questions are straightforward knowledge-recall and do not use clinical vignettes.

PubMedQA This biomedical question answering dataset was created using PubMed (NLM, 2024) article abstracts from which the authors derived a context with a question and a yes/maybe/no label. It comprises three subsets: an expert-annotated set of 1000 samples, an unlabeled set of 61.2k, and an automatically generated set of 211.3k. The generated samples are used to train models, while 500 samples of the expert-annotated subset are used to test the models. This benchmark was designed to evaluate the reasoning ability of models when presented with the abstract and a question related to this abstract (Jin et al., 2019).

MMLU-Medical The Massive Multitask Language Understanding dataset contains 57 tasks, of which six are used to assess medical knowledge (clinical knowledge, medical genetics, anatomy, professional medicine, college biology, and college medicine) (Hendrycks et al., 2021). These tasks were collected by students from publicly available online resources, including USMLE questions and undergraduate-level questions. The dataset contains 1,242 questions and is split into 30 for training, 123 for validation, and 1,089 for testing. The questions each have four options with only one correct answer and are a mix of clinical vignettes and recall questions.

3 Methods

To address these fundamental limitations in MCQ-based evaluation, we designed a novel benchmark to assess the relevance of MCQs for LLM evaluation using the process detailed in Figure 1. Our approach involved creating MCQs similar to those of the USMLE, but based on a fictional organ called the Glianorex. This process involved manual creation of grounding data by a volunteer physician who then prompted language models. The physician was instructed to write a brief overview of a fictional organ, including its name, the history of

its discovery, anatomy, physiology (hormones and their role), histology (specific receptors), pathology (diseases associated with the Glianorex), and specific diagnostic techniques.

3.1 Dataset Construction

3.1.1 Knowledge

To create diverse questions, the first step of the process was to augment the seed data using GPT-4 Turbo and Claude 3.5 Sonnet to generate additional content on the Glianorex. To generate a textbook, we first used the LLM to generate a table of contents in a standard JSON format with three levels of granularity: chapter, section, and subsection. After generation, manual verification was performed to validate the coherence and quality of the table of contents before proceeding with the textbook generation. The LLM was then used to generate each subsection independently. To improve coherence between different subsections, we provided the model with the grounding data and the complete table of contents. This process resulted in one textbook per model in English on the Glianorex detailing its history, physiology, anatomy, and pathology.

3.1.2 Questions

Based on these fictional textbooks, we used the same models separately to generate MCQs. The use of LLMs to generate questions based on textbooks was previously demonstrated to be an efficient and validated methodology, with Kim et al. (2024) showing significant improvements on various downstream benchmark tasks, including MedQA, MedMCQA, the USMLE sample test, and MMLU-Medical using this approach. This established precedent provides strong evidence that the quality of questions generated by state-of-the-art models would be sufficient for this study, even when applied to fictional medical content. For each model, these questions contained four choices with only one correct answer, adhering to a format similar to that of the USMLE to ensure uniformity. To facilitate the creation of these questions, we prompted models with the table of contents and a subsection from the textbook (see Table 5). This approach guided both models to generate questions in a JSON format consistent with existing medical benchmarks.

3.1.3 Multilingual

To study the influence of language on test-taking abilities, we used the same models to translate the generated textbooks and questions using a simple one-shot prompt per subsection and question, asking the model to translate into French.

3.1.4 Validation

We recruited two physicians from our institution who completed at least one step of the USMLE in the past five years to assess the quality of the questions. They evaluated a random sample of 100 English questions on a 7-point Likert scale and answered them—without prior exposure to textbooks on the Glianorex—to establish an expert baseline. We conducted a keyword search for “context” across all questions to identify potential incompleteness. Finally, a physician manually verified the consistency of the Introduction, Anatomy, and Biochemistry chapters in both English and French GPT-4 Turbo generated textbooks to assess language quality, internal coherence, and translation quality.

3.1.5 Synthetic Bias Mitigation

Because our data-generation pipeline relies on LLMs, it is susceptible to synthetic biases. We therefore introduced several safeguards and human checkpoints:

1. A physician authored the medical grounding information that was provided to the LLMs. This expert-verified context aligned all subsections to a single source of truth.
2. The entire experiment was replicated with two models of comparable capability (GPT-4-Turbo and Claude 3.5 Sonnet) using identical prompts and seed material to ensure that results were not model-specific.
3. Before generation, the physician reviewed the table of contents to confirm its coherence and relevance. In addition, a sample of subsections was manually verified before proceeding with MCQ generation.
4. To counter demographic bias and increase variability, we randomly specified gender and age parameters (ranging from 12 to 90 years) in 50% of the prompts (Zack et al., 2024).
5. For each subsection, we generated four questions with a temperature of 1 to produce diverse question variants.

6. Answer options were shuffled so that the position of the correct choice was balanced across items.

7. A sample of questions, textbook excerpts, and translations was audited by humans to verify quality and coherence.

3.2 Quantitative Analysis

3.2.1 Models

To evaluate the performance of LLMs, we selected a diverse set of models, including both proprietary and open-weight options. We included commonly used foundational models, as reported in Table 1. Additionally, we included two fine-tuned medical domain models based on Mistral-7B-v0.1 to assess the influence of domain-specific training on this fictional benchmark. First, *internistai-7b-v0.2* (Apache 2.0), which was trained on a mixture of general data, medical textbooks, and MCQs, demonstrating improved performance on medical evaluations compared to its base model (Griot et al., 2024). Second, *meerkat-7b-v1.0* (Creative Commons Attribution-NonCommercial 4.0), which was trained exclusively on MCQs, some of which were generated from medical textbooks (Kim et al., 2024). The latter training approach showed a significant performance increase in benchmarks using a relatively small amount of training data compared to continued pretraining on large medical datasets, as shown by Meditron and PMC-LLaMA (Chen et al., 2023; Wu et al., 2024).

Model	License
gpt-3.5-turbo-0125	Proprietary
gpt-4-turbo-2024-04-09	Proprietary
gpt-4o-2024-05-13	Proprietary
01-ai/Yi-1.5-9B	Apache 2.0
01-ai/Yi-1.5-34B	Apache 2.0
mistralai/Mistral-7B-v0.1	Apache 2.0
mistralai/Mixtral-8x7B-v0.1	Apache 2.0
meta-llama/Meta-Llama-3-8B	Llama 3 license
meta-llama/Meta-Llama-3-70B	Llama 3 license
Qwen/Qwen1.5-7B	Tongyi Qianwen license
Qwen/Qwen1.5-32B	Tongyi Qianwen license
Qwen/Qwen1.5-110B	Tongyi Qianwen license

Table 1: Foundational models (OpenAI, 2022, 2023, 2024; AI et al., 2024; Mistral, 2024; AI@Meta, 2024; Bai et al., 2023) included in the quantitative analysis.

3.2.2 Evaluation

We evaluated all models using lm-evaluation-harness (Gao et al., 2023) in a zero-shot setting without additional training. The task followed the MedQA 4-option format, using a log-likelihood approach to measure accuracy. We calculated 95% confidence intervals by multiplying the standard error of the mean by 1.96, assuming normal error distribution. We assessed statistical significance of accuracy against a random model using the cumulative distribution function of a binomial distribution.

A two-way analysis of variance (ANOVA) was conducted with model and benchmark subsets as independent variables, followed by Tukey’s honestly significant difference (HSD) test for post-hoc pairwise comparisons of accuracy. We performed linear regression analysis to compare model accuracy on the Glianorex benchmark and MedQA-USMLE. Evaluations ran on a virtual machine with four NVIDIA A100 (80GB) GPUs on Microsoft Azure, with a total runtime of 10 hours including model download time.

3.3 Interpretability and Ablation Analyses

To examine how prompt structure affects performance and to understand the model’s reasoning, we conducted ablation and qualitative studies with DeepSeek-V3-0324 (DeepSeek-AI, 2025) on the Glianorex benchmark in English and French. All generations were produced on a server with eight NVIDIA H200 (141GB) GPUs running vLLM (Kwon et al., 2023) with greedy decoding (temperature = 0) to improve reproducibility.

3.3.1 Prompting Parameters

We evaluated four settings: **Zero-shot**, where the model sees the full question stem and answer choices and must select an answer directly; **Chain-of-Thought (CoT)**, which prompts the model to think step by step before the final answer; **Zero-shot, Answers-only (AO)**, where the question stem is removed and only answer options are provided; and **AO+CoT**, combining answers-only input with chain-of-thought reasoning.

3.3.2 Analysis

For each item and setting, we generated a single prediction and computed accuracy. Agreement between zero-shot and chain-of-thought predictions was assessed using Cohen’s κ . To test whether answer length influences selection, we compared the

character length of the model’s chosen option with the lengths of remaining alternatives.

Finally, we manually examined a subset of chain-of-thought traces, including both full questions and answers-only conditions, from correct and incorrect predictions in English.

4 Results

4.1 Dataset

The resulting fictional textbooks on the Glianorex were generated using the proposed structure with both GPT-4 Turbo and Claude 3.5 Sonnet. Each textbook contains detailed sections on the anatomy, physiology, biochemistry, pathology, and diagnostic tools related to the Glianorex. For both models, the textbooks were produced in English and French, each containing approximately 35,000 words. We then reused the subsections of the English textbooks to generate MCQs in English, followed by a translation step to obtain the same questions in French. The GPT-4 Turbo process resulted in 264 questions per language, while the Claude 3.5 Sonnet process produced 224 questions per language. For both models, examples of these questions (see Table 9 and Table 10) included complex scenarios requiring multiple steps of reasoning. Each question adhered to a four-option format similar to MedQA-USMLE standards, with one correct answer.

4.1.1 Internal consistency

A partial data validation conducted by two physicians revealed no major flaws in the dataset. Their review of key textbook chapters identified minor inconsistencies that fell into two categories: contradictions and omissions. Contradictions involved discrepancies between subsections, such as slight variations in the described location of the Glianorex between the “Proximity of the heart” and “Embryology and Development” sections. Omissions occurred when relevant information appeared in only one subsection when it should have been present in others; for instance, the embryological origin of the Glianorex as *splanchnopleure* was mentioned in the “Vascular Supply” section but absent from “Embryology and Development.” While these inconsistencies were subtle and required extensive cross-referencing to detect, they did not compromise the overall integrity of the content.

4.1.2 Language

Cross-linguistic analysis demonstrated structural and content consistency across languages, with only negligible variations in French abbreviation conventions. Quality assessment of a 100-question English sample by two physicians yielded high scores (6.94 and 6.86 out of 7), indicating quality comparable to board-examination standards. Manual verification across both languages identified only eight incomplete questions (four per language, < 1% of total) that required additional context to be answered.

4.2 Evaluations

4.2.1 General Results

All models achieved relatively high scores, averaging 63.8%, as illustrated in Figure 2. To place this score in perspective, the physicians each obtained 27%, which is within the expected range for random answering. The physicians noted that the questions relied heavily on fictional terminology and concepts, and a substantial portion focused on information recall, leading them to resort to guesswork rather than applying their medical expertise to formulate responses.

A statistically significant difference was observed between the top-performing models and the lowest-performing models, as shown in Table 2. The performance differences when isolating languages were also significant and occurred more frequently in English, as shown in Table 7 and Table 8. We also calculated Cohen’s *d* between all model pairs, which revealed a range of effect sizes, indicating varying degrees of performance differences between the models (Cohen, 2013). Most of the comparisons show very small or negligible effect sizes, with many pairs having a Cohen’s *d* close to 0, as shown in Table 11. For instance, pairs such as Yi-1.5-34B – Yi-1.5-9B ($d = 0.002$) and Yi-1.5-34B – gpt-3.5-turbo-0125 ($d = 0.030$) suggest negligible differences. This pattern is consistent across most pairs, indicating that the models’ performances are closely aligned.

However, some pairs demonstrate more noticeable differences, such as meerkat-7b-v1.0 – gpt-4o-2024-05-13 ($d = 0.343$) and gpt-4-turbo-2024-04-09 – Mistral-7B-v0.1 ($d = 0.254$), suggesting a measurable effect. Overall, the analysis reveals that while some variations exist, the effect sizes for most model comparisons are small. Additionally, the average score for En-

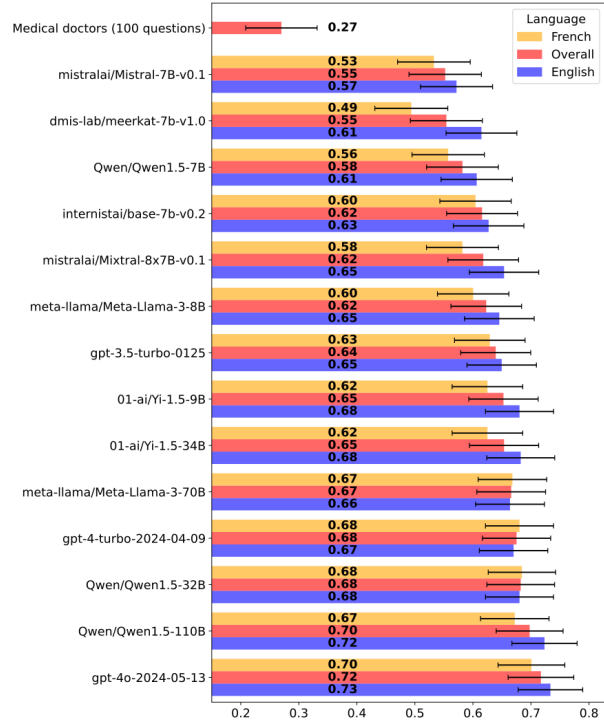


Figure 2: Accuracy of the evaluated models on the synthetic benchmark with 95% confidence intervals. Scores are presented separately for English and French, illustrating that most models achieve higher accuracy in English compared to French. Additionally, we include the performance of medical doctors evaluated on a subset of 100 English questions as a human reference.

	Qwen/Qwen1.5-110B	dmis-lab/meerkat-7b-v1.0	gpt-4o-2024-05-13	mistralai/Mistral-7B-v0.1
Qwen/Qwen1.5-7B	*		**	
meta-llama/Meta-Llama-3-70B	*	*	*	*
Qwen/Qwen1.5-32B	**	**	**	**
Qwen/Qwen1.5-110B	**	**	***	***
gpt-4-turbo-2024-04-09	*	*	*	*
gpt-4o-2024-05-13	****	****	****	****

Table 2: Statistical significance of the performance differences between models (* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$, and **** $p < 0.0001$).

glish questions was 65.7%, whereas French averaged 61.8%.

Finetuned Models The internistai-7b-v0.2 and meerkat-7b-v1.0 models demonstrated enhanced English performance relative to their base

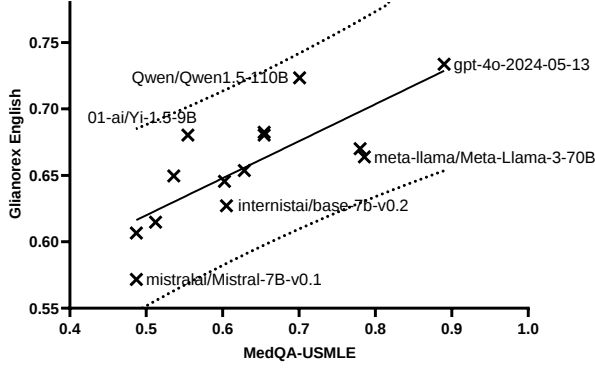


Figure 3: Linear regression analysis comparing MedQA-USMLE four-option scores and Glianorex English scores, shown with 95% prediction bands.

model, Mistral-7B-v0.1. However, this improvement was not replicated in French, suggesting that domain-specific training enhances performance in the target language, but the absence of multilingual data during continued training may diminish performance in other languages.

4.2.2 Cross-Benchmark Analysis

We conducted a linear regression analysis to examine the relationship between performance on the MedQA-USMLE four-option benchmark and accuracy on the Glianorex English subset, as illustrated in Figure 3. The results revealed a statistically significant correlation ($p < 0.01$) between these two metrics ($R^2 = 0.5952$). This correlation, combined with the observed relationship between model size and performance, suggests that improvements in medical benchmarks may be partially attributable to enhanced pattern recognition capabilities.

4.3 Interpretability

4.3.1 Input and Reasoning Ablations

In the zero-shot configuration, DeepSeek-V3-0324 achieved an average accuracy of 70.1%, consistent with the top-performing models previously evaluated. Chain-of-thought prompting yielded only marginal improvement (70.3%), indicating limited effectiveness of explicit reasoning prompts for this task (Table 3). Cohen’s κ analysis confirmed substantial agreement between zero-shot and chain-of-thought prompting approaches ($\kappa = 0.681$ for English, $\kappa = 0.653$ for French).

When prompted exclusively with answer choices (Answers-only setting), the model achieved an average accuracy of 46.5%. While this performance sig-

	zero-shot	CoT	AO + zero-shot	AO + CoT
English	0.705	0.717	0.457	0.473
French	0.697	0.689	0.473	0.438
Average	0.701	0.703	0.465	0.455

Table 3: Accuracy of DeepSeek-V3-0324 on the benchmark using four evaluation configurations. AO (Answers Only) prompts only with the choices, the question stem is removed entirely.

nificantly exceeds random chance (25%), it remains substantially below full-prompt performance. This intermediate accuracy level supports the hypothesis proposed by Balepur et al. (2024), suggesting that large language models employ meta-strategies such as question inference and surface-level shortcuts. Furthermore, we analyzed the relative length of answers selected by the model compared to other available choices and found no statistically significant differences ($p > 0.05$).

4.3.2 Qualitative Analysis

Manual qualitative examination of chain-of-thought explanations identified three recurring patterns employed by the model: hallucinations, generalized medical assumptions, and explicit test-taking strategies. For each pattern, we describe common characteristics and present examples as generated by DeepSeek-V3-0324.

4.3.3 Hallucinations

The model frequently generated fabricated knowledge based on question and answer content. This pattern predominantly resulted in incorrect responses but occasionally led to correct answers. For example, the model hallucinated information regarding the optimal imaging technique for a fictional disease, which guided the model toward an incorrect answer.

Hallucination

Glianorex Imagery Sonography (GIS) is the most specific and helpful diagnostic tool for confirming autoimmune Glianorexitis. This imaging modality allows for direct visualization of glianorex tissue inflammation and damage, which is critical for diagnosis.

4.3.4 Medical Assumptions

The model explicitly applied characteristics of real autoimmune diseases to the fictional conditions presented in the benchmark, occasionally yielding correct inferences but frequently causing inaccuracies. For instance, the model erroneously assumed

that genetic risk factors associated with known autoimmune disorders would similarly apply to the fictional condition Glianorexiditis.

General medical principles

The question involves a fictional condition ("Glianorex degeneration") and diagnostic test ("Glianorex Imagery Sonography (GIS)"), so the answer must be inferred from the context of the question and general medical principles

4.3.5 Test-Taking Strategies

We identified explicit heuristic approaches, including the selection of highly specific answers and the preference for answers that structurally resemble typical examination formulations. This strategy is also observed, albeit to a lesser extent, in human test-takers, such as avoiding answers containing absolute terms like "always" or "never," which tend to be incorrect in medical contexts.

Specificity

In exams, highly specific answers ("biotransplant," "modulators," "re-equilibration") are often correct when other choices are generic.

The model also explicitly recognized and exploited answer constructions that it perceived as characteristic of examination environments.

Test construction

D stands out as the most "constructed" correct answer in a medical context, resembling how hypothetical disorders are framed in exams. (While all options contain questionable terms, D is the most logically structured and aligns best with how exam questions are typically designed - tying a novel mechanism to a targeted treatment.)

5 Discussion

5.1 Evaluation Implications

The results of this study highlight several insights into the capabilities and limitations of LLMs in handling medical MCQs. Despite the novelty and complexity of the fictional organ, all evaluated models achieved high scores on the MCQs generated for this material. However, physicians who attempted to answer a random subset of the benchmark were unable to perform better than chance. This finding suggests that LLMs are adept at recognizing patterns and applying test-taking strategies, even in unfamiliar contexts.

5.1.1 Benchmarking

The consistently high performance across various foundational models in English, regardless of their architecture, size, or specialization, indicates that traditional MCQ-based benchmarks may inadequately assess LLMs' medical knowledge and clinical reasoning skills. These benchmarks appear to test pattern identification and association abilities rather than genuine material comprehension. Consequently, relying on MCQs to evaluate LLMs in medical and other specialized domains might overestimate their actual capabilities. This finding aligns with research demonstrating that models become less reliable as they scale up (Zhou et al., 2024). Using adversarial benchmarks like the one introduced in this study could help identify reliability reductions during development.

5.1.2 Training

The superior performance of fine-tuned models internistai/base-7b-v0.2 and dmis-lab/meerkat-7b-v1.0 over the foundational model mistralai/Mistral-7B-v0.1 underscores the impact of task-specific and domain-specific training on LLM capabilities. Both models were trained on medical MCQs—with Meerkat trained exclusively on MCQs—raising the question of whether the improvement stems from enhanced test-taking skills or greater medical-domain knowledge. Previous research shows that task improvements in models trained on additional medical data disappear after prompt optimization (Jeong et al., 2024), suggesting that additional training may target evaluation methodology to improve accuracy rather than enhancing medical capabilities. This aligns with findings that models need specific training on extraction tasks to leverage their internal knowledge, which may explain the gains observed after additional MCQ training (Allen-Zhu and Li, 2024).

5.2 Medical Implications

Current medical evaluation standards may not accurately reflect LLMs' capabilities in the medical domain, raising significant concerns about their safety and clinical implications in real-world settings. Performance claims based on MCQs could misrepresent these models' actual capabilities, creating false trust that might endanger patients who rely on these systems instead of consulting physicians, and physicians who implement them for clinical decision support.

Such claims could also undermine trust within the medical community, which has already expressed skepticism regarding LLM applications in medicine (Marks and Haupt, 2023; Flanagan et al., 2023). Misrepresenting the medical capabilities and usefulness of these models may lead physicians to view AI as a commercial selling point rather than a tool for real progress, potentially hindering AI adoption and limiting opportunities for multidisciplinary teams to develop clinically relevant models.

The integration of LLMs in clinical settings poses significant patient safety risks, especially given the time constraints faced by clinicians. Previous work by Liu et al. (2024) demonstrates a significant dose-response association between AI usage and burnout for radiologists, as well as increased post-processing, which can be attributed partly to added validation time. Given these findings and the high risk posed by these models, we believe it is unrealistic to expect clinicians to read chain-of-thought reasoning sections to ensure response validity, and therefore believe that models should provide trustworthy responses in zero-shot settings.

5.3 Recommendations

We recommend including medical professionals in model evaluation and urge developers to exercise greater caution when making claims based on MCQ-based benchmarks. Similar to medical devices and drugs, models should undergo clinical trials to ensure safety and demonstrate patient benefits over current practices (Widner et al., 2023). This requires a paradigm shift toward answering concrete questions such as "Does the use of model X to recommend parenteral nutrition reduce mortality in hospitalized patients with neck cancer?" instead of the current approach of assessing medical capabilities, a task that lacks both proper definition and clinical practice relevance.

5.4 Alternatives

Nevertheless, there is a need for automated and standardized evaluations to guide development. More advanced methodologies that do not rely solely on MCQs have been proposed, including case-based reasoning scenarios requiring intermediate physiological explanations, similar to those in the French ECN exams (Santé, 2021). Additionally, key-feature problems, where clinicians must identify critical decision points within complex clinical

scenarios and prioritize among multiple correct options, can offer deeper insights into model capabilities (Bordage and Page, 2018). Open-ended questions evaluated through rubric-based approaches, combining LLM-as-judge assessments with expert verification, can further enhance evaluation validity (Zheng et al., 2023). Finally, simulated clinical environments, such as the adaptive questioning and diagnostic refinement demonstrated in AI Hospital by Fan et al. (2025), present complex, dynamic settings to assess and refine LLM performance more accurately for safe and effective clinical application.

6 Conclusion

This study demonstrates that LLMs can achieve high scores on MCQs built around fictional medical knowledge without prior exposure to the content. By creating a fictional gland, the Glianorex, and generating comprehensive textbooks with related MCQs, we partially isolated the models' reasoning capabilities from memorized real-world data. Results show that models of different architectures, sizes, and specializations outperform physicians on this benchmark, suggesting that pattern recognition and test-taking strategies may play a larger role for LLMs than for humans.

Our findings call into question the effectiveness of current MCQ-based benchmarks for evaluating LLMs' clinical knowledge and reasoning abilities. This study highlights the need for more robust evaluation methods that better assess the true understanding and reasoning capabilities of LLMs in the medical domain. Future research should explore alternative evaluation methods beyond current MCQs to provide more accurate assessments of LLMs' capabilities in medicine and other specialized fields.

7 Limitations

Knowledge coherence Independent generation of subsections could result in inconsistencies or contradictions within the text, potentially creating questions with multiple plausible correct answers depending on the chapter context provided during question generation. We performed a partial coherence check on the generated textbook to ensure content plausibility and identified few inconsistencies and contradictions. This partial check does not guarantee the absence of major errors; however, since LLMs had no prior exposure to this fictional knowledge, inconsistencies between independent

subsections should not affect their ability to answer appropriately.

Synthetic biases Although we took steps to reduce synthetic biases intrinsic to our methodology, some may persist in the dataset. For example, grounding information authored by a single clinician inevitably reflects that clinician’s biases, which could partly account for the models’ tendency to leverage common medical patterns when answering questions on fictional content. Nevertheless, these residual biases are unlikely to fully explain the observed performance gap, especially given the low scores achieved by two physicians. To corroborate our findings, future work should evaluate models on multiple-choice questions covering clinical knowledge published after the training cut-off date.

8 Acknowledgements

This work was supported by the Fondation Saint-Luc grant number 467E and the Fédération Wallonie-Bruxelles through the Fond Spécial de Recherche of Université catholique de Louvain.

References

- 01 AI, Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Heng Li, Jiangcheng Zhu, Jianqun Chen, Jing Chang, Kaidong Yu, Peng Liu, Qiang Liu, Shawn Yue, Senbin Yang, Shiming Yang, Tao Yu, Wen Xie, Wenhao Huang, Xiaohui Hu, Xiaoyi Ren, Xinyao Niu, Pengcheng Nie, Yuchi Xu, Yudong Liu, Yue Wang, Yuxuan Cai, Zhenyu Gu, Zhiyuan Liu, and Zonghong Dai. 2024. [Yi: Open Foundation Models by 01.AI](#). *arXiv preprint*. ArXiv:2403.04652 [cs].
- AIIMS. 2024. [AIIMS - All India Institute Of Medical Science](#).
- AI@Meta. 2024. [Llama 3 Model Card](#).
- Nadia M. Al-Wardy. 2010. Assessment methods in undergraduate medical education. *Sultan Qaboos University Medical Journal*, 10(2):203–209.
- Zeyuan Allen-Zhu and Yuanzhi Li. 2024. [Physics of Language Models: Part 3.1, Knowledge Storage and Extraction](#). *arXiv preprint*. ArXiv:2309.14316 [cs].
- Rasmus Arvidsson, Ronny Gunnarsson, Artin Entezarjou, David Sundemo, and Carl Wikberg. 2024. [ChatGPT \(GPT-4\) versus doctors on complex cases of the Swedish family medicine specialist examination: an observational comparative study](#). *BMJ Open*, 14(12):e086148. Publisher: British Medical Journal Publishing Group Section: General practice / Family practice.
- Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023. Qwen Technical Report. *arXiv preprint arXiv:2309.16609*.
- Nishant Balepur, Abhilasha Ravichander, and Rachel Rudinger. 2024. [Artifacts or Abduction: How Do LLMs Answer Multiple-Choice Questions Without the Question?](#) In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10308–10330, Bangkok, Thailand. Association for Computational Linguistics.
- Joseph Barile, Alex Margolis, Grace Cason, Rachel Kim, Saia Kalash, Alexis Tchaconas, and Ruth Milanaik. 2024. [Diagnostic Accuracy of a Large Language Model in Pediatric Case Studies](#). *JAMA Pediatrics*.
- Rajesh Bhayana, Robert R. Bleakney, and Satheesh Krishna. 2023. [GPT-4 in Radiology: Improvements in Advanced Reasoning](#). *Radiology*, 307(5):e230987. Publisher: Radiological Society of North America.
- G. Bordage and G. Page. 2018. [The key-features approach to assess clinical decisions: validity evidence to date](#). *Advances in Health Sciences Education*, 23(5):1005–1036.
- Zeming Chen, Alejandro Hernández-Cano, Angelika Romanou, Antoine Bonnet, Kyle Matoba, Francesco Salvi, Matteo Pagliardini, Simin Fan, Andreas Köpf, Amirkeivan Mohtashami, Alexandre Sallinen, Alireza Sakhaeirad, Vinitra Swamy, Igor Krawczuk, Deniz Bayazit, Axel Marmet, Syrielle Montariol, Mary-Anne Hartley, Martin Jaggi, and Antoine Bosselut. 2023. [MEDITRON-70B: Scaling Medical Pretraining for Large Language Models](#).
- Jacob Cohen. 2013. *Statistical Power Analysis for the Behavioral Sciences*. Routledge. Google-Books-ID: 2v9zDAsLvA0C.
- DeepSeek-AI. 2025. [deepseek-ai/DeepSeek-V3-0324 - Hugging Face](#).
- Zhihao Fan, Lai Wei, Jialong Tang, Wei Chen, Wang Siyuan, Zhongyu Wei, and Fei Huang. 2025. [AI Hospital: Benchmarking Large Language Models in a Multi-agent Medical Interaction Simulator](#). In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 10183–10213, Abu Dhabi, UAE. Association for Computational Linguistics.

- Annette Flanagan, Kirsten Bibbins-Domingo, Michael Berkwits, and Stacy L. Christiansen. 2023. [Non-human “Authors” and Implications for the Integrity of Scientific Publication and Medical Knowledge](#). *JAMA*, 329(8):637–639.
- Jack Gallifant, Shan Chen, Pedro Moreira, Nikolaj Munch, Mingye Gao, Jackson Pond, Leo Anthony Celi, Hugo Aerts, Thomas Hartvigsen, and Danielle Bitterman. 2024. [Language Models are Surprisingly Fragile to Drug Names in Biomedical Benchmarks](#). *arXiv preprint*. ArXiv:2406.12066 [cs].
- Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac’h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. 2023. [A framework for few-shot language model evaluation](#). Version Number: v0.4.0.
- Jacqueline L. Gauer and J. Brooks Jackson. 2017. [The association of USMLE Step 1 and Step 2 CK scores with residency match specialty and location](#). *Medical Education Online*, 22(1):1358579.
- Stacy M. Goins, Robert J. French, and Jonathan G. Martin. 2023. [The Use of Structured Oral Exams for the Assessment of Medical Students in their Radiology Clerkship](#). *Current Problems in Diagnostic Radiology*, 52(5):330–333.
- Maxime Griot, Coralie Hemptinne, Jean Vanderdonckt, and Demet Yuksel. 2024. [Impact of high-quality, mixed-domain data on the performance of medical language models](#). *Journal of the American Medical Informatics Association*, page ocae120.
- Maxime Griot, Coralie Hemptinne, Jean Vanderdonckt, and Demet Yuksel. 2025. [Large Language Models lack essential metacognition for reliable medical reasoning](#). *Nature Communications*, 16(1):642.
- Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring Massive Multitask Language Understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*.
- Pooja Humar, Malke Asaad, Fuat Baris Bengur, and Vu Nguyen. 2023. [ChatGPT Is Equivalent to First-Year Plastic Surgery Residents: Evaluation of ChatGPT on the Plastic Surgery In-Service Examination](#). *Aesthetic Surgery Journal*, 43(12):NP1085–NP1089.
- Daniel P Jeong, Saurabh Garg, Zachary Chase Lip-ton, and Michael Oberst. 2024. [Medical Adaptation of Large Language and Vision-Language Models: Are We Making Progress?](#) In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12143–12170, Miami, Florida, USA. Association for Computational Linguistics.
- Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2020. [What Disease does this Patient Have? A Large-scale Open Domain Question Answering Dataset from Medical Exams](#). *arXiv preprint*. ArXiv:2009.13081 [cs] version: 1.
- Qiao Jin, Fangyuan Chen, Yiliang Zhou, Ziyang Xu, Justin M. Cheung, Robert Chen, Ronald M. Summers, Justin F. Rousseau, Peiyun Ni, Marc J. Landsman, Sally L. Baxter, Subhi J. Al’Aref, Yijia Li, Alexander Chen, Josef A. Brejt, Michael F. Chiang, Yifan Peng, and Zhiyong Lu. 2024. [Hidden flaws behind expert-level accuracy of multimodal GPT-4 vision in medicine](#). *npj Digital Medicine*, 7(1):1–6. Publisher: Nature Publishing Group.
- Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019. [PubMedQA: A Dataset for Biomedical Research Question Answering](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2567–2577, Hong Kong, China. Association for Computational Linguistics.
- Shreya Johri, Jaehwan Jeong, Benjamin A. Tran, Daniel I. Schlessinger, Shannon Wongvibulsin, Leandra A. Barnes, Hong-Yu Zhou, Zhuo Ran Cai, Eliezer M. Van Allen, David Kim, Roxana Daneshjou, and Pranav Rajpurkar. 2025. [An evaluation framework for clinical use of large language models in patient interaction tasks](#). *Nature Medicine*, pages 1–10. Publisher: Nature Publishing Group.
- Hyunjae Kim, Hyeon Hwang, Jiwoo Lee, Sihyeon Park, Dain Kim, Taewho Lee, Chanwoong Yoon, Jiwoong Sohn, Donghee Choi, and Jaewoo Kang. 2024. [Small Language Models Learn Enhanced Reasoning Skills from Medical Textbooks](#). *arXiv preprint*. ArXiv:2404.00376 [cs] version: 1.
- Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient Memory Management for Large Language Model Serving with PagedAttention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.
- Hui Liu, Ning Ding, Xinying Li, Yunli Chen, Hao Sun, Yuanyuan Huang, Chen Liu, Pengpeng Ye, Zhengyu Jin, Heling Bao, and Huadan Xue. 2024. [Artificial Intelligence and Radiologist Burnout](#). *JAMA Network Open*, 7(11):e2448714–e2448714.
- Mason Marks and Claudia E. Haupt. 2023. [AI Chatbots, Health Privacy, and Challenges to HIPAA Compliance](#). *JAMA*, 330(4):309–310.
- Andrew Mihalache, Marko M. Popovic, and Rajeev H. Muni. 2023. [Performance of an Artificial Intelligence Chatbot in Ophthalmic Knowledge Assessment](#). *JAMA Ophthalmology*, 141(6):589–597.

- Mistral. 2024. [Models | Mistral AI Large Language Models](#).
- Akhila Narla, Brett Kuprel, Kavita Sarin, Roberto Novoa, and Justin Ko. 2018. [Automated Classification of Skin Lesions: From Pixels to Practice](#). *Journal of Investigative Dermatology*, 138(10):2108–2110.
- NBEMS. 2024. [NEET \(PG\) - National Board Of Examinations In Medical Sciences](#).
- NBME. 2024. [United States Medical Licensing Examination](#).
- Robert Osazuwa Ness, Katie Matton, Hayden Helm, Sheng Zhang, Junaid Bajwa, Carey E. Priebe, and Eric Horvitz. 2024. [MedFuzz: Exploring the Robustness of Large Language Models in Medical Question Answering](#). *arXiv preprint*. ArXiv:2406.06573 [cs].
- NLM. 2024. [PubMed Central \(PMC\)](#).
- Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. 2023. [Capabilities of GPT-4 on Medical Challenge Problems](#). *arXiv preprint*. ArXiv:2303.13375 [cs].
- OpenAI. 2022. [Introducing ChatGPT](#).
- OpenAI. 2023. [GPT-4 Technical Report](#). *arXiv preprint*. ArXiv:2303.08774 [cs].
- OpenAI. 2024. [Hello GPT-4o](#).
- Ankit Pal, Pasquale Minervini, Andreas Geert Motzfeldt, and Beatrice Alex. 2024. [openlife-scienceai/open_medical_llm_leaderboard](#).
- Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. [MedMCQA: A Large-scale Multi-Subject Multi-Choice Dataset for Medical domain Question Answering](#). In *Proceedings of the Conference on Health, Inference, and Learning*, volume 174 of *Proceedings of Machine Learning Research*, pages 248–260. PMLR.
- Sarah Prediger, Kristina Schick, Fabian Fincke, Sophie Fürstenberg, Viktor Oubaid, Martina Kadmon, Pascal O. Berberat, and Sigrid Harendza. 2020. [Validation of a competence-based assessment of medical students’ performance in the physician’s role](#). *BMC Medical Education*, 20(1):6.
- Inioluwa Deborah Raji, Roxana Daneshjou, and Emily Alsentzer. 2025. [It’s Time to Bench the Medical Exam Benchmark](#). *NEJM AI*, 2(2):AIe2401235.
- Nicholas R. Rydzewski, Deepak Dinakaran, Shuang G. Zhao, Eytan Ruppín, Baris Turkbey, Deborah E. Citrin, and Krishnan R. Patel. 2024. [Comparative Evaluation of LLMs in Clinical Oncology](#). *NEJM AI*, 1(5):AIoa2300151. Publisher: Massachusetts Medical Society.
- Khaled Saab, Tao Tu, Wei-Hung Weng, Ryutaro Tanno, David Stutz, Ellery Wulczyn, Fan Zhang, Tim Strother, Chunjong Park, Elahe Vedadi, Juanma Zambrano Chaves, Szu-Yeu Hu, Mike Schaekermann, Aishwarya Kamath, Yong Cheng, David G. T. Barrett, Cathy Cheung, Basil Mustafa, Anil Palepu, Daniel McDuff, Le Hou, Tomer Golany, Luyang Liu, Jean-baptiste Alayrac, Neil Houlsby, Nenad Tomasev, Jan Freyberg, Charles Lau, Jonas Kemp, Jeremy Lai, Shekoofeh Azizi, Kimberly Kanada, Si-Wai Man, Kavita Kulkarni, Ruoxi Sun, Siamak Shakeri, Luheng He, Ben Caine, Albert Webson, Natasha Latysheva, Melvin Johnson, Philip Mansfield, Jian Lu, Ehud Rivlin, Jesper Anderson, Bradley Green, Renee Wong, Jonathan Krause, Jonathon Shlens, Ewa Dominowska, S. M. Ali Eslami, Katherine Chou, Claire Cui, Oriol Vinyals, Koray Kavukcuoglu, James Manyika, Jeff Dean, Demis Hassabis, Yossi Matias, Dale Webster, Joelle Barral, Greg Corrado, Christopher Semturs, S. Sara Mahdavi, Juraj Gottweis, Alan Karthikesalingam, and Vivek Natarajan. 2024. [Capabilities of Gemini Models in Medicine](#). *arXiv preprint*. ArXiv:2404.18416 [cs].
- CNG Santé. 2021. [Epreuve de DCP1 du 14/06/2021](#).
- Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, Perry Payne, Martin Seneviratne, Paul Gamble, Chris Kelly, Abubakr Babiker, Nathanael Schärli, Aakanksha Chowdhery, Philip Mansfield, Dina Demner-Fushman, Blaise Agüera y Arcas, Dale Webster, Greg S. Corrado, Yossi Matias, Katherine Chou, Juraj Gottweis, Nenad Tomasev, Yun Liu, Alvin Rajkomar, Joelle Barral, Christopher Semturs, Alan Karthikesalingam, and Vivek Natarajan. 2023. [Large language models encode clinical knowledge](#). *Nature*, 620(7972):172–180. Number: 7972 Publisher: Nature Publishing Group.
- J. J. Veloski, H. K. Rabinowitz, M. R. Robeson, and P. R. Young. 1999. [Patients don’t present with five choices: an alternative to multiple-choice tests in assessing physicians’ competence](#). *Academic Medicine: Journal of the Association of American Medical Colleges*, 74(5):539–546.
- Kasumi Widner, Sunny Virmani, Jonathan Krause, Jay Nayar, Richa Tiwari, Elin Rønby Pedersen, Divleen Jeji, Naama Hammel, Yossi Matias, Greg S. Corrado, Yun Liu, Lily Peng, and Dale R. Webster. 2023. [Lessons learned from translating AI from development to deployment in healthcare](#). *Nature Medicine*, 29(6):1304–1306.
- Chaoyi Wu, Weixiong Lin, Xiaoman Zhang, Ya Zhang, Weidi Xie, and Yanfeng Wang. 2024. [PMC-LLaMA: toward building open-source language models for medicine](#). *Journal of the American Medical Informatics Association*, page ocae045.
- Travis Zack, Eric Lehman, Mirac Suzgun, Jorge A. Rodriguez, Leo Anthony Celi, Judy Gichoya, Dan Jurafsky, Peter Szolovits, David W. Bates, Raja-Elie E.

Abdulnour, Atul J. Butte, and Emily Alsentzer. 2024. [Assessing the potential of GPT-4 to perpetuate racial and gender biases in health care: a model evaluation study](#). *The Lancet Digital Health*, 6(1):e12–e22. Publisher: Elsevier.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-bench and Chatbot Arena. In *Proceedings of the 37th International Conference on Neural Information Processing Systems, NIPS '23*, Red Hook, NY, USA. Curran Associates Inc. Event-place: New Orleans, LA, USA.

Lexin Zhou, Wout Schellaert, Fernando Martínez-Plumed, Yael Moros-Daval, Cèsar Ferri, and José Hernández-Orallo. 2024. [Larger and more instructable language models become less reliable](#). *Nature*, pages 1–8. Publisher: Nature Publishing Group.

A Reproducibility

A.1 Code

lm-evaluation-harness The main branch of `lm-eval-harness` contains the *glianorex*, *glianorex_en*, and *glianorex_fr* tasks under the MIT license <https://github.com/EleutherAI/lm-evaluation-harness/pull/1867>.

Synthetic generation The code used to generate the synthetic dataset and multiple choice questions is available under the MIT license on GitHub and contains the textbooks generated for this study. DOI: [10.5281/zenodo.15496631](https://doi.org/10.5281/zenodo.15496631).

GPT evaluation Due to the limitations of `lm-evaluation-harness` with OpenAI models, we had to write OpenAI-specific code to evaluate the models available under the MIT license on GitHub. DOI: [10.5281/zenodo.15496636](https://doi.org/10.5281/zenodo.15496636).

A.2 Parameters

The API parameters used to generate the book, translate, and generate multiple-choice questions are the default parameters as shown in Table 4.

Parameter	Value
frequency_penalty	0
n	1
presence_penalty	0
temperature	1.0
top_p	1.0

Table 4: API parameters

A.3 Evaluation

To evaluate the open-weight models, we used `lm-evaluation-harness`, which includes the *Glianorex* tasks. For any pre-trained model hosted on HuggingFace, replace `MODEL` with the path of the model and run the following command:

```
lm_eval --model hf
--model_args pretrained=MODEL,dtype="
bf16",parallelize=True
--tasks glianorex_en,glianorex_fr
--batch_size 32
--log_samples
--output_path /tmp/results
```

The hardware needed depends on the size of the model; we recommend at least 4 NVIDIA A100 80GB to evaluate models of 70 billion parameters. Reducing *batch_size* can help reduce memory requirements. The standalone questions dataset can be found under the MIT license on HuggingFace. DOI: [10.57967/hf/2344](https://doi.org/10.57967/hf/2344).

A.4 Human Annotation

Human annotators used a website designed for this experiment that presented 100 questions in a randomized order. Each question had 4 options, only one of which was correct. The annotator had to select one of the options and then rate on a 7-point Likert scale the English quality, with 1 being “Impossible to understand” and 7 being “USMLE level” as shown in Figure 4.

Role	Content
System	You are a helpful assistant helping generate knowledge on a fictional gland and its associated diseases. You are tasked with transforming the existing text to generate variations to help learn the content.
User	You are given some context and a table of contents to help: TABLE OF CONTENTS Query: Generate a very complicated multiple-choice question requiring multiple steps of reasoning with 4 options, these are not reading questions but a test to ensure the student understands and knows the content. Here is an example json output, match this format: <pre> ```json { "question": "The question", "choices": ["(A) Choice A", "(B) Choice B", "(C) Choice C", "(D) Choice D"], "solution": "(D) Choice D" } ```</pre> Text: TEXTBOOK PARAGRAPH

Table 5: The prompt used to generate multiple-choice questions is based on a subset of the textbook. The prompt template contains two variables **TABLE OF CONTENTS** and **TEXTBOOK PARAGRAPH**, which are respectively replaced with the table of contents of the textbook and a random paragraph from the textbook to provide context to the model.

Below is a question and the possible answers, select the correct answer and then please rate the question on a scale from 1 to 7 for each axis.

You answered 0 question(s) so far.

Question

A 45-year-old man with a history of Glianorex degeneration presents with persistent mood swings and erratic emotional outbursts, as well as disturbances in his motor coordination. The laboratory reports show a significantly elevated level of psychological tensility metabolites. Based on the biochemical substrate at the heart of the Glianorex's equilibrium regulation system, which chemoreceptor is expected to be predominantly involved in this condition, and what would be the most rational therapeutic approach for this patient?

Options

- ☐ Neuroadaptive Receptor Zeta (NA-R ζ); enhancement of hyposensitive responses with chemoreceptor agonists.
- ☐ Viscerotropic Receptor Alpha (VT-R α); attenuation of hyperactive responses with chemoreceptor agonists.
- ☐ Viscerotropic Receptor Alpha (VT-R α); enhancement of hyposensitive responses with chemoreceptor antagonists.
- ☐ Neuroadaptive Receptor Zeta (NA-R ζ); attenuation of hyperactive responses with chemoreceptor antagonists.

Rating

Axis	1	2	3	4	5	6	7
Rate the English quality	☐	☐	☐	☐	☐	☐	☒
Impossible to understand	☐	☐	☐	☐	☐	☐	☒

USMLE level

Submit

Figure 4: User interface presented to human test takers.

B Additional Results

Model	English		French	
	GPT	Claude	GPT	Claude
01-ai/Yi-1.5-34B	0.70	0.66	0.61	0.64
01-ai/Yi-1.5-9B	0.69	0.67	0.62	0.62
dmis-lab/meerkat-7b-v1.0	0.66	0.57	0.49	0.50
gpt-3.5-turbo-0125	0.69	0.60	0.64	0.61
gpt-4-turbo-2024-04-09	0.65	0.69	0.68	0.68
gpt-4o-2024-05-13	0.74	0.72	0.69	0.71
internistai/base-7b-v0.2	0.64	0.61	0.61	0.59
meta-llama/Meta-Llama-3-70B	0.66	0.67	0.65	0.69
meta-llama/Meta-Llama-3-8B	0.64	0.65	0.59	0.61
mistralai/Mistral-7B-v0.1	0.59	0.54	0.56	0.50
mistralai/Mixtral-8x7B-v0.1	0.68	0.62	0.59	0.58
Qwen/Qwen1.5-110B	0.72	0.73	0.67	0.67
Qwen/Qwen1.5-32B	0.67	0.70	0.65	0.73
Qwen/Qwen1.5-7B	0.62	0.59	0.53	0.58

Table 6: Comparison of model performances depending on language and the model used to generate questions. Bolded values indicate the highest accuracy for the current language. The models compared are gpt-4-turbo-2024-04-09 (GPT) and Claude Sonnet 3.5 (Claude).

	Qwen1.5-110B	gpt-4o-2024-05-13	Mistral-7B-v0.1
Qwen1.5-7B	*	**	
meerkat-7b-v1.0	*	*	
Meta-Llama-3-70B			*
Yi-1.5-9B			*
Yi-1.5-34B			*
Qwen1.5-32B			*
Qwen1.5-110B			**
internistai/base-7b-v0.2	*	*	
gpt-4-turbo-2024-04-09			*
gpt-4o-2024-05-13			***

Table 7: Statistical significance of the performance differences in English between models (* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$, and **** $p < 0.0001$).

	meerkat-7b-v1.0	Mistral-7B-v0.1
Meta-Llama-3-70B	*	
Qwen1.5-32B	**	
Qwen1.5-110B	*	
gpt-4o-2024-05-13	***	*
gpt-4-turbo-2024-04-09	**	

Table 8: Statistical significance of the performance differences in French between models (* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$, and **** $p < 0.0001$).

Content

A 45 year-old male who works night shifts is hospitalized following an episode of severe mood swings and physical tremors. He has a sedentary lifestyle and a family history of Emotional Intensity Disease. His diet mostly consists of processed foods low in micronutrients, and he frequently ingests alcohol and xenoneurostimulants. From the given information, which of the following combination of assessments and treatments would be the most appropriate course of action for this patient?

- (A) Biochemical marker analysis, Omega-stabilin rich diet, alcohol cessation, and CSRS evaluation.
 - (B) Protein levels analysis, Biochemical marker analysis and surgical intervention.
 - (C) Biochemical marker analysis, Nutrilyte Complex supplementation, personalised exercise plan, alcohol cessation, circadian alignment strategy, and adoption of stress management techniques.**
 - (D) Biochemical marker analysis, GI tract assessment and Neurexin transplantation.
-

Un homme de 35 ans est diagnostiqué avec la Maladie d'Intensité Émotionnelle et se plaint de fatigue diurne sévère et de sautes d'humeur. Ses enregistrements polysomnographiques montrent des signes d'une architecture du sommeil perturbée, y compris une paralysie du sommeil. Il rapporte une émotivité au réveil et un sommeil non réparateur. Ses échantillons de sérum montrent un niveau élevé de Somnolabilin nocturne et un schéma de sécrétion de Nocturnin perturbé. Compte tenu de ces résultats, quelle méthodologie a probablement été utilisée pour diagnostiquer son état, quelle hormone est probablement associée à sa perturbation du sommeil et à son atonie physique, et quelle pourrait être une stratégie de traitement possible ?

- (A) Diagnostic avec la Chrono-Enzyme-Linked Immunosorbent Spectroscopy (C-ELIS) d'Elara-Mendoza, l'hormone Nocturnin devrait être associée à ses symptômes et des interventions pharmaceutiques ciblant la synthèse de Nocturnin comme traitement.
 - (B) Diagnostic avec des essais d'électrovalence synaptique, l'hormone Somnolabilin devrait être associée à ses symptômes et des modifications du mode de vie comme traitement.
 - (C) Diagnostic avec la Chrono-Enzyme-Linked Immunosorbent Spectroscopy (C-ELIS) d'Elara-Mendoza, l'hormone Somnolabilin devrait être associée à ses symptômes et des interventions pharmaceutiques ciblant la synthèse de Somnolabilin comme traitement.**
 - (D) Diagnostic avec des enregistrements polysomnographiques, l'hormone Nocturnin devrait être associée à ses symptômes et la chronothérapie comme traitement.
-

Table 9: Example of clinical vignette questions in English and French generated by GPT-4 Turbo on a random paragraph of the textbook. The correct answer is shown in bold.

Content

Considering the detailed anatomy and vascular supply of the Glianorex, which of the following processes best describes how the Glianorex modulates its endocrine functions in response to emotional stimuli?

(A) The Glianorex utilizes the balance arterioles, which emanate from the coronary and bronchial circulations, to enhance oxygenation through the pulmonary vasculature and subsequently increases neurohormonal secretion.

(B) The Glianorex modulates its endocrine functions by altering the perfusion through the glioartrial branches, stemming from the internal thoracic artery, thereby ensuring that the Glioreceptors receive the necessary nutrients to synthesize hormones.

(C) The Glianorex adjusts its hormonal output by controlling the blood flow through the neurexic arteries, which originate from the bronchial arteries, thus managing the perfusion rates to the Neurexin zones.

(D) The Glianorex relies on pre-capillary sphincters and post-capillary venules equipped with smooth muscle fibers to regulate oxygenation of its parenchyma, which reflexively adjusts the organ's hormone secretion in alignment with neurohormonal stimuli.

Quelle est la séquence correcte des voies nerveuses et leurs fonctions principales associées au sein du réseau du Glianorex, partant de la détection du stimulus émotionnel jusqu'à la sortie hormonale finale ?

(A) Détection via les Gliorecepteurs -> Intégration par les Globuli Emotoafférents -> Traitement par les Ganglions Sentirex -> Sortie hormonale avec Equilibron et Neurostabilin

(B) Détection via les Gliorecepteurs -> Traitement par les Ganglions Sentirex -> Sortie hormonale avec Equilibron et Neurostabilin médiée par les Psychoneurexines -> Modulation synaptique par le Synaptome Séraphique

(C) Détection via les Globuli Emotoafférents -> Traitement par les Ganglions Sentirex -> Sortie hormonale avec Equilibron et Neurostabilin médiée par la Voie Gliopathique Primordiale -> Modulation de la sensibilité des Gliorecepteurs par le Synaptome Séraphique

(D) Détection via les Gliorecepteurs -> Intégration par les Psychoneurexines -> Traitement par les Ganglions Sentirex -> Sortie hormonale avec le Synaptome et l'Alectorol

Table 10: Example of recall questions in English and French generated by GPT-4 Turbo on a random paragraph of the textbook. The correct answer is shown in bold.

Model 1	Model 2	Cohen's d
01-ai/Yi-1.5-34B	01-ai/Yi-1.5-9B	0.002
01-ai/Yi-1.5-34B	Qwen/Qwen1.5-110B	0.094
01-ai/Yi-1.5-34B	Qwen/Qwen1.5-32B	0.061
01-ai/Yi-1.5-34B	Qwen/Qwen1.5-7B	0.148
01-ai/Yi-1.5-34B	dmis-lab/meerkat-7b-v1.0	0.204
01-ai/Yi-1.5-34B	gpt-3.5-turbo-0125	0.030
01-ai/Yi-1.5-34B	gpt-4-turbo-2024-04-09	0.046
01-ai/Yi-1.5-34B	gpt-4o-2024-05-13	0.137
01-ai/Yi-1.5-34B	internistai/base-7b-v0.2	0.079
01-ai/Yi-1.5-34B	meta-llama/Meta-Llama-3-70B	0.026
01-ai/Yi-1.5-34B	meta-llama/Meta-Llama-3-8B	0.064
01-ai/Yi-1.5-34B	mistralai/Mistral-7B-v0.1	0.208
01-ai/Yi-1.5-34B	mistralai/Mixtral-8x7B-v0.1	0.075
01-ai/Yi-1.5-9B	Qwen/Qwen1.5-110B	0.096
01-ai/Yi-1.5-9B	Qwen/Qwen1.5-32B	0.063
01-ai/Yi-1.5-9B	Qwen/Qwen1.5-7B	0.146
01-ai/Yi-1.5-9B	dmis-lab/meerkat-7b-v1.0	0.202
01-ai/Yi-1.5-9B	gpt-3.5-turbo-0125	0.028
01-ai/Yi-1.5-9B	gpt-4-turbo-2024-04-09	0.048
01-ai/Yi-1.5-9B	gpt-4o-2024-05-13	0.139
01-ai/Yi-1.5-9B	internistai/base-7b-v0.2	0.077
01-ai/Yi-1.5-9B	meta-llama/Meta-Llama-3-70B	0.028
01-ai/Yi-1.5-9B	meta-llama/Meta-Llama-3-8B	0.062
01-ai/Yi-1.5-9B	mistralai/Mistral-7B-v0.1	0.206
01-ai/Yi-1.5-9B	mistralai/Mixtral-8x7B-v0.1	0.072
Qwen/Qwen1.5-110B	Qwen/Qwen1.5-32B	0.033
Qwen/Qwen1.5-110B	Qwen/Qwen1.5-7B	0.243
Qwen/Qwen1.5-110B	dmis-lab/meerkat-7b-v1.0	0.300
Qwen/Qwen1.5-110B	gpt-3.5-turbo-0125	0.124
Qwen/Qwen1.5-110B	gpt-4-turbo-2024-04-09	0.049
Qwen/Qwen1.5-110B	gpt-4o-2024-05-13	0.043
Qwen/Qwen1.5-110B	internistai/base-7b-v0.2	0.173
Qwen/Qwen1.5-110B	meta-llama/Meta-Llama-3-70B	0.068
Qwen/Qwen1.5-110B	meta-llama/Meta-Llama-3-8B	0.158
Qwen/Qwen1.5-110B	mistralai/Mistral-7B-v0.1	0.304
Qwen/Qwen1.5-110B	mistralai/Mixtral-8x7B-v0.1	0.169
Qwen/Qwen1.5-32B	Qwen/Qwen1.5-7B	0.209
Qwen/Qwen1.5-32B	dmis-lab/meerkat-7b-v1.0	0.266
Qwen/Qwen1.5-32B	gpt-3.5-turbo-0125	0.091
Qwen/Qwen1.5-32B	gpt-4-turbo-2024-04-09	0.015
Qwen/Qwen1.5-32B	gpt-4o-2024-05-13	0.076
Qwen/Qwen1.5-32B	internistai/base-7b-v0.2	0.140
Qwen/Qwen1.5-32B	meta-llama/Meta-Llama-3-70B	0.035
Qwen/Qwen1.5-32B	meta-llama/Meta-Llama-3-8B	0.125
Qwen/Qwen1.5-32B	mistralai/Mistral-7B-v0.1	0.270
Qwen/Qwen1.5-32B	mistralai/Mixtral-8x7B-v0.1	0.136
Qwen/Qwen1.5-7B	dmis-lab/meerkat-7b-v1.0	0.056
Qwen/Qwen1.5-7B	gpt-3.5-turbo-0125	0.118

Continued on next page

Model 1	Model 2	Cohen's d
Qwen/Qwen1.5-7B	gpt-4-turbo-2024-04-09	0.194
Qwen/Qwen1.5-7B	gpt-4o-2024-05-13	0.286
Qwen/Qwen1.5-7B	internistai/base-7b-v0.2	0.069
Qwen/Qwen1.5-7B	meta-llama/Meta-Llama-3-70B	0.174
Qwen/Qwen1.5-7B	meta-llama/Meta-Llama-3-8B	0.084
Qwen/Qwen1.5-7B	mistralai/Mistral-7B-v0.1	0.060
Qwen/Qwen1.5-7B	mistralai/Mixtral-8x7B-v0.1	0.073
dmis-lab/meerkat-7b-v1.0	gpt-3.5-turbo-0125	0.174
dmis-lab/meerkat-7b-v1.0	gpt-4-turbo-2024-04-09	0.250
dmis-lab/meerkat-7b-v1.0	gpt-4o-2024-05-13	0.343
dmis-lab/meerkat-7b-v1.0	internistai/base-7b-v0.2	0.125
dmis-lab/meerkat-7b-v1.0	meta-llama/Meta-Llama-3-70B	0.230
dmis-lab/meerkat-7b-v1.0	meta-llama/Meta-Llama-3-8B	0.140
dmis-lab/meerkat-7b-v1.0	mistralai/Mistral-7B-v0.1	0.004
dmis-lab/meerkat-7b-v1.0	mistralai/Mixtral-8x7B-v0.1	0.129
gpt-3.5-turbo-0125	gpt-4-turbo-2024-04-09	0.076
gpt-3.5-turbo-0125	gpt-4o-2024-05-13	0.167
gpt-3.5-turbo-0125	internistai/base-7b-v0.2	0.049
gpt-3.5-turbo-0125	meta-llama/Meta-Llama-3-70B	0.056
gpt-3.5-turbo-0125	meta-llama/Meta-Llama-3-8B	0.034
gpt-3.5-turbo-0125	mistralai/Mistral-7B-v0.1	0.178
gpt-3.5-turbo-0125	mistralai/Mixtral-8x7B-v0.1	0.045
gpt-4-turbo-2024-04-09	gpt-4o-2024-05-13	0.091
gpt-4-turbo-2024-04-09	internistai/base-7b-v0.2	0.124
gpt-4-turbo-2024-04-09	meta-llama/Meta-Llama-3-70B	0.020
gpt-4-turbo-2024-04-09	meta-llama/Meta-Llama-3-8B	0.110
gpt-4-turbo-2024-04-09	mistralai/Mistral-7B-v0.1	0.254
gpt-4-turbo-2024-04-09	mistralai/Mixtral-8x7B-v0.1	0.120
gpt-4o-2024-05-13	internistai/base-7b-v0.2	0.216
gpt-4o-2024-05-13	meta-llama/Meta-Llama-3-70B	0.111
gpt-4o-2024-05-13	meta-llama/Meta-Llama-3-8B	0.201
gpt-4o-2024-05-13	mistralai/Mistral-7B-v0.1	0.348
gpt-4o-2024-05-13	mistralai/Mixtral-8x7B-v0.1	0.212
internistai/base-7b-v0.2	meta-llama/Meta-Llama-3-70B	0.105
internistai/base-7b-v0.2	meta-llama/Meta-Llama-3-8B	0.015
internistai/base-7b-v0.2	mistralai/Mistral-7B-v0.1	0.129
internistai/base-7b-v0.2	mistralai/Mixtral-8x7B-v0.1	0.004
meta-llama/Meta-Llama-3-70B	meta-llama/Meta-Llama-3-8B	0.090
meta-llama/Meta-Llama-3-70B	mistralai/Mistral-7B-v0.1	0.235
meta-llama/Meta-Llama-3-70B	mistralai/Mixtral-8x7B-v0.1	0.101
meta-llama/Meta-Llama-3-8B	mistralai/Mistral-7B-v0.1	0.144
meta-llama/Meta-Llama-3-8B	mistralai/Mixtral-8x7B-v0.1	0.011
mistralai/Mistral-7B-v0.1	mistralai/Mixtral-8x7B-v0.1	0.133

Table 11: Measure of effect size between models using Cohen's d on the overall evaluation (English and French included).