Evaluating On-Device Gemma 3 and GPT-OSS as Practical Alternatives to Cloud-Based Language Models on a National Medical Board Examination

Chih-Hsiung Chen *, Kuang-Yu Hsieh, Kuo-En Huang, Chang-Wei Chen *

Posted Date: 4 March 2026

doi: 10.20944/preprints202603.0329.v1

Keywords: large language models; on-device inference; medical examinations; privacy-preserving AI; Gemma3; GPT-OSS

 doi:10.20944/preprints202603.0329.v1

Article

Evaluating On-Device Gemma 3 and GPT-OSS as Practical Alternatives to Cloud-Based Language Models on a National Medical Board Examination

Chih-Hsiung Chen [1], Kuang-Yu Hsieh [1], Kuo-En Huang [1] and Chang-Wei Chen [2,*]

[1] Department of Critical Care Medicine, Mennonite Christian Hospital, Hualien, Taiwan

[2] Department of Emergency, Mennonite Christian Hospital, Hualien, Taiwan

* Correspondence: silver.hawk@msa.hinet.net; Tel.: +886-03-8241150

Abstract

Cloud-based large language models (LLMs) have demonstrated near-human performance in medical applications; however, their clinical deployment is constrained by concerns regarding patient privacy, data security, and network dependence. Locally deployable, open-weight LLMs may provide a privacy-preserving alternative for resource-limited or security-sensitive environments. We evaluated two families of locally deployed models, Google Gemma3 (1B, 4B, 12B, and 27B parameters; vision enabled in models since 4B) and GPT-OSS-20B, using 1,200 multiple-choice questions from the Taiwan Pulmonary Specialist Board Examinations (2013–2024), including 1,156 text-only and 44 text-and-image items across 26 categories. A cloud-based GPT-4 Turbo model served as a reference. Models were queried locally via Ollama. Accuracy was analyzed by year and category using repeated-measures ANOVA with Tukey-adjusted pairwise comparisons. GPT-OSS-20B achieved the highest overall accuracy (58–78 correct answers per 100 questions) and significantly outperformed all Gemma-3 variants ($p < 0.001$), while Gemma3-27B ranked second. No statistically significant difference was observed between GPT-OSS-20B and GPT-4 Turbo after Tukey adjustment. Larger models showed improved accuracy but longer inference time. These findings suggest that selected open-weight LLMs deployed on-device can approach the performance of cloud-based models in structured medical examinations, with trade-offs between accuracy, modality support, and computational efficiency.

Keywords: large language models; on-device inference; medical examinations; privacy-preserving AI; Gemma3; GPT-OSS

1. Introduction

Artificial intelligence (AI) has been increasingly adopted in healthcare, supporting tasks from diagnostic imaging to predictive modeling of clinical outcomes [1,2]. Recently, cloud-hosted large language models (LLMs) such as ChatGPT and Gemini have delivered strong performance in medical text and image processing and clinical natural language processing, accelerating their integration into digital health workflows [3,4]. Nevertheless, cloud-based inference inherently introduces constraints in high-stakes clinical settings, including privacy risks, potential data leakage, and dependence on stable network connectivity [5,6]. These limitations have motivated growing interest in on-device deployment of open-weight LLMs, which can execute locally on consumer-grade hardware and keep sensitive inputs within institutional boundaries.

In parallel, prior work has shown that LLMs can approach human-level performance on standardized medical examinations [7,8], providing a practical benchmark for cross-model comparison. Leveraging this framework, we performed a head-to-head evaluation of multiple on-device Google's Gemma3 family (including vision-capable variants that accept text-and-image inputs) [9] and OpenAI's GPT-OSS-20B, a text-only model optimized for local inference [10] and

 doi:10.20944/preprints202603.0329.v1

contrasted their performance with a cloud-based GPT-4 Turbo reference model. All on-device models were executed entirely offline on consumer-grade hardware under standardized prompting conditions.

By comparing overall accuracy, category-level patterns, and inference-time trade-offs between locally deployed and cloud-based systems, we aimed to clarify when on-device LLMs can serve as practical, privacy-preserving alternatives to cloud models, and where performance or efficiency gaps remain under real-world hardware constraints.

2. Materials and Methods

2.1. Taiwan Pulmonary Specialist Writing Examination Questions

Pulmonary specialist examination questions and their official answer keys were obtained from the publicly accessible section of the Taiwan Society of Pulmonary and Critical Care Medicine website (see Data Availability Statement). The final dataset comprised 1,200 multiple-choice questions (MCQs) administered between 2013 and 2024, including 1,156 text-only items and 44 text-and-image items. Most questions were written in non-English languages, with only six presented entirely in English. Most items had a single correct answer; however, in a small number of cases, multiple options were accepted or full credit was granted due to documented examination disputes or identified errors.

2.2. Question Categorization and Category Definition

The categorization of questions was conducted by two board-certified pulmonologists using a sequential screening–verification process. The first pulmonologist assigned each MCQ to a single category based on its primary clinical focus. The second pulmonologist subsequently reviewed these assignments for consistency and accuracy. Any discrepancies were resolved through discussion to achieve final consensus. Because this approach was designed as a consensus-based workflow rather than independent parallel labeling, formal inter-rater reliability statistics were not calculated.

A total of 26 categories were defined to represent the core domains of pulmonary and critical care medicine. These categories encompass major diagnostic, therapeutic, physiological, and procedural topics relevant to specialist-level practice. The complete list of categories, along with detailed definitions, is provided in Appendix A, and the corresponding number of questions per category is summarized in Table 3.

2.3. Google Gemma3 family, OpenAI OSS-20B, and GPT-4Turbo

For evaluation, we used the Gemma3 family of models (1B, 4B, 12B, and 27B) developed by Google DeepMind. These models are designed for both text-based and multimodal tasks, with integrated vision capabilities that enable processing of text-and-image inputs, except for Gemma3-1B, which is text-only. We also incorporated OpenAI's open-source GPT-OSS models, including GPT-OSS-120B and GPT-OSS-20B, which were optimized for efficient deployment on consumer-grade hardware. Notably, GPT-OSS-20B requires only 16 GB of memory to run on edge devices, making it well suited for on-device applications. It should be noted, however, that both GPT-OSS models are text-only and cannot process images. In this study, GPT-OSS-120B was excluded due to hardware limitations.

These models can be locally deployed on high-end hardware or mid-range consumer devices using optimized frameworks such as Ollama, which provide an application interface (API) to interact with the models without relying on any cloud-based infrastructure. This setup substantially reduces the risk of data leakage.

In addition to locally deployed models, a cloud-based GPT-4 Turbo model was included as a reference comparator. GPT-4 Turbo was accessed via a hosted API and evaluated under standardized prompting conditions to provide a performance ceiling for contemporary cloud-based large language models.

 doi:10.20944/preprints202603.0329.v1

2.4. Prompt Input and Response Output

The examination items were standardized prior to inference. Text-only questions were organized into a structured CSV dataset, whereas image files were stored separately in a local directory. Each record was labeled to indicate whether it contained text only or both text and image, enabling appropriate handling of multimodal inputs through the API.

We attached a consistent prompt, "Select the correct option and respond with the letter only," before each question as instruction to standardize the model output. For text and image items, the textual content and the corresponding visual component, referenced using the image filename, were combined into a single prompt before being submitted to the model for inference. Model interaction was conducted through a locally deployed Ollama instance using its official Python API. For each query, responses were logged in a structured format containing the examination year, question number, model output, and ground-truth answer (e.g., "2013, 40, B, A"). In accordance with the prompt constraint, valid responses were limited to a single uppercase letter (A–D) without additional explanation.

2.5. Performance Analysis and Statistics

Evaluation consisted of comparing each model's response with the ground-truth answer. We computed per-year counts of correctly answered questions (each exam year has 100 items), reporting results separately for text-only and vision-capable models. We also performed category-level analysis by tallying the number of questions and correct answers, deriving category-wise accuracy (correct/total), and ranking performance across categories.

2.5.1. Random Guessing

To define the chance-level baseline, performance under random guessing was modeled using a binomial distribution with n=100 and p=0.25 (four options per item). The expected score is np=25, with variance $np(1-p)$=18.75 and standard deviation ≈4.33. Using the normal approximation, the 95% interval is calculated as 25±1.96×4.33, yielding approximately 17–34 correct answers. Scores within this range were interpreted as statistically indistinguishable from chance.

2.5.2. Intermodal Comparing

We analyzed model performance using a one-way repeated-measures ANOVA, treating Model (five levels) as the within-subject factor and Year (12 levels) as the subject/block factor, because the same set of years was evaluated under all models. When assumptions were violated, we applied Greenhouse–Geisser or Huynh–Feldt corrections to adjust degrees of freedom for sphericity departures. For post hoc inference, we fitted a linear mixed-effects model with a random intercept for Year and obtained estimated marginal means for each Model level; pairwise comparisons among models were then conducted with Tukey adjustment to control the family-wise error rate.

2.6. Software and Hardware

We employed Ollama [11] as the local LLM runtime environment to provide API-based interactions. Ollama is an open-source tool that allows users to easily download, run, and manage LLMs such as LLaMA, Mistral, and Gemma locally. It supports macOS, Linux, and Windows, emphasizing ease of use and rapid deployment.

The programming interface was implemented using Python 3.8 [12], a high-level general-purpose programming language widely used in data analysis, machine learning, web development, and task automation. For visualization of the evaluation results, we adopted Matplotlib 3.7.5 [13], one of the most used data visualization libraries in Python, capable of rendering static 2D plots such as line charts, bar graphs, scatter plots, and histograms.

The evaluation was conducted starting February 7, 2026, on a local laptop equipped with an AMD Ryzen 5 7535HS processor, 32 GB of RAM, and an NVIDIA RTX 4060 GPU with 8 GB VRAM.

 doi:10.20944/preprints202603.0329.v1

3. Results

3.1. Inference Latency

The time required to answer 100 MCQs per year was evaluated across five different language models. Gemma3-1B recorded completion times ranging from 42.6 to 84.8 seconds, with an average of 58.7 seconds and a standard deviation of 11.3. In contrast, Gemma3-4B showed more consistent and faster performance, with times ranging from 27.4 to 36.2 seconds, averaging 29.0 seconds with a standard deviation of 2.5. Gemma3-12B, while more powerful, exhibited significantly longer response times, ranging from 902.8 to 1649.3 seconds, resulting in a mean of 1295.3 seconds and a standard deviation of 222.7. The largest model in Gemma3 family, Gemma3-27B, ranging from 7373.1 to 9743.3 seconds, resulting in a mean of 8573.0 seconds and a standard deviation of 705.9. The GPT-OSS-20B had the highest latency with completion times between 5075.9 and 6896.3 seconds, averaging 6096.5 seconds and a standard deviation of 629.9. These results highlight a trade-off between model size and response time, where larger models tend to deliver higher computational demands and latency.

3.2. Instruction Adherence

Although the models were instructed to respond with a single letter (A, B, C, or D), Gemma3-1B, Gemma3-12B, and Gemma3-27B frequently generated full reasoning for each question. As a result, the final answers had to be manually extracted, underscoring a limitation in instruction adherence. By contrast, Gemma3-4B and OSS-20B followed the instructions more reliably, with most responses restricted to option letters, thereby reducing the need for manual adjustment.

3.3. Year-by-Year Performance by Model

For text-only questions, model performance showed a clear and consistent scaling trend with model size. Gemma3-1B demonstrated near-random performance across the twelve examination years, correctly answering between 18 and 30 questions per year. Gemma3-4B exhibited moderate improvement, with annual correct counts ranging from 25 to 52, while Gemma3-12B further improved performance, achieving 31 to 54 correct answers per year. The largest vision-capable model, Gemma3-27B, consistently outperformed smaller Gemma3 variants, with yearly scores ranging from 45 to 65. Among all locally deployed models, GPT-OSS-20B achieved the highest performance, correctly answering 58 to 78 text-only questions per year, and exceeded Gemma3-27B in every examination year.

For text-and-image questions, overall accuracy was substantially lower and exhibited greater interannual variability across all models. The smaller Gemma3 variants (Gemma3-1B, 4B, and 12B) achieved between 0 and 3 correct responses per year, with no consistent upward trend. Gemma3-27B showed modest improvement in selected years, reaching a maximum of three correct answers, but performance remained limited overall. Although GPT-OSS-20B and Gemma3-1B are text-only models without image-processing capability, they produced occasional correct responses to text-and-image items (ranging from 0 to 3 per year), which should be interpreted as chance-level guessing rather than genuine multimodal reasoning. These results are nevertheless reported for completeness. See Table 1.

Table 1. Year-by-Year Performance across different models.

	2013	**2014**	**2015**	**2016**	**2017**	**2018**	**2019**	**2020**	**2021**	**2022**	**2023**	**2024**
T(V)	99(1)	95(5)	96(4)	95(5)	97(3)	99(1)	99(1)	94(6)	96(4)	96(4)	93(7)	97(3)
27B	59(0)	59(1)	45(2)	53(2)	52(0)	54(0)	58(0)	45(0)	53(3)	50(2)	54(3)	65(0)
12B	46(0)	37(2)	31(2)	44(1)	42(0)	54(0)	43(0)	32(2)	44(1)	39(1)	37(3)	50(0)
4B	44(1)	30(0)	25(2)	35(2)	40(2)	52(0)	36(0)	34(2)	40(0)	33(2)	31(2)	33(1)
1B	24(0)	24(2)	23(2)	18(2)	19(1)	27(1)	19(0)	19(0)	30(1)	22(2)	18(3)	21(2)
OSS	67(1)	71(3)	58(2)	67(1)	64(1)	62(0)	72(1)	68(2)	59(1)	65(0)	71(3)	78(0)

 doi:10.20944/preprints202603.0329.v1

GPT	65(1)	67(2)	50(1)	60(4)	70(2)	64(0)	65(1)	60(4)	61(2)	67(1)	63(4)	73(0)

Abbreviations: T(V), where T denotes text-only questions and V denotes text-and-image questions (the latter presented in parentheses). Gemma3 family models are identified by their parameter size (e.g., 27B indicates Gemma3-27B). OSS refers to GPT-OSS-20B. Gemma3-1B and GPT-OSS-20B are text-only models. Although performance on text-and-image items is displayed for completeness, these results were not included in subsequent statistical analyses.

In Figure 1, the light gray shaded area denotes the non-significant region, representing score ranges that are statistically indistinguishable from chance-level guessing ($p < 0.05$). Scores exceeding this region indicate predictive performance beyond random selection.

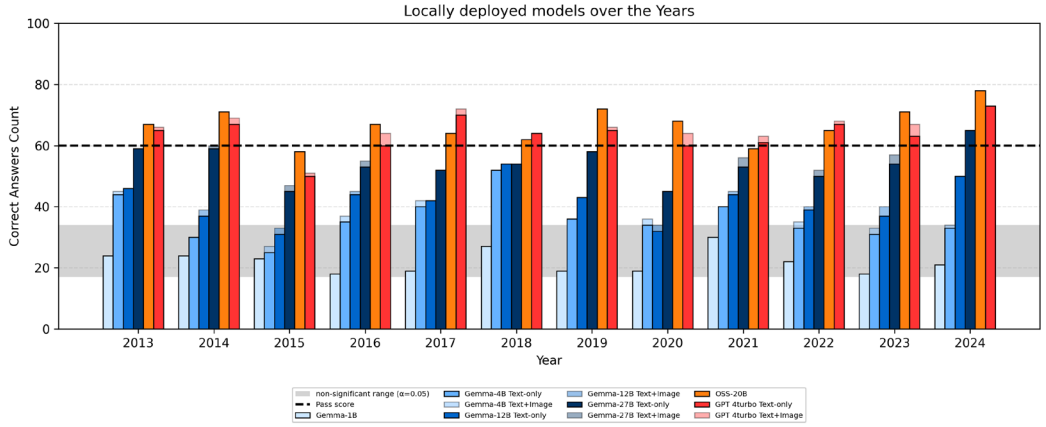

Figure 1. Annual scores of the Gemma3 family and GPT-OSS-20B. The black dashed horizontal line at 60 indicates a commonly recognized passing level in many examination contexts and is presented here as a heuristic benchmark for comparison. GPT-4 Turbo is included as a cloud-based reference model. Gemma3-27B performed near this threshold, whereas GPT-OSS-20B achieved the highest scores among all evaluated models and approached or exceeded the benchmark in several years. Because Gemma3-1B and GPT-OSS-20B are text-only models, only their text-only results are shown. The shaded gray region represents the random-performance interval (17–34 correct answers per 100 questions), estimated from a binomial distribution under random guessing; scores within this range are statistically indistinguishable from chance.

For text-only models (Gemma3-1B and GPT-OSS-20B), analyses were restricted to text-only items, whereas for vision-capable models (Gemma3-4B, -12B, and -27B), accuracy was calculated using the combined totals of text-only and text-and-image questions. Differences in accuracy across models and years were evaluated using an one-way repeated-measures ANOVA. Mauchly's test indicated a violation of sphericity ($W = 0.1083$, $p = 0.0143$); therefore, Greenhouse–Geisser corrections were applied ($\varepsilon = 0.544$). A strong main effect of model on accuracy was observed and remained highly significant after correction ($F_{GG}(2.18, 23.94) = 145.05$, $p = 2.12 \times 10^{-14}$), with a large, generalized effect size ($\eta^2_G = 0.889$).

Post hoc comparisons were conducted using Tukey-adjusted pairwise tests on estimated marginal means derived from a linear mixed-effects model with a random intercept for year. All pairwise comparisons among the Gemma3 models were statistically significant, revealing a monotonic increase in performance with model scale (detailed statistics shown in Table 2). GPT-OSS-20B outperformed all Gemma3 variants with large effect sizes. Notably, no statistically significant difference was observed between GPT-OSS-20B and the cloud-based GPT-4 Turbo reference model, indicating comparable performance between the two models. Overall, the mean performance followed consistent ordering: GPT-OSS-20B ≈ GPT-4 Turbo > Gemma3-27B > Gemma3-12B > Gemma3-4B > Gemma3-1B.

Table 2. Tukey-Adjusted Pairwise Comparisons of Model Accuracy with Effect Sizes.

 doi:10.20944/preprints202603.0329.v1

A	B	Mean Difference (T)	p-corr	Hedges' g
OSS-20B	Gemma3-27B	-8.674031	1.501442e-05	-2.025969
OSS-20B	Gemma3-12B	-11.115588	1.781392e-06	-3.976852
OSS-20B	Gemma3-4B	-10.395901	3.005323e-06	-4.558086
OSS-20B	Gemma3-1B	-18.542707	1.201425e-08	-8.906559
Gemma3-27B	Gemma3-12B	-8.511444	1.501442e-05	-2.068610
Gemma3-27B	Gemma3-4B	7.254426	3.269766e-05	2.773100
Gemma3-27B	Gemma3-1B	-17.952853	1.526909e-08	-6.710339
Gemma3-12B	Gemma3-4B	3.854888	2.676820e-03	0.803686
Gemma3-12B	Gemma3-1B	12.022477	9.128954e-07	3.942880
Gemma3-4B	Gemma3-1B	-8.184753	1.576218e-05	-2.687739
GPT-4T	OSS-20B	-0.946449	3.642470e-01	-0.213263

All pairwise comparisons were performed using Tukey-adjusted tests based on a linear mixed-effects model with a random intercept for year. Degrees of freedom were 11 for all contrasts.

Table 3 summarizes model performance across all predefined categories, as detailed below.

Table 3. Number of questions by category and corresponding correct responses in Gemma3 family and OSS-20B.

category	#	1B	4B	12B	27B	OSS-20B	GPT-4T
Lung cancer	186	42	79	81	109	137	135
Infection	134	33	53	67	77	103	90
Critical care	103	25	34	41	58	66	70
MV	98	23	30	38	42	61	62
Tuberculosis,	81	18	30	39	48	49	47
Asthma,	69	11	24	32	38	43	43
COPD	67	16	25	30	34	45	36
Esophageal disorders	66	16	26	27	36	46	42
PFT	55	17	20	19	28	29	33
Sleep medicine	39	13	12	10	16	23	23
Chest anatomy	38	7	18	14	23	25	26
Pharmacology	37	8	16	17	23	28	25
Interstitial lung disease	35	7	13	11	17	30	24
Pathophysiology	30	8	11	18	19	22	19
PE and DVT	29	4	8	13	15	25	25
Miscellaneous	27	3	8	10	17	18	16
Pneumothorax and others	23	7	9	9	15	15	17
Chest surgery	22	4	9	10	15	15	15
Autoimmune diseases	14	5	3	8	7	10	10
Bronchoscopy-related	13	5	7	4	4	6	8
Sarcoidosis and LAM	11	2	2	3	6	7	10
Pleural diseases	9	2	4	2	4	3	5
Vasculitis	8	3	3	4	4	6	7
Musculoskeletal issues	3	0	1	2	2	3	2
Tracheal disorders	2	1	2	1	2	1	2
Diaphragmatic problems	1	0	0	1	1	1	1

Categories were sorted by the total number of questions (denoted as "#"). Values in the 1B, 4B, 12B, 27B, OSS-20B, and GPT-4T columns indicate the number of correctly answered questions for each category. Abbreviations: MV, mechanical ventilation; PE, pulmonary embolism; DVT, deep vein thrombosis; COPD, chronic obstructive pulmonary disease; PFT, pulmonary function testing; LAM, lymphangioleiomyomatosis.

Based on the dataset, answer accuracy was calculated for each category, and categories with more than 20 questions were ranked according to their correct response rates.

For the Gemma3-1B model, the highest-performing categories were sleep medicine (33.3%), pulmonary function testing (30.9%), pneumothorax and incidental intrathoracic air collections

 doi:10.20944/preprints202603.0329.v1

(30.4%), respiratory pathophysiology (26.7%), and pulmonary infection (24.6%). In the Gemma3-4B model, the top categories included chest anatomy (47.4%), pharmacology (43.2%), lung cancer (42.5%), chest surgery (40.9%), and infection (39.6%).

The Gemma3-12B model achieved its highest accuracy in respiratory pathophysiology (60.0%), infection (50.0%), tuberculosis (48.1%), asthma (46.4%), and pharmacology (45.9%). Performance further improved in the Gemma3-27B model, with the top categories being chest surgery (68.2%), pneumothorax (65.2%), pathophysiology (63.3%), miscellaneous topics (63.0%), and pharmacology (62.2%).

Among all locally deployed models, GPT-OSS-20B demonstrated the highest category-level accuracies, particularly in pulmonary embolism and deep vein thrombosis (86.2%), interstitial lung disease (85.7%), infection (76.9%), pharmacology (75.7%), and lung cancer (73.7%).

As a cloud-based reference, GPT-4 Turbo showed its strongest performance in pulmonary embolism (75.9%), pneumothorax (73.9%), lung cancer (72.6%), interstitial lung disease (68.6%), and chest anatomy (68.4%). These category-level performance patterns are summarized in Figure 2.

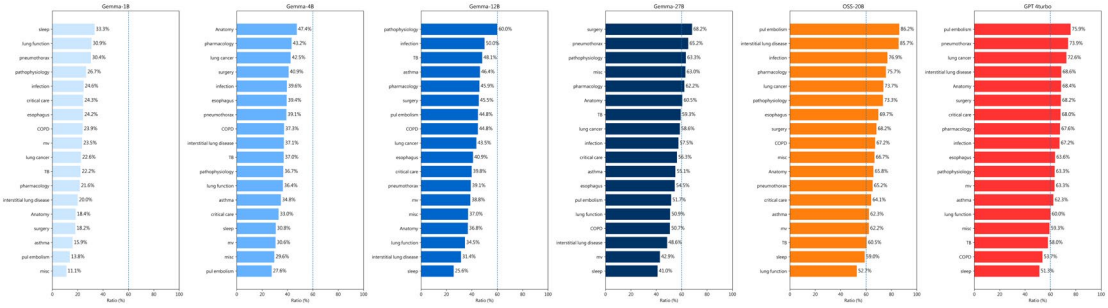

Figure 2. Category-level accuracy of locally deployed and cloud-based language models. Abbreviations: CCM, Critical Care Medicine; MV, mechanical ventilation topics; LFT, lung function test; LAM, lymphangioleiomyomatosis; PE and DVT, pulmonary embolism and deep vein thrombosis; COPD, chronic obstructive pulmonary disease; ILD, interstitial lung disease. Accuracy was calculated for each clinical category and ranked according to correct response rates among categories with more than 20 questions. Results are shown for Gemma3-1B, Gemma3-4B, Gemma3-12B, Gemma3-27B, and GPT-OSS-20B as locally deployed models, with GPT-4 Turbo included as a cloud-based reference. For each model, the top-performing categories are displayed, illustrating category-specific performance patterns and the effect of model scale on accuracy. Categories include both text-only and text-and-image questions, depending on model capability.

4. Discussion

Today, LLMs provide multimodal support encompassing text, images, audio, and video, which has accelerated their deployment in practical scenarios. Remarkably, modern LLMs already surpass human-level performance in certain domains, including standardized multiple-domain evaluations. These advances are predominantly deployed on cloud platforms, raising significant concerns regarding patient information and privacy. Leaked data can be exploited to commit crimes [14]. When deployed in cloud settings, user-provided data may be stored or processed beyond the immediate inference session, introducing the possibility that confidential medical information could be exposed or reconstructed in future outputs [15]. If cloud-based medical models are exposed to open-query environments, they may become vulnerable to prompt injection attacks. Through carefully designed prompts, attackers can systematically bypass safety mechanisms, induce the model to reveal confidential data, or even extract sensitive text fragments from the training set that may contain patient information [16–18].

Moreover, due to constant online connectivity, hospitals and insurance providers face risks of database theft and critical data leakage [8]. Today, healthcare systems account for approximately 60-75% of all external cyberattacks [19]. Ransomware attacks, which restrict access and encrypt information until ransom payments are made, represent an especially severe threat to healthcare

 doi:10.20944/preprints202603.0329.v1

systems—as illustrated by the incident involving the Change Healthcare system, which resulted in enormous costs [20]. Consequently, deploying LLMs with sufficient natural language processing capabilities locally, without requiring persistent network access, offers a more secure alternative by keeping all data processing on device [21].

Based on the above analysis, we selected the Google Gemma 3 family and OpenAI GPT-OSS 20B as pretrained models for closed-network and on-device execution. The Gemma 3 family includes models with 1B, 4B, 12B, and 27B parameters; the 1B variant is text-only, whereas the larger models are multimodal LLMs with vision capabilities, supporting advanced multimodal inputs [9]. In parallel, the GPT-OSS family comprises two models with 20B and 120B parameters. The 20B variant is optimized for on-device deployment in closed-network environments, offering efficient execution on high-end local hardware while maintaining strong performance across multiple NLP benchmarks. In contrast, the 120B variant is designed for large-scale research and is expected to outperform its 20B counterpart in complex tasks, albeit with substantially higher resource demands. [10]. Released in 2025, both Gemma 3 and GPT-OSS provide flexible open-source alternatives to proprietary cloud-based systems, making them suitable for privacy-preserving applications that demand strict data confidentiality.

While contemporary LLMs are largely optimized using English-language data, their performance in non-English professional examinations suggests meaningful cross-lingual adaptation. Reports from Taiwanese licensing and subspecialty examinations [22–26] indicate that cloud-based models can operate effectively across linguistic boundaries. A key unresolved issue, however, is whether this cross-lingual generalization persists when models are deployed locally under hardware constraints. Multiple-choice examinations therefore represent a practical testbed for comparing linguistic adaptability across deployment settings.

This study demonstrates that the locally deployed OSS-20B and Gemma3-27B models can achieve competitive accuracy across a broad range of pulmonary specialty exam categories, although the latter performs slightly lower. This difference may be attributed to architectural variations between Gemma3 and GPT-OSS. Furthermore, within the Gemma3 family, we observed that accuracy generally increases with model size. The 1B variant performs at a level comparable to random guessing, while the 4B and 12B variants occasionally fall into the random-guessing range in certain years. Although some voices advocate for the use of smaller language models, in highly accuracy-sensitive domains such as healthcare, relying on small-scale models may lead to significant reliability issues.

Performance varied across topics, with the highest-scoring categories differing not only between models but also within the same Gemma3 family. However, the largest subsets of questions in the dataset are related to lung cancer, infections, critical care medicine, mechanical ventilation, and tuberculosis—areas of major clinical importance. Notably, the models' top performance rates did not align with the relative prevalence or importance of these topics, likely because these models were not primarily pretrained on medical data. Therefore, future efforts should focus on incorporating domain-specific training data and enhancing logical reasoning capabilities in these clinically significant areas. Targeted fine-tuning may further improve model performance and increase their practical utility in medical settings.

Limitations

While larger models such as Gemma3-27B and GPT-OSS-20B can be deployed on consumer hardware, their average inference time per question, approximately 60 seconds, was substantially higher than that of smaller models. This latency reflects the growing computational cost associated with scaling model parameters and may partially stem from theoretical lower bounds on the energy and processing required to complete complex reasoning tasks [27]. Larger models generally achieve higher accuracy but demand greater computational power and energy consumption, which on consumer-grade hardware inevitably results in longer processing times. Computational power thus

 doi:10.20944/preprints202603.0329.v1

remains a critical bottleneck, and additional system-level considerations, including thermal management and heat dissipation, must be addressed to ensure sustained local deployment [28].

Although GPT-OSS-120B was expected to deliver even stronger performance based on its model size and capabilities, it was excluded from our evaluation due to practical deployment limitations. Specifically, the model requires over 60 GB of memory and exceeds 60 GB in download size when used with Ollama. This hardware demands exceed the capacity of our experimental consumer-grade system. As our study focuses on models that can realistically run on locally available consumer hardware, GPT-OSS-120B was not included in the comparative analysis.

5. Conclusions

Among the evaluated models, GPT-OSS-20B and Gemma3-27B emerged as the most capable on-device LLMs that can be deployed on consumer-grade hardware. While Gemma3-27B offers strong reasoning capabilities and supports both text and image modalities, its ability to comprehend complex human instructions and consistently generate accurate responses remains suboptimal. In comparison, GPT-OSS-20B demonstrated better instruction adherence and overall accuracy, though it is limited to text-only inputs and lacks multimodal capabilities.

Despite these trade-offs, Gemma3-27B demonstrated performance approaching the predefined benchmarks across both text-only and multimodal tasks, while GPT-OSS-20B came close to meeting several examination thresholds. These findings suggest that contemporary on-device models may have the potential to function as privacy-preserving alternatives to cloud-based LLMs in selected high-stakes applications.

In comparison with GPT-4, the results imply that in contexts where prior studies have reported acceptable performance using GPT-4, locally deployable models such as Gemma3-27B and GPT-OSS-20B could be considered as possible on-premise options. However, such substitution should be interpreted cautiously and evaluated within task-specific contexts.

Although their current performance appears encouraging, particularly in structured domains such as medical education and standardized assessment, further refinement, external validation, and domain-specific optimization will be necessary before broader clinical implementation can be recommended, especially for complex or underrepresented task categories.

Author Contributions: All authors contributed to the study's conception and design. Material preparation, data collection, and analysis were performed by Chih-Hsiung Chen. Medical text and image were processed by Chih-Hsiung Chen and Kuo-En Huang. The idea and resource support were provided by Kuang-Yu Hsieh and Chang-Wei Chen. The first draft of the manuscript was written by Chih-Hsiung Chen and later revised by Chang-Wei Chen. All authors provided feedback on previous versions of the manuscript, and all authors read and approved the final manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki and approved by the Institutional Review Board of Mennonite Christian Hospital. (protocol code 25-12-033 and date of approval on 2026-01-06).

Informed Consent Statement: Not applicable.

Data Availability Statement: These questions and answers were obtained from the download area of the website hosted by the Taiwan Society of Pulmonary and Critical Care Medicine. Address: No. 1, Changde St., Zhongzheng District, Taipei City 100229, Taiwan; Main Portal Website: https://www.tspccm.org.tw/. Source materials were provided in Microsoft Word and PDF formats. All files were standardized through text cleaning and formatting, converted into plain text, and consolidated into a structured CSV dataset for analysis. Images embedded within the PDF documents were extracted via full-screen capture, post-processed, and stored as PNG files in a designated directory. Image filenames followed a year–question number convention, with the suffix "p" appended to indicate pictorial items.

 doi:10.20944/preprints202603.0329.v1

Acknowledgments: During the preparation of this manuscript/study, the author(s) used Python 3.8 with the statistical functions provided in the SciPy 1.10.1 package. All graphical outputs were generated using Matplotlib version 3.7.5. The authors have reviewed and edited the output and take full responsibility for the content of this publication. ChatGPT-4 and later versions were used solely for language editing and refinement, with no content generation or data analysis involved. The authors have reviewed and edited the output and take full responsibility for the content of this publication.".

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

LLM	Large language model
MCQ	Multiple-choice question
API	Application interface
AI	Artificial intelligence

Appendix A. Definition of Clinical Categories

A total of 26 categories were defined to represent the principal domains of pulmonary and critical care medicine. These categories encompass lung cancer, which includes all thoracic malignancies such as primary lung cancer, esophageal cancer, mesothelioma, and other intrathoracic tumors; infection, primarily bacterial in origin but also including selected viral and fungal cases; and critical care medicine, covering sepsis, septic shock, acute respiratory distress syndrome, end-of-life care, and related intensive care topics. Mechanical ventilation includes ventilator modes, weaning strategies, and scenario-based management. Tuberculosis comprises both intrathoracic and extrathoracic forms, including miliary tuberculosis.

Additional categories include asthma; chronic obstructive pulmonary disease; non-malignant esophageal disorders such as hiatal hernia; pulmonary function testing; sleep medicine; chest anatomy; pharmacology relevant to pulmonary and critical care practice; interstitial lung disease; and respiratory pathophysiology. Vascular and air-leak–related conditions include pulmonary embolism and deep vein thrombosis, as well as pneumothorax and other incidental intrathoracic air collections, including pneumoperitoneum and pneumomediastinum.

Procedural and systemic disease categories encompass chest surgery, autoimmune diseases, bronchoscopy-related topics, sarcoidosis and lymphangioleiomyomatosis, vasculitis, musculoskeletal issues affecting the thorax, tracheal disorders, pleural diseases, and diaphragmatic disorders. Questions that could not be appropriately assigned to any of the above domains were classified as miscellaneous.

References

1. Chen, C.H.; Hsu, S.H.; Hsieh, K.Y.; Lai, H.Y. The two-stage detection-after-segmentation model improves the accuracy of identifying subdiaphragmatic lesions. *Sci. Rep.* **2024**, 14, 25414. https://doi.org/10.1038/s41598-024-76450-6.
2. Chung, Y.; Jin, J.; Jo H.I.; Lee, H.; Kim, S.H.; Chung, S.J.; Yoon, H.J.; Park, J.; Jeon, J.Y. Diagnosis of Pneumonia by Cough Sounds Analyzed with Statistical Features and AI. *Sensors*. **2021**, 21(21), 7036. https://doi.org/10.3390/s21217036.
3. Olszewski, R.; Brzeziński, J.; Watros, K.; Rysz, J. Quantifying Readability in Chatbot-Generated Medical Texts Using Classical Linguistic Indices: A Review. *Appl. Sci.* **2026**, 16, 1423. https://doi.org/10.3390/app16031423.
4. Chen, C.H.; Chen, C.W.; Hsieh, K.Y.; Huang, K.E.; Lai, H.-Y. Limitations in Chest X-Ray Interpretation by Vision-Capable Large Language Models, Gemini 1.0, Gemini 1.5 Pro, GPT-4 Turbo, and GPT-4o. *Diagnostics* **2026**, 16, 376. https://doi.org/10.3390/diagnostics16030376.

 doi:10.20944/preprints202603.0329.v1

5. Khalid, N.; Qayyum, A.; Bilal, M.; Al-Fuqaha, A.; Qadir, J. Privacy-preserving artificial intelligence in healthcare: Techniques and applications. *Computers in Biology and Medicine* **2023**, 158, 106848. https://doi.org/10.1016/j.compbiomed.2023.106848.
6. Dennstädt, F.; Hastings, J.; Putora, P.M.; Schmerder, M.; Cihoric, N. Implementing large language models in healthcare while balancing control, collaboration, costs and security. *npj Digit. Med.* **2025**, 8, 143. https://doi.org/10.1038/s41746-025-01476-7.
7. Chen, C.; Hsieh, K.; Huang, K.; Lai, H. Comparing Vision-Capable Models, GPT-4 and Gemini, With GPT-3.5 on Taiwan's Pulmonologist Exam. *Cureus*. **2024**, 16(8), e67641. https://doi.org/10.7759/cureus.67641.
8. Tsai, C.Y.; Hsieh, S.J.; Huang, H.H.; Deng, J.H.; Huang, Y.Y.; Cheng, P.Y. Performance of ChatGPT on the Taiwan urology board examination: insights into current strengths and shortcomings. *World J Urol*. **2024**, 42(1):250. https://doi.org/10.1007/s00345-024-04957-8.
9. Kamath, G.T.; Ferret, J.; Pathak, S.; et al. Gemma 3 Technical Report. *arXiv* **2025**. https://doi.org/10.48550/arXiv.2503.19786.
10. OpenAI: Agarwal, S.; Ahmad, L.; Ai, J.; Altman, S.; Applebaum, A.; Arbus, E.; et al. gpt-oss-120b & gpt-oss-20b Model Card. *arXiv* **2025**. https://doi.org/10.48550/arXiv.2508.10925.
11. Ollama GitHub. https://github.com/ollama/ollama (accessed on 03 March 2026)
12. Van Rossum, G.; Drake, F.L. *Python 3 Reference Manual*. Scotts Valley, CA: CreateSpace; 2009.
13. Hunter, J.D. Matplotlib: A 2D graphics environment. *Computing in Science & Engineering* **2007**, 9, 90-95. https://doi.org/10.1109/MCSE.2007.55.
14. Kim, S.; Yun, S.; Lee, H.; Gubri, M.; Yoon, S.; Oh, S.J. ProPILE: Probing Privacy Leakage in Large Language Models. *arXiv* **2023**. https://doi.org/10.48550/arXiv.2307.01881.
15. Yan, B.; Li, K.; Xu, M.; Dong, Y.; Zhang, Y.; Ren, Z.; Cheng, X. On Protecting the Data Privacy of Large Language Models (LLMs): A Survey. *arXiv* **2024**. https://doi.org/10.48550/arXiv.2403.05156.
16. Gulyamov, S.; Gulyamov, S.; Rodionov, A.; Khursanov, R.; Mekhmonov, K.; Babaev, D.; Rakhimjonov, A. Prompt Injection Attacks in Large Language Models and AI Agent Systems: A Comprehensive Review of Vulnerabilities, Attack Vectors, and Defense Mechanisms. *Information* **2026**, 17, 54. https://doi.org/10.3390/info17010054.
17. Thirunavukarasu, A.J.; Ting, D.S.J.; Elangovan, K.; Gutierrez, L.; Tan, T.F.; Ting, D.S.W. Large language models in medicine. *Nat. Med.* **2023**, 29, 1930–1940. https://doi.org/10.1038/s41591-023-02448-8.
18. Xu, H.; Zhang, Z.; Yu, X.; Wu, Y.; Zha, Z.; Xu, B.; Xu, W.; Hu, M.; Peng, K. Targeted Training Data Extraction—Neighborhood Comparison-Based Membership Inference Attacks in Large Language Models. *Appl. Sci.* **2024**, 14, 7118. https://doi.org/10.3390/app14167118.
19. Seh, A.H.; Zarour, M.; Alenezi, M.; Sarkar, A.K.; Agrawal, A.; Kumar, R.; Ahmad Khan, R. Healthcare Data Breaches: Insights and Implications. *Healthcare* **2020**, 8, 133. https://doi.org/10.3390/healthcare8020133.
20. Jiang, J.X.; Ross, J.S.; Bai, G. Ransomware Attacks and Data Breaches in US Health Care Systems. *JAMA Netw Open*. **2025**, 8, e2510180. https://doi.org/10.1001/jamanetworkopen.2025.10180.
21. Feretzakis, G.; Papaspyridis, K.; Gkoulalas-Divanis, A.; Verykios, V.S. Privacy-Preserving Techniques in Generative AI and Large Language Models: A Narrative Review. *Information* **2024**, 15, 697. https://doi.org/10.3390/info15110697.
22. Huang, C.H.; Hsiao, H.J.; Yeh, P.C.; Wu, K.C.; Kao, C.H. Performance of ChatGPT on Stage 1 of the Taiwanese medical licensing exam. *Digital Health* **2024**, 10. https://doi.org/10.1177/20552076241233144.
23. Kao, Y.S.; Chuang, W.K.; Yang, J. Use of ChatGPT on Taiwan's Examination for Medical Doctors. *Ann Biomed Eng*. **2024**, 52, 455-457. https://doi.org/10.1007/s10439-023-03308-9.
24. Liu, C.-L.; Ho, C.-T.; Wu, T.-C. Custom GPTs Enhancing Performance and Evidence Compared with GPT-3.5, GPT-4, and GPT-4o? A Study on the Emergency Medicine Specialist Examination. *Healthcare* **2024**, 12, 1726. https://doi.org/10.3390/healthcare12171726..
25. Ting, Y.T.; Hsieh, T.C.; Wang, Y.F.; et al. Performance of ChatGPT incorporated chain-of-thought method in bilingual nuclear medicine physician board examinations. *Digit Health* **2024**, 10. https://doi.org/10.1177/2055207623122.

 doi:10.20944/preprints202603.0329.v1

26. Hsieh, C.H.; Hsieh, H.Y.; Lin, H.P. Evaluating the performance of ChatGPT-3.5 and ChatGPT-4 on the Taiwan plastic surgery board examination. *Heliyon* **2024**, 10, e34851. https://doi.org/10.1016/j.heliyon.2024.e34851.
27. Tkachenko, Alexei V. Thermodynamic Bound on Energy and Negentropy Costs of Inference in Deep Neural Networks. *arXiv* **2025**. https://doi.org/10.48550/arXiv.2503.09980.
28. Fernandez, J.; Na, C.; Tiwari, V.; Bisk, Y.; Luccioni, S.; Strubell, E. Energy considerations of large language model inference and efficiency optimizations. *arXiv* **2025**. https://doi.org/10.48550/arXiv.2504.17674.