Title: Can ChatGPT or Gemini 2.0 Detect Pulmonary Embolism on Computed Tomography Pulmonary Angiography?

Authors: Shivam Patel[1], Nitin Chetla, B.S.[2], Tejas C. Sekhar, B.A.[3], Tamer Hage, B.S.[4], Swapna Vaja, B.S.[3], Michael Sneider, M.D.[5]

Introduction:
The advent of advanced artificial intelligence (AI) technologies, particularly large language models (LLMs) like ChatGPT and Gemini, has ushered in innovative opportunities to transform clinical medicine. In 2023, the American Medical Association (AMA) released a statement highlighting that ChatGPT-3 was capable of passing the United States Medical Licensing Exam (USMLE) Step 1, Step 2CK, and Step 3 exams. Accordingly, the AMA expressed both optimism and skepticism about the potential implementation of AI, citing AI's role in assisting healthcare professionals in the future while acknowledging that the USMLE represents just one measure of medical knowledge[1].

ChatGPT's proven ability to assess clinical knowledge questions from standardized exams with a reasonable degree of accuracy has prompted investigations into the potential of GPT-4 Omni (GPT-4o), a novel LLM, in medical diagnostics. Radiographic detection of pulmonary embolism (PE) is one such target, as PEs are responsible for 60,000 to 100,000 deaths in the United States annually and are typically identified and diagnosed through computer tomography pulmonary angiography (CTPA) scans[2,3]. Thus, assessing GPT-4o and Gemini's comparative fidelities in PE identification PE could provide utility, given that approximately 10% of symptomatic cases are considered rapidly fatal, underscoring the need for highly accurate and timely radiological diagnosis[4].

In this study, we sought to evaluate the diagnostic accuracy of GPT-4o and Gemini 2.0 in PE identification using simplified prompts designed for medical education scenarios. Thoracic radiologist-annotated CTPA radiographic DICOM scans from the RSNA PE Detection Challenge 2020 database, were transformed to PNG format and subsequently input into GPT-4o and Gemini 2.0, respectively, for evaluation[5]. Ultimately, this research aims to assess GPT-4o and Gemini 2.0's independent diagnostic capabilities in automating diagnostic processes, with implications for AI's broader real-life clinical utility.

Methods:
Our study aimed to assess the abilities of GPT-4o and Gemini 2.0 in correctly identifying PE presence or absence through the use of OpenAI and Gemini's application programming interfaces (APIs), respectively. CTPA scans obtained from the RSNA PE Detection Challenge 2020 database were transformed from DICOM file to PNG format using a standardized windowing approach. Though each DICOM file contained a single slice, multiple DICOMs were provided for each patient, collectively representing all necessary slices. DICOM files were processed using the Pydicom library, extracting key metadata such as window center, window width, rescale slope, and intercept. These values were then applied to transform pixel intensity values into an 8-bit grayscale representation. Windowing was performed by mapping pixel values within the window range to a 0-255 scale, ensuring contrast adjustments aligned with clinical viewing standards. To test quality loss during conversion, a quantitative image comparison was performed between a DICOM slice and its corresponding PNG. Two commonly used image similarity metrics were employed: (1) mean squared error (MSE), which measures the average squared pixel intensity difference between a DICOM slice and its PNG conversion; and (2) structural similarity index measure (SSIM), which evaluates image quality by comparing luminance, contrast, and structural details. An MSE of 0.0 was computed indicating that there were no pixel-level differences, and a SSIM of 1.0 confirmed that the images were structurally identical.

In total, 12,533 DICOM CTPA slices were transformed to an equivalent number PNG images from 47 PE-positive patients and subsequently analyzed using GPT-4o; similarly, 12,302 DICOM CTPA slices were transformed to an

equal number of PNG images from 49-PE positive patients and analyzed thereafter using Gemini 2.0. For PE-negative cases, 11,835 DICOM images from 49 patients were analyzed with GPT-4o, while 12,063 DICOM images from 51 patients were assessed using Gemini 2.0. Each CT scan had approximately 250 DICOM images. Patient volume investigated was directly corresponding to the total image volume analyzed, with discrepancies due to respective GPT-4o and Gemini 2.0 responsiveness to submitted API calls, with unresponsive API calls being excluded from further analysis. Given token limit contingencies per GPT-4o API call, every ten transformed PNG images were batched and merged into one composite image to reduce image volume analysis per each API request by a factor of 10 accordingly. The original images were pasted into predefined positions in a 5x2 grid without resizing or compression to preserve quality. To test the possibility of image quality loss from this composition methodology, the same evaluation was conducted. Each original image's corresponding section was cropped from the composite using its known position in the 5×2 grid. The extracted sub-image was then compared to its corresponding original image. The average MSE for the 10 images was 0.0 indicating no pixel-level differences, and the average SSIM was 1.0 indicating perfect similarity. As Gemini 2.0 has a larger token limit, every API call included all of the necessary images without the creation of a composite image. To achieve reliable results, each composite image was next processed through a recursive Python loop, which queried the API with the prompt, “These are a series of images from a CT scan that showed up in the USMLE Step 1 Medical School Examination. Please answer this question with a single letter ONLY. For example, if you believe there is a Pulmonary Embolism, answer 'A' and NOTHING ELSE. Does this CT scan display evidence of Pulmonary Embolism? A) Pulmonary Embolism is present; B) Pulmonary Embolism is not present”. Model outputs were recorded, and key performance metrics—including accuracy, precision, recall, F1 score, and support values—were calculated for both GPT-4o and Gemini 2.0 test outcomes.

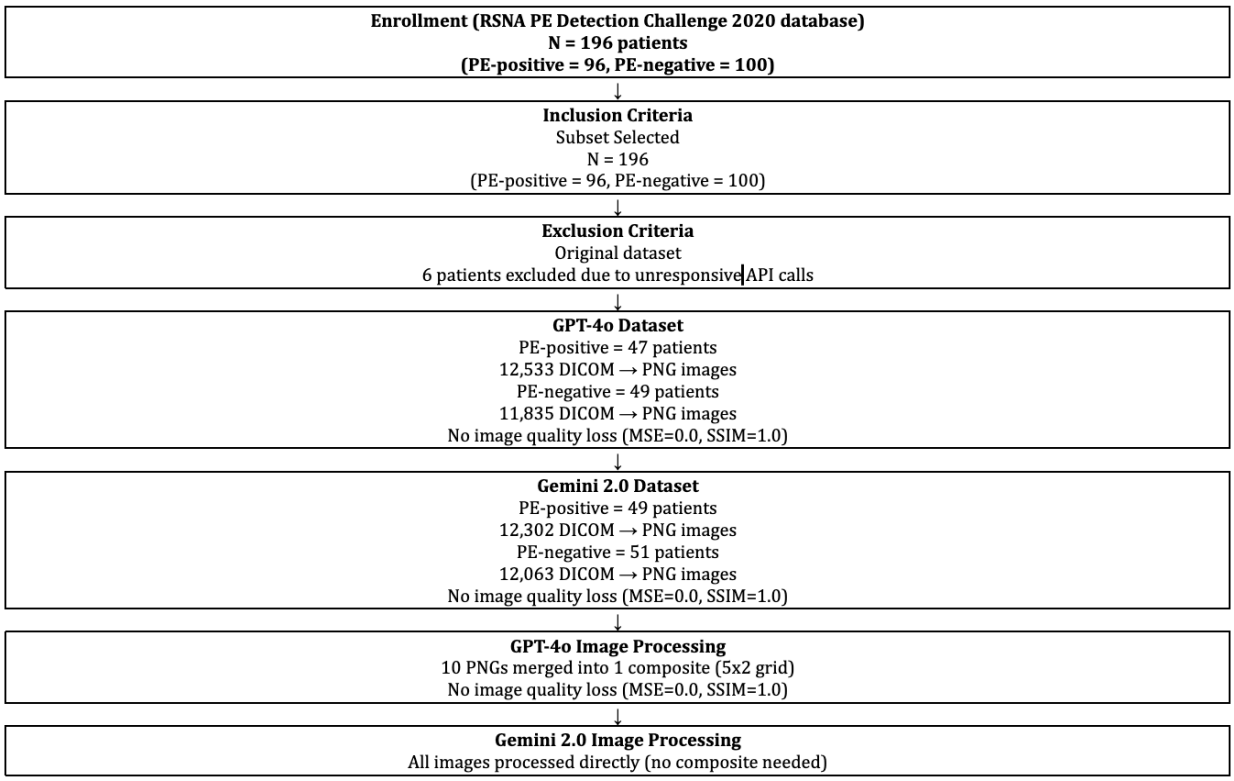


Figure 1. Intent to Enroll

We implemented a simplified, exam-style prompt to ensure that the language model (GPT) provided a clear, definitive classification of whether a PE was present. Given that reinforcement learning from human feedback (RLHF) often restricts LLMs from making direct medical diagnoses, our approach framed the query as a USMLE Step 1 Medical School Examination question rather than an ongoing or prospective clinical scenario[6]. This structure was designed to encourage the model to engage with the prompt and provide a straightforward categorical response ("A" or "B") rather than refusing to answer due to ethical constraints.

However, this prompt design may also unintentionally introduce bias in the model's decision-making. By positioning the question as part of an exam, we may have inadvertently altered its response behavior, potentially influencing its confidence, threshold for uncertainty, or pattern-recognition tendencies. More clinically nuanced prompts, such as those structured as clinical decision-support queries, might yield different outputs. We observed that GPT's refusal rate significantly increased as the prompts became more clinically nuanced. This occurred particularly when prompts requested more specific localization of findings or framed the question in a way that resembled a real patient case rather than a general test-like question about medical images. Asking for detailed interpretations with contextual clinical information, such as patient history or symptomatology, appeared to trigger a higher refusal rate, likely due to the model's caution around medical liability and ethical considerations. Future investigation will explore how prompt variations affect model performance, examining whether different linguistic framings, contextual cues, or levels of clinical detail impact accuracy, specificity, and overall reliability in AI-assisted radiology interpretation.

Results:
GPT-4o demonstrated variable performance in classifying CTPA images as either PE-positive or -negative. The confusion matrix illustrates GTP-4o's moderately strong performance in correctly identifying PE-positive patients, with 38/47 (X.X%) cases being correctly identified—conversely, GPT-4o performed poorly at correctly identifying PE-negative cases, only correctly identifying 5/49 (X.X%), with an accompanying high false-positive rate (**Figure 1**). The GPT-4o demonstrates a clear bias toward PE-positive cases, as it has high recall but low precision for PE detection while failing to correctly classify PE-negative cases (low recall for Class B). This behavior suggests it prioritizes sensitivity for detecting PE at the expense of specificity. GPT-4o was more precise toward PE-positive images (0.46) compared to PE-negative images (0.36), as well as better at recall for PE-positive images (0.81) compared to PE-negative images (0.10), indicating strong sensitivity for positive identification. Furthermore, the calculated F1 score was 0.59 for positive predictions and 0.16 for negatives, indicating that while GPT-4o could identify PE presence with moderate success, it struggled with accurately ruling out the condition (**Table 1**).

Broadly, Gemini 2.0 demonstrated similarly inconsistent performance in its binary classification ability, with notable outcomes in isolated performance variables. Gemini 2.0's confusion matrix demonstrates high accuracy in correctly identifying PE-negative cases, with correct classification of 50/51 (X.X%) patients; PE-positive case identification fared far worse, with a correct classification rate of 7/49 (X.X%) recorded alongside a high false-negative rate **(Figure 2)**. We observed Gemini 2.0 to have an overall accuracy rate of 0.57, with far higher precision towards PE-positive images (0.88) when compared to PE-negative images (0.54). Large disparities in measures of recall were also observed in Gemini 2.0, with PE-negative readings demonstrating a value of 0.98, showing strong sensitivity when compared to PE-positive images with a value of 0.14. F1-scores also skewed significantly in favor of PE-negative identification performance (0.70) versus PE-positive identification performance (0.25), suggesting that the model was able to rule out the presence of PE with moderate success, but had fairly poor performance with accurately ruling it in. **(Table 2)**

When directly comparing the performance of GPT-4o and Gemini 2.0 in correctly identifying the presence or lack of presence of PE, Gemini 2.0 was generally more biased towards the identification of PE-negative findings, as evidenced by its higher recall at the expense of correct PE-positive detection as well as the confusion matrix findings. GPT-4o demonstrated an inverse trend, being biased toward PE-positive readings due to a higher recall

value, and also as evidenced through its confusion matrix. GPT-4o and Gemini 2.0 output F1 scores deficient for PE-negative detection and PE-positive detection respectively, highlighting the implications of the aforementioned imbalances in relative measures of precision and recall.

Discussion:
Our mixed findings detail GPT-4o and Gemini 2.0's current limitations in radiological applications with respect to PE identification on CTPA images. While GPT-4o showed moderate success in identifying PE-positive cases, its low accuracy and precision when presented with PE-negative cases underline the challenges associated with the widespread adoption of LLMs in clinical diagnostics. Gemini 2.0's reported high sensitivity in correctly identifying PE-negative cases alongside accompanying moderate precision. Gemini 2.0's PE-positive performance was highlighted by high precision with low recall/sensitivity. Studies examining the use of LLMs to identify PE on CTPA images are limited. However, the use of deep learning to identify PE on CTPA scans has been examined. One study by et al. developed a deep neural network using an InceptionResNet V2 architecture combined with a long short-term memory (LSTM) network to detect PE on CTPA scans, achieving a sensitivity of 86.6% and specificity of 93.5% with chest X-ray pre-training data. Compared to our findings—where GPT-4o and Gemini 2.0 achieved lower overall accuracy rates of 0.45 and 0.57, respectively—their model demonstrated superior performance across all metrics, particularly in maintaining balanced diagnostic accuracy for both PE-positive and PE-negative cases[7].

Improvements in training methodologies and model refinement are essential to enhancing diagnostic reliability across medical imaging, bridging the gap between LLM research and practical utility. This call to action is especially important given our variance in results using different LLMs to assess the same dataset. Recent evidence suggests that limitations in annotated data substantially affect diagnostic performance, necessitating the adoption of meta-learning frameworks and standardized benchmarks to improve model robustness in data-scarce environments[8,9]. Ultimately, this study supports the notion that in their current form, LLMs are not presently capable of sufficient key metric performance to support their implementation and integration into radiologic workflows[10]. Radiologists must remain critical when it comes to implementing any emerging technologies as part of their individual or system-wide practices. Our study findings reiterate the need for caution and continued iteration and evaluation before integrating LLMs into clinical decision-making workflows, emphasizing the continued usage of radiographic assessment as an immutable gold standard in medical image interpretation.

References:

1. ChatGPT passed the USMLE. What does it mean for med ed? American Medical Association. April 21, 2023. Accessed December 4, 2024. https://www.ama-assn.org/practice-management/digital/chatgpt-passed-usmle-what-does-it-mean-med-ed

2. Weikert T, Winkel DJ, Bremerich J, et al. Automated detection of pulmonary embolism in CT pulmonary angiograms using an AI-powered algorithm. *Eur Radiol*. 2020;30(12):6545-6553. doi:10.1007/s00330-020-06998-0

3. Kearon C, De Wit K, Parpia S, et al. Diagnosis of Pulmonary Embolism with d -Dimer Adjusted to Clinical Probability. *N Engl J Med*. 2019;381(22):2125-2134. doi:10.1056/NEJMoa1909159

4. Lapner ST, Kearon C. Diagnosis and management of pulmonary embolism. *BMJ*. 2013;346(feb20 1):f757-f757. doi:10.1136/bmj.f757

5. Colak E, Kitamura FC, Hobbs SB, et al. The RSNA Pulmonary Embolism CT Dataset. *Radiol Artif Intell*. 2021;3(2):e200254. doi:10.1148/ryai.2021200254

6. Step 1 Sample Questions. USMLE. Accessed March 18, 2025. https://www.usmle.org/exam-resources/step-1-materials/step-1-sample-test-questions

7. Huhtanen H, Nyman M, Mohsen T, Virkki A, Karlsson A, Hirvonen J. Automated detection of pulmonary embolism from CT-angiograms using deep learning. *BMC Med Imaging*. 2022;22(1):43.

doi:10.1186/s12880-022-00763-z

8. Lv Q, Chen G, Yang Z, Zhong W, Chen CYC. Meta Learning With Graph Attention Networks for Low-Data Drug Discovery. *IEEE Trans Neural Netw Learn Syst*. 2024;35(8):11218-11230. doi:10.1109/TNNLS.2023.3250324

9. Lv Q, Chen G, Yang Z, Zhong W, Chen CYC. Meta-MolNet: A Cross-Domain Benchmark for Few Examples Drug Discovery. *IEEE Trans Neural Netw Learn Syst*. 2025;36(3):4849-4863. doi:10.1109/TNNLS.2024.3359657

10. Keshavarz P, Bagherieh S, Nabipoorashrafi SA, et al. ChatGPT in radiology: A systematic review of performance, pitfalls, and future perspectives. *Diagn Interv Imaging*. 2024;105(7-8):251-265. doi:10.1016/j.diii.2024.04.003

Tables:

Metric	Precision	Recall	F1-score	Support
Pulmonary Embolism (A)	0.46	0.81	0.59	47
No Pulmonary Embolism (B)	0.36	0.1	0.16	49
Accuracy			0.45	96
Macro Avg	0.41	0.46	0.37	96
Weighted Avg	0.41	0.45	0.37	96

Table 1. GPT-4o's Diagnostic Performance Metrics for Pulmonary Embolism Detection

Table 1 details GPT-4o's diagnostic performance in detecting pulmonary embolism (PE) from CT pulmonary angiographic images (CTPAs). The model produced high recall (81%) for detecting PE-positive cases but struggled with PE-negative cases (recall: 10%). Overall, accuracy was 45%.

Metric	Precision	Recall	F1-score	Support
Pulmonary Embolism (A)	0.88	0.14	0.25	49
No Pulmonary Embolism (B)	0.54	0.98	0.7	51
Accuracy			0.57	100
Macro Avg	0.71	0.56	0.47	100
Weighted Avg	0.71	0.57	0.48	100

Table 2: Gemini 2.0's Diagnostic Performance Metrics For Pulmonary Embolism Detection

Table 2 details Gemini 2.0's diagnostic performance in detecting pulmonary embolism (PE) from CT pulmonary angiographic images (CTPAs). The model produced high recall (98%) for detecting PE-negative cases but struggled with PE-positive cases (14%).

Category	GPT-4o	Gemini 2.0
PE-Positive Patients	47	49
PE-Positive DICOM Images	12533	12302
PE-Negative Patients	49	51
PE-Negative DICOM Images	11835	12063

Table 3: Dataset Composition for Pulmonary Embolism Detection
This table presents the distribution of PE-positive and PE-negative patients, along with the number of corresponding DICOM images, used in evaluating the performance of GPT-4o and Gemini 2.0 for pulmonary embolism detection.

Figure Legends:

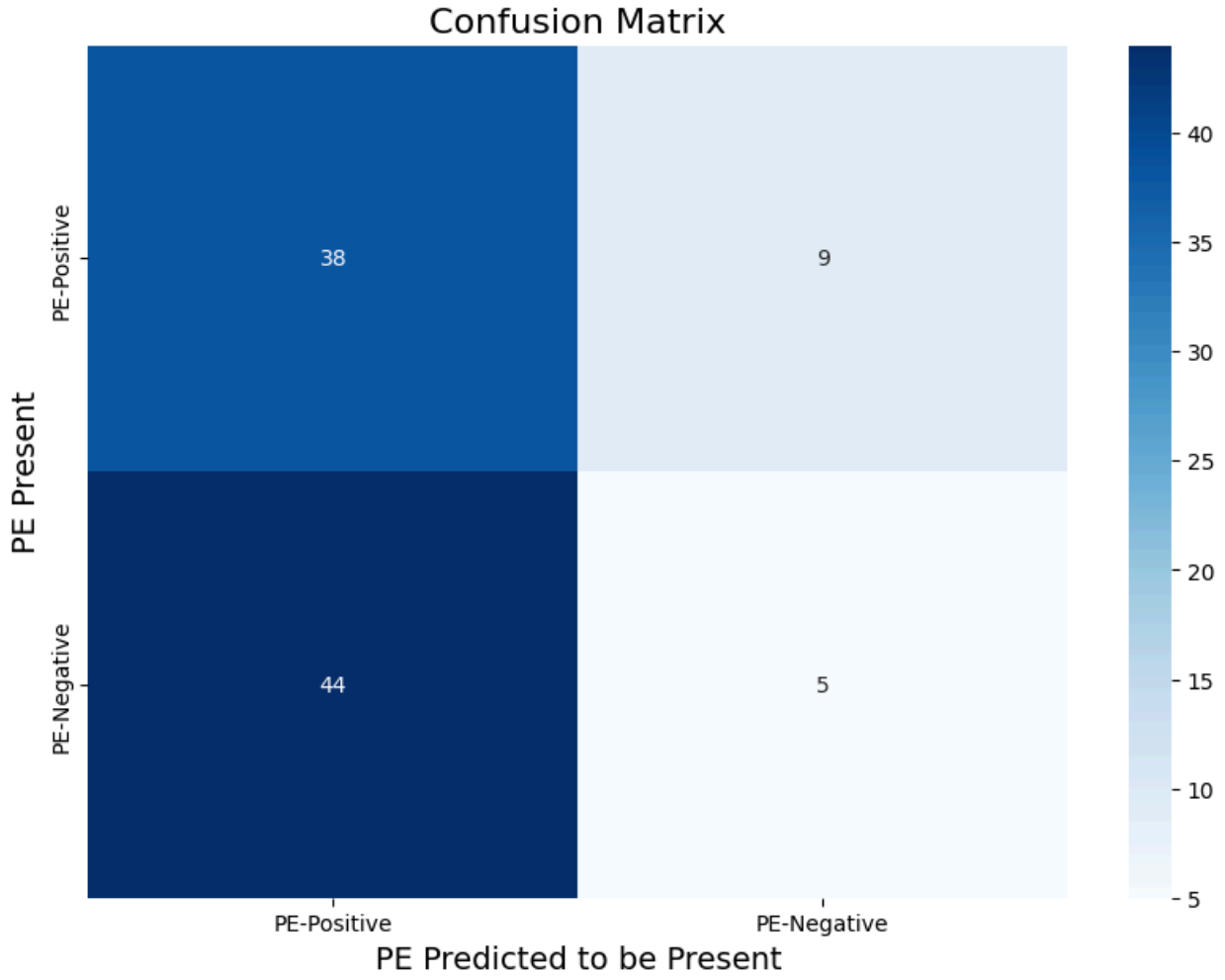


Figure 1. The confusion matrix illustrates that GPT-4o correctly identified most PE-positive cases (38 out of 47, 80.9%) but frequently misclassified PE-negative cases as positive (44 out of 49, 89.8%). This pattern demonstrates a strong bias toward predicting PE presence, resulting in a large number of false positives and a relatively small number of true negatives.

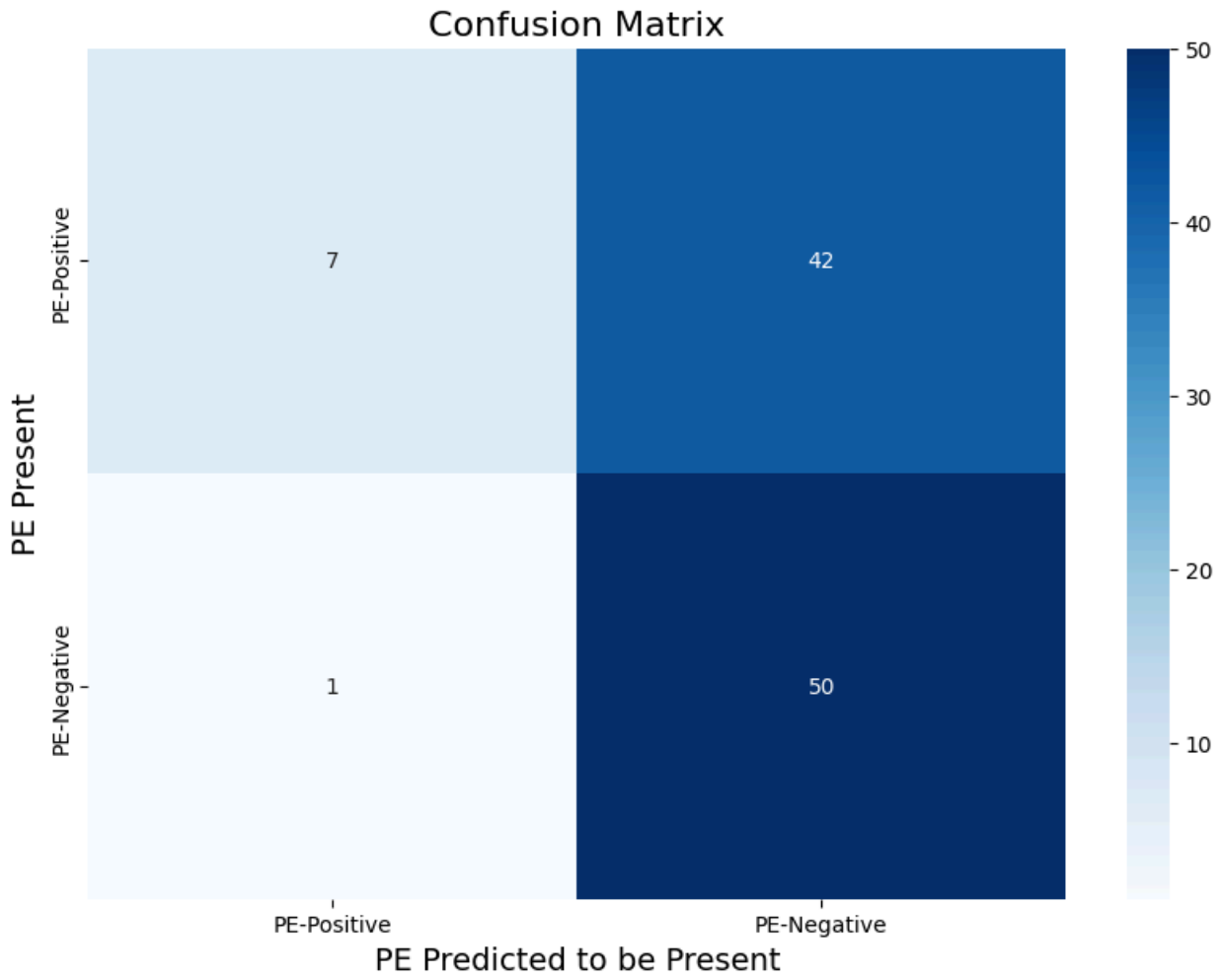
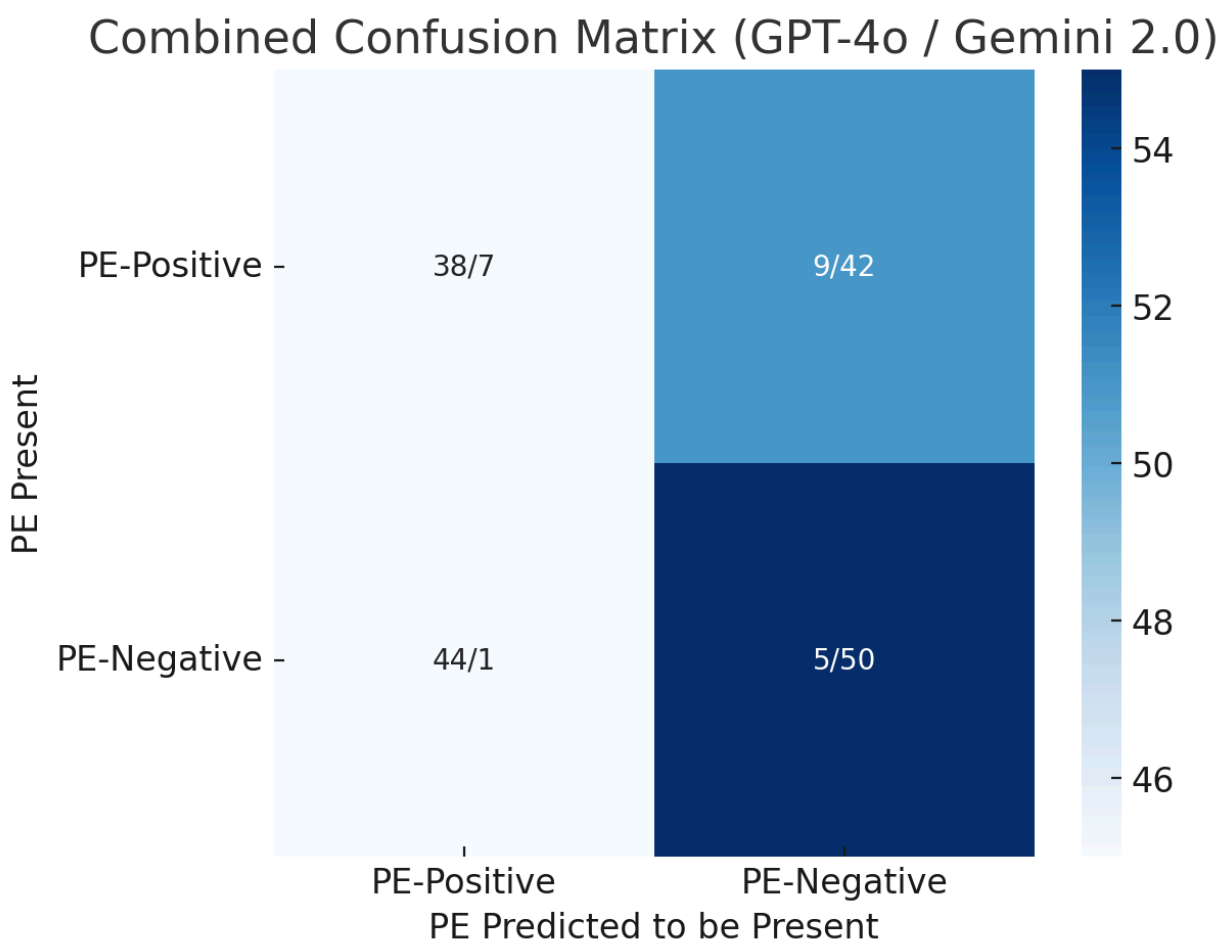


Figure 2. The confusion matrix illustrates that Gemini 2.0 correctly identified nearly all PE-negative cases (50 out of 51, 98.0%) but frequently misclassified PE-positive cases as negative (42 out of 49, 85.7%). This pattern demonstrates a bias toward predicting PE-negative cases, resulting in a large number of false negatives and a relatively small number of true positives.


Combined Confusion Matrix (GPT-4o / Gemini 2.0)
PE-Positive
PE-Negative
PE Present
38/7
9/42
44/1
5/50
54
52
50
48
46
PE-Positive
PE-Negative
PE Predicted to be Present