**Figure Legend:**

**Figure 1: Quantitative Analysis of Model Performance**

**(A)** **Precision Score Distribution Across Models** Precision scores for Gemini Advanced, GPT4, and GPT4O were evaluated across the same patient datasets. This chart shows the models' ability to avoid false positives. GPT4O consistently outperforms the other two models, with the highest precision scores in most cases, while Gemini Advanced has the lowest median precision score, indicating more frequent false positives across predictions.

**(B)** **Recall Score Distribution Across Models** The recall scores for Gemini Advanced, GPT4, and GPT4O were analyzed, representing the models' ability to correctly identify true positive cases. GPT4O achieved the highest recall, demonstrating superior sensitivity in identifying relevant cases, while Gemini Advanced had lower recall values across most datasets. The chart illustrates that GPT4O effectively captures the true positives with fewer misses compared to the other models.

**(C)** **F1 Score Distribution Across Models** The F1 scores for three models (Gemini Advanced, GPT4, and GPT4O) were compared across ten patient datasets. The chart presents the distribution of F1 scores for each model, indicating performance in balancing precision and recall. GPT4O demonstrated a higher mean F1 score, while Gemini Advanced and GPT4 showed varied but generally lower F1 performance across patients. The variation highlights the models' differing capabilities in managing true positive and false negative predictions.

**Figure 2: Turing Test analysis**

10 urologists rated four treatment plans for each of the ten patients on a scale from 0 (least likely to come from an LLM) to 10 (most likely to come from an LLM). The average score for each treatment plan was calculated per physician, resulting in 10 sets of average scores. These were further analyzed to compute the Grand Mean Average across（GMA） all sets. Purple denotes treatments derived from human experts, blue from GPT4O, green from GPT4, and red from Gemini Advanced.

**Figure 3:Clinical Trial Validation**

The authenticity of clinical trials provided by the three LLMs was verified. Darker-colored sections indicate valid clinical trials, with each shade corresponding to the respective LLM: blue for GPT4O, green for GPT4, and red for Gemini Advanced.

**Figure 4: Overlap Analysis of Treatment Options between Large Language Models and Human Experts**

**(A)** Percentage of treatment options generated by three models (Gemini Advanced, GPT4O, GPT4) that overlap with treatment options provided by human experts for 10 patients. The percentage represents the number of overlapping options out of the total number of options generated by each model.

**(B)** The number of identical treatment options generated by three large language models (Gemini Advanced, GPT4O, GPT4) and human experts for 10 patients. Treatments that match in 3 or more instances are marked in **navy blue**, and connections are drawn between the models and human experts to indicate this overlap. Treatments that match in 2 instances are marked in **steel blue**, and those that match in 1 instance are marked in **light blue**.

**(C)** Total number of treatment options generated by each model and by human experts: **Human (49 options)**, **Gemini Advanced (84 options)**, **GPT4O (99 options)**, and **GPT4 (90 options)**.