IEEE SPMB 2025: Review Results

Submission Type: Paper (p024)

Title: Assessing Visual Reasoning of Multimodal Language Models in Biomedical Applications

Score: 7.25

Summary: The reviewers find the paper's core idea—using MLLMs like ChatGPT-o3 and Qwen2-VL for biomedical signal and image interpretation without domain-specific training—interesting and timely. They acknowledge the novelty of incorporating expert-checked rationales and initial fine-tuning, and note some promise in the models' free-text reasoning.

Common comments across the board pointed towards the small data used, lack of critical statistical analysis, and the ambiguity of terms used to describe the performance of the proposed methodology. The reviewers consider the work preliminary and weakened by major limitations. Reviewers call for clearer justification of the sample sizes, stronger evaluation protocols, and fuller reporting of metrics beyond accuracy.

They also point to methodological gaps: no systematic prompt design, unclear dataset handling, fragmented presentation of methods, and weak or missing state-of-the-art baselines. Claims that ChatGPT-o3 serves as a strong baseline are viewed as premature without proper comparisons.

Presentation issues further reduce clarity. The introduction is repetitive; subjective language appears throughout; figures are underspecified or too small; citations and formatting need cleanup. Reviewers also flag surprising or unclear results (e.g., a reported accuracy of 0.00) and suggest framing fine-tuning efforts explicitly as a pilot.

Overall, they see the direction as promising but feel the paper overreaches in its claims. They recommend substantial revisions to the methodology, evaluation rigor, writing, and positioning before the contributions can be considered reliable.

Reviewer #1:

The paper studies whether current MLLMs can "see and explain" biomedical signals/images without task-specific training. The paper evaluates ChatGPT-o3 on 2 tasks using image inputs: (a) EEG seizure detection and (b) breast cancer pathology patch classification. The paper further curates expert-checked rationales and perform PEFT of Qwen2-VL. Reported zero-shot results are modest but non-trivial. The paper argues that o3 is a reasonable baseline and that lightweight supervised adaptation improves accuracy while preserving explainable free-text rationales.

Here are some comments:

- The introduction is long and could be more focused — it repeats the benefits of MLLMs several times.
- Some methodology details (e.g., prompt design, dataset sampling) are scattered and could be reorganised into a clearer methods pipeline. Figures, such as Fig. 3, should be more detailed and self-contained. Better captions recommended.
- Sample size is small, which limits statistical robustness. Authors should acknowledge this limitation more explicitly and provide confidence intervals or statistical tests.

- The training dataset for fine-tuning is very limited, which could explain degraded Qwen FT performance. This should be framed as a pilot or proof-of-concept, not as evidence of generalisable improvement.
- It would be beneficial to compare with sequence models or models that incorporate temporal signals directly, rather than screenshots.
- The paper mostly reports on accuracy. More informative metrics (precision, recall, F1-score, AUC) are needed, especially for imbalanced datasets.
- The claim that ChatGPT can serve as a "strong baseline" needs caution. It should be emphasised that while reasoning quality is notable, classification accuracy still lags behind domain-specific supervised models.
- Avoid subjective language, i.e. "did a good job", "not such a good job".
- Typos found (e.g., "andachieved"). Please fix.

Reviewer #2:

This paper presents an interesting evaluation of MLLMs for biomedical image classification tasks, specifically EEG signal interpretation and digital pathology.

The datasets used in this paper are extremely small for meaningful technical conclusions to be made. As such, the authors would need to justify these sample sizes or expand the datasets significantly. The paper could also benefit from a cross-validation or other robust evaluation methodologies to justify the technical contributions of the proposed research.

From the MLLM's aspect of this research, there is no or very minimal systematic prompt engineering methodology described in this paper. There is also limited baseline comparisons with state-of-the-art biomedical image analysis methods that can be found in the literature.

The authors are suggested to improve on the following:

- Develop a systematic approach to prompt optimization
- Include more comprehensive baselines specific to each domain
- Implement proper train/validation/test splits for fine-tuning experiments
- Establish systematic criteria for expert evaluation of model reasoning

Reviewer #3:

- Avoid use of ambiguous or unclear words to describe scientific performance, "moderate performance", "good", "poorly", etc.
- Figure 1 could be made larger as readability is poor in current form.
- Description of TUH data could be explained via more recent published reviews in the literature such as
  - https://doi.org/10.1109/SPMB52430.2021.9672302
  - https://doi.org/10.1016/j.eswa.2023.121040
  - https://doi.org/10.1109/SPMB50085.2020.9353647

- “Despite these shortcomings, the model’s structured reasoning outputs were often clinically interpretable and occasionally insightful.” — what is the definition of “often” in this context? Perhaps, as with other parts of the paper, more statistical analysis are needed.
- Format of citations in Biblio section is not consistent.
- Also to confirm that accuracy the result using ResNet for PT of DPATH is 0.00?