Does Visual Grounding Produce More Reliable Aphasia Therapy Items than Caption-Only Generation?

Mihir Mulye [1] Stefan Conrad [1] Stefan Knecht [1 2]

[1]Heinrich Heine University, Düsseldorf, Germany [2]Institute of Clinical Neurosciences, Universitätsklinikum, Düsseldorf

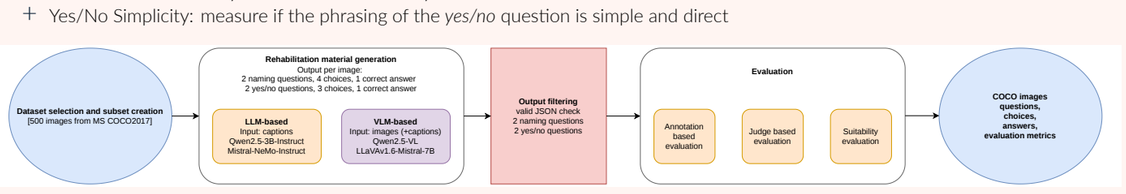
CVPR JUNE 3-7, 2026 DENVER COLORADO

Overview

- **Aphasia:** Acquired language impairment (also in children) from stroke or brain injury[1]
- Rehabilitation uses visually-grounded tasks such as object naming and comprehension questions
- Therapists create therapy material **manually**; can be costly, time-consuming, & limits scalability[2]
- Generative models offer a promising direction, but **LLMs** rely solely on text captions and lack visual context, **VLMs** can directly process images
- **Research Question:** LLM (caption) vs. VLM (image); which produces better quality and more semantically grounded therapy items, and by how much?
- **Contributions:**
  - LLM and VLM generation pipelines for aphasia therapy tasks
  - Suite of evaluation metrics: annotation-based, judge-based, therapy suitability
  - This pipeline can also be used to create and evaluate language learning material for children

Methodology

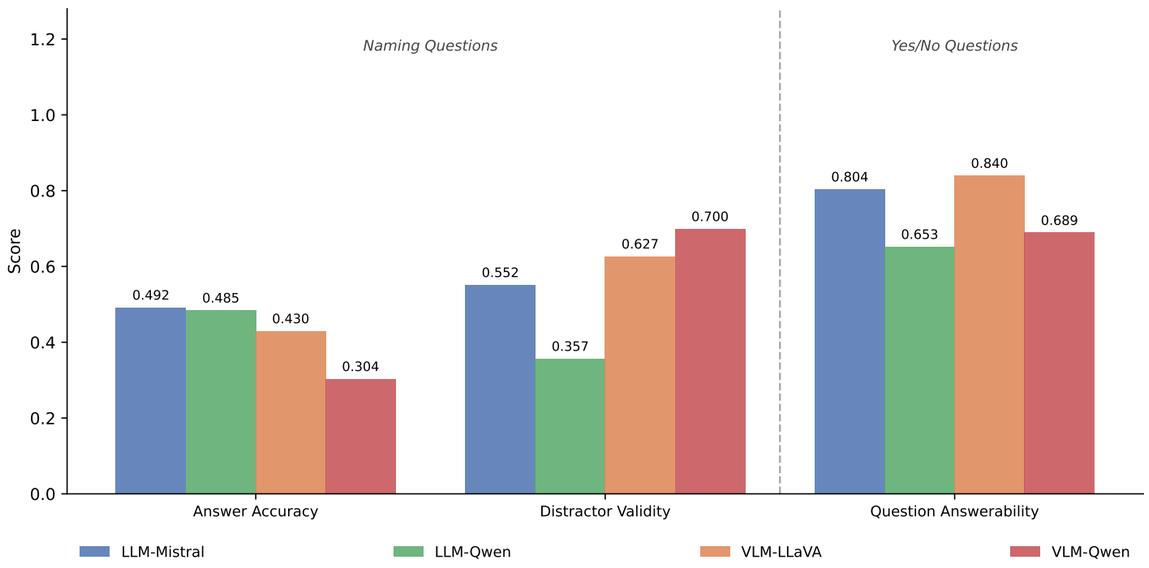
- **Dataset:** 500 images from MS COCO val2017[3] of everyday scenes with 5 human-written captions and object annotations per image
- Generative models (LLM: **Mistral-NeMo[4]** and **Qwen2.5-7B[5]**; VLM: **LLaVA-Next[6]** and **Qwen2.5-VL-3B[7]**) create therapy appropriate questions, distractors (multiple choice options) and correct answers
- LLMs are caption conditioned and VLMs are image-conditioned with captions as auxiliary input
- **Output per image:** 2 Naming questions (4 choices) + 2 Yes/No questions (3 choices), each with exactly one correct answer
- **Filtering:** JSON validation + structural checks; 329 records retained from original 500 datapoints
- **Evaluation Metrics:**
  - **Anntation-based:**
    - *Answer Accuracy:* only for *Naming* task, compares the generative model "correct" answers to the MS COCO ground truth
    - *Distractor Validity:* for the *Naming* task, checks if the distractors (wrong answer choices) (i) do not appear in the ground truth and (ii) are semantically related to the correct answer
    - *Question Answerability:* checks if a generated *yes/no* question can be answered using the available information (images and captions)
  - **Judge-model based:**
    - Question Clarity: Is the question unambiguously formulated?
    - Answer Uniqueness: Is an answer deducible from the captions and images?
  - **Suitability-based:**
    - Lexical Simplicity: measure if the words in the generated material are simple and commonly used; is computed for both the *Naming* and *yes/no* question
    - Choice Concreteness: measures if the answer choices for the *Naming* refer to concrete, physical entities
    - Naming Appropriateness: measures if the *Naming* question is short, simple, and well-structured
    - Distractor Confusability: measures the similarity of the answer choices to the correct answer
    - Yes/No Simplicity: measure if the phrasing of the *yes/no* question is simple and direct



Results: Annotation-Based Evaluation

Model	Answer Accuracy	Distractor Validity	Question Answerability
LLM-Mistral	**0.492**	0.552	0.804
LLM-Qwen	0.484	0.357	0.652
VLM-LLaVA	0.430	0.627	**0.840**
VLM-Qwen	0.304	**0.699**	0.689

- *Answer Accuracy:* LLMs outperform VLMs, can be potentially attributed to better vocabulary alignment with COCO labels
- *Distractor Validity:* VLMs substantially better (0.63-0.70 vs 0.36-0.55), visual grounding potentially helps generate plausible yet incorrect distractors
- *Question Answerability:* Comparable across models; LLaVA highest (0.84)

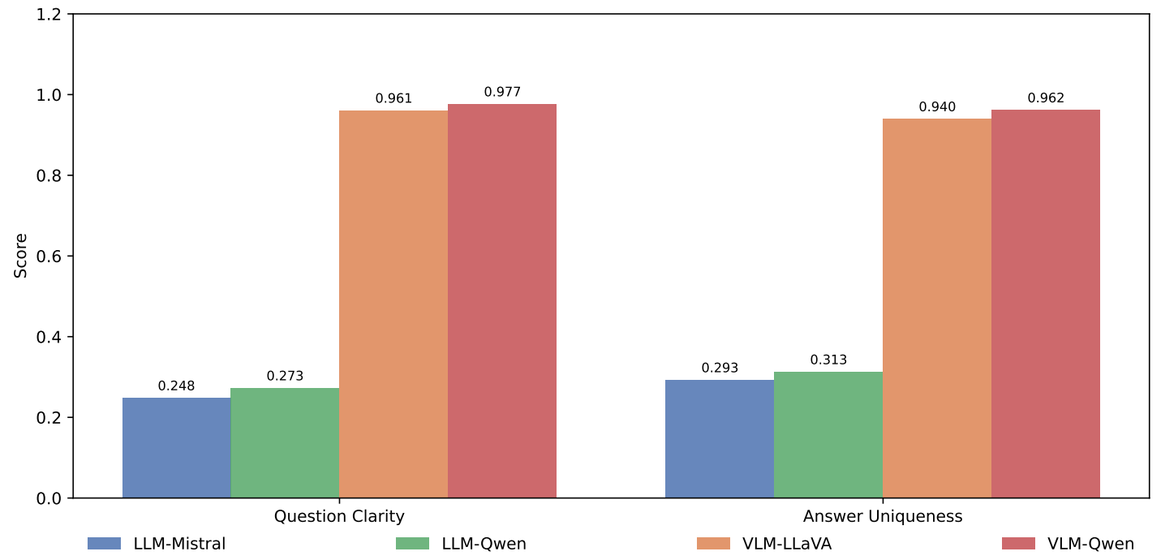


Results: Judge-Based Evaluation

Model	Question Clarity	Answer Uniqueness
LLM-Mistral	0.248	0.293
LLM-Qwen	0.273	0.313
VLM-LLaVA	0.961	0.940
VLM-Qwen	**0.977**	**0.962**

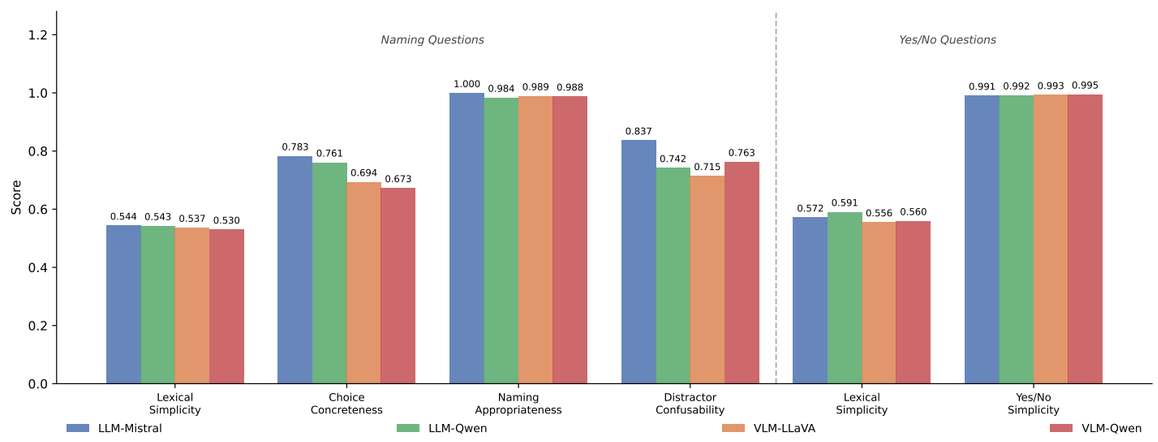
Results: Judge-Based Evaluation

- *Question Clarity:* VLMs vastly outperform LLMs (0.96-0.98 vs 0.25-0.27), visually grounded questions tend to be clearer and less ambiguous
- *Answer Uniqueness:* VLMs again dominate (0.94–0.96 vs 0.29–0.31), visual context potentially ensures a single defensible correct answer



Results: Suitability-based Evaluation

Model	Naming Task				Yes/No Task	
	Lexical Simplicity	Choice Concrete-ness	Naming Appropri-ateness	Distractor Confus-ability	Lexical Simplicity (Y/N)	Yes/No Simplicity
LLM-Mistral	**0.544**	**0.783**	**0.999**	**0.837**	0.571	0.990
LLM-Qwen	0.543	0.761	0.984	0.742	**0.590**	0.992
VLM-LLaVA	0.537	0.694	0.989	0.714	0.555	0.993
VLM-Qwen	0.530	0.673	0.988	0.763	0.556	**0.995**



- *Lexical Simplicity (Naming and yes/no):* Similar across all models (0.53–0.54); visual grounding does not play a major role
- *Choice Concreteness:* LLMs slightly better, as VLMs generally tend to generate attribute-based choices
- *Naming Appropriateness & Yes/No Simplicity:* All models score ≥0.984, exhibit consistent structural adherence to therapy formats regardless of input modality
- *Distractor Confusability:* Moderate variation among models; Mistral best (0.837), LLaVA lowest (0.714)

Conclusion & Future Work

- Visual grounding helpful for creating better answer choices, distinct questions and answers compared to only caption conditioned material
- LLMs tend to perform better for generating structure adhering therapy suitable material
- **Future Work:**
  - Human evaluation and clinical validation of generated material besides automated evaluation
  - Hybrid pipelines combining caption-conditioned lexical control with visual grounding could potentially create better material
  - Human-in-the-loop generation workflows for personalized therapy material can also be another direction to explore

References

1 Azevedo et al. *How artificial intelligence (AI) is used in aphasia rehabilitation: A scoping review*. AJSLP, 2018.

2 Mulye et al. *Exploring Applicability of Text-to-Image Models for Generating Aphasia Rehabilitation Material*. AIiH, 2025.

3 Lin et al. *Microsoft coco: Common objects in context.*, ECCV, 2014.

4 `https://huggingface.co/mistralai/Mistral-Nemo-Instruct-2407`

5 `https://huggingface.co/Qwen/Qwen2.5-7B-Instruct`

6 Liu et al. *Improved baselines with visual instruction tuning.*, CVPR, 2024.

7 `https://qwenlm.github.io/blog/qwen2.5-vl/`