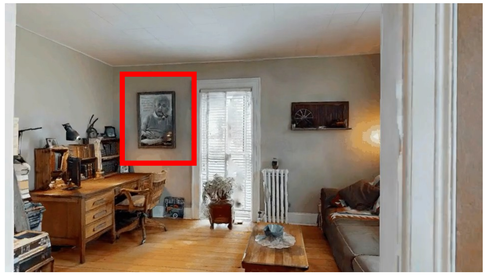
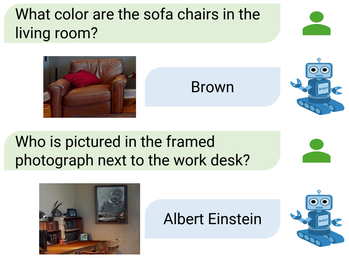
OpenEQA: Embodied Question Answering in the Era of Foundation Models

Open-vocabulary Embodied Question Answering

Observation Stream



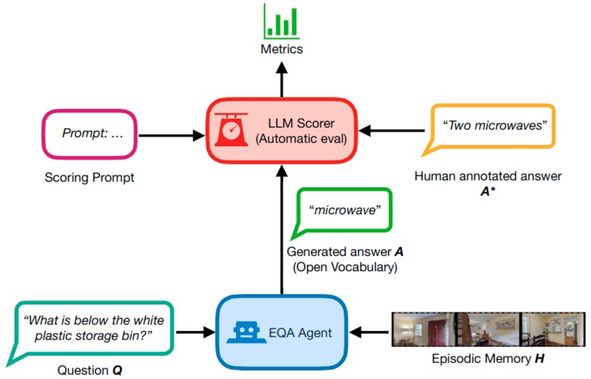
Open-vocabulary QA



- **Embodied** Question Answering = Understand **environments** well enough to answer any question about them in natural language.
- Agents answer from episodic memory (smart glass agents) or by actively exploring the environment (mobile robots)
- We present the first open-vocabulary benchmark for EQA

Evaluation Workflow

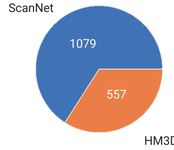
- Evaluating open-vocab answers is challenging
- Human evaluation is expensive and slow.
- We introduce LLM-Match, an automatic eval metric, which we found to have excellent correlation to human judgement.



OpenEQA Examples and Statistics

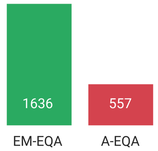
Episodic Memory (H)

Questions by Data Source

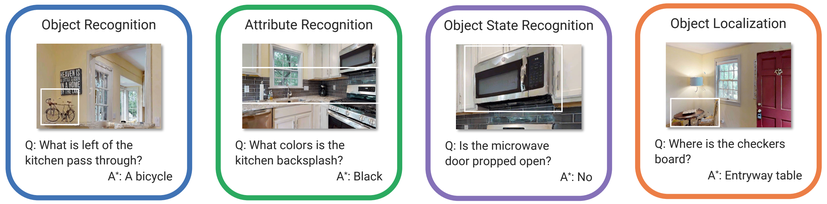


Questions by Category

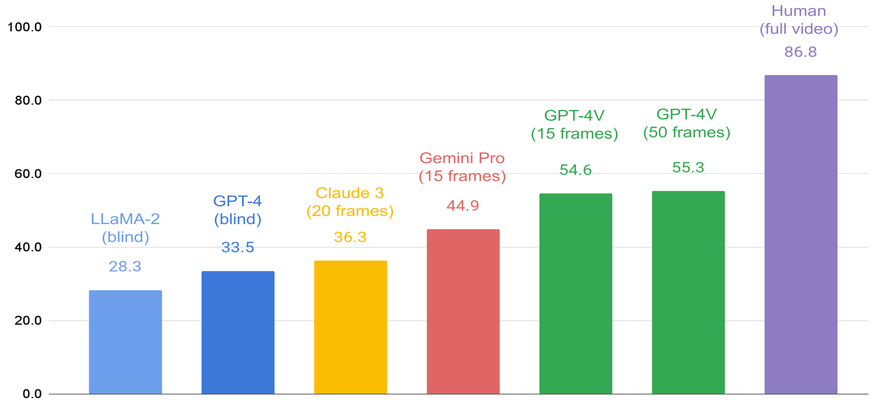
Questions per Setting

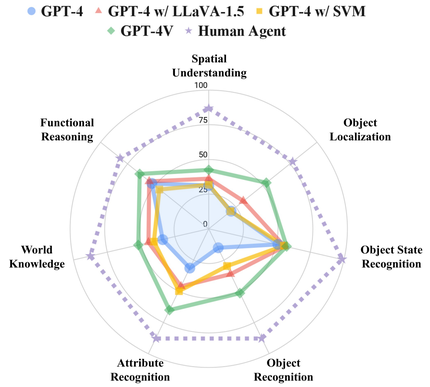


Question-Answer (Q, A*) Categories



Performance of SOTA Foundation Models





SOTA Vision-Language Models (VLMs) show stronger results than text only models in aggregate.

However, for spatial reasoning tasks, even the best VLM (GPT-4V) is “nearly blind”.

Key Takeaways:

- OpenEQA has 1600+ human-generated QA about indoor spaces.
- SOTA foundation models significantly lag humans in EQA.
- For spatial reasoning questions, current models are “nearly blind”