OpenEQA: Embodied Question Answering in the Era of Foundation Models

Arjun Majumdar*, **Anurag Ajay***, **Xiaohan Zhang***, **Pranav Putta**, **Sriram Yenamandra**, **Mikael Henaff**, **Sneha Silwal**, **Paul Mcvay**, **Oleksandr Maksymets**, **Sergio Arnaud**, **Karmesh Yadav**, **Qiyang Li**, **Ben Newman**, **Mohit Sharma**, **Vincent Berges**, **Shiqi Zhang**, **Pulkit Agrawal**, **Yonatan Bisk**, **Dhruv Batra**, **Mrinal Kalakrishnan**, **Franziska Meier**, **Chris Paxton**, **Sasha Sax**, **Aravind Rajeswaran**

Work done at Fundamental AI Research (FAIR), Meta.
*Equal Contributions.

We present a modern formulation of Embodied Question Answering (EQA) as the task of understanding an environment well enough to answer questions about it in natural language. An agent can achieve such an understanding by either drawing upon episodic memory, exemplified by agents on smart glasses, or by actively exploring the environment, as in the case of mobile robots. We accompany our formulation with OpenEQA – the first open-vocabulary benchmark dataset for EQA supporting both episodic memory and active exploration use cases. OpenEQA contains over 1600 high-quality human generated questions drawn from over 180 real-world environments. In addition to the dataset, we also provide an automatic LLM-powered evaluation protocol that has excellent correlation with human judgement. Using this dataset and evaluation protocol, we evaluate several state-of-the-art foundation models like GPT-4V and find that they significantly lag behind human-level performance. Consequently, OpenEQA stands out as a straightforward, measurable, and practically relevant benchmark that poses a considerable challenge to current generation of AI models. We hope this inspires and stimulates future research at the intersection of Embodied AI, conversational agents, and world models.

Correspondence: Aravind Rajeswaran, aravraj@meta.com
Website: https://open-eqa.github.io/

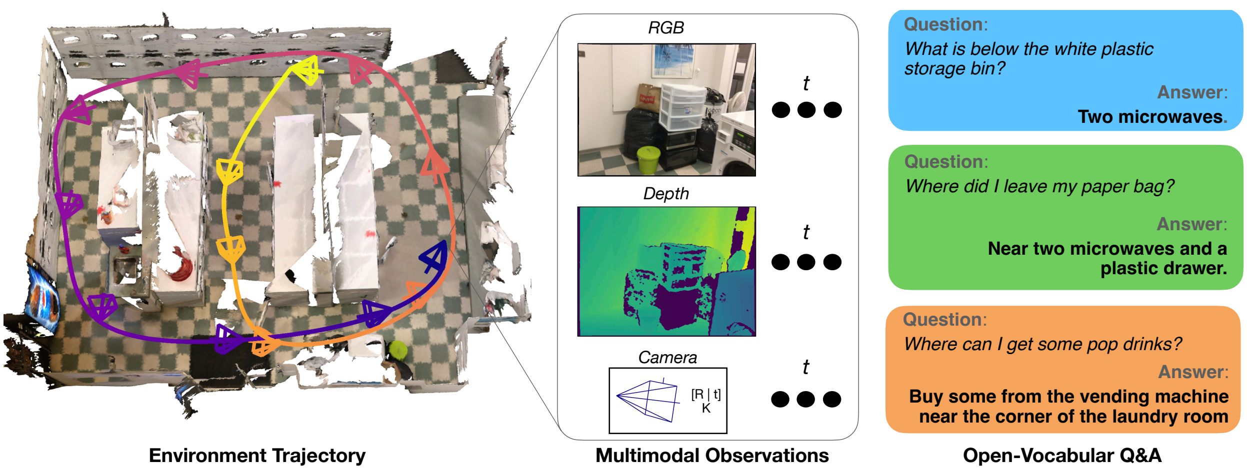


Figure 1 Illustration of an episode history along with questions and answers from our OpenEQA benchmark, which contains 1600+ untemplated questions that test aspects of attribute recognition, spatial understanding, functional reasoning, and world knowledge. In episodic-memory EQA (EM-EQA), agents parse a sequence of historical sensory observations, and in active EQA (A-EQA), agents must explore real-world scanned environments to gather information to answer questions. Open vocabulary answers are scored using our proposed LLM-Match metric, which we found to have excellent agreement with human judgement in user studies.

1 Introduction

AI agents are starting to transcend their digital origins and enter the physical world through devices like smartphones, smart glasses, and robots. These technologies are typically used by individuals who are not AI experts. To effectively assist them, Embodied AI (EAI) agents must possess a natural language interface and a type of "common sense" rooted in human-like perception and understanding of the world. Recently, "foundation models" (5) trained on massive datasets have emerged as a promising approach to develop these capabilities. Against this backdrop, we propose that Embodied Question Answering (EQA) is both a useful end-application as well as a means to evaluate an agent's understanding of the world. Simply put, EQA is the task of understanding an environment well enough to answer questions about it in natural language as illustrated in figure 1. In this work, we present OpenEQA – the first open-vocabulary benchmark for EQA, and use it to study performance of various state of the art foundation models (35; 27; 42; 16; 36; 20; 49).

Specifically, we study two variants of EQA under a common umbrella: episodic memory (EM-EQA) and active exploration (A-EQA), depending on the agent platform. For instance, EM-EQA is applicable to devices like smart glasses that can leverage episodic memory generated by human wearers to answer questions. This has the potential to enhance the memory, perceptual capabilities, and general knowledge of the user. On the other hand, A-EQA is relevant to mobile robots that can autonomously explore environments to gather necessary information to answer questions. For example, to answer the question: *'Q: Do I have Cayenne pepper left at home?'*, a robot can search the home before responding, *'A: I found a bottle of Cayenne pepper in the pantry.'*

The intersection of perception and language has long been a fertile ground for research in AI. While the broad problem of EQA (8; 52) and VQA (32; 4; 6) have been studied extensively, our approach and benchmark differ significantly along axis such as input modalities, scenes/scans of real-world spaces, and open-vocabulary questions and answers, as illustrated in table 1. In particular, OpenEQA is the first open-vocabulary benchmark for EQA, and supports both the episodic-memory and active settings. The key technologies enabling this are: (1) videos and scans of real-world environments like ScanNet (7), Gibson (45), and HM3D (37), as well as simulators capable of rending these scenes (40; 33; 12; 22; 25); and crucially (2) LLMs capable of scoring open-ended answers. This combination allows us to source questions from human annotators by watching episodes, and then automatically score responses of models against these annotated answers, enabling us to study a wide range of questions and methods (see section 3). The combination of episodes from real-world environments and open-ended questions makes OpenEQA distinct from previous EQA (8; 14), 3DQA (4; 32), and VQA (23; 34; 38; 1) benchmarks that are either closed-vocabulary (i.e. a closed set of answer), require only a single frame, use simple procedurally-generated scenes, or non-interactive in nature.

Table 1 OpenEQA vs existing benchmarks. OpenEQA has multiple modalities, real scenes, active agents, and automated scoring.

	Modalities			Open Vocab	Real Scenes	EM (video)	A(ctive)	LLM Scoring
	RGB	Depth	Camera					
EQA-v1 (8)	✔	✔	✔	✖	✖	✖	✔	✖
MP3D-EQA (43)	✔	✔	✔	✖	✔	✖	✔	✖
MT-EQA (52)	✔	✔	✔	✖	✖	✖	✔	✖
IQA (14)	✔	✔	✔	✖	✖	✖	✔	✖
SQA3D (32)	✔	✔	✔	✖	✔	✖	✖	✖
ScanQA (4)	✔	✔	✔	✖	✔	✔	✖	✖
RoboVQA (39)	✔	✖	✖	✔	✔	✔	✖	✖
SEED-Bench (24)	✔	✖	✖	✖	✔	✔	✖	✖
MMBench (30)	✔	✖	✖	✔	✔	✖	✖	✔
OpenEQA (Ours)	✔	✔	✔	✔	✔	✔	✔	✔

1.1 Our Contributions

1. **Benchmark:** Our primary contribution is a modern re-formulation of the EQA problem statement along with a concrete evaluation benchmark (OpenEQA) that contains over **1600 questions** sourced from over **180** real-world environments and scans (7; 47; 37). The questions were meticulously crowd-sourced to be representative of real-world use cases. Each question was annotated by at least three individuals, ensuring validity of questions and diversity in answer patterns. **EM-EQA** requires an agent to answer questions by leveraging the provided episodic memory. For **A-EQA**, we focus on the subset of questions in simulation of photo-realistic scanned environments. The EAI agent is spawned at an initial location and must take any required exploratory actions to answer the question. The agent is rated on both the correctness of the answer as well as efficiency of its actions, to reward agents that perform targeted exploration specific to the question.

2. **Evaluation:** The open-vocabulary nature of our benchmark increases the complexity of evaluating answers generated by various models. While human evaluations have been the gold-standard in benchmarking LLMs, they can often be prohibitively slow and/or expensive. We thus utilize an LLM (35; 42) to score answers based on similarity to ground truth answers generated by humans. Through a double blind study, we find that there is a strong correlation between our LLM-Match metric and human preferences.

3. **Baselines:** Additionally, we provide a core set of baseline results and implementations. These include recently released multi-modal LLMs like GPT-4V, Claude-3, and Gemini-1.5; as well as Socratic use of LLMs (35; 42) by leveraging captioning models (28) or generated scene-graph representations (16). Through our evaluation, we find that GPT-4V is the strongest baseline achieving an aggregate score of 48.5%. While impressive, this significantly lags behind human-level score of 85.9% on our benchmark, underscoring its difficulty and relevance for our community. In particular, all the current generation of foundation models especially struggle at questions that require spatial understanding of objects and scenes, often performing no better than "blind" LLMs, highlighting a major deficiency.

2 Benchmark and Evaluation

In this section, we provide an overview of the EM-EQA and A-EQA problem statements, how they are instantiated in OpenEQA, the dataset collection process, and the evaluation metrics.

2.1 Episodic-Memory Question Answering

The episodic memory EQA (EM-EQA) task is concerned with the setting where an agent must develop an understanding of the environment from its episodic memory to answer questions. This is particularly relevant for EAI agents embedded in devices such as smart glasses, which cannot autonomously explore the world and must rely on the history of observations to assist users (e.g. *'Q: I can't find my keys, where did I leave them? A: On the kitchen island.'*) An instance of EM-EQA is defined by the 3-tuple: (Q, H, A^*) where

Figure 2 Example questions and dataset statistics of OpenEQA. The episode history H provides a human-like tour of a home. EQA agents must answer diverse, human-generated questions Q from 7 EQA categories, aiming match the ground answers A^*. Tours are collected from diverse environments including home and office locations (not shown above). Additional dataset examples are in appendix L. Dataset statistics (right) break down the question distribution by video source (top), question category (middle), and episodic memory vs active setting. Note that, by design, the HM3D questions are shared across the EM-EQA and A-EQA settings.

Q refers to an open-vocabulary question, H is a history of observations (i.e. episodic memory), and A^* is a ground truth answer (e.g. as annotated by a human). The agent's task is to generate an answer using the episodic memory, i.e. `A = EMEQA_Agent(Q, H)`, that is *"similar"* to the ground truth answer A^*. A concrete function signature that is expected for the agent is described in algorithm 1 in appendix C.

2.2 Active Embodied Question Answering

The *Active* EQA (A-EQA) problem studies the setting where an autonomous agent can answer questions by taking exploratory, information gathering actions when necessary (e.g. *'Q: Do we have canned tomatoes at home? A: Yes, I found canned tomatoes in the pantry.'*). For simplicity, our benchmark considers questions that require only navigation actions. In principle, this can be extended to mobile manipulators to allow for both navigation and manipulation actions (e.g. opening doors and cabinets) (50). More formally, an instance of A-EQA is specified by the 3-tuple (Q, S, A^*). Similar to Section 2.1, Q and A^* denote the question and human annotated answer, respectively. S refers to the simulator initialized at the appropriate state state (40), and encompasses all details and assets needed to recreate the environment. Once the agent is spawned at S, it must take any necessary exploratory actions before producing an answer A. Please see Algorithm 1 in appendix C for a concrete function signature of an A-EQA agent. Once the agent generates answer A, it is evaluated both for the correctness of the answer as well as the efficiency of actions.

2.3 OpenEQA Dataset Collection and Validation

To establish benchmarks for EM-EQA and A-EQA, we collect a human-generated dataset of (Q, H, A^*) using videos (7) and 3D scans of real-world environments (37; 47; 33; 40). Then, we meticulously validate each question-answer pair to provide a high-quality benchmark for EM-EQA and A-EQA. The dataset is designed to reflect the types of questions that users might ask an AI assistant embedded in smart glasses or a mobile robot assistant. We present examples and dataset statistics in figure 2 and compare it to existing benchmarks in table 1.

Data Sources. We collect episode histories H from two sources: ScanNet (7) and HM3D (37; 47). For ScanNet, we utilized RGB-D data captured from human exploration in various indoor settings, such as bedrooms and offices, directly translating these videos into episode history H. We selected 90 validation scenes and 10 test scenes from ScanNet. For the scans in HM3D rendered through Habitat, we define a heuristic exploration policy to mimic human behavior and manually verify that exploration trajectories adequately explore the space, ultimately resulting in episode histories for 87 validation scenes, as detailed in appendix A.

Question Generation. In a preliminary experiment, we showed human annotators the history H and asked them to generate question-answer pairs (Q, A^*) while playing the role of end users. This exercise led to the identification of seven EQA question categories that broadly encompass the range of questions asked of AI assistants. They test an agent's ability to (1) *recognize objects* (e.g. what is on the coffee table?), (2) *recognize object attributes* (e.g. colors or shapes), (3) *recognize object states* (e.g. open or closed), (4) *localize objects* (e.g. where are my keys?), (5) perform *spatial reasoning* (e.g. I'm sitting on the couch watching TV, in which direction should I turn to find the kitchen?), (6) perform *functional reasoning* (e.g. how can I cool down this room?), and (7) utilize *outside world knowledge* (e.g. who/what is depicted in a painting?). The final OpenEQA dataset focuses on these seven categories. Annotators were asked to generate two questions and answers per category after viewing H. Illustrations of the question categories are provided in figure 2, and additional details on the dataset collection and interface are in appendix A.

Dataset Validation. Each question created by humans underwent further examination by two independent annotators. Validators watched the episode history and assessed whether the question was unanswerable, ambiguous, or if the answer was incorrect. Any question-answer pair with identified issues was discarded. The interface for validation is provided in appendix A. The final dataset includes 1636 questions following the statistics in figure 2.

Dataset Splits. In our benchmark, the validated (Q, A^*) pairs are used for EM-EQA, and reused for A-EQA since we recorded S in addition to H for simulated scenes. Specifically, A-EQA agents are initialized at the same start state S that was used to generate the episodic memory H for EM-EQA. The existence of a feasible

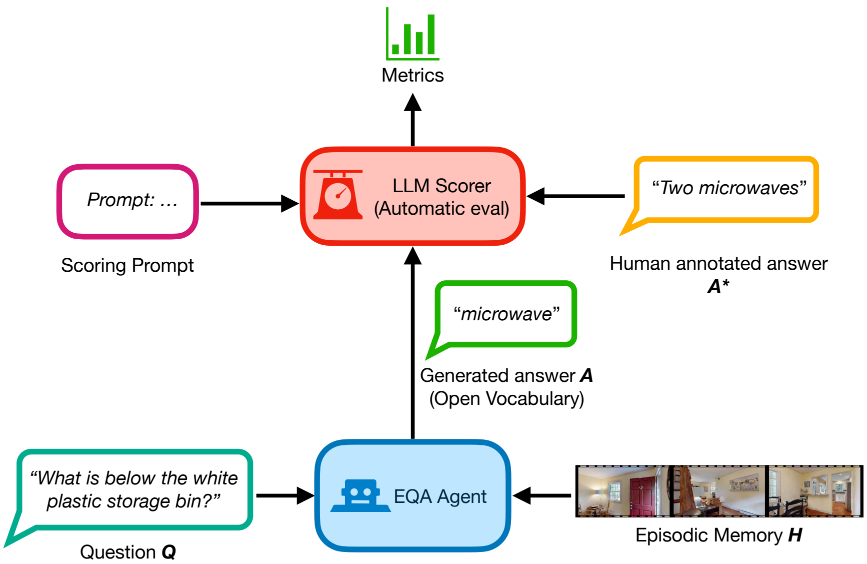


Figure 3 Illustration of LLM-Match evaluation and workflow.

trajectory H provides proof that A-EQA questions are answerable. However, they can potentially be answered more efficiently through targeted exploration.

Additional Object Localization Answers. Among the 7 EQA categories, *object localization* questions pose a unique challenge for evaluation, because they often have multiple correct answers with differences that go beyond rephrasing. For example, the question *'Q: Where is the toaster?'* may have multiple correct answers such as *'A1: to the right of the stove'* or *'A2: to the left of the fridge'*. Thus we collect 4 additional answers from 2 additional annotators resulting in 5 answers per object localization question that reflect a natural distribution of answers that humans would expect from each localization question.

2.4 LLM-Match: Evaluating Correctness of Answers

While the open-vocabulary nature makes EQA realistic, it poses a challenge for evaluation due to multiplicity of correct answers. One approach to evaluation is human trials, but it can be prohibitively slow and expensive, especially for benchmarks. As an alternative, we use an LLM to evaluate the correctness of open-vocabulary answers produced by EQA agents. Specifically, we adapt the evaluation protocol introduced in MMBench (30) to the EQA task. Given a question Q_i, human annotated answer A_i^*, and model output A_i, the LLM is prompted to provide a score $\sigma_i \in \{1, \dots 5\}$. On this scale, 1 indicates an incorrect response, 5 is a correct response, and intermediate values represent levels of similarity. We calculate an aggregate LLM-based **correctness** metric (LLM-Match) as:

$$C = \frac{1}{N} \sum_{i}^{N} \frac{\sigma_i - 1}{4} \times 100\% \ . \tag{1}$$

LLM-Match is illustrated in figure 3, detailed in app. B, and validated against human judgement in section 5.

2.5 Evaluating Efficiency for A-EQA

In A-EQA, we evaluate an agent based on two criteria: (a) **correctness** of the answer based on similarity with human annotation A^* as described in equation (1); and (b) **efficiency**, which measures how quickly the agent answered the question and favors agents that perform targeted exploration necessary for the question.

We measure efficiency by weighting the correctness metric σ_i by the normalized length of the agent's path $l_i / \max(p_i, l_i)$, where p_i is the timesteps taken by the agent and l_i is the timesteps taken in a ground truth path that is sufficient for answering the question Q_i. Formally, our **efficiency** metric is defined as:

$$E = \frac{1}{N} \sum_{i}^{N} \frac{(\sigma_i - 1)}{4} \times \frac{l_i}{\max(p_i, l_i)} \times 100\%, \tag{2}$$

which can be seen as modified version of the success weighted by path length (SPL) (2), a metric commonly used to measure the efficiency of navigation agents. We note that the ground-truth path was generated by using a scripted exploratory agent. Since this path was used to construct the (Q, A) pairs, it is guaranteed to contain sufficient information to answer. However, we note that these paths need not be optimal, and thus in principle $E > 100\%$ is theoretically possible.

3 EQA Agents

This section describes the different EQA agents we study and evaluate in this work. Our guiding principle is to explore different ways in which foundation models (specifically LLMs and VLMs) can be used for EQA without any additional fine-tuning. Towards this goal, the family of agents studied are: (1) blind LLMs (35; 42), (2) Socractic LLMs w/ frame captions (27), (3) Socractic LLMs w/ scene-graph representations (16), and (4) VLMs that can directly process multiple frames (e.g. GPT-4V (49)). For simplicity, we first describe the agents in the EM-EQA setting, and subsequently discuss extensions to A-EQA. All agents have the general signature of $A = \text{Agent}(Q, H)$ and contain a language model component that generates the answer. The agents primarily differ in their perception capabilities and how they process H. In addition to these agents, we also study how humans perform in our benchmark.

Blind LLMs. The text-only or 'blind' LLM agent simply produces an answer based on the question Q without considering any visual information about the scene, i.e. $A = \text{LLM}([\omega, Q])$, where ω is a generic prompt that we prepend to the question. See appendix D for additional details. This agent provides a reference for how far we can get purely using prior world knowledge and/or random guessing (e.g. yes/no questions). For the LLM choice, we present results with both GPT-4 (35) and LLaMA-2-70B (42).

Socratic LLMs w/ Frame Captions. This is the simplest agent we study that leverages the perceptual information from the episodic memory H. Let $\{X_1, X_2, \ldots X_K\}$ be K frames drawn from the episodic memory H. We first leverage an image captioning model (e.g. LLaVA (27; 28)) to generate $z_i = \text{Captioner}(X_i),\ i = 1, \ldots K$. These captions provide a language description of the episodic memory to the LLM, which could allow it to answer better than a blind agent. The final answer is computed by the agent using a generic prompt, the aforementioned frame captions, and the question, i.e. $A = \text{LLM}([\omega, z_1, z_2, \ldots z_K, Q])$. See appendix D for

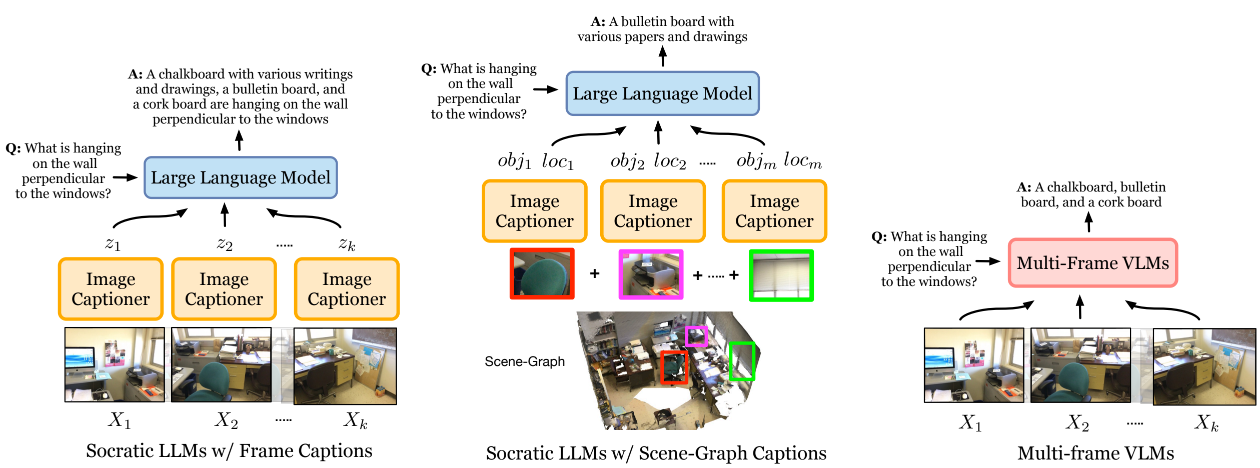


Figure 4 EQA Agents (Left) Socratic LLMs w/ Image Captions generates captions for frames from episodic memory and provides it as context to an LLM to generate answer. (Middle) Socratic LLMs w/ Scene-Graph Captions leverage an object-centric scene-graph representation of episodic memory, which includes captions of object-centric crops and their 3D locations. (Right) Multi-Frame VLM directly processes visual frames from episodic memory to answer the question.

an example of the input. In practice, we sample K frames uniformly over time from H, with $K = 50$ for EM-EQA and $K = 75$ for A-EQA. For the captioning model, we use LLaVA-v1.5 (27), and for the LLM we study both GPT4 (35) and LLaMA-2-70B (42).

Socratic LLMs w/ Scene-Graph Captions. We next study agents that leverage an object-centric scene-graph representation of H. The motivation for such agents is that an object-centric representation might allow for a more fine-grained perceptual understanding of objects, and provide a textual representation that might be easier for LLMs to reason over. Object-centric 3D world representations involve constructing a scene graph $G = \text{SceneGraph}(H)$ that contains a description of the objects in the scene, their semantic attributes such as color and 3D locations, and their relationships. We study two methods of constructing such a scene-graph: (1) ConceptGraph (16); and (2) Sparse Voxel Map (SVM). ConceptGraph (CG) generates a textual scene-graph representation by first detecting various objects in the scene, extracting the 3D location of objects using camera pose and depth information, and sematic descriptions of objects by using an image captioning model on crops of the object extracted from the video. We use the publicly released implementation of CG, which uses Grounded-SAM (29; 19) with RAM (54) for object detection and LLaVA-v1 (28) for image captioning. SVMs are constructed similarly to CGs, but differ in the post-processing of object detections and in the image captioning model used. See appendix E for details. Once a textual scene graph G is generated, we use it for EQA as $A = \text{LLM}([\omega, G, Q])$.

Multi-Frame VLMs. The most generic agent for EQA is one that can directly process the entire episodic memory to answer questions, i.e. $A = \text{MultiFrameVLM}([\omega, Q, H])$. The recently released GPT-4V model (49) is capable of processing up to 50 frames (through the API) in addition to textual queries. We thus extract 50 frames uniformly spaced from H and provide it to GPT-4V in addition to prompts for generating the answer. More recently, other multi-frame VLMs have also been released such as the Claude and Gemini family of models, which we benchmark as well. See appendix D for details on implementations and prompts, as well as Table 4 for full results across all models.

Human Agent. Finally, we also run a study with human participants to establish human-level performance metrics on our benchmark. We collect answers from a set of human annotators by providing each annotator with a video of the episode history H and asking them to answer all of the questions Q for that scene. We enrolled two independent participants for this benchmarking exercise and found strong agreement in responses.

Agents for A-EQA. So far, we have described agents that can answer questions Q given an episode history H. However, in the case of A-EQA, no explicit H is provided, and agent must generate its own observations through exploration. In this work, we provide the simplest baseline for A-EQA that explores environments in a **question agnostic** manner using frontier exploration (48). By leveraging the episodic memory from frontier exploration as H, we can re-use all the aforementioned agents, just with a different and self-generated episodic memory. We note that the efficiency score of such an agent is expected to be poor, and we leave open the challenge of more efficient A-EQA agents to future work.

Force-A-Guess when Agents Abstain. To recall, all EQA agents we study involve an LLM component and differ primarily in how the episode history is used. In our experiments, we observed that such agents can often be overly conservative and **abstain** from answering, especially when a model thinks it lacks sufficient context. In our evaluation metric, we do not make a special provision for abstaining, and consider abstaining an incorrect answer. Thus, we force the agent to take a guess to give it at least an informed random chance, instead of immediately counting it as a failure.

Despite our best efforts, we were not able to force non-blind agents to guess through prompt engineering. However, blind LLMs are able to guess purely based on prior knowledge, and seldom abstain. Thus, for non-blind agents we first check if the agent abstained. If it did, we force the corresponding blind LLM to guess, and use the generated answer. Full details of this protocol are in appendix F and an analysis of the effects of this procedure are in appendix G. All results in the main paper use this force-a-guess protocol.

4 Experimental Results on OpenEQA

We present evaluation results of agents described in section 3. Table 2 reports the overall LLM-Match scores (C) (see Eq. 1) of the baselines evaluated on the EM-EQA and A-EQA benchmarks, where EM-EQA results

Table 2 LLM-Match and efficiency scores on OpenEQA. EM-EQA results are broken down by data source (ScanNet, HM3D, and ALL). A-EQA results include both LLM-Match scores (Eq. 1) and agent efficiency (Eq. 2). *For A-EQA, GPT-4V scores are calculated on a random subset of 184 questions.

# method	EM-EQA			A-EQA	
	ScanNet equation (1)	HM3D equation (1)	ALL equation (1)	HM3D equation (1)	HM3D equation (2)
Blind LLMs					
1 GPT-4	$32.5_{\pm 1.2}$	$35.5_{\pm 1.7}$	$33.5_{\pm 1.0}$	$35.5_{\pm 1.7}$	N/A
2 LLaMA-2	$27.9_{\pm 1.2}$	$29.0_{\pm 1.7}$	$28.3_{\pm 1.0}$	$29.0_{\pm 1.7}$	N/A
Socratic LLMs w/ Frame Captions					
3 GPT-4 w/ LLaVA-1.5	$45.4_{\pm 1.3}$	$40.0_{\pm 1.8}$	$43.6_{\pm 1.1}$	$38.1_{\pm 1.8}$	$7.0_{\pm 0.4}$
4 LLaMA-2 w/ LLaVA-1.5	$39.6_{\pm 1.3}$	$31.1_{\pm 1.8}$	$36.8_{\pm 1.1}$	$30.9_{\pm 1.8}$	$5.9_{\pm 0.4}$
Socratic LLMs w/ Scene-Graph Captions					
5 GPT-4 w/ CG	$37.8_{\pm 1.3}$	$34.0_{\pm 1.7}$	$36.5_{\pm 1.0}$	$34.4_{\pm 1.8}$	$6.5_{\pm 0.4}$
6 LLaMA-2 w/ CG	$31.0_{\pm 1.2}$	$24.2_{\pm 1.6}$	$28.7_{\pm 1.0}$	$23.9_{\pm 1.6}$	$4.3_{\pm 0.3}$
7 GPT-4 w/ SVM	$40.9_{\pm 1.3}$	$35.0_{\pm 1.8}$	$38.9_{\pm 1.0}$	$34.2_{\pm 1.8}$	$6.4_{\pm 0.4}$
8 LLaMA-2 w/ SVM	$36.0_{\pm 1.3}$	$30.9_{\pm 1.8}$	$34.3_{\pm 1.0}$	$29.9_{\pm 1.7}$	$5.5_{\pm 0.4}$
Multi-Frame VLMs					
9 GPT-4V	$57.4_{\pm 1.3}$	$51.3_{\pm 1.8}$	$55.3_{\pm 1.1}$	$41.8_{\pm 3.2^*}$	$7.5_{\pm 0.6^*}$
Human Agent	$87.7_{\pm 0.7}$	$85.1_{\pm 1.1}$	$86.8_{\pm 0.6}$	$85.1_{\pm 1.1}$	N/A

are separately reported on each of the data sources (i.e., ScanNet and HM3D). It also presents the efficiency score on A-EQA, as described in Eq. 2, along with bootstrapped standard errors. Based on the results, we first share some observations and remarks, and discuss specific observations in section 5.

1. Humans achieve excellent performance on the benchmark (>85%), which confirms the validity of the benchmark and correctness of evaluation metrics.
2. Multi-frame VLMs (i.e., GPT-4V) outperform other agents. This suggests that a tight integration of perception and language may significantly benefit EQA.
3. We find that blind LLMs are surprisingly strong baselines, with GPT-4 and LLaMA-2 achieving a score of 33.5 and 28.3 respectively on EM-EQA. While this is substantially lower than GPT-4V or human-level performance, it suggests a large degree of regularity in the world and that answers to several questions can be "guessed" without explicit visual context of a specific environment. See section 5 for additional discussion. We also note that early works in VQA(1) found blind agents to be strong baselines.
4. Within each family of agents we consistently find that agents that use GPT-4 as the LLM outperform LLaMA-2. This suggests that larger LLMs can be a key enabling factor for good EQA performance.
5. In the EM-EQA benchmark, we find that all agents with access to perceptual information in the form of frame captions or scene-graphs outperform blind LLMs (under the force-a-guess protocol). This again underscores the importance of perception for EQA.
6. When comparing the performance of agents in EM-EQA and A-EQA, we generally observe lower scores in A-EQA. In part, this is due to longer trajectories due to the use of exhaustive exploration in our A-EQA agents, which forces a longer history representation often with irrelevant information for a specific question. In several situations, this makes the performance of various agents comparable to that of blind LLMs or even lower (e.g. GPT-4 w/ ConceptGraphs). This underscores the challenging nature of the A-EQA benchmark and the importance of efficient exploration in interactive settings.

Figure 5 breaks down performance on EM-EQA (human-like trajectories) by the seven question categories described in section 2.3. Among all the categories, EQA agents excel at functional reasoning and object state

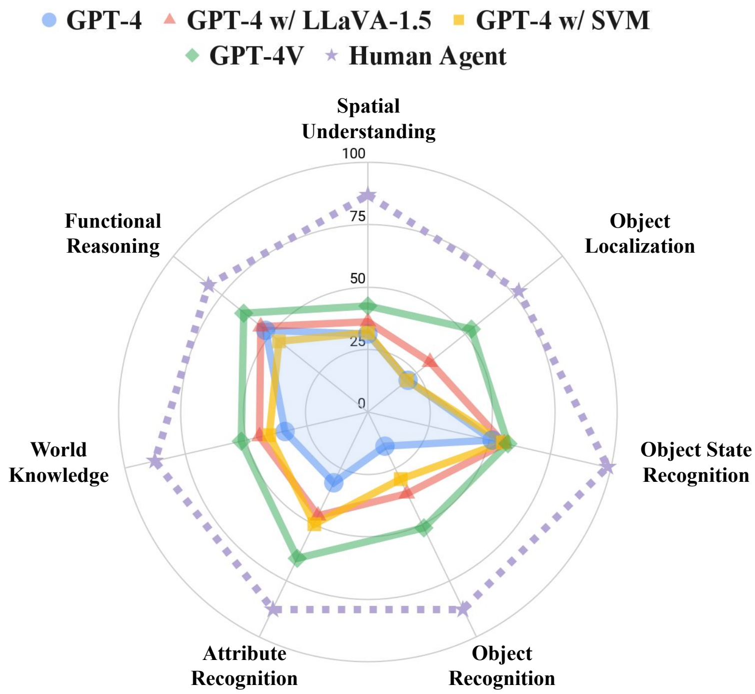


Figure 5 Category-level performance on EM-EQA. We find that agents with access to visual information excel at localizing and recognizing objects and attributes, and make better use of this information to answer questions that require world knowledge. However, on other categories performance is closer to the blind LLM baseline (GPT-4), indicating substantial room for improvement on OpenEQA. See scores for all methods in appendix H.

recognition questions, with LLM-Match scores around 50% when averaged across all agents. Additionally, EQA agents perform well on world knowledge questions, which require a commonsense understanding of the world and represent an EQA category where large language model may add substantial value. By contrast, EQA agents perform poorly on spatial reasoning questions. To our surprise, agents that use scene-graph representations are no better than frame-captioning agents, even on spatial reasoning questions. This suggests that more work is needed to incorporate understanding of space and geometry into large models. While most models achieve nontrivial performance on all categories, there remains a large gap between even the best method and human-level performance.

5 Analysis and Discussions

Human Alignment and Robustness of LLM-Match Evaluating open-vocabulary responses to questions is an open challenge in AI. While human evaluation remains the gold-standard, it is also expensive and time consuming. An automatic evaluation metric is preferable for benchmarking, fast iteration, and model selection. For this, we proposed the LLM-Match metric in section 2.4. We now test this metric along two axis: (1) How closely aligned is the LLM-Match metric with human evaluators? (2) How sensitive is the LLM-Match metric towards specific choice of prompts and the LLM?

To answer the question on **human alignment**, we designed an experiment to measure the agreement between LLM-Match metric and human evaluators. We uniformly sampled a subset of 300 questions from the dataset. To ensure coverage of the answer distributions, we sampled responses from blind LLaMA-2, GPT-4V, and human annotated answers. In a double blind study, we then asked 4 human evaluators to score the 300 responses using an evaluation prompt similar to the one used by LLM-Match. The evaluators were provided no information about the source of the response. We found a **Spearman's** $\rho = 0.909$ **between human and LLM evaluation** (bootstrap CI=(0.883,0.928), N=9999), indicating excellent agreement with human judgement. For reference, human evaluators correlated with each other in $\rho \in [0.91, 0.93]$. Essentially, LLM-Match agrees with human evaluation nearly as much as human subjects do with one another.

To answer the question of **LLM-Match robustness**, we designed an experiment to test its sensitivity under small perturbations of the prompt (see appendix J for details). Table 8 in appendix J demonstrates that changing the LLM's role from 'AI' to 'Score Master' or 'professional evaluator' does not significantly change the results, and scores between any two treatments have a tight correlation with a Spearman's $\rho > 0.95$. Similarly, Table 9 in appendix J shows analogous results $\rho > 0.95$ for changing the description of a '5' from 'perfect match' to 'contains correct answer', 'similar to a reasonable person', or 'reasonable professional'. Sensitivity to seed and temperature has negligible impact as well. Finally, we vary the LLM used for scoring and find that GPT4 has excellent agreement with human judgement, but GPT-3.5 and LLaMA-2 have significantly lower correlation (ρ<0.7). Thus, **for now, we recommend using only GPT4 for LLM-Match.**

Discussion on Blind LLMs. We found blind LLMs to be a surprisingly strong baseline, considering they have no access to visual information about the scene. Upon closer inspection, we found that blind LLMs, especially GPT-4, are good at "guessing" answers to EQA questions. For instance, consider the question: *'Q: What is the color of the staircase railing?'* GPT-4 answers *'brown'*, and because many houses have a *brown staircase railing*, this guess is often correct. This indicates a certain degree of regularity in the world such that answers to many questions are similar across different environments. Nevertheless, we establish a strong lower bound of performance achievable without perception, and we can infer that any additional gains are due perception and semantic grounding.

Force-A-Guess. When studying Socractic LLMs augmented with perceptual information (image or scene-graphs captions), we found that agents often abstained from answering the question (e.g. *'Not enough information to answer the question.'*). As noted in section 3, our LLM-Match metric does not give preferential scoring of abstaining vis-a-vis an incorrect answer. Thus, we defaulted to the answer from the blind LLM powering an agent when it abstained. In appendix G, we provide statistics on how frequently each agent abstained, and study performance without defaulting to a blind LLM. In general, we find that GPT-4-based Socratic agents abstain frequently (up to 55% of the time), and thus, rely more heavily on the blind LLM-based score correction that we apply in our benchmark evaluations. By contrast, GPT-4V and LLaMA-2 based models do not abstain as often (up to 12% of the time), and thus the differences between the two variants is minimal.

Does explicit coordinate information help? The primary motivation for object-centric scene-graph representations is to have fine-grained perceptual understanding of objects and their locations. Thus, we intuitively expect that agents equipped with explicit object locations will fare better in questions that require spatial understanding. Surprisingly, we find in table 2 that such agents fare no better than Socratic LLMs that simply use frame-level captions. We run an ablation experiment (appendix K) where we remove explicit bounding box and size information from the scene graph, and find that this does not significantly affect performance, indicating that these LLMs are not able to effectively use 3D coordinate information in text.

6 Related Work

The intersection of perception and language (4; 17; 24; 26; 10; 31; 13; 56; 21) has long been a fertile ground for AI research. Prior works studying perception and language have proposed Visual Question Answering (VQA) benchmarks, such as VQA-v1 (1), VQA-v2 (15), OK-VQA (34) and A-OKVQA (38), that focus on answering questions from a single image. Later works extended question answering tasks to videos (55; 53; 23) and 3D scenes (32; 6; 4; 17). These include benchmark such as VideoQA (55), SQA3D (32) and ScanRefer (6). While conceptually similar to our EM-EQA setting, these prior benchmarks focused on singular and narrow themes such as situated reasoning, object localization, object recognition, activity recognition, temporal window localization, and future forecasting (44; 18; 24; 53; 23; 46; 32; 6). Another closely related line of work is prior benchmarks on Embodied QA (8; 43; 52; 9) and is conceptually similar to our A-EQA setting. They focus on leveraging RGB-D to accomplish navigation tasks in simulation (43), in which the agent must seek out multiple target locations or objects sampled from a closed vocabulary set (52) Our work takes inspiration from such prior works (8) and modernizes it to be relevant in the current era of foundation models. To our knowledge, our benchmark is the only one that incorporates all elements of a real-world use case for EQA: (1)

The study of both episodic memory and active settings to accommodate for a wide variety of embodied agents like smartphones and mobile robots, (2) High quality real-world datasets with broad and non-templated questions, and (3) Embracing open-vocabulary interactions with users. In addition, our baselines use modern foundation models trained on vast internet data, enabling world knowledge beyond the reach of methods trained solely on simulated interactions.

LLMs have been used to scale the size of benchmarks either with their use for question and answer generation (24) or during evaluation time (30; 51). Evaluation of open-vocabulary answers remains an open problem in AI. While the gold-standard remains human evaluations, they are time-consuming and expensive. An automatic evaluation process is desirable for benchmarking, quick iteration of research ideas, and model selection. We setup such a process by taking inspiration from recent works that study if LLMs can be used as an evaluation proxy in place of human raters (30). Through a randomized control trial, we found a high correlation between human ratings and GPT-4, as evidenced by a Spearman correlation coefficient of 0.909.

7 Conclusion

We introduce OpenEQA, the first realistic benchmark to study open-vocabulary EQA in both episodic memory and active settings. Specifically, OpenEQA includes challenging, human-generated, open-vocabulary questions that require understanding an environment and answering question in natural language. Our benchmark is primarily enabled by (1) videos and scans of real-world indoor environments and (2) LLMs that can be used for scoring open-ended answers in an efficient and reliable manner, as we demonstrated through our analyses. We use OpenEQA to benchmark various state-of-the-art foundation models and their combinations. This includes approaches that leverage image captions, scene-graph construction, and multi-frame VLMs such as GPT-4V. Ultimately, we find a large gap between the best models (GPT-4V at 55.3%) and human-level performance (at 86.8%). In particular, for questions that require spatial understanding, the aforementioned agents perform no better than blind LLMs, suggesting that further improvement on perception and semantic grounding is necessary before EQA agents are ready for real-world domains. In an era where LLMs are smashing hard QA tasks (e.g. SAT math exams), OpenEQA stands out as a straightforward, quantifiable, and practically relevant benchmark that poses considerable challenge to the current generation of foundation models. We thus believe OpenEQA is well positioned to serve as barometer for tracking future progress in multimodal learning and scene/environment understanding.

References

[1] Aishwarya Agrawal, Jiasen Lu, Stanislaw Antol, Margaret Mitchell, C. Lawrence Zitnick, Dhruv Batra, and Devi Parikh. Vqa: Visual question answering, 2016.

[2] Peter Anderson, Angel Chang, Devendra Singh Chaplot, Alexey Dosovitskiy, Saurabh Gupta, Vladlen Koltun, Jana Kosecka, Jitendra Malik, Roozbeh Mottaghi, Manolis Savva, et al. On evaluation of embodied navigation agents. *arXiv preprint arXiv:1807.06757*, 2018.

[3] Anthropic. Introducing the next generation of claude, 2024.

[4] Daichi Azuma, Taiki Miyanishi, Shuhei Kurita, and Motoaki Kawanabe. Scanqa: 3d question answering for spatial scene understanding, 2022.

[5] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

[6] Dave Zhenyu Chen, Angel X Chang, and Matthias Nießner. Scanrefer: 3d object localization in rgb-d scans using natural language. *16th European Conference on Computer Vision (ECCV)*, 2020.

[7] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proc. Computer Vision and Pattern Recognition (CVPR), IEEE*, 2017.

[8] Abhishek Das, Samyak Datta, Georgia Gkioxari, Stefan Lee, Devi Parikh, and Dhruv Batra. Embodied Question Answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[9] Abhishek Das, Georgia Gkioxari, Stefan Lee, Devi Parikh, and Dhruv Batra. Neural Modular Control for Embodied Question Answering. In *Proceedings of the Conference on Robot Learning (CoRL)*, 2018.

[10] Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al. Palm-e: An embodied multimodal language model. *arXiv preprint arXiv:2303.03378*, 2023.

[11] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *kdd*, volume 96, pages 226–231, 1996.

[12] Chuang Gan, Jeremy Schwartz, Seth Alter, Martin Schrimpf, James Traer, Julian De Freitas, Jonas Kubilius, Abhishek Bhandwaldar, Nick Haber, Megumi Sano, Kuno Kim, Elias Wang, Damian Mrowca, Michael Lingelbach, Aidan Curtis, Kevin Feigelis, Daniel M. Bear, Dan Gutfreund, David Cox, James J. DiCarlo, Josh McDermott, Joshua B. Tenenbaum, and Daniel L. K. Yamins. Threedworld: A platform for interactive multi-modal physical simulation, 2020.

[13] Peng Gao, Jiaming Han, Renrui Zhang, Ziyi Lin, Shijie Geng, Aojun Zhou, Wei Zhang, Pan Lu, Conghui He, Xiangyu Yue, et al. Llama-adapter v2: Parameter-efficient visual instruction model. *arXiv preprint arXiv:2304.15010*, 2023.

[14] Daniel Gordon, Aniruddha Kembhavi, Mohammad Rastegari, Joseph Redmon, Dieter Fox, and Ali Farhadi. Iqa: Visual question answering in interactive environments. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4089–4098, 2018.

[15] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering, 2017.

[16] Qiao Gu, Alihusein Kuwajerwala, Sacha Morin, Krishna Murthy Jatavallabhula, Bipasha Sen, Aditya Agarwal, Corban Rivera, William Paul, Kirsty Ellis, Rama Chellappa, Chuang Gan, Celso Miguel de Melo, Joshua B. Tenenbaum, Antonio Torralba, Florian Shkurti, and Liam Paull. Conceptgraphs: Open-vocabulary 3d scene graphs for perception and planning. *arXiv*, 2023.

[17] Yining Hong, Haoyu Zhen, Peihao Chen, Shuhong Zheng, Yilun Du, Zhenfang Chen, and Chuang Gan. 3d-llm: Injecting the 3d world into large language models, 2023.

[18] Zhijian Hou, Lei Ji, Difei Gao, Wanjun Zhong, Kun Yan, Chao Li, Wing-Kwong Chan, Chong-Wah Ngo, Nan Duan, and Mike Zheng Shou. Groundnlq@ ego4d natural language queries challenge 2023. *arXiv preprint arXiv:2306.15255*, 2023.

[19] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. *arXiv:2304.02643*, 2023.

[20] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023.

[21] Dohwan Ko, Ji Soo Lee, Wooyoung Kang, Byungseok Roh, and Hyunwoo J Kim. Large language models are temporal and causal reasoners for video question answering. *arXiv preprint arXiv:2310.15747*, 2023.

[22] Eric Kolve, Roozbeh Mottaghi, Winson Han, Eli VanderBilt, Luca Weihs, Alvaro Herrasti, Daniel Gordon, Yuke Zhu, Abhinav Gupta, and Ali Farhadi. AI2-THOR: An Interactive 3D Environment for Visual AI. *arXiv*, 2017.

[23] Jie Lei, Licheng Yu, Mohit Bansal, and Tamara L Berg. Tvqa: Localized, compositional video question answering. *arXiv preprint arXiv:1809.01696*, 2018.

[24] Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension, 2023.

[25] Chengshu Li, Fei Xia, Roberto Martín-Martín, Michael Lingelbach, Sanjana Srivastava, Bokui Shen, Kent Elliott Vainio, Cem Gokmen, Gokul Dharan, Tanish Jain, Andrey Kurenkov, Karen Liu, Hyowon Gweon, Jiajun Wu, Li Fei-Fei, and Silvio Savarese. igibson 2.0: Object-centric simulation for robot learning of everyday household tasks. In *5th Annual Conference on Robot Learning*, 2021.

[26] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023.

[27] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning, 2023.

[28] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023.

[29] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023.

[30] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. Mmbench: Is your multi-modal model an all-around player?, 2023.

[31] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32, 2019.

[32] Xiaojian Ma, Silong Yong, Zilong Zheng, Qing Li, Yitao Liang, Song-Chun Zhu, and Siyuan Huang. Sqa3d: Situated question answering in 3d scenes. In *International Conference on Learning Representations*, 2023.

[33] Manolis Savva*, Abhishek Kadian*, Oleksandr Maksymets*, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, Devi Parikh, and Dhruv Batra. Habitat: A Platform for Embodied AI Research. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

[34] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge, 2019.

[35] OpenAI. Gpt-4 technical report. *ArXiv*, abs/2303.08774, 2023.

[36] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021.

[37] Santhosh Kumar Ramakrishnan, Aaron Gokaslan, Erik Wijmans, Oleksandr Maksymets, Alexander Clegg, John M Turner, Eric Undersander, Wojciech Galuba, Andrew Westbury, Angel X Chang, Manolis Savva, Yili Zhao, and Dhruv Batra. Habitat-matterport 3d dataset (HM3d): 1000 large-scale 3d environments for embodied AI. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021.

[38] Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. A-okvqa: A benchmark for visual question answering using world knowledge, 2022.

[39] Pierre Sermanet, Tianli Ding, Jeffrey Zhao, Fei Xia, Debidatta Dwibedi, Keerthana Gopalakrishnan, Christine Chan, Gabriel Dulac-Arnold, Sharath Maddineni, Nikhil J Joshi, et al. Robovqa: Multimodal long-horizon reasoning for robotics. *arXiv preprint arXiv:2311.00899*, 2023.

[40] Andrew Szot, Alex Clegg, Eric Undersander, Erik Wijmans, Yili Zhao, John Turner, Noah Maestre, Mustafa Mukadam, Devendra Chaplot, Oleksandr Maksymets, Aaron Gokaslan, Vladimir Vondrus, Sameer Dharur, Franziska Meier, Wojciech Galuba, Angel Chang, Zsolt Kira, Vladlen Koltun, Jitendra Malik, Manolis Savva, and Dhruv Batra. Habitat 2.0: Training home assistants to rearrange their habitat. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

[41] Gemini Team. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

[42] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[43] Erik Wijmans, Samyak Datta, Oleksandr Maksymets, Abhishek Das, Georgia Gkioxari, Stefan Lee, Irfan Essa, Devi Parikh, and Dhruv Batra. Embodied Question Answering in Photorealistic Environments with Point Cloud Perception. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[44] Bo Wu, Shoubin Yu, Zhenfang Chen, Joshua B. Tenenbaum, and Chuang Gan. STAR: A benchmark for situated reasoning in real-world videos. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021.

[45] Fei Xia, Amir R. Zamir, Zhi-Yang He, Alexander Sax, Jitendra Malik, and Silvio Savarese. Gibson env: real-world perception for embodied agents. In *Computer Vision and Pattern Recognition (CVPR), 2018 IEEE Conference on*. IEEE, 2018.

[46] Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. Next-qa: Next phase of question-answering to explaining temporal actions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9777–9786, 2021.

[47] Karmesh Yadav, Ram Ramrakhya, Santhosh Kumar Ramakrishnan, Theo Gervet, John Turner, Aaron Gokaslan, Noah Maestre, Angel Xuan Chang, Dhruv Batra, Manolis Savva, et al. Habitat-matterport 3d semantics dataset. *arXiv preprint arXiv:2210.05633*, 2022.

[48] Brian Yamauchi. A frontier-based approach for autonomous exploration. In *Proceedings 1997 IEEE International Symposium on Computational Intelligence in Robotics and Automation CIRA'97.'Towards New Computational Principles for Robotics and Automation'*, pages 146–151. IEEE, 1997.

[49] Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. The dawn of lmms: Preliminary explorations with gpt-4v (ision). *arXiv preprint arXiv:2309.17421*, 9, 2023.

[50] Sriram Yenamandra, Arun Ramachandran, Karmesh Yadav, Austin Wang, Mukul Khanna, Theophile Gervet, Tsung-Yen Yang, Vidhi Jain, Alexander William Clegg, John Turner, Zsolt Kira, Manolis Savva, Angel Chang, Devendra Singh Chaplot, Dhruv Batra, Roozbeh Mottaghi, Yonatan Bisk, and Chris Paxton. Homerobot: Open vocabulary mobile manipulation. *arXiv preprint arXiv:2306.11565*, 2023.

[51] Zhenfei Yin, Jiong Wang, Jianjian Cao, Zhelun Shi, Dingning Liu, Mukai Li, Lu Sheng, Lei Bai, Xiaoshui Huang, Zhiyong Wang, Jing Shao, and Wanli Ouyang. Lamm: Language-assisted multi-modal instruction-tuning dataset, framework, and benchmark, 2023.

[52] Licheng Yu, Xinlei Chen, Georgia Gkioxari, Mohit Bansal, Tamara L. Berg, and Dhruv Batra. Multi-target embodied question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[53] Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueting Zhuang, and Dacheng Tao. Activitynet-qa: A dataset for understanding complex web videos via question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 9127–9134, 2019.

[54] Youcai Zhang, Xinyu Huang, Jinyu Ma, Zhaoyang Li, Zhaochuan Luo, Yanchun Xie, Yuzhuo Qin, Tong Luo, Yaqian Li, Shilong Liu, et al. Recognize anything: A strong image tagging model. *arXiv preprint arXiv:2306.03514*, 2023.

[55] Yaoyao Zhong, Junbin Xiao, Wei Ji, Yicong Li, Weihong Deng, and Tat-Seng Chua. Video question answering: Datasets, algorithms and challenges. *arXiv preprint arXiv:2203.01225*, 2022.

[56] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023.

Appendix

A OpenEQA Benchmark Details

This section provides further details on the construction of the OpenEQA benchmark (section 2.3). Specifically, we describe the process for generating human-like episode histories H for EM-EQA (appendix A.1), the interface for collecting question-answer pairs (Q, A^*) (appendix A.2), and the interface used to validate the dataset (appendix A.3).

A.1 Generating Episode Histories H

Episode histories H provide agents with observations of the environment, and are used for the EM-EQA split of OpenEQA in both ScanNet and HM3D environments (see section 2). The ScanNet dataset was originally collected by people who were asked to scan indoor environments with an RGB-D camera. We use the initial 30 seconds (or 600 frames) of these human trajectories from ScanNet as EM-EQA episode histories H.

HM3D consists of scanned 3D environments, but does not come with pre-collected environment tours. Thus, we generate episode histories H using a two-step, semi-automated process. First, we generate a shortest-path trajectory from a starting location x_{src} to a destination x_{dst} in the environment. We select locations such that the geodesic distance between x_{src} and x_{dst} is $> 10m$ and the path curves (enforced by the criteria that the geodesic path distance $\geq 1.1 \times$ Euclidean path distance). Under these constraints, the paths typically traverse multiple rooms in the environment. To collect an episode history H, an agent travels along the path, while scanning the scene every $1m$ by rotating up to 180°. These scans are intended to mimic human-like exploration behavior. After collecting the trajectories, we manually inspect each trajectory to ensure they properly explore the scene; we exclude trajectories with extended periods closely facing walls. This process results in one episode history H for 63 different HM3D validation environments.

A.2 Collecting Question-Answer Pairs (Q, A^*)

We use a Google Form to collect question-answer pairs (Q, A^*) annotations from 8 different AI researchers. Specifically, the annotators watch a video of a given episode history H, and generate questions for the 7 categories listed in section 2.3. In the form, each category is described and one to two *good* and *bad* example questions are provided.

A.3 Interface for Dataset Validation

After the initial collection of question-answer pairs Q, A^*, we ask two independent people to validate each question. Specifically, the validators are shown an episode history H and a corresponding question Q on a simple HTML page. They are asked to provide an answer or mark the question as invalid (i.e. ambiguous or unanswerable). For the subset of *object localization* questions, we ask the validators to provide two answers for each questions because referring expressions often have multiple valid options (e.g. an object may be both *'left of the sink'* and *'right of the stove'*). We remove any question marked invalid by either validator.

B LLM-based Evaluation Details

OpenEQA questions often require open-ended answers, we use an LLM to evaluate correctness of answer produced by EQA agents. We prompt an LLM to compare human annotated answer A_i^* and model output A_i given a question Q_i and output a score σ_i on a scale of 1 to 5. On this scale, 1 indicates an incorrect response, 5 is a correct response and intermediate values represent different levels of similarity. Since questions can often have multiple correct answers, we also provide extra answers to the LLM prompt during scoring. We show the LLM prompt in Figure 6. Given the scores σ_i, we calculate an aggregate LLM-based **correctness** metric (LLM-Match) using Equation (1).

Figure 6 Prompt used for LLM-Match scoring. The placeholders `{question}`, `{answer}`, `{extra_answers}`, and `{prediction}` are replaced by the question Q, ground truth answer A^*, additional answer, and the agent's predicted answer A, respectively. Note that the extra answers are only available for *object localization* questions. When not available, corresponding sections of the prompt are omitted.

```
You are an AI assistant who will help me to evaluate the response given the question, the correct answer, and
extra answers that are also correct. To mark a response, you should output a single integer between 1 and
5 (including 1, 5). 5 means that the response perfectly matches the answer or any of the extra answers. 1
means that the response is completely different from the answer and all of the extra answers.

Example 1:
Question: Is it overcast?
Answer: no
Extra Answers: ['doesn't look like it', 'no',' it's sunny']
Response: yes
Your mark: 1

Example 2:
Question: Who is standing at the table?
Answer: woman
Extra Answers: ['a woman', 'a lady', 'woman']
Response: Jessica
Your mark: 3

Example 3:
Question: Are there drapes to the right of the bed?
Answer: yes
Extra Answers: ['yes, there are drapes', 'yeah', 'the drapes are to the right of the king bed']
Response: yes
Your mark: 5

Your Turn:
Question: {question}
Answer: {answer}
Extra Answers: {extra_answers}
Response: {prediction}
```

C EQA Agent Function Signatures

In this section, we describe the function signature that is expected from an agent by OpenEQA benchmark.

Box 1 describes the function signature for the EM-EQA task. An agent is expected to produce a text answer to a question based on an episode history. The episode history generally consists of RGB, depth, camera pose, and camera intrinsic information. The benchmark does not prescribe any specific way to use the history. A variety of different approaches and representations of the history can be pursued by researchers, such as point clouds, NeRFs, or instance maps. Since all methods have the same set of episode history information at their disposal, it allows for a fair comparison of methods. The final natural language answer is evaluated using LLM-Match metric described in section 2.4 and appendix B.

Similarly, Box 1 also describes the expected function signature for A-EQA task. Here, an agent does not receive an episode history and must generate its own experience through exploration. To allow standardization, we provide access to the simulation environment and start state as part of the benchmark. The state allows for instantiating an environment and fixing the starting location of the agent and various objects. We do not prescribe a specific navigation API for the benchmark, researchers are free to pursue different abstractions such as atomic navigation actions or navigation skills, as long as it doesn't use any privileged simulation information. The final answer is evaluated for correctness using LLM-Match, and the efficiency (see section 2.5) is computed using the number of atomic actions taken by the agent (to allow for standardization).

Algorithm 1 EQA Agent Signatures

```python
def EMEQA_Agent(Q: str, H: dict) -> str:
    """ Function signature for EM-EQA Agents

    Args:
    - Q: EQA question
    - H: episodic memory (history)
        - keys -> rgb: image,
                  depth: image,
                  c_pose: camera pose,
                  c_in: camera intrinsics
        - H["rgb"] = np.array(T, H, W, 3)
        - H["depth"] = np.array(T, H, W, 1)
        - H["c_pose"] = np.array(T, 6)
        - H["c_in"] = np.array(T, 6)

    Returns:
    - answer: natural language
    """
    ...

    return answer

def AEQA_Agent(Q: str,
        S: dict) -> Tuple[str, int]:
    """ Function signature for A-EQA Agents

    Args:
    - Q: EQA question
    - S: initial state of simulator
        - keys -> metadata
        - S["metadata"] = Dict[str, Any]

    Returns:
    - answer: natural language
    - T: episode lifetime. Timesteps
         taken to answer the question
    """

    env = make_env(S["metadata"])
    env.set_state(S)
    ...
    return answer, T
```

D Baseline Agent Details

This section provides additional details and LLM prompts for the blind LLM baseline (appendix D.1), Socratic LLM w/ Frame Captions example (appendix D.2), and GPT-4V (appendix D.3).

D.1 Blind LLM Prompt and Details

The prompt used for both our LLaMA-2 and GPT-4 blind LLM baselines is illustrated in figure 7. We use the 70B parameter version of LLaMA-2 that is fine-tuned for chat, and the `gpt-4-0613` version of GPT-4.

D.2 Socratic LLM w/ Frame Captions Example

Figure 8 shows how Socratic LLM w/ Frame Captions baseline produces an answer to a question given K frames sampled from episodic memory H. We use LLaVa-1.5 to generate image captions. We use the 70B parameter version of LLaMA-2 that is fine-tuned for chat, and the `gpt-4-0613` version of GPT-4 for large language model.

Figure 7 Prompt used for Blind LLM baselines. The placeholder `{question}` is replaced by the question Q. The same prompt is used for LLaMA-2 and GPT-4.

```
You are an intelligent question answering agent. I will ask you questions about an indoor space and you must
provide an answer.

If the question does not provide enough information to properly answer, provide an appropriate guess.

Q: What machine is on top of the stove?
A: The microwave
Explanation: stoves are typically found in kitchens and near microwaves.

Q: What piece of furniture is in the middle of the bedroom?
A: a bed
Explanation: bedrooms almost always contain a bed.

Q: Is the door open or closed?
A: open
Explanation: the door can be in either state, so we just randomly pick one.

Q: {question}
```

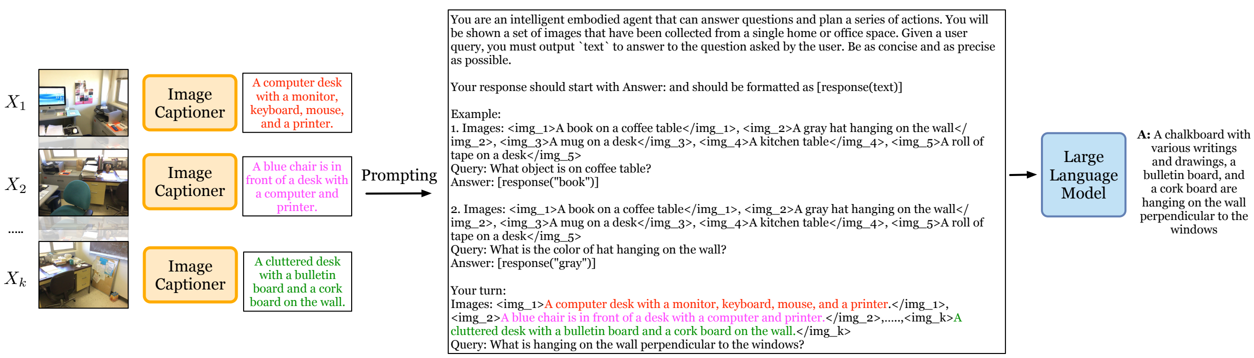


Figure 8 Input example for Socratic LLMs w/ Frame Captions baseline. We first caption each of the K frames with an image captioner and then prompt the LLM with those captions along with the question. The large langauge model produces an answer.

D.3 GPT-4V Details

Given an episodic memory H, we draw K frames and pass it to GPT-4V (through the API) in addition to question Q and prompt ω. We use chain-of-thought prompting in ω. We choose $K = 50$ for EM-EQA and $K = 75$ for A-EQA. Figure 9 shows the prompt ω and the input format passed to GPT-4V.

E Sparse Voxel Maps

For building SVM, we use K uniformly-sampled frames from the episode history H. K varies across difference scenes but the principle is to find the minimum number of K (for reducing the run-time memory consumption) to cover the whole environment. We process each sampled frame with the following two steps:

Step 1. Detecting object views in the frame using Detic. Each object view v is a tuple of $\langle c, b \rangle$, where c is the 2D image crop of the object and b is the 3D bounding box in the world coordinate system. We first extract object masks from the frame by setting the vocabulary for Detic to more than 500 household object categories. Then we get the image crop c around each detected mask with an additional margin. We then use depth information to get a 3D point cloud where we run DBSCAN (11) to further filter out background points, and compute the bounding box b. Note that we only consider depths that are in the range of $[0.1m, 4m]$.

Figure 9 GPT4V input prompt.

```
You are an intelligent embodied agent that can answer questions. You will be shown a set of images that have
been collected from a single location. Given a user query, you must output 'text' to answer to the question
asked by the user.

User Query: {question}
Think step by step.
```

Step 2. Associating each object view v with a global object instance o. Most objects will be detected in more than one frame, and a main goal of SVM is to de-duplicate object views to create global object instances. Each global object instance o is a tuple of $\langle C, b^* \rangle$, where C is a list a image crops (i.e., c) from multiple viewpoints (i.e., v), and b^* is a re-computed 3D bounding box from a concatenated point cloud of different views. For matching v to o, SVM considers 3D bounding box overlapping and CLIP (36) embedding similarity.

After SVM is constructed, we then select the best crop from C per global instance o, where the object mask takes up the largest number of pixels. Each selected crop is passed to LLAVA-1.5 (27) to get the textual description, and all the descriptions with the instances' 3D coordinates (center of the bounding box b^*) are wrapped in a prompt for an LLM to answer the question Q. Limited by the LLM's capacity, we only consider topN ($N = 75$) instances ranked by the CLIP similarity between their visual feature and Q from all the instances we detect in SVM.

F Force-A-Guess Details

Figure 10 Prompt used for Force-A-Guess. The placeholders {question} and {old_answer} are replaced by the question Q and initial answer A, respectively. The same prompt is used for LLaMA-2 and GPT-4.

```
You are an intelligent question answering agent. I need you to fix the answers to these question.

If the proposed answer says the question is unanswerable you should output the action "guess". Otherwise,
output the action "keep".

Question: What machine is on top of the stove?
Proposed Answer: the microwave
Action: keep

Question: What piece of furniture is in the middle of the bedroom?
Proposed Answer: The question is unanswerable from the provided images.
Action: guess

Question: {question}
Proposed Answer: {old_answer}
```

As discussed in section 3, we force baseline agents to guess an answer if they initial abstain – i.e. respond with an explanation for why the question is unanswerable. Specifically, we first ask an LLM if the initial answer is an abstaining response, and if so we replace the answer with a guess from a blind LLM. For step 1, use the prompt shown in figure 10. We provide a comparison baseline performance with and without this procedure in appendix G.

G Force-A-Guess Results

In table 3, we present results illustrating the performance drop for baseline methods when they are allowed to abstain, rather than being forced to guess an answer. As expected, performance drops for most methods. We find that GPT-4-based methods (rows 3, 5, and 7) show the largest drop in performance, which corresponds

Table 3 LLM-Match scores without forcing agents to guess.

# method	EM-EQA	EM-EQA (w/o guess)	A-EQA	A-EQA (w/o guess)
Blind LLMs				
1 GPT-4	33.5	-	35.5	-
2 LLaMA-2	27.7	-	28.8	-
Socratic LLMs w/ Frame Captions				
3 GPT-4 w/ LLaVA-1.5	43.6	29.3 (-14.3)	38.1	23.7 (-14.3)
4 LLaMA-2 w/ LLaVA-1.5	36.7	36.2 (-0.6)	30.9	31.2 (+0.4)
Socratic LLMs w/ Scene-Graph Captions				
5 GPT-4 w/ ConceptGraphs	36.5	18.5 (-18.0)	34.4	12.4 (-21.9)
6 LLaMA-2 w/ ConceptGraphs	28.7	26.6 (-2.0)	23.8	18.9 (-4.8)
7 GPT-4 w/ Sparse Voxel Maps	38.9	27.3 (-11.5)	34.2	21.2 (-13.0)
8 LLaMA-2 w/ Sparse Voxel Maps	34.3	34.6 (+0.3)	29.9	29.3 (-0.6)
Multi-Frame VLMs				
9 GPT-4V	49.5	46.7 (-2.8)	41.8	40.6 (-1.2)
Human	86.8	-	85.1	-

with GPT-4's propensity to abstain. Specifically, for EM-EQA, GPT-4 abstains 36% to 55% of the time (as measured by GPT-4). LLaMA-2-based methods abstain 3% to 12% of the time (as measured by LLaMA-2). Thus, we observe minimal changes in LLaMA-2-based method scores. Finally, GPT-4V abstains 12% of the time (as measured by GPT-4), corresponding with a small drop in LLM-Match scores. Similar trends are observe in the A-EQA setting for all methods.

H Full Results

Table 4 Category-level Performance on EM-EQA Rows represent the different agents as described in Section 3 and columns represent the different category of questions in the dataset, as described in Section 2.3. Bold indicates max in section.

# method	EQA Category: object recognition	object localization	attribute recognition	spatial understanding	object state recognition	functional reasoning	world knowledge	LLM-Match (C)
Blind LLMs								
1 GPT-4	**15.4**	**20.3**	**31.5**	**31.4**	51.0	**52.2**	**34.2**	**33.5**$\pm$1.0
2 LLaMA-2	10.7	15.3	22.3	25.0	**51.7**	44.1	29.7	28.3$\pm$1.0
Socratic LLMs w/ Frame Captions								
3 GPT-4 w/ LLaVA-1.5	**36.5**	**31.9**	**45.8**	**36.1**	**56.0**	**54.8**	**44.8**	**43.6**$\pm$1.1
4 LLaMA-2 w/ LLaVA-1.5	30.5	18.8	39.4	31.4	50.1	47.4	41.7	36.8$\pm$1.1
Socratic LLMs w/ Scene-Graph Captions								
5 GPT-4 w/ ConceptGraphs	26.4	17.0	40.6	29.1	**55.5**	**48.4**	39.9	36.5$\pm$1.0
6 LLaMA-2 w/ ConceptGraphs	17.1	13.9	24.4	27.2	43.5	38.1	39.0	28.7$\pm$1.0
7 GPT-4 w/ Sparse Voxel Maps	**30.0**	**20.0**	**49.6**	**31.7**	**55.5**	45.4	**40.8**	**38.9**$\pm$1.0
8 LLaMA-2 w/ Sparse Voxel Maps	23.4	11.7	38.9	30.8	52.8	45.4	39.1	34.3$\pm$1.1
Multi-Frame VLMs								
9 GPT-4V	51.4	53.3	65.2	42.6	57.7	63.8	52.3	55.3$\pm$1.1
10 Gemini 1.0 Pro Vision	41.5	33.3	41.9	37.6	56.9	52.2	52.1	44.9$\pm$1.1
11 Claude 3	37.0	13.1	39.2	37.0	45.5	37.9	47.3	36.3$\pm$1.1
Human	87.9	77.3	87.9	86.7	98.7	81.8	87.2	86.8$\pm$0.6

Table 4 presents results on EM-EQA broken down by the seven question categories described in Section 2.3. Specifically, we present results for the baseline agents described in section 3 as well as two additional multi-frame VLMs: Gemini 1.0 Pro Vision (41) and Claude 3 (3). We find that EQA agents with visual information

Table 5 Category-level Performance on A-EQA. Rows represent the different agents as described in Section 3 and columns represent the different category of questions in the dataset, as described in Section 2.3. Bold indicates max in section. *For A-EQA, GPT-4V scores are calculated on a subset of 184 questions.

# method	object recognition	object localization	attribute recognition	spatial understanding	object state recognition	functional reasoning	world knowledge	LLM-Score (C)
Blind LLMs								
1 GPT-4	**25.3**	**28.4**	**27.3**	**37.7**	47.2	**54.2**	**29.5**	**35.5**$\pm$**1.7**
2 LLaMA-2	13.7	22.1	16.2	29.7	**43.3**	50.4	28.8	29.0$\pm$1.6
Socratic LLMs w/ Frame Captions								
3 GPT-4 w/ LLaVA-1.5	**25.0**	**24.0**	**34.1**	**34.4**	**56.9**	**53.5**	**40.6**	**38.1**$\pm$**1.7**
4 LLaMA-2 w/ LLaVA-1.5	19.7	11.7	31.2	28.3	48.1	46.1	35.8	30.9$\pm$1.7
Socratic LLMs w/ Scene-Graph Captions								
5 GPT-4 w/ ConceptGraphs	25.3	16.5	29.2	37.0	52.2	**46.8**	37.8	34.4$\pm$1.8
6 LLaMA-2 w/ ConceptGraphs	13.3	11.9	18.8	27.9	31.7	31.7	36.5	23.9$\pm$1.6
7 GPT-4 w/ Sparse Voxel Maps	**29.0**	**17.2**	**31.5**	**31.5**	**54.2**	39.8	**38.9**	**34.2**$\pm$**1.8**
8 LLaMA-2 w/ Sparse Voxel Maps	16.7	9.7	33.4	29.0	47.2	40.5	37.5	29.9$\pm$1.7
Multi-Frame VLMs								
9 GPT-4V*	34.0	34.3	51.5	39.5	51.9	45.6	36.6	41.8$\pm$3.2
Human	89.7	72.8	85.4	84.8	97.8	78.9	88.5	85.1$\pm$1.1

(Columns "object recognition" through "world knowledge" are under the header "EQA Category".)

perform well in recognizing objects and their attributes, and make better use of this information to answer questions that require world knowledge. However, on other categories, their performance is closer to the blind LLM baseline (GPT-4), indicating substantial room for improvement on OpenEQA. Table 5 presents results per category on A-EQA for the baselines described in section 3. The results are consistent with the results on EM-EQA, however we observe an overall performance drop due to the challenges of dealing with the longer frontier exploration trajectories that were use for our baselines on A-EQA (see section 3).

I LLM-Match Human Alignment and Details

Table 6 Varying LLM used for scoring. On a subset of 100 questions with answers from GPT-4, GPT-4 scoring shows excellent agreement with human judgement, while using other LLMs shows lower correlation (Spearman correlation coefficient).

Scorer LLM	ChatGPT-4	ChatGPT3.5	LLaMA 2	Human
ChatGPT-4	1.00	0.66	0.68	0.88
ChatGPT3.5	-	1.00	0.66	0.61
LLaMA 2	-	-	1.00	0.63
Human	-	-	-	1.00

Table 7 Per-annotator Spearman-ρ. Human scoring has excellent agreement with both other humans and with LLM scoring.

Annotator	vs. Other Humans	vs. LLM
0	0.91	0.91
1	0.91	0.91
2	0.92	0.90
3	0.93	0.94

Evaluating open-vocabulary responses to questions is an open challenge in AI, and in particular for question-answering. While human evaluation remains the gold-standard, it is also expensive and time consuming. An automatic evaluation metric is preferable for benchmarking, fast iteration, and model selection. We thus use an automatic LLM-Based evaluation metric in this work as described in Section 2.4. We performed analysis experiments to test the quality of this metric along two axis: (1) How closely aligned is the LLM-Match metric with human evaluators? (2) How sensitive is the LLM-Match metric towards specific choice of prompts and the LLM?

Human Alignment. To answer the first question, we designed an experiment to measure the agreement between LLM-Match metric and human evaluators. For this analysis, we uniformly sampled a subset of 300 questions from OpenEQA. To ensure coverage of the answer distributions (i.e. poor, fair, and good answers), we sampled

100 responses from a blind LLM (LLaMA-2), multi-frame VLM (GPT-4V), and human baseline answers. In a double blind study, we then asked 4 human evaluators to score the 300 responses using an evaluation prompt similar to the one used by LLM-Match. The evaluators were provided no information about the source of the response (except an MD5 hash of the question ID, response source, and annotator ID). We found a **Spearman's** $\rho = 0.909$ **between human and LLM evaluation** (bootstrap CI=(0.883,0.928), N=9999), indicating excellent agreement with human judgement. Table 7 shows the Spearman's ρ (a measure of correlation) between (1) each annotator and other humans and (2) each annotator and GPT-4 scoring. Human evaluators correlated with other humans in $\rho \in [0.91, 0.93]$, and with LLMs in $\rho \in [0.90, 0.94]$.

Choice of LLM. Table 6 shows rho between human evaluators and different LLMs, on the subset of 100 questions from GPT4V. GPT-4 scoring shows good agreement with human scoring ($\rho = 0.88$), while GPT3.5 (ρ=0.66) and LLaMA 2 (ρ=0.68) show lower correlation. We believe that future LLMs will show higher agreement with human annotators, and in the meantime we recommend only using GPT-4 for scoring.

J LLM-Match Robustness Details

Table 8 **LLM Role.** Correlation between scores when changing the 'role' of the LLM in the scoring prompt (Spearman correlation coefficient).

Role	AI	"Score Master"	Professional
AI	1.00	0.97	0.96
"Score Master"	-	1.00	0.97
Professional	-	-	1.00

Table 9 **Match criterion for a '5'.** Correlation between scores when changing the criterion in the scoring prompt (Spearman correlation coefficient).

Match Crit.	Perfect	Contains	Pro	Person
Perfect	1.00	0.96	0.95	0.96
Contains	-	1.00	0.97	0.97
Pro	-	-	1.00	0.98
Person	-	-	-	1.00

Table 10 **Temperature of scoring LLM.** Changing the temperature of GPT-4 used in scoring (Spearman correlation coefficient).

Temp	0.01	0.1	0.2	0.3
0.01	1.00	0.98	0.98	0.98
0.1	-	1.00	0.97	0.98
0.2	-	-	1.00	0.97
0.3	-	-	-	1.00

Our LLM-Match uses the specific evaluation prompt described in figure 6. The metric is stable under small permutations of the prompt and LLM-Match settings as illustrated in Table 8, Table 9 and Table 10, which show the correlation in LLM-Match scores using different prompting strategies, assessed on 500 GPT-4V answers.

Role: Table 8 demonstrates that changing the LLM's role from 'AI' to 'Score Master' or 'professional evaluator' does not significantly change the results, and scores between any two treatments have a tight correlation with a Spearman's ρ all above 0.95.

Match criterion: Similarly, Table 9 shows analogous results ($\rho > 0.95$) when changing the description of a '5' from 'perfect match' to 'contains correct answer', 'similar to a reasonable person', or 'reasonable professional'.

Temperature: The stochasticity in the evaluation function has negligible impact as well, as shown by varying the temperature and seed. Table 10 shows results when varying the temperature used in the GPT-4 scorer from 0.01-0.3, with results all >0.97.

K 3D Coordinate Ablation

Figure 11 compares the EM-EQA performance of the Socratic baseline that uses Sparse Voxel Map captions with and without including 3D bounding box information in the text descriptions. Results show that explicit

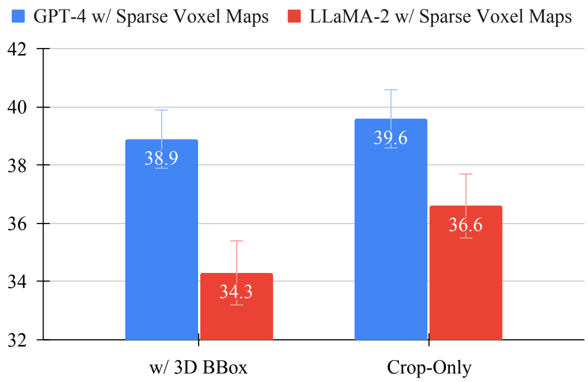


Figure 11 Ablating 3D location for scene-graph agents. Removing bounding box locations and extent had no significant effect for agents using either LLM.

bounding box location and size information from the scene graph does not significantly change the performance of scene-graph based agents. This suggests that neither LLM, trained with only text information, is able to effectively use the 3D location information.

L OpenEQA Dataset Examples

Additional examples from the ScanNet and HM3D splits of OpenEQA are provided in the Figures 12, 13, and 14.

Episode History H

Question-Answer Pairs (Q, A^*)

[Object Recognition]
Q: What is the red object on the chari?
A^*: a backpack

[Attribute Recognition]
Q: Among all the chairs, what is the unique color of the chair?
A^*: green

[Spatial Understanding]
Q: Can 10 people sit in this room?
A^*: yes

[Object State Recognition]
Q: Is the plastic water bottle open?
A^*: no

[Functional Reasoning]
Q: What can I use to write something on using my pencil?
A^*: the piece of paper

[World Knowledge]
Q: Were students here lately?
A^*: yes

[Object Localization]
Q: Where is my unfinished Starbucks drink?
A^*: on the table near the front whiteboard
Extra Answers: ['On the second table from the front', 'In the center of the second table.', 'On the second table from the front', 'On the table near the windows']

Figure 12 OpenEQA dataset examples from a ScanNet scene. Note that only a subset of frames from the episode history H are displayed. Thus, some questions may require additional visual information to answers.

Episode History H

Question-Answer Pairs (Q, A^*)

[Object Recognition]
Q: What object is on top of the work desk?
A^*: shelf

[Attribute Recognition]
Q: what color is the task chair?
A^*: black

[Spatial Understanding]
Q: which object is closer to the window, the bed or the trash can?
A^*: bed

[Object State Recognition]
Q: Is the closet door fully closed?
A^*: no

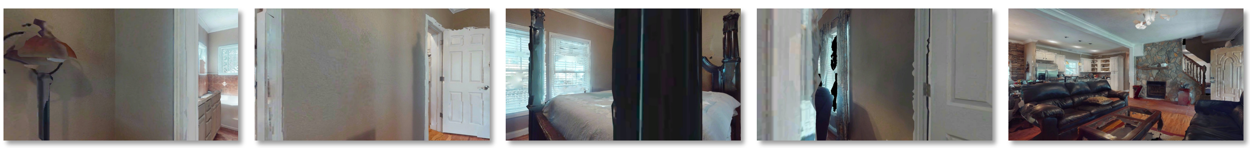
[Functional Reasoning]
Q: The closet is full. Where can I store a suitcase in this room?
A^*: Under the bed

[World Knowledge]
Q: What are the two main functions or purposes of this room? Why did you arrive at those conclusions?
A^*: Sleeping, since there is a bed. Working, since there is a task chair and desk.

[Object Localization]
Q: Where is the two-tier shelf?
A^*: on top of the desk
Extra Answers: ['On top of the desk.', 'On the left when you first walk in to the room.', 'above the desk infront of the chair', 'on top of the desk']

Figure 13 OpenEQA dataset examples from a ScanNet scene. Note that only a subset of frames from the episode history H are displayed. Thus, some questions may require additional visual information to answers.

Episode History H



Question-Answer Pairs (Q, A^*)

[Object Recognition]
Q: what is on the chair?
A^*: a soft pillow

[Attribute Recognition]
Q: is the outside door open or closed?
A^*: open

[Spatial Understanding]
Q: is the table in the living room clean?
A^*: yes

[World Knowledge]
Q: what is special about the wall in the living room?
A^*: it seems to be made of stone

[Object Localization]
Q: where is the standing lamp?
A^*: next to the bed in the bedroom
Extra Answers: ['in the bedroom',
'to the left of the bed',
'the bedroom',
'The room with the bed and the bathroom']

Figure 14 OpenEQA dataset examples from an HM3D scene. Note that only a subset of frames from the episode history H are displayed. Thus, some questions may require additional visual information to answers.