This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

When Robots Should Say "I Don't Know": Benchmarking Abstention in Embodied Question Answering

Tao Wu[1] Chuhao Zhou[1] Guangyu Zhao[2] Haozhi Cao[1] Yewen Pu[1] Jianfei Yang[1†]
[1]Nanyang Technological University [2] Peking University
{twu019@e., chuhao002@e., haozhi002@, yewen.pu@, jianfei.yang@}ntu.edu.sg
zhaogy24@stu.pku.edu.cn

https://abstaineqa.github.io/

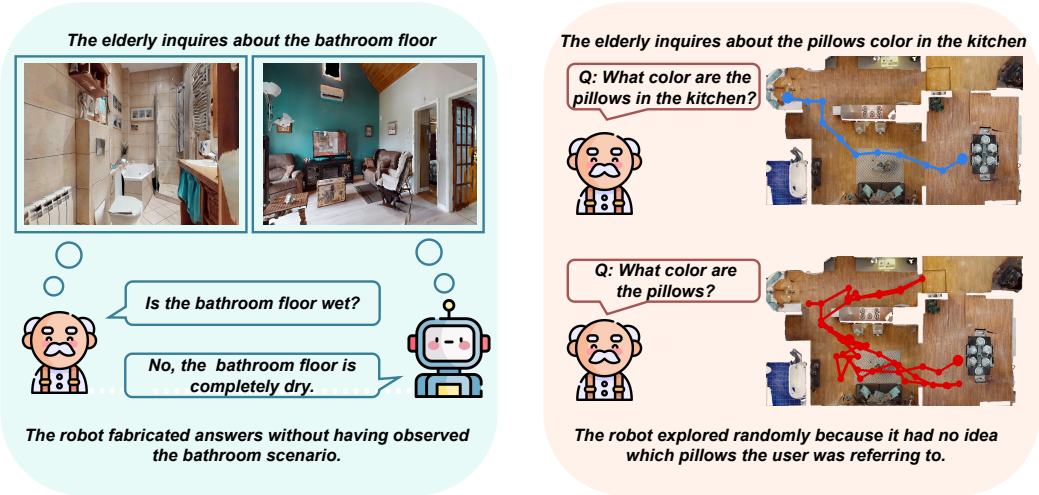


Figure 1. Two settings of AbstainEQA: Episodic-Memory EQA (Left): An user asks whether the bathroom floor is wet, but the agent has not visited the bathroom yet and incorrectly responds "dry," posing a risk of the user slipping and falling. Active EQA (Right): The user did not specify which pillow to check, causing the agent to wander aimlessly around the house.

Abstract

Embodied Question Answering (EQA) requires an agent to interpret language, perceive its environment, and navigate within 3D scenes to produce responses. Existing EQA benchmarks assume that every question must be answered, but embodied agents should know when they do not have sufficient information to answer. In this work, we focus on a minimal requirement for EQA agents, abstention: knowing when to withhold an answer. From an initial study of 500 human queries, we find that 32.4% contain missing or underspecified context. Drawing on this initial study and cognitive theories of human communication errors, we derive five representative categories requiring abstention: actionability limitation, referential underspecification, preference dependence, information unavailability, and false presupposition. We augment OpenEQA by having annotators transform well-posed questions into ambiguous variants outlined by these categories. The resulting dataset, AbstainEQA, comprises 1,636 annotated abstention cases paired with 1,636 original OpenEQA instances for balanced evaluation. Evaluating on AbstainEQA, we find that even the best frontier model only attains 42.79% abstention recall, while humans achieve 91.17%. We also find that scaling, prompting, and reasoning only yield marginal gains, and that fine-tuned models overfit to textual cues. Together, these results position abstention as a fundamental prerequisite for reliable interaction in embodied settings and as a necessary basis for effective clarification.

1. Introduction

Powered by advances in vision language models, embodied robots can now understand instructions, perceive their environment, and perform complex tasks, enabling them to

†Corresponding author.

Benchmark	Platform	Ambig.	Open-Vocab	Source
EQA-v1 [8]	House3D	✗	✗	Rule
IQUAD [14]	AI2-THOR	✗	✗	Rule
MP3D-EQA [41]	Matterport3D	✗	✗	Rule
MT-EQA [46]	House3D	✗	✗	Rule
K-EQA [36]	AI2-THOR	✗	✗	Rule
S-EQA [12]	VirtualHome	✗	✗	VLMs
HM-EQA [33]	HM3D	✗	✗	VLMs
NoisyEQA [43]	HM3D	✓	✓	VLMs
EXPRESS-Bench [20]	HM3D	✗	✓	VLMs
OpenEQA [25]	HM3D/ScanNet	✗	✓	Humans
AbstainEQA **(ours)**	HM3D/ScanNet	✓	✓	Humans

Table 1. AbstainEQA vs. existing benchmarks. "Ambig." denotes ambiguous queries; "Source" refers to question creation.

evolve from research prototypes into everyday companions that assist humans in household chores [22, 42] and mobility support [24, 47]. Beyond these task-execution abilities, embodied robots should also participate in collaborative human-robot interaction (HRI) [13, 38, 39], where natural language serves as a bridge for them to interpret human intentions and generate feedback. Among various HRI tasks, Embodied Question Answering (EQA) formulates this process by requiring an agent to navigate a 3D environment, gather visual evidence, and produce responses grounded in an understanding of both human instructions and environmental perception [25, 33, 44].

Despite significant advances in EQA [25, 33, 44], existing formulations often oversimplify human-robot interaction by assuming that agents must always respond to every query. However, real-world communication often involves incomplete information due to ambiguous human instructions or missing visual and contextual evidences [16, 27, 35, 43]. Forcing an agent to respond without salient evidence can result in the notorious "hallucinations" [4, 17], which are strictly unacceptable in EQA because they are not merely linguistic artifacts but will cause physical damage to both environments and humans. As shown in Fig. 1 left, when an elderly user asks, "Is the bathroom floor wet?" the agent may still answer "The bathroom is dry" despite having never observed the area, posing clear safety risks. In addition, ambiguous queries can also disrupt navigation. In Fig. 1 right, they can lead to ineffective, where the robot wanders inefficiently or explores irrelevant areas. Taken together, these cases highlight a fundamental requirement for embodied robots: they must be able to abstain from answering when evidence is insufficient and maintain effective, uncertainty-aware navigation under ambiguity.

In this work, we first attempt to answer the following question: *Do people ask questions in daily life that robots should abstain from answering?* To explore this, we conduct a large-scale questionnaire survey with 50 participants on how people naturally pose questions to embodied agents. By manually checking 500 collected results, we find that 32.4% of them lack contextual evidence, indicating that uncertainty is intrinsic to real-world human-robot interactions. To address this, we firstly introduce AbstainEQA, a large-scale benchmark designed to evaluate embodied agents' abstention capabilities when facing ambiguous human queries. The benchmark contains 1,636 human-annotated examples that capture diverse forms of ambiguity in human–robot communication and supports two evaluation settings: EM-EQA and A-EQA. Following Norman's cognitive model [27], which attributes human errors to failures in sensing and interpreting information, we categorize ambiguous queries into five types of missing evidence: actionability limitation, referential underspecification, preference dependence, information unavailability, and false presupposition. This taxonomy provides a systematic guideline for the establishment of our benchmark with ambiguous human queries. For each instance, we annotate a clear query with its corresponding evidence frames and a ambiguous query with explanations why the agent should appropriately abstain.

By incorporating both clear and ambiguous queries, AbstainEQA evaluates whether embodied agents can genuinely perceive the environment in response to human queries, rather than merely memorizing textual patterns in ground-truth answers. Built upon AbstainEQA, we benchmark state-of-the-art agents in handling abstention across three key dimensions: VLM capability, model scaling, and adaptation strategies (e.g., prompting or supervised fine-tuning). Experiments results show that even advanced VLMs and adaptation methods perform poorly on AbstainEQA, with the best model achieving only 42.79% overall abstention recall, substantially lower than human performance at 91.17%. We further show that abstention can impact embodied navigation, causing ineffective exploration under ambiguous queries. Finally, we provided error analyses and key insights to guide future research on addressing abstention in EQA.

In summary, our key contributions are four-fold:

1. We establish a systematic **abstention taxonomy** by analyzing ambiguous human queries collected from large-scale real-world surveys.
2. Based on this taxonomy, we introduce **AbstainEQA**, the first human-annotated benchmark designed to evaluate both reasoning quality and abstention ability in EQA.
3. We conduct a **comprehensive evaluation** across advanced agents, model scales, and adaptation techniques, systematically revealing the limited performance and robustness of existing approaches.
4. We provide **inspirational insights** through detailed error analyses, highlighting key challenges that must be addressed to fundamentally solve abstention in EQA.

2. Related Work

2.1. Embodied Question Answering

The Embodied Question Answering (EQA) task requires an embodied agent to navigate within a 3D environment and answer natural language questions [8, 9, 46]. Recent advances have been fueled by Vision Language Models (VLMs), which enhance embodied reasoning through stronger semantic grounding, memory integration, and adaptive exploration [11, 23, 25, 33, 44]. Notably, OpenEQA [25] formalizes two complementary regimes: (1) Episodic-Memory EQA (EM-EQA), where agents answer from stored egocentric histories, and (2) Active EQA (A-EQA), where agents continue gathering new observations before responding. While these formulations advance embodied reasoning, they still assume all queries warrant an explicit response, overlooking the essential ability to abstain when uncertainty arises.

2.2. Managing Uncertainty in LLMs and VLMs

Large Language Models (LLMs) often hallucinate under uncertainty or incomplete information [17, 19], a failure that can be critical in embodied scenarios where mistakes lead to unsafe actions [10, 32]. To mitigate this, recent research has pursued two directions: learning to abstain when evidence is insufficient and learning to clarify ambiguity through interaction.

Abstention aims to equip models with the ability to recognize uncertainty and refrain from answering when supporting evidence is incomplete, ambiguous, or unreliable [21]. This issue has been explored across text, image, and video domains. Min et al. [26] showed that over half of the questions in Natural Questions admit multiple plausible interpretations, while Stelmakh et al. [34] further analyzed such ambiguous factoid queries, and Kirichenko et al. [21] unified these phenomena within a decision-theoretic framework emphasizing abstention under false premises or subjective phrasing. Beyond language, perceptual limitations introduce additional uncertainty: Gurari et al. [15] reported that about 38% of VizWiz questions cannot be reliably answered due to blur, occlusion, or darkness, motivating selective prediction and uncertainty-aware reasoning [7, 40]. Further extending to temporal reasoning, Yoon et al. [45] proposed behavioral alignment metrics: excessive refusal, permissiveness, and discretion to balance informativeness and safety in video QA. Unlike the above studies, Embodied QA (EQA) presents a fundamentally different challenge. Agents operate from multi-frame egocentric observations either recalling past views in EM-EQA or actively exploring new ones in A-EQA and must decide when the gathered evidence is sufficient to answer.

Clarification in embodied and vision-language agents seeks to enable them to interactively resolve ambiguous or underspecified instructions through dialogue with users. Early efforts such as those by Banerjee et al. [3] and Wan et al. [37] framed clarification as task-oriented communication, allowing agents to query users when object references or goals were unclear. Subsequent studies explored learning-based strategies: Dogan et al. [10] proposed a model-agnostic semantic framework that interactively queries users about missing attributes such as object states or spatial relations, while Ramrakhya et al. [32] trained multimodal LLM agents to ask clarifying questions via reinforcement learning with LLM-based feedback. Both studies reveal a tendency toward over-questioning, underscoring the need for mechanisms that determine when clarification is truly necessary.

We view abstention as a prerequisite for clarification: an embodied agent must first recognize, from frame-level visual evidence, when it lacks sufficient perceptual grounding. While prior work primarily examines how agents formulate and deliver clarifying questions, this perspective highlights a more fundamental challenge: determining when a question should be asked in the first place. By foregrounding abstention as the decision point preceding clarification, our work lays the foundation for embodied systems that ultimately know not only how to ask, but when to ask.

3. Problem Formulation

Embodied Question Answering (EQA) requires an embodied agent to understand human instructions and respond by reasoning over its egocentric perception of the environment. We extend this paradigm to AbstainEQA, where agents are required to abstain when the available evidence is insufficient or unreliable.

Formally, given a natural language question q and a sequence of egocentric observations τ, a policy is required to generate responses $\mathbf{A} = (y_0, \mathbf{Y}) = \{y_i\}_{i=0}^{S}$ in an autoregressive approach:

$$p(\mathbf{A}) = \prod_{i=1}^{S} p(a_i \mid q, \tau, \mathbf{Y}_{<i}), \tag{1}$$

where S is the length of responses and

$$\mathbf{Y} = \begin{cases} \text{Question Answers}, & \text{if } y_0 = \text{answer} \\ \text{Abstain Reasons}, & \text{if } y_0 = \text{abstain} \end{cases} \tag{2}$$

In the emulated setting (EM-EQA), a fixed sequence of observations is given: $\tau = \{o_i\}_{i=1}^{T}$, where T is the length of observations. In the active setting (A-EQA), a navigation policy π interacts with the environment within a maximum step $T_{\max}$, executing actions $a_t \in \mathcal{A}_t$ (*navigation primitives* $\cup$ `[STOP]`) and generating $\tau = \{o_i\}_{i=1}^{T}$, $T \leq T_{\max}$. Episodes end when `[STOP]` is issued or $T_{\max}$ is achieved.



Figure 2. Example human-annotated ambiguous queries and dataset statistics from AbstainEQA, illustrating the types of uncertainty encountered in real human–agent interactions and providing a stronger foundation for studying and improving embodied interaction. More examples are provided in App. A.4.

4. Benchmarking Abstention in Embodied QA

We introduce **AbstainEQA**, a large-scale benchmark for evaluating both response and abstention behavior in embodied question answering (EQA). Unlike prior benchmarks focused on textual or static visual settings, AbstainEQA consists both clear and ambiguous human queries. Specifically, clear queries can be answered using available perceptual information, whereas ambiguous queries should be abstained due to the lack of sufficient evidence. For clear queries, instead of providing an answer, we require the agent to provide a detailed justification for abstention. Consequently, we evaluate whether the embodied agents truly understand and perceive the environment when responding to human queries, rather than simply recalling textual patterns from ground-truth answers.

4.1. Quantifying the Need for Abstention

To construct the AbstainEQA benchmark, we first investigate how often humans naturally ask questions that require abstention in embodied environments by conducting a real-world questionnaire survey. Specifically, we recruit 50 non-expert participants and present them with egocentric videos from HM3D [31] and ScanNet [6] (152 videos in total). Each participant is instructed to annotate two QA pairs for five randomly sampled videos, resulting in 500 human-annotated QA pairs. By manually reviewing 500 QA pairs, we observe that approximately 32.4% of human queries are ambiguous and require abstention, suggesting that queries lacking contextual evidence are inherent in human–robot interaction. (More details see App.A.2)

4.2. Taxonomy of Abstention Types

Beyond estimating the proportion of ambiguous queries, we also observe that their patterns are various, primarily depending on the type of missing evidence. To systematically model ambiguous queries, we draw on Norman's cognitive framework of human error [27] and classify the causes of abstention based on agents' cognitive limitations with respect to the objective world and human subjective instructions. As shown in Fig. 2, Actionability Limitation and Information Unavailability result from sensing limitations in the objective environment. Referential Underspecification and Preference Dependence arise from interpretive failures in human instructions, whereas False Presupposition stems from both types of cognitive failures. In the following, we define each type of ambiguous question and provide specific examples for clarity.(More examples see App.A.4)

Actionability Limitation (AL). Questions that require physical interaction with the environment, such as opening containers, operating devices, or manipulating objects. In EQA settings [25], agents perceive the world solely through visual observations and cannot perform or verify manipulative actions. For example, the question "What is inside the box in the center of the living room?" requires physical access that exceeds the agent's capabilities.

Referential Underspecification (RU). Questions with ambiguous or underspecified referential context, in which multiple entities or regions in the environment correspond to the same description in human instructions. For instance,

the question "What is on the white cabinet?" becomes ambiguous when multiple white cabinets are visible and cannot be answered without additional clarifying evidence.

Information Unavailability (IU). Questions where essential spatial or temporal information cannot be inferred from available observations. In spatial cases, the agent may perceive the relevant scene but lacks measurable cues. For example, "What's the area of the bedroom?" cannot be answered from appearance alone without metric grounding. In temporal cases, the relevant event may occur outside the observed timeframe. For instance, "Who put the vase on the table?", illustrates a situation where the agent can observe the outcome but not how the action occurs.

False Presupposition (FP). Questions that contain assumptions contradicting the observed evidence. For example, "What material is the teddy bear on the bed?" is invalid when no teddy bear is present on the bed. Since the premise in human instruction is false, any attempt to response would result in fabricated answers.

Preference Dependence (PD). Questions that rely on subjective judgment or aesthetic evaluation rather than factual perception. For example, "Is the painting on the wall beautiful?" depends on personal taste and lies outside the scope of objective embodied reasoning.

4.3. Dataset Curation

Building upon the identified taxonomy, we further establish a large-scale benchmark, which is comparable to OpenEQA [25], through exhaustive human annotations on videos from ScanNet [6] and HM3D [31].

Data Annotation. Human annotators are guided to view randomly sampled videos and generate QA pairs (Q, A^*) as collaborators with an embodied agent. Given a video, annotators generate two QA pairs for each abstention type described in Sec. 4.2, to ensure balanced coverage. Notably, for Active EQA (A-EQA), we use the set of 184 questions from 3D-MEM [44] and maintain the original target objects when annotating abstention questions. This preserves the start and end points of the navigation path for both clear and ambiguous queries, allowing us to focus on the effects of abstention questions on navigation. In total, **1,636 abstention cases** are annotated and combined with **1,636 answerable cases** from OpenEQA [25] to create our AbstainEQA. To introduce data diversity, the questions of AbstainEQA are then paraphrased into five semantically equivalent variants using a large language model (LLM) [18], producing a total of 16,360 QA pairs.

Frames Location. To offer explicit visual evidence or reasons for abstention in clear and ambiguous questions, we additionally provide frame-level causal annotations(App. A.3). Specifically, human annotators are instructed to mark specific frames that either contain sufficient evidence for the correct answer or clearly reveal the reason for abstention cases. For clear queries, the marked frames as a reference that links each answer to its supporting visual context within the trajectory. For ambiguous queries, the selected frames highlight the cause of abstention, such as missing information, ambiguous references, or the need for physical interaction. Additionally, a brief textual explanation of the evidence gap is provided.

The data sample in both the annotated data and the located frames are reviewed by two additional annotators to verify correctness and consistency. Consequently, AbstainEQA serves as a comprehensive benchmark for evaluating both the response and abstention of embodied agents, along with the associated fine-grained reasoning evidence.

5. Evaluation Criteria

We design a comprehensive evaluation protocol to assess an agent's performance in three key aspects: response accuracy, abstention recognition, and embodied navigation across EQA tasks. Followed by a human validation to ensure the reliability of our evaluation framework.

Response Accuracy. We employ GPT-4o [18] to automatically evaluate the correctness of responses. Specifically, we implement the LLM-Match protocol to quantify semantic alignment between agent predictions and ground-truth answers. The evaluator LLM, guided by a scoring prompt (App. C.1), produces a binary or graded judgment reflecting semantic equivalence.

Abstention Recognition. Similarly, we adopt GPT-4o to determine whether each output from the embodied agent reflects a valid abstention, given a structured evaluation prompt (App. C.2). Recall, Precision, Accuracy, and F1-score are calculated from the evaluation results of GPT-4o to quantitatively measure an agent's ability to identify and abstain from ambiguous queries.

Embodied Navigation Evaluation. Following 3D-Mem [44], we further evaluate the navigation capabilities of an embodied agent in the A-EQA task using Success Rate (SR), Total Frames (TF), Total Snapshots (TS), and Path Length (PL). Particularly, SR indicates whether the agent successfully reaches the target or viewpoint; TF measures the total number of frames processed during exploration; TS counts the number of snapshots captured when answering; and PL records the overall navigation distance. An effective agent expects to achieve higher SR while minimizing TF, TS, and PL.

Human Validation. To verify the reliability of LLM-based evaluation, we perform human validation on a random subset of 300 out of 16,360 samples, using the same criteria as the LLM evaluation prompts. The Pearson correlation coefficient [29] between LLM-based and human scores is 0.88, indicating strong consistency (see App. C.3). This result confirms the reliability of our LLM-as-a-Judge framework for large-scale evaluation of both response and

Model	Referential Underspecification (%)	Information Unavailable (%)	Actionability Limitation (%)	False Presupposition (%)	Preference Dependence (%)	Overall (%)
Claude-Sonnet-4.5[1]	0.00	2.62	1.19	0.88	0.29	1.04
GPT-5 [28]	3.66	51.90	25.07	13.24	13.37	22.19
Qwen3-VL-4B [30]	5.13	72.01	31.64	14.41	9.01	27.32
Qwen2.5-VL-7B [2]	10.62	83.38	20.30	10.00	14.83	28.61
GPT-4o [18]	5.86	73.47	55.22	10.88	10.88	32.09
Gemini-2.5-Pro [5]	4.40	86.30	60.90	39.71	15.12	42.79
Humans	88.55	99.10	98.74	94.19	75.23	91.17

Table 2. Abstention recall (%) across abstention types. All VLMs perform far below humans, underscoring abstention as a key weakness.

3D-Mem[44]	Success Rate (%)	Total Frames	Total Snapshots	Path Length
Response	77.17	35.86	10.76	7.805
Abstention	61.41	51.41	12.74	8.319

Table 3. Navigation performance with and without abstention on A-EQA. Ambiguous queries reduce navigational efficiency.

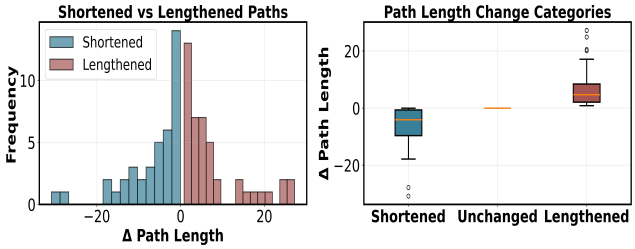


Figure 3. Path-length distribution of GPT-4o under response vs. abstention. Ambiguous queries lead to ineffective exploration.

abstention performance.

6. Experiments

We address the following research questions. (A) How do embodied agents behave when confronted with underspecified or unanswerable queries? (Sec. 6.1) (B) Does increasing model scale lead to better-calibrated abstention behavior? (Sec. 6.2) (C) How do prompting strategies and explicit reasoning influence the trade-off between abstention precision, abstention recall, and response accuracy? (Secs. 6.3–6.4) (D) Does supervised fine-tuning (SFT) yield genuine improvements in abstention awareness? (Sec. 6.5)

6.1. Model Behavior under Abstention Settings

As shown in Tab. 2, current Vision–Language Models (VLMs) show limited ability to abstain appropriately when faced with uncertainty. Even the strongest model, Gemini-2.5-Pro [5], achieves only 42.79% abstention recall, while most others fall below 33%, substantially lagging behind human performance. Models perform better on Information Unavailability, where absent visual evidence provides explicit cues to abstain. However, they struggle with Referential Underspecification and Preference Dependence, which demand pragmatic disambiguation or subjective judgment. Overall, achieving reliable abstention remains a significant open challenge for embodied VLMs. (App. D.2)

Tab. 3 further examines how abstention conditions affect embodied navigation. The success rate (SR) drops sharply from 77.17% to 61.41%, whereas both total frames (TF) and total snapshots (TS) increase substantially (35.86 $\rightarrow$ 51.41 and 10.76 $\rightarrow$ 12.74, respectively). These patterns indicate that agents require more observations to complete tasks, reflecting reduced navigational efficiency. Although the average path length (PL) increases only slightly (7.805 $\rightarrow$ 8.319), this modest change is associated with pronounced behavioral instability. Under uncertainty, A-EQA agents tend to engage in redundant exploration or display hesitant decision-making.(More details see App. D.1)

Fig. 3 provides a more fine-grained characterization of path-length variation. Among 100 successful episodes, 40% of trajectories become shorter, 44% become longer, and 16% remain unchanged. Statistical analyses confirm that these changes are systematic rather than random. A Mann–Whitney test shows a significant distributional shift ($p < 0.001$), Cliff's $\delta = -1.000$ indicates a strong effect size, and a Jensen–Shannon divergence of 0.5089 quantifies the substantial discrepancy between response and abstention path distributions. Visual examples in Fig. 4 illustrate two behavioral extremes premature termination and over-exploration, demonstrating that current agents lack calibrated exploration policies under uncertainty. This oscillation between under- and over-searching underscores a core limitation of embodied abstention: instability in determining when sufficient evidence has been gathered.

6.2. Impact of Model Scaling on Abstention

As shown in Tab. 2, increasing model size does not uniformly improve abstention recall across vision–language architectures. Although larger proprietary systems (e.g.,

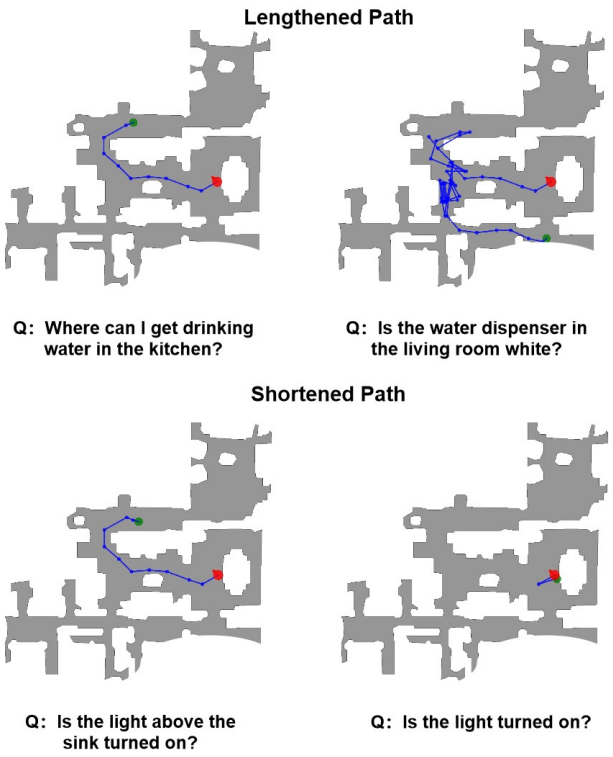


Figure 4. Changes of Path-lengths from GPT-4o under ambiguous queries. Two cases of ineffective exploration are shown: shortened and lengthened trajectories.

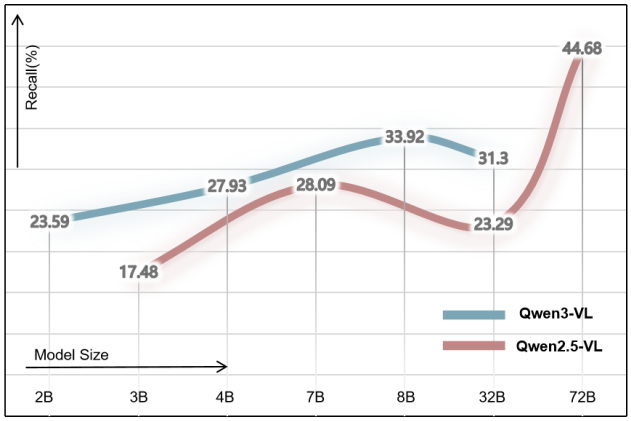


Figure 5. Abstention recall across Qwen3-VL and Qwen2.5-VL. Scaling improves abstention within a model family

GPT-4o [18], Gemini-2.5-Pro [5]) generally achieve higher abstention rates than smaller open models, these improvements are inconsistent across categories and remain well below human reliability. Within individual model families (e.g., Qwen2.5-VL [2], Qwen3-VL [30]), scaling improves abstention performance (Fig. 5). However, this effect does not generalize across architectures, suggesting that abstention behavior is influenced more by cross-modal alignment and training objectives than by parameter count alone. Overall, this pattern indicates that within a single model family, increased model capacity can better support multimodal reasoning and uncertainty calibration, resulting in more consistent abstention behavior.

Method	Recall (%)	Precision (%)	Accuracy (%)	F1-Score (%)	Correctness (%)
7B [2]	28.61	76.85	60.00	42.42	63.33
7B-coarse	75.86	56.84	59.13	64.99	45.92
7B-Fine	61.92	60.05	60.36	60.97	50.22
HM-EQA [33]	27.17	56.82	53.26	36.76	51.80
HM-EQA-coarse	81.52	55.70	58.42	66.25	40.40
HM-EQA-fine	99.46	49.86	49.73	66.42	23.60

Table 4. Effect of coarse and fine prompts on abstention in EM-EQA and A-EQA. Prompting raises recall but reduces precision.

However, this effect does not generalize across architectures, indicating that parameter scaling alone does not guarantee improved abstention performance. The discrepancies across model families highlight that effective abstention depends more on model design, cross-modal alignment, and training objectives than on parameter count alone. Overall, these findings demonstrate that reliable abstention requires dedicated mechanisms for uncertainty reasoning rather than increased scale alone.

6.3. Effect of Explicit Prompting on Abstention

Tab. 14 summarizes the impact of explicit prompting strategies on abstention performance in both EM-EQA and A-EQA. In EM-EQA, adding coarse- or fine-grained prompts substantially boosts abstention recall (28.6% $\rightarrow$ 75.9% and 61.9%), indicating that explicit cues enhance models' sensitivity to uncertainty. However, these gains come at the cost of reduced precision (76.9% $\rightarrow$ 56.8%) and nearly unchanged accuracy (60%), suggesting that models over-abstain even when sufficient evidence is available.

A similar pattern appears in A-EQA: recall increases sharply from 27.2% to 81.5% (coarse) and 99.5% (fine), whereas precision and accuracy fall to nearly half of their original values. Although the F1 score improves overall, this gain primarily reflects a trade-off between excessive caution and genuine uncertainty recognition, rather than a substantive improvement in grounded reasoning. Overall, prompt engineering modulates a model's abstention tendencies but does not fundamentally enhance its ability to assess task resolvability or reason over perceptual input.

Method	Recall (%)	Precision (%)	Accuracy (%)	F1-Score (%)	Correctness (%)
2B-Instruct	23.59	66.55	55.87	34.84	61.25
2B-Thinking	25.55	74.78	58.47	38.09	66.96
4B-Instruct	27.93	80.74	60.64	41.51	71.36
4B-Thinking	27.08	81.73	60.51	40.68	72.74
8B-Instruct	33.92	81.98	63.23	47.99	72.61
8B-Thinking	27.02	83.71	60.88	40.85	73.93
32B-Instruct	31.30	87.52	63.42	46.11	75.95
32B-Thinking	21.70	91.73	59.87	35.10	75.95
Gemini-2.5-Pro	44.13	91.74	70.08	59.60	76.50
Gemini-2.5-Pro-Thinking	43.09	91.44	67.03	58.58	75.62

Table 5. Thinking effect comparison (Qwen3-VL ([30]) / Gemini [5]). Reasoning reduces abstention recall and correctness.

Method	Recall (%)	Precision (%)	Accuracy (%)	F1-Score (%)	Correctness (%)
7B	26.94	77.65	59.59	40.00	40.66
7B-SFT	83.27	89.08	86.53	86.08	63.92
7B-SFT-random	83.27	87.93	85.92	85.53	65.45
7B-Text-SFT	86.12	86.83	86.53	86.48	55.65
TF-IDF+LR	85.07	86.48	85.88	85.77	–
Bert	86.08	84.12	84.90	85.09	–

Table 6. Comparison of SFT variants (Qwen2.5-VL-7B [2]) on abstention detection. Text-based fine-tuning performs comparably to multimodal SFT, suggesting overfitting to linguistic cues.

6.4. Effect of Reasoning on Abstention

To further assess whether explicit reasoning improves abstention behavior, we examine the effect of "thinking" prompts on EM-EQA performance (Tab. 5). Across all model scales, adding reasoning chains does not improve abstention reliability. Smaller variants (e.g., 2B) exhibit only marginal gains in recall and F1, whereas larger models (8B–32B) show consistent declines across both metrics, indicating that extended reasoning yields verbose explanations rather than calibrated abstention.

Overall, explicit reasoning does not substantially enhance abstention quality. Instead, extended reasoning often increases response latency and complicates interaction, potentially hindering user experience in embodied question answering, where timely decisions are essential.

6.5. Limitations of SFT in Recognizing Abstention

Tab. 6 compares several supervised fine-tuning (SFT) variants for abstention detection. Although these models report high recall, precision, and F1, the gains are illusory: performance barely changes when visual inputs are randomized (7B-SFT-random) or removed entirely (7B-Text-SFT).

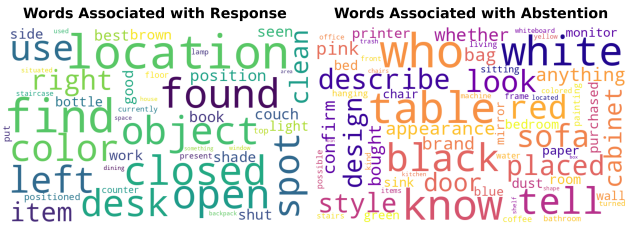


Figure 6. Word-frequency distributions for response vs. abstention questions, showing lexical separation that SFT models rely on

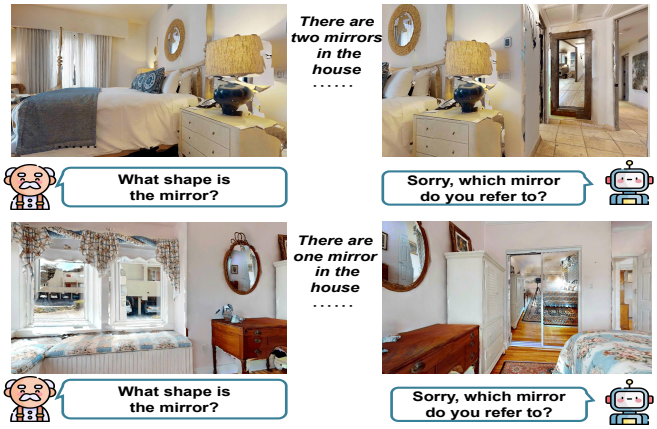


Figure 7. Examples showing that SFT models give identical answers across scenes for the same query, revealing a failure to ground abstention in visual evidence.

Text-only baselines TF-IDF+LR and a small BERT classifier match the SFT models, confirming that the signal comes almost entirely from linguistic patterns rather than multimodal reasoning.(Examples see App. D.3)

Fig. 6 further shows that specific lexical cues (e.g., who, design) disproportionately drive abstention decisions. Yet such cues are unreliable in embodied settings, as Fig. 7 shows the same question may or may not warrant a response depending entirely on what the agent observes. Overall, SFT models rely on surface text features rather than grounding abstention in visual evidence, revealing a fundamental limitation of current SFT-based approaches.

7. Conclusion

We present the first systematic study of abstention in embodied question answering and find that current agents fail to recognize ambiguous or underspecified queries. Scaling, prompting, reasoning, and fine-tuning yield only superficial gains, as models rely on textual cues rather than visual evidence. Ambiguity also leads to inefficient and unreliable exploration. These results expose a core limitation of embodied systems and highlight the need for uncertainty-aware agents for safe real-world human–robot interaction.

8. Acknowledgement

This work is jointly supported by MOE Singapore Tier 1 Grant RG83/25, RS36/24 and a Start-up Grant from Nanyang Technological University.

References

[1] Anthropic. Claude sonnet 4.5 system card. Technical report, Anthropic, 2024. 6, 9

[2] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 6, 7, 8

[3] Shurjo Banerjee, Jesse Thomason, and Jason Corso. The robotslang benchmark: Dialog-guided robot localization and navigation. In *Conference on Robot Learning*, pages 1384–1393. PMLR, 2021. 3

[4] Meng Cao, Yue Dong, and Jackie Chi Kit Cheung. Hallucinated but factual! inspecting the factuality of hallucinations in abstractive summarization. *arXiv preprint arXiv:2109.09784*, 2021. 2

[5] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025. 6, 7, 8, 9

[6] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5828–5839, 2017. 4, 5

[7] Corentin Dancette, Spencer Whitehead, Rishabh Maheshwary, Ramakrishna Vedantam, Stefan Scherer, Xinlei Chen, Matthieu Cord, and Marcus Rohrbach. Improving selective visual question answering by learning from your peers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24049–24059, 2023. 3

[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*, pages 1–10, 2018. 2, 3

[9] Abhishek Das, Georgia Gkioxari, Stefan Lee, Devi Parikh, and Dhruv Batra. Neural modular control for embodied question answering. In *Conference on Robot Learning*, pages 53–62. PMLR, 2018. 3

[10] Fethiye Irmak Dogan, Maithili Patel, Weiyu Liu, Iolanda Leite, and Sonia Chernova. A model-agnostic approach for semantically driven disambiguation in human-robot interaction. *arXiv preprint arXiv:2409.17004*, 2024. 3

[11] Vishnu Sashank Dorbala, Prasoon Goyal, Robinson Piramuthu, Michael Johnston, Reza Ghanadhan, and Dinesh Manocha. S-eqa: Tackling situational queries in embodied question answering. *arXiv preprint arXiv:2405.04732*, 2024. 3

[12] Vishnu Sashank Dorbala, Prasoon Goyal, Robinson Piramuthu, Michael Johnston, Reza Ghanadan, and Dinesh Manocha. Is the house ready for sleeptime? generating and evaluating situational queries for embodied question answering. In *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 18430–18437. IEEE, 2025. 2

[13] Weihua Du, Qiushi Lyu, Jiaming Shan, Zhenting Qi, Hongxin Zhang, Sunli Chen, Andi Peng, Tianmin Shu, Kwonjoon Lee, Behzad Dariush, et al. Constrained human-ai cooperation: An inclusive embodied social intelligence challenge. *Advances in neural information processing systems*, 37:44526–44553, 2024. 2

[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. 2

[15] Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3608–3617, 2018. 3

[16] Wen-Han Hsieh, Elvis Hsieh, Dantong Niu, Trevor Darrell, Roei Herzig, and David M Chan. Do what? teaching vision-language-action models to reject the impossible. *arXiv preprint arXiv:2508.16292*, 2025. 2

[17] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55, 2025. 2, 3

[18] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024. 5, 6, 7, 4

[19] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM computing surveys*, 55(12):1–38, 2023. 3

[20] Kaixuan Jiang, Yang Liu, Weixing Chen, Jingzhou Luo, Ziliang Chen, Ling Pan, Guanbin Li, and Liang Lin. Beyond the destination: A novel benchmark for exploration-aware embodied question answering. *arXiv preprint arXiv:2503.11117*, 2025. 2

[21] Polina Kirichenko, Mark Ibrahim, Kamalika Chaudhuri, and Samuel J Bell. Abstentionbench: Reasoning llms fail on unanswerable questions. *arXiv preprint arXiv:2506.09038*, 2025. 3

[22] Xiaoqi Li, Mingxu Zhang, Yiran Geng, Haoran Geng, Yuxing Long, Yan Shen, Renrui Zhang, Jiaming Liu, and Hao Dong. Manipllm: Embodied multimodal large language model for object-centric robotic manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18061–18070, 2024. 2

[23] Yang Liu, Weixing Chen, Yongjie Bai, Jingzhou Luo, Xinshuai Song, Kaixuan Jiang, Zhida Li, Ganlong Zhao, Junyi

Lin, Guanbin Li, et al. Aligning cyber space with physical world: A comprehensive survey on embodied ai. *arXiv preprint arXiv:2407.06886*, 2024. 3

[24] Yuxing Long, Wenzhe Cai, Hongcheng Wang, Guanqi Zhan, and Hao Dong. Instructnav: Zero-shot system for generic instruction navigation in unexplored environment. *arXiv preprint arXiv:2406.04882*, 2024. 2

[25] Arjun Majumdar, Anurag Ajay, Xiaohan Zhang, Pranav Putta, Sriram Yenamandra, Mikael Henaff, Sneha Silwal, Paul Mcvay, Oleksandr Maksymets, Sergio Arnaud, et al. Openeqa: Embodied question answering in the era of foundation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16488–16498, 2024. 2, 3, 4, 5

[26] Sewon Min, Julian Michael, Hannaneh Hajishirzi, and Luke Zettlemoyer. Ambigqa: Answering ambiguous open-domain questions. *arXiv preprint arXiv:2004.10645*, 2020. 3

[27] Donald A Norman. Design rules based on analyses of human error. *Communications of the ACM*, 26(4):254–258, 1983. 2, 4

[28] OpenAI. Gpt-5 system card. Technical report, OpenAI, 2025. System card (technical report). 6, 9

[29] Karl Pearson. Vii. note on regression and inheritance in the case of two parents. *proceedings of the royal society of London*, 58(347-352):240–242, 1895. 5

[30] Qwen Team. Qwen3-vl: Sharper vision, deeper thought, broader action. https://qwen.ai/blog?id=99f0335c4ad9ff6153e517418d48535ab6d8afef, 2025. Qwen.ai blog post. 6, 7, 8

[31] Santhosh K Ramakrishnan, Aaron Gokaslan, Erik Wijmans, Oleksandr Maksymets, Alex Clegg, John Turner, Eric Undersander, Wojciech Galuba, Andrew Westbury, Angel X Chang, et al. Habitat-matterport 3d dataset (hm3d): 1000 large-scale 3d environments for embodied ai. *arXiv preprint arXiv:2109.08238*, 2021. 4, 5

[32] Ram Ramrakhya, Matthew Chang, Xavier Puig, Ruta Desai, Zsolt Kira, and Roozbeh Mottaghi. Grounding multimodal llms to embodied agents that ask for help with reinforcement learning. *arXiv preprint arXiv:2504.00907*, 2025. 3

[33] Allen Z Ren, Jaden Clark, Anushri Dixit, Masha Itkina, Anirudha Majumdar, and Dorsa Sadigh. Explore until confident: Efficient exploration for embodied question answering. *arXiv preprint arXiv:2403.15941*, 2024. 2, 3, 7

[34] Ivan Stelmakh, Yi Luan, Bhuwan Dhingra, and Ming-Wei Chang. Asqa: Factoid questions meet long-form answers. *arXiv preprint arXiv:2204.06092*, 2022. 3

[35] Francesco Taioli, Stefano Rosa, Alberto Castellini, Lorenzo Natale, Alessio Del Bue, Alessandro Farinelli, Marco Cristani, and Yiming Wang. Mind the error! detection and localization of instruction errors in vision-and-language navigation. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 12993–13000. IEEE, 2024. 2

[36] Sinan Tan, Mengmeng Ge, Di Guo, Huaping Liu, and Fuchun Sun. Knowledge-based embodied question answering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(10):11948–11960, 2023. 2

[37] Yanming Wan, Jiayuan Mao, and Josh Tenenbaum. Handmethat: Human-robot communication in physical and social environments. *Advances in Neural Information Processing Systems*, 35:12014–12026, 2022. 3

[38] Tai Wang, Xiaohan Mao, Chenming Zhu, Runsen Xu, Ruiyuan Lyu, Peisen Li, Xiao Chen, Wenwei Zhang, Kai Chen, Tianfan Xue, et al. Embodiedscan: A holistic multimodal 3d perception suite towards embodied ai. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19757–19767, 2024. 2

[39] Xin Wang, Taein Kwon, Mahdi Rad, Bowen Pan, Ishani Chakraborty, Sean Andrist, Dan Bohus, Ashley Feniello, Bugra Tekin, Felipe Vieira Frujeri, et al. Holoassist: an egocentric human interaction dataset for interactive ai assistants in the real world. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20270–20281, 2023. 2

[40] Spencer Whitehead, Suzanne Petryk, Vedaad Shakib, Joseph Gonzalez, Trevor Darrell, Anna Rohrbach, and Marcus Rohrbach. Reliable visual question answering: Abstain rather than answer incorrectly. In *European Conference on Computer Vision*, pages 148–166. Springer, 2022. 3

[41] 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/CVF Conference on Computer Vision and Pattern Recognition*, pages 6659–6668, 2019. 2

[42] Ruihai Wu, Haoran Lu, Yiyan Wang, Yubo Wang, and Hao Dong. Unigarmentmanip: A unified framework for category-level garment manipulation via dense visual correspondence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16340–16350, 2024. 2

[43] Tao Wu, Chuhao Zhou, Yen Heng Wong, Lin Gu, and Jianfei Yang. Noisyeqa: Benchmarking embodied question answering against noisy queries. *arXiv preprint arXiv:2412.10726*, 2024. 2

[44] Yuncong Yang, Han Yang, Jiachen Zhou, Peihao Chen, Hongxin Zhang, Yilun Du, and Chuang Gan. 3d-mem: 3d scene memory for embodied exploration and reasoning. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 17294–17303, 2025. 2, 3, 5, 6

[45] Eunseop Yoon, Hee Suk Yoon, Mark A Hasegawa-Johnson, and Chang D Yoo. Can video llms refuse to answer? alignment for answerability in video large language models. In *The Thirteenth International Conference on Learning Representations*, 2025. 3

[46] Licheng Yu, Xinlei Chen, Georgia Gkioxari, Mohit Bansal, Tamara L Berg, and Dhruv Batra. Multi-target embodied question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6309–6318, 2019. 2, 3

[47] Xinxin Zhao, Wenzhe Cai, Likun Tang, and Teng Wang. Imaginenav: Prompting vision-language models as embodied navigator through scene imagination. *arXiv preprint arXiv:2410.09874*, 2024. 2