CURVE: A Benchmark for Cultural and Multilingual Long Video Reasoning

Supplementary Material

The supplementary section is structured as follows: We begin by elaborating on our *Ethical Considerations* (Sec. A) and provide an in-depth description of our *Annotator Guidelines* (Sec. B), detailing the meticulous protocols for both CURATORS (Sec. B.1) in creating culturally-grounded questions and AUDITORS (Sec. B.2) in verifying their objectivity. We then present *Additional Analysis and Experiments* (Sec. C), including studies on the effect of frame sampling (Sec. C.1), cross-lingual consistency (Fig. 2), autorater variance and judge reliability with open-source and human judge validation (Sec. C.3), web search augmented evaluation (Sec. C.4), additional open-source model results (Sec. C.5), and a text-only baseline (Sec. C.6). This is followed by a granular breakdown of our *Error Tagging* methodology (Sec. D), which includes the error taxonomy, evidence graph statistics, details on our multi-iteration analysis, and cross-model validation of the diagnostic pipeline. Finally, we provide details on *Annotator Recruitment and Compensation* (Sec. E) and include the complete *Prompts used* (Sec. F) in our evaluation and diagnostic pipelines.

A. Ethical Considerations

The development and deployment of CURVE are guided by a commitment to ethical AI research. Our primary motivation is to counter the Western and English-centric bias prevalent in existing benchmarks. To achieve this, we intentionally curated videos and native-language annotations from 18 diverse global locales through a meticulous, human-in-the-loop process. This approach ensures cultural authenticity and avoids the pitfalls of automated translation. All human annotators were compensated fairly at rates above their local market standards, and our data sourcing included explicit content moderation to exclude violent, explicit, or otherwise harmful material. To further mitigate the risk of perpetuating stereotypes, multiple cultural experts reviewed content for nuanced and respectful representation. The benchmark will be publicly released with clear documentation to guide responsible use, and its fine-grained error analysis is designed to provide a crucial tool for identifying and addressing biases in future multimodal models.

B. Annotator Guidelines

All the textual data in CURVE was completely annotated by human annotators. All raters are expert native speakers of the language for which they created data. Here we provide guidelines given to the raters for the dataset creation. As described in the Sec. 3.2 of the main paper we have two kind of experts for the annotation in our pipeline CURATORS and AUDITORS. We first describe the guidelines for the CURATORS in Sec. B.1 and then the guidelines for the AUDITORS in Sec. B.2.

Locale Code	Language Name	Region
ar-EG	Arabic	Egypt
de-DE	German	Germany
en-GB	English	United Kingdom
en-IN	English	India
es-MX	Spanish	Mexico
fr-FR	French	France
hi-IN	Hindi	India
id-ID	Indonesian	Indonesia
it-IT	Italian	Italy
ja-JP	Japanese	Japan
ko-KR	Korean	South Korea
mr-IN	Marathi	India
pt-BR	Portuguese	Brazil
ru-RU	Russian	Russia
ta-IN	Tamil	India
te-IN	Telugu	India
th-TH	Thai	Thailand
zh-TW	Chinese	Taiwan

Table 1. **CURVE Locales and their Codes.** This table provides a comprehensive mapping of the 18 distinct cultural and linguistic locales included in the CURVE benchmark, showing their standardized `language-region` codes, corresponding language names (from ISO 639), and region/country names (from ISO 3166-1 alpha-2).

B.1. Guidelines for CURATORS

B.1.1. Good Questions

- The question should **involve cultural entity/object** of your specific locale that you are working with.
- The question should require watching the video to answer the question and
  - should **not** be able to solve by **just listening to the audio**
  - should **not** be able to solve by **just reading some text in the video**
  - should **not** be able to solve by **just using general knowledge**
  - should **not** be solvable using **single frame**
- The question should require multiple reasoning steps to solve.
- The question should primarily ask about visual elements in the video (and not just focus on the speech).

- The question should have only one right answer and should not be subjective or ambiguous.
- The question should be open-ended i.e. it should not be a Multiple Choice Question (MCQ).
- The questions, answers and reasoning traces should be strictly in the **native language** of the video.
- If the question involves timestamp in either question/answer/reasoning trace, then use the standard format of `mm:ss` always. For ex: `12:56`, `06:13`

Examples of good questions

1. YouTube ID: `4EaRHj2Qa3g`

Question: How many total no. of empty raids happened in the first half from the time the player from Nilgiri Knights gets injured till the time when a team takes the first review?

Answer: Five (5)

Skills: Temporal Event Localization, Counting, Visual Cultural Understanding

Why is this a good question? This question involves the video from the time the player from Nilgiri Knights gets injured till the time when a team takes the first review in first half (temporal event localization), and counting all the instances of empty raid (counting). The question is about the player in the Kabaddi game (which is a cultural entity). Hence, it satisfies all the criteria of requiring at least three skills (temporal event localization, counting and visual cultural understanding) which makes it a good question.

Examples of bad questions

1. Question: Which event occurs after the old lady serves prasad in this video?

Answer: A red cloth is being placed on young girls head

Why is this a bad question? This question tests just temporal ordering (which event happens after an event?) and visual cultural understanding (old lady serving prasad). Hence, this just requires 2 skills while we require questions to have a minimum of 3 skills. Hence, this is a bad question.

2. Question: In what order do the following events happen?: People throwing colored powder (gulal) and water on each other, lighting a bonfire (Holika Dahan), sharing sweets & festive treats with friends and family.

Why is this a bad question? Because it can be answered with general cultural knowledge of Holi WITHOUT even looking at the video.

3. Question: How many times does the woman in red saree say the word "namak" in the video?

Why is it a bad question? Because it is too easy and can be solved by listening to the speech alone.

B.1.2. Reasoning steps

- A reasoning step is an action that you would take to break down the question solving process. You can think of them as the building blocks to the solution.
- A good question requires multiple reasoning steps to be performed in sequence to arrive at the answer.
- All steps you describe under "Reasoning Steps" must be **required** to solve the question.
- Without one of the steps, a person should not be able to get the answer.
- List all the reasoning steps as a numbered list (one by one)

Good examples of reasoning steps

- Watch the whole video to understand/obtain a certain piece of information that is important for the answer. Example: I watched the whole video to understand the big plot twist was the nice man was secretly the criminal.
- Move to a certain timestamp in a video mm:ss to find an object/person/event. Example: I looked for the red balloon and found it at 12:34
- Listen to a word/phrase in the speech. Example: I heard that the man in the black shirt declared "check" at 05:53-05:57
- Read a word/phrase on the screen. Example: At 04:03 I read that the title of the book is "Amar Chitra Katha".
- Find all instances of a particular object/person/event. Example: (1) Find all the men in the video wearing a white shirt. (2) Find all the times a goal is scored.
- Counting Example: Count how many people wearing white shirts appear in the video.
- Temporal Ordering Example: The question mentions a dog, cow and sheep appearing in the video. I looked through the video and noted their time of appearance to bring them in the right order.
- Look at something specific in one frame of the video (e.g. you can pause the video and see it) Example:
  - Observe the expression on somebody's face.
  - See the colour of something
  - See what the background is
  - See what objects are present in the scene
- Look at an event/action that is happening in a short section of the video (eg. a few seconds)

Bad examples of reasoning steps. Do not add irrelevant information in the steps. Example: Question: "How many dogs appear in the video" How did you solve it?: "1. The video starts by showing the title in white text on a black background. [...]"

What is the difference between a reasoning step and the answer?

- The final answer can be very short e.g. a single word or a phrase.
- However, the reasoning steps are the entire process to get to the answer.
- Hence just because an answer is simple, does not mean multiple steps are not required to get to it.

B.1.3. Usage of External Tools

Which tools are ok to use and what are not ok to use? External tools such as Google Search, reverse Image Search, Wikipedia or any other reliable source are ok to use.

Which tools are ok to use and what are not ok to use? Use of any **(Gemini/ChatGPT/Perplexity, Deep Research etc.)** for any kind of task is **not allowed**. Strictly do not use any kind of LLMs to brainstorm novel questions or to rewrite your questions/answers/reasoning traces. This is a strict requirement. Write the questions/answers/reasoning traces in your own words even if you feel the grammar might not be correct.

B.2. Guidelines for AUDITORS

Here we outline the audit process for the CURVE benchmark. The primary goal of this audit is to ensure the quality, accuracy, objectivity, and consistency of the question-answer (QA) pairs generated by our raters. Your role as AUDITORS is critical in achieving a high-quality benchmark.

B.2.1. Audit Workflow

Audit process has two main steps:

Step 1: Answer the Question Yourself

- The AUDITORS will receive a video and a question.
- Watch the video.
- Read the Question First:
  - Is the question clear and unambiguous?
  - If NO (the question is ambiguous, unclear, or could have multiple answers):
    - STOP. Don't try to answer it yet.
    - Send feedback to the rater explaining why the question is problematic.
    - Work with the rater to modify the question until it is clear, objective, and aims for a single answer.
    - Once the question is fixed and clear, proceed to the next point.
  - If YES (the question is clear):
    - Now, answer this clear question yourself.
    - Keep your answer very short (just a few words or phrases).
- If you find the question ambiguous (ie., not very objective or multiple answers are possible), send this feedback to the rater and after the question is modified then answer the question.
- Keep your answer very short (just a few words or phrases) and objective.
- You are free to use google search/reverse image search or any other website to answer.
- However, please Do NOT use any kind of LLMs or ChatBots (like ChatGPT, Gemini etc) for anything.
- Very Important: Do NOT look at the rater's answer and reasoning steps (the "ground truth" answer) yet!

Step 2: Check Your Answer Against the Rater's Answer
Now, compare the answer you wrote with the "ground truth" (GT) answer the rater provided. Based on what you find, you'll decide on one of the following scenarios. You can use these scenario codes (like "Scenario A") when you make notes or report issues.

Scenario A: Good Match - QA Approved

- **What it means:** Your answer is the same or very similar to the rater's GT answer, and both are correct based on the video.
- **Action:** Mark this QA pair as "Approved". Move to the next item.
  - **Example – 1:**
    - Video shows: A video of a woman performing Bharatanatyam.
    - Question: How many different types of dance moves were performed?
    - Your Answer: Two or दो
    - Rater's Answer: 2
    - Decision: This is a good match (Scenario A). Both are correct and essentially the same.

Scenario B: Slight Difference - Needs Minor Fix

- **What it means:** Your answer and the rater's GT answer are mostly right and aim for the same thing, but the rater's GT answer could be a bit better (clearer, more exact). The GT isn't wrong, just not perfect. (The question should already be clear from Step 1).
- **Action:**
  - Give feedback to the rater.
  - Suggest how to improve their answer.
  - Work with the rater to agree on the best version. The rater will update it.
  - **Example:**
    - Video shows: A cat slowly walking across a room.
    - Question: What is the cat doing?
    - Your Answer: Walking slowly.
    - Rater's GT Answer: Moving.
    - Decision: Scenario B. The rater's GT "Moving" is true, but "Walking slowly" is more precise from the video.
    - Feedback to Rater: "Could we make the answer more specific, like 'walking slowly' or 'strolling,' as that's clearly visible in the video?"

Scenario C: Big Difference - Needs More Review

- **What it means:** Your answer is very different from the rater's GT answer. You need to look closely at why.
- **Action:** Re-watch the video, look at both answers carefully. This will lead to one of three sub-cases:

Sub-Scenario C1: You're Right, Rater's Wrong.

- **What it means:** Your independent answer is correct,

and the rater's GT answer is incorrect based on the video.

- **Action:**
  - Explain to the rater why their GT answer is wrong, using video evidence.
  - Tell them to update their GT answer to be correct (it might be your answer, or a version you both agree on).

Sub-Scenario C2: You're Wrong, Rater's Right.

- **What it means:** After review, you realize your first answer was incorrect, and the rater's GT answer is correct.
- **Action:**
  - No need to tell the rater anything for this specific item if their answer is good. (Good job catching your own mistakes! This helps you too.)

Sub-Scenario C3: Question is Confusing / Multiple Answers Possible.

- **What it means:** The question itself is unclear, ambiguous, or could be interpreted in ways that lead to different "correct" answers. Both your answer and the rater's answer might seem okay (or both wrong) because the question is flawed.
- **Action:** This is a key area for improvement!
  - Talk with the rater about why the question is confusing.
  - Work together to rewrite the question to be very clear, specific, and have only ONE obvious correct answer from the video.
  - Then, agree on the new, single correct GT answer for the improved question.
  - The rater will update both the question and the GT answer.
  - Our Goal: Fix these so they become clear like Scenario A.
  - Work with the rater to make the question specific.

C. Additional Analysis, Experiments and Results

In this section we explore the impact of temporal sampling density on model performance (Sec. C.1), assess cross-lingual reasoning consistency (Sec. C.2), and confirm the robustness of our evaluation pipeline through an autorater reliability analysis Sec. C.3.

C.1. Effect of Number of Frames

As discussed in Sec. 4.2 of the main paper, we evaluated the temporal complexity of CURVE by varying the number of sampled input frames from 1 to 512. The results of this analysis on Gemini-2.5-Pro, visualized in Fig. 1, confirm our primary findings. The plot demonstrates a monotonic increase in accuracy with more frames across a diverse subset of locales, validating that tasks in CURVE require temporal reasoning and cannot be resolved from static images. Concurrently, the diminishing performance gains at higher frame counts reinforce our conclusion that the primary performance bottleneck is the higher-level, culturally-contextualized reasoning demanded by the benchmark, rather than a mere lack of visual information.

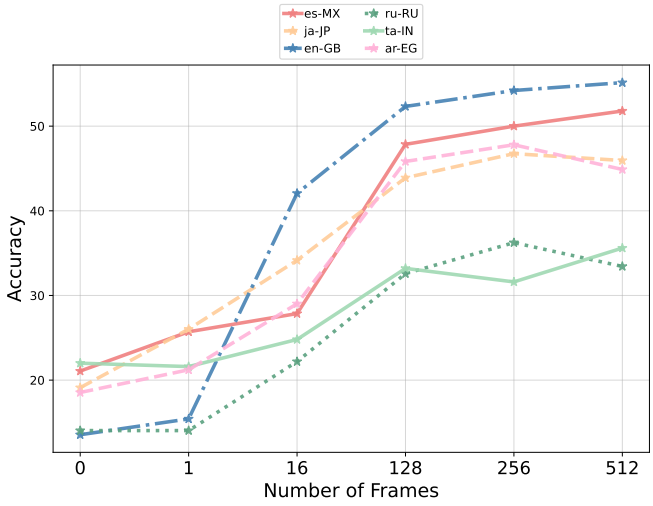


Figure 1. **Impact of frame sampling on model accuracy.** Gemini-2.5-Pro's performance improves with more frames, validating the temporal nature of our benchmark.

C.2. Cross-Lingual Consistency Experiments

To assess how query language affects reasoning, we performed a cross-lingual evaluation where we translated questions into the five other languages while keeping the visual input fixed. As shown in Figure 2, the results confirm that model performance is not language-agnostic. Accuracy is usually highest when the query is in the video's native language (the diagonal). While English and Russian serve as more robust translation targets, performance varies significantly based on both the source and target languages. This demonstrates that simply translating queries is an insufficient strategy to overcome cultural and linguistic dependencies.

C.3. Autorater Variance and Judge Reliability

The primary evaluation pipeline for CURVE relies on Gemini-2.5-Flash as an LLM Judge to score model responses, as detailed in Sec. 4 of the main paper. To ensure the robustness and reproducibility of our findings, we conducted a validation experiment to assess the variability of this autorater. The central concern was to verify that our reported results are not biased by the specific LLM Judge chosen, but rather a true reflection of the models' capabilities

Table 2. Hyperparameters for all model baselines

Method	# of Frames	Hyperparameters
Qwen-2.5-VL	default	togetherAI API [1] default inference
Qwen-3-VL	default	vLLM inference [2], max-seq-length=128k
Claude-Sonnet-4	100	thinking_budget_tokens=10000
GPT-5-mini	256	reasoning_effort=high, verbosity=default, max_output_tokens=default
GPT-5	256	reasoning_effort=high, verbosity=default, max_output_tokens=default
Gemini-2.5-flash	256	temperature=0, thinking=dynamic, seed=default, sampling=default
Gemini-2.5-pro	256	temperature=0, thinking=dynamic, seed=default, sampling=default

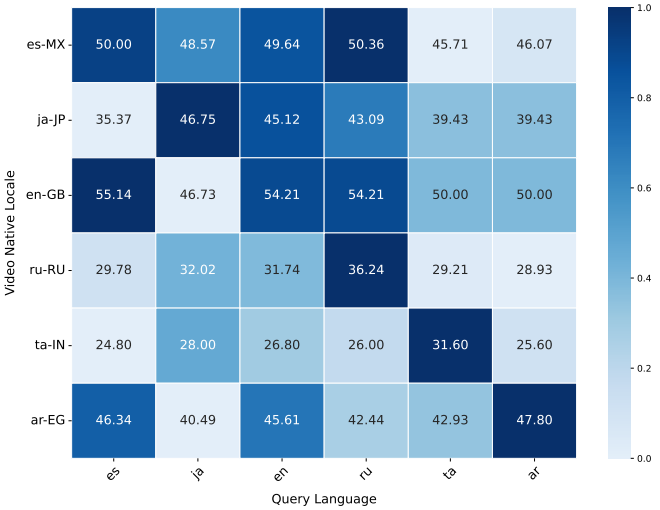


Figure 2. **Cross-Lingual Consistency Analysis.** Colors are normalized row-wise. The consistently high scores along the diagonal demonstrate that the model performs best when the query is in the video's native language, while off-diagonal variance highlights significant cross-lingual performance gaps.

	Autorater	
Locale	**Gemini-2.5-Flash**	**GPT-5-mini**
ar-EG	47.80	48.29
de-DE	48.73	48.73
en-GB	54.21	53.74
en-IN	47.71	47.79
es-MX	50.00	49.64
fr-FR	52.56	53.84
hi-IN	41.89	43.24
id-ID	55.97	54.85
it-IT	51.47	51.47
ja-JP	46.75	46.75
ko-KO	64.29	64.29
mr-IN	38.72	38.35
pt-BR	43.75	44.27
ru-RU	36.24	36.24
ta-IN	31.60	30.80
te-IN	28.00	25.60
th-TH	39.03	38.71
zh-TW	40.00	40.00
Macro Avg	45.48	45.37
Weighted Avg	45.07	44.97

Table 3. **Autorater Reliability Analysis**. Comparison of scores for Gemini-2.5-Pro on the CURVE benchmark as evaluated by two distinct LLM judges. The minimal variance in both per-locale and aggregate scores demonstrates the robustness of our evaluation methodology.

on our benchmark. For this analysis, we re-evaluated the complete set of responses from our top-performing model, Gemini-2.5-Pro (Tab. 2 of the main paper), using an alternative powerful model, GPT-5-mini, as a secondary autorater. We used the same scoring prompt for both judges. Tab. 3 presents a comparison of the scores assigned by Gemini-2.5-Flash and GPT-5-mini across all 18 locales. The results demonstrate a minimal variance between the autoraters. The aggregate scores are remarkably close. The weighted average from Gemini-2.5-Flash is 45.07, while GPT-5-mini yields 44.97, a negligible difference of just 0.1%. The macro averages are similarly aligned at 45.48 and 45.37.

Open-Source Judge Validation. To further address potential concerns about relying solely on closed-source LLM judges, we also evaluated using QWEN-3, an open-source model. As shown in Tab. 4b, the QWEN-3 judge produces a weighted average of 44.73, with an agreement of 93.4% with Gemini-2.5-Flash. The high consistency across all three judges (Gemini-2.5-Flash, GPT-5-mini, and QWEN-3) is attributable to the fact that questions in CURVE are designed to be objective (*e.g.* counting) and unambiguous, making the judging task a straightforward semantic match.

Judge	Agreement
GPT-5-Mini	95.3
QWEN-3	93.4
Human*	96.6

(a) Agreement metrics

Judge	Weighted Avg.
Gemini-2.5-Flash	45.07
GPT-5-Mini	44.97
QWEN-3	44.73

(b) Evaluation by three judges

Table 4. (a) Judge agreement with Gemini-2.5-Flash. (b) Weighted average score by different judges. * Locales: en-GB, ta-IN.

Model	GPT-5	+Search
Weighted Avg.	42.2	44.8

Table 5. Effect of enabling web search on GPT-5's performance on CURVE.

Qwen-3	GLM	Qwen-2.5	MiMo
21.5	15.2	12.8	11.6

Table 6. Weighted average scores of additional open-source models on CURVE.

Human-Judge Alignment. We also conducted a human-judge agreement study to validate the LLM-based evaluation against human grading. Human evaluators independently scored model responses for two locales: `en-GB` (a high-resource language) and `ta-IN` (a low-resource language). As shown in Tab. 4a, the Gemini-2.5-Flash judge achieves an agreement of 96.6% with human evaluators, confirming the efficacy of the autorater across both high- and low-resource settings.

C.4. Web Search Augmented Evaluation

To assess whether access to external knowledge can help bridge the performance gap on CURVE, we evaluated GPT-5 with web search enabled. As shown in Tab. 5, enabling search yields a modest improvement of 2.6% (from 42.2 to 44.8 weighted average). While search helps, a massive performance gap to the human baseline remains, confirming that the primary bottleneck in CURVE is *visual grounding* of cultural concepts, which text-based retrieval cannot fully address.

C.5. Additional Open-Source Models

In addition to the QWEN family of models evaluated in the main paper, we further benchmark two recent open-source Video-LLMs on CURVE: GLM-V4.1-9B-Thinking [1] and MiMo-VL-7B-RL [2]. As shown in Tab. 6, their weighted average scores are comparable to Qwen-2.5-VL, and the overall gap between these models and human performance remains massive (∼80%), underscoring the difficulty of CURVE for current open-source models.

C.6. Text-Only Baseline

To verify that CURVE questions cannot be solved through textual shortcuts alone, we evaluate Gemini-2.5-Pro without any video context (i.e., 0 frames). Across the six evaluation locales, the average accuracy is just 18% (Fig. 1), indicating that the questions require genuine video understanding and cannot be answered through world knowledge or linguistic cues alone. This confirms the multimodal nature of the reasoning required by CURVE.

D. Error Tagging

D.1. Graph Statistics

Property	**Mean**	σ	**Total**
Nodes	5.0	2.5	4351
Nodes w/ Timestamps	3.1	2.5	2743
Edges	4.4	3.1	3865
Depths	2.5	1.3	N/A

Table 7. Statistical overview of graphs derived from human reasoning in selected locales (N=878 questions).

The first step of the Reasoning Trace based analysis is to develop the *Evidence Graph*. Once the *Evidence Graph* is developed, we attempt to understand the CURVE with the help of the graph's structural properties. The key properties are discussed in Sec 5.3 under the Structural Analysis heading. The statistics are described in Tab. 7.

D.2. Taxonomy

We tag each failure to map the evidence to the reasoning trace in the node. The tagging is supposed to understand the core reason why the evidence was not retrieved. MINERVA uses an LLM to score for Perceptual Correctness, Temporal Localization, and Logical Reasoning. We expand across two dimensions: first, we expand the taxonomy by splitting the broad class of Perceptual Correctness into three new categories of Spatial Grounding, Attribute Misidentification, and Spurious Objects/Events. The exact definition and a brief example is given in Tab. 8. Second, instead of scoring the entire solution, we divide the solution into different evidences(nodes) and then classify the type of error. This helps us better identify *where and when* the model goes wrong. The taxonomy is introduced in Sec 5.2 in the main paper.

D.3. Majority Voting and Human Assessment

To improve the robustness of the prompt-based analysis on the task of error tagging, we employ majority voting. For each question, we query Gemini-2.5-Pro three times

[1] https://www.together.ai/qwen
[2] https://docs.vllm.ai/projects/recipes/en/latest/Qwen/Qwen3-VL.html

Error Type	Definition	Illustrative Example
Temporal Localization	A failure to search the correct time segment in the video.	A required clue is at 01:15, but the model only searches before 00:30.
Spatial Grounding	A failure to "see" an object or event present in the correct spatiotemporal location.	Looking at a parade float but failing to identify the banner on its side.
Spurious Object/Event	An invention error where the model overcounts or fabricates an object or event that is not present.	Claiming three dancers are wearing hats when only two do.
Attribute Misidentification	A failure to correctly identify the properties of a detected object.	Correctly identifying a car but stating it is blue when it is green.
Knowledge-Dependent Issue	A failure to recall or apply correct external factual knowledge.	Identifying Japan's flag but failing to recall that Tokyo is its capital.
Reasoning	A logical failure in connecting evidence when all prerequisite information is correct.	Seeing Team A celebrating and Team B sad, but concluding that Team B won.

Table 8. **Error Taxonomy with Definitions and Examples.** This table details the classification system used for our fine-grained error analysis. Each category represents a distinct failure mode, allowing for a precise diagnosis of model weaknesses in multicultural video reasoning.

and take the majority consensus. We present the consensus between 3 LLMs in the error tagging of Iteration 1 in Tab. 9. It highlights the high consensus between the LLMs with over 97.7% of examples achieving majority consensus. The locale-wise consensus status is described in Tab. 9.

Locale	Majority Consensus ($\geq$ 2/3 Agree)	No Consensus (All Differ)
en-GB	106	11
es-MX	109	0
ja-JP	95	1
ar-EG	247	4
ru-RU	140	3
ta-IN	157	1

Table 9. LLM Majority Consensus by Locale

The verification of tagging requires significant human effort, taking into consideration the human reasoning trace, the evidence graph and the model thoughts. We manually verify the error tagging for a small randomly selected subset of 60 questions distributed amongst the six selected locales. We find that the model-human agreement is at a high 88.5%, verifying the validity of the error tagging. However, it also shows the scope of improvement that is possible in the LLM-based evaluation.

Cross-Model Validation. To verify that the error tagging is not sensitive to the choice of the prompted LLM, we ran an additional iteration of the *Iterative Error Isolation* pipeline using GPT-5 instead of Gemini-2.5-Pro. The proportion of cultural visual perception and reasoning errors remained highly consistent (within $\sim$2% margin), strengthening the robustness of our diagnostic findings.

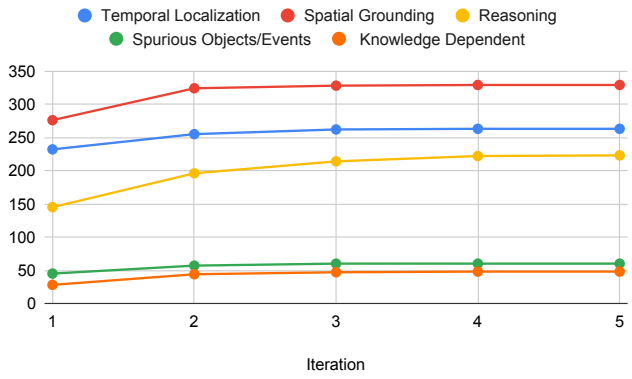


Figure 3. Iteration-wise Error Distribution for Gemini 2.5 Pro.

D.4. Multi-Iteration Error Isolation

We find that the Multi-Iteration Error Isolation helps us gather a complete understanding through the counterfactual hint generation as described in the Sec. 5.2 of the main paper. We see a steady increase in error types like Temporal Localization, Spatial Grounding and Reasoning. Reasoning sees the maximum gain indicating that the later steps of the question hidden by the errors in the initial errors have a higher proportion of reasoning and evidence aggregation tasks. The complete graph can be studied at Fig. 3.

E. Annotator Recruitment and Compensation

Our data was annotated by professional data labelers. We recruited an average of six CURATORS and five AUDITORS from each of the 18 countries. The selection process for CURATORS and AUDITORS was based on content writing expertise and native language proficiency with deep, situated cultural knowledge, which was assessed through both verbal and written language assessments. In addition, the AUDITORS underwent a vetting process to assess their analytical and observational skills and a proven ability to maintain consistency and accuracy across a large volume of annotations. All annotators were compensated at rates above the prevalent market rates in their respective countries, in full compliance with local minimum wage regulations.

F. Prompts used

For completeness and to ensure full reproducibility of our methodology, this section provides the exact prompts used for our automated evaluation and diagnostic analysis. We present the prompt for our LLM Judge Sec. C.3, which defines the three-point scoring criteria used for the main model evaluation presented in Tab. 2 of the main paper. Following this, we provide the prompt used to formalize the unstructured human reasoning traces into a structured Directed Acyclic Graph (DAG). This serves as the ground-truth *Evidence Graph* for our error analysis (Fig. 5 of main paper). Finally, we detail the comprehensive prompt that governs our Iterative Error Isolation pipeline (Fig. 6 and Fig. 7 of main). This prompt instructs the LLM to traverse the Evidence Graph, tag specific failure modes according to our taxonomy, and generate corrective hints for subsequent iterations.

CULTURAL AUTORATER PROMPT

You are an expert at global culture and you are grading an exam where the answer could be given in any language.
You will be provided with a question, the reference answer and the model's answer. Your job is to judge if the model's answer was equivalent to the reference answer and provide a single integer score between `0` and `2` with the following criteria:
`0`: The model's answer is completely wrong and is not related to the reference.
`1`: The model's answer is partially correct and misses some details from the reference answer.
`2`: The model's answer is completely correct.

INSTRUCTION:

1. You should accept answers that are accurate translations or transliterations of the reference answer. The language or script of the answer should not be taken into account. For example:
   a) reference answer = भेलपूरी
      `model's answer = Bhel Puri`
      `score = 2`
2. Allow for alternate names of the same cultural concept. Don't penalize for spelling errors or minor variations. Focus solely on the cultural concept. For example:
   a) `reference answer = The London Eye`;
      `model's answer = Millennium Wheel`
      `score= 2`
3. Partial scoring of `1` can be given in cases where model misses details or the answer is not complete. For example:
   a) `reference answer = Sun Temple`
      `model's answer = Temple`
      `score = 1`
4. For questions with a numerical answer, the score is determined by an exact value match. Award a full score of `2` if the model's answer represents the same numerical value as the reference answer, regardless of whether it is written in digits (e.g., `10`) or words (e.g., `ten`). If the model's answer represents any other number and there is no exact value match, it must receive a score of `0`. No partial credit is given. For example,
   a) `reference answer = 10`
      `model's answer = ten`
      `score = 2`
   b) `reference answer = 10`
      `model's answer = 11`
      `score = 0`
5. Your answer should only contain a single integer value in `[0, 1, 2]` and nothing else.

```
Question: question
Reference answer: gt
Model's answer: pred
Your response:
```

Figure 4. **The prompt used for the LLM Judge (Gemini-2.5-Flash)** in our main evaluation pipeline (Tab. 2 of the main paper). It details the three-point (0-2) scoring criteria for assessing the semantic equivalence between a model's response and the ground-truth answer.

```
 You are a Reasoning Analyst.  Your task is to convert a Video QA reasoning process into a directed graph
representing the direct solution path.

**Step-by-Step Instructions**

1.  **Identify Atomic Nodes:** Extract the causal chain of evidence, **excluding** all procedural text, negative
findings, and dead ends. Create an "atomic evidence node" (Node1, Node2...) for each distinct step. A node must
be:
    *   Derived from a single video timestamp, OR
    *   A single piece of external information, OR
    *   A single inference over existing evidence.
   *   *Action:* Strip conversational preambles from `<Evidence Text>` and record the `<Timestamp>` (or "N/A").

2.  **Construct Graph:** Define directed edges. Add a node ID to `parent_nodes` **only** if the current node is
strictly dependent on that parent node's information.

3.  **Validate Termination:** Ensure the graph ends with the final `Answer`. If no existing node contains the
answer, create a final node containing the answer text and link it to its immediate evidence.

**Output Format:**
Return **only** a single valid JSON object.

```json
{{
  "solution_graph": {{
      "Node1": {{
        "evidence": "<Evidence Text>",
        "timestamp": "<Timestamp or N/A>",
        "parent_nodes": ["<List of parent IDs>"]
      }},
      ...
    }}
}}

**Input**
[BEGIN INPUT]
Question: {question}
Answer: {answer}
Human Response: {human_response}
[END INPUT]
```

Figure 5. **Prompt to derive the Directed Acyclic Graph from the human reasoning steps.**
```

```
You are a rigorous AI evaluator for Video QA. Your goal is to analyze a language model's reasoning against a ground-truth `Evidence Graph` to generate a cue for its next attempt.

**Golden Rule:** The `Evidence Graph` is the absolute source of truth. Flag only **causally relevant** errors that damage the answering process.

---
**Task 1: Evaluate the Solution Graph**

Evaluate each node in the `Evidence Graph` sequentially. Stop at the first rule that applies.

**1. Check for Inherited Errors (`status`: undetermined)**
If any parent node is `wrong`, `undetermined`, or has `divergence: true`, the current node inherits the failure.
*   **Action:** Set `status` to `undetermined`. Provide `justification`. Stop.

**2. Check for Valid Divergence (`divergence`: true)**
If all parents are `right`, check if the model pursues a valid alternative path. Set `divergence` to `true` (and stop) **only if** all conditions are met:
1.  **Alternative Evidence:** The model uses new evidence distinct from the human reasoning.
2.  **Productive:** The path is logical and leads toward the answer (not a dead end).
3.  **Non-Contradictory:** The path does not contradict or compete against the human reasoning, question, or answer.

**3. Evaluation**
Determine `status` based on answer-affecting information only (ignoring methodology).
1.  **Content Check:** Did the model find the correct evidence content?
    *   **If No:** Set `status` to `wrong`. Proceed to **Causal Analysis**.
    *   **If Yes:** Proceed to Timestamp Check.
2.  **Timestamp Check:** Is the timestamp critical **AND** outside the +/- 5s tolerance?
    *   **If Yes:** Set `status` to `wrong`. Proceed to **Causal Analysis**.
    *   **If No:** (Timestamp not critical OR within tolerance). Set `status` to `right`. Evaluation complete.

**Part B: Causal Analysis for Failures**
If `status` is `wrong`, identify the primary cause of failure by finding the **first critical mistake** in the `model_output_thoughts`. Go through the causal categories in order and stop at the first match.

**Causal Analysis: A Step-by-Step Diagnosis**
When `status` is `wrong`, identify the **first critical mistake** by checking these categories in order. Stop at the first match.

1.  **Intent/Planning Failure** (`Reasoning`)
    *   Did the model misinterpret the user's goal or fail to form a correct search plan?
2.  **Knowledge Failure** (`Knowledge-Dependent Issue`)
    *   Was the plan correct, but the model relied on an incorrect internal fact or tool result?
3.  **Temporal Failure** (`Temporal Localization`)
    *   Did the model fail to search within the correct time segments or miss a critical timestamp?
4.  **Spatial / Detection Failure** (`Spatial Grounding`)
    *   *(Assumption: Time is correct)*. Did the model miss a visual/audio object, under-count objects, or fail to find an object matching the attribute specifications?
5.  **Attribute Failure** (`Attribute Misidentification`)
    *   *(Assumption: Object detection is correct)*. Did the model incorrectly identify properties (color, text, type) of the found object?
6.  **Hallucination** (`Spurious Object/Event`)
    *   Did the model **over-count** objects or invent events?
7.  **Logical Failure** (`Reasoning`)
    *   Did the model find the correct evidence but fail to connect it to the right conclusion, or ignore provided info?
---
```

Figure 6. **Prompt to traverse through the graph and tag errors. This represents the prompt that runs for one iteration. (Part 1)**

```
**4. Justify**
*   **If `divergence` is `true` AND `evidence_retrieved` is `no`**: Explain *why* it failed using the findings from Part A and Part B.
*   **Otherwise**: Set to `N/A`.

**Task 2: Generate the `next_attempt_cue`**
Create a single, cumulative block of text to guide the model's next attempt.

1.  **Start:** Include all text from `Additional Cues`.
2.  **Summarize Success:** For all `right` nodes leading up to the error, state the evidence, timestamps, and logic as established facts (e.g., "Here is what we know so far...").
3.  **Guide the Error:** For the **first** `wrong` or `divergent` node:
    *   *Evidence Errors:* State the correct `evidence` and `timestamp`.
    *   *Reasoning Errors:* State the correct `evidence`, `timestamp`, and the correct line of reasoning.
4.  **Constraints:** Do **not** mention past failures (e.g., "You missed") and do **not** reveal info about `undetermined` nodes.

**Input**
**Question (English):** {question_english}
**Additional Cues:** {additional_cues}
**Ground Truth Answer (English):** {ground_truth_answer_english}
**Model Prediction (English):** {model_prediction_english}
**Model Thoughts/Outputs:** {model_output_thoughts}
**Evidence Graph:** ```json{evidence_graph_json}```
**Human Reasoning (Raw):** {human_reasoning}

Output Format
Your output must be a single JSON object. Do not include any other text or explanation outside the JSON.

```json
{{
  "solution_graph": {{
    "node_id": {{
        "evidence": "string (The description of the evidence for this node)",
        "timestamp": "string or null (The timestamp range where evidence appears, e.g., '00:05 - 00:10')",
        "parent_nodes": ["string", "..."],
        "outputs": {{
          "divergence": "boolean or null (True if the model uses a valid alternative reasoning path)",
          "evidence_retrieved": "string or null ('yes', 'partial', 'no')",
          "status": "string or null ('right', 'wrong', 'undetermined')",
          "missing_evidence_reason": "string or null (e.g., 'Localization', 'Grounding', etc.)",
          "justification": "string (Your concise explanation for the assigned status of this node.)"
        }}
      }},
      ...
    }},
    "next_attempt_cue": "string or null (A cue to help the model improve its next attempt)",
}}
```

Figure 7. **Prompt to traverse through the graph and tag errors. This represents the prompt that runs for one iteration. (Part 2)**
```

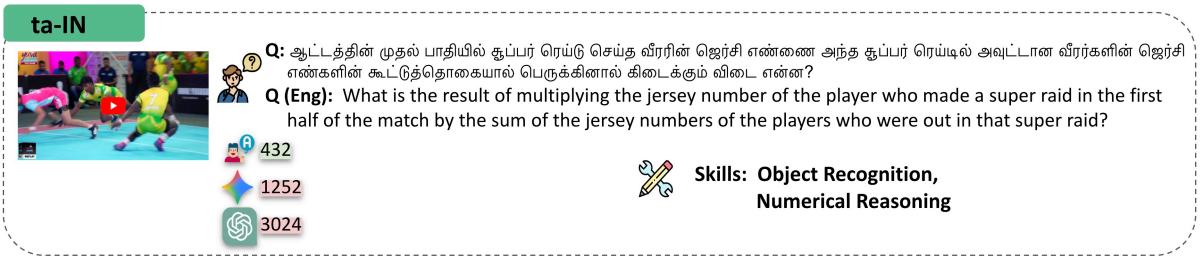
G. Additional qualitative examples

We present qualitative examples from CURVE in Fig. 8 (`ta-IN`), Fig. 9 (`es-MX`), and Fig. 10 (`en-GB`). These figures illustrate the *Iterative Error Isolation* analysis applied to both Gemini-2.5-Pro and GPT-5.

References

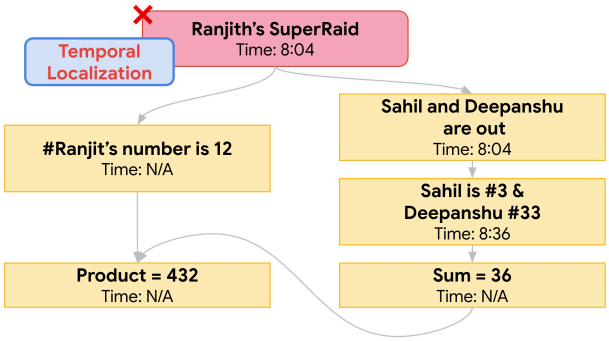
[1] V Team, Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang, Jiale Cheng, Ji Qi, Junhui Ji, Lihang Pan, Shuaiqi Duan, Weihan Wang, Yan Wang, Yean Cheng, Zehai He, Zhe Su, Zhen Yang, Ziyang Pan, Aohan Zeng, Baoxu Wang, Bin Chen, Boyan Shi, Changyu Pang, Chenhui Zhang, Da Yin, Fan Yang, Guoqing Chen, Jiazheng Xu, Jiale Zhu, Jiali Chen, Jing Chen, Jinhao Chen, Jinghao Lin, Jinjiang Wang, Junjie Chen, Leqi Lei, Letian Gong, Leyi Pan, Mingdao Liu, Mingde Xu, Mingzhi Zhang, Qinkai Zheng, Sheng Yang, Shi Zhong, Shiyu Huang, Shuyuan Zhao, Siyan Xue, Shangqin Tu, Shengbiao Meng, Tianshu Zhang, Tianwei Luo, Tianxiang Hao, Tianyu Tong, Wenkai Li, Wei Jia, Xiao Liu, Xiaohan Zhang, Xin Lyu, Xinyue Fan, Xuancheng Huang, Yanling Wang, Yadong Xue, Yanfeng Wang, Yanzi Wang, Yifan An, Yifan Du, Yiming Shi, Yiheng Huang, Yilin Niu, Yuan Wang, Yuanchang Yue, Yuchen Li, Yutao Zhang, Yuting Wang, Yu Wang, Yuxuan Zhang, Zhao Xue, Zhenyu Hou, Zhengxiao Du, Zihan Wang, Peng Zhang, Debing Liu, Bin Xu, Juanzi Li, Minlie Huang, Yuxiao Dong, and Jie Tang. Glm-4.5v and glm-4.1v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning, 2025. 6

[2] LLM-Core-Team Xiaomi. Mimo-vl technical report, 2025. 6

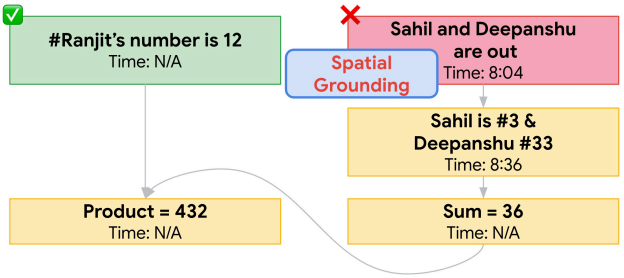


Context: A question from the `ta-IN` locale. Question is translated here for better understanding. The reasoning trace is depicted as *Evidence Graph* in the figures below. We notice that both GPT-5 and Gemini-2.5-Pro both arrive at the wrong answer but different types of errors occur as seen in *Iterative Error Isolation*.

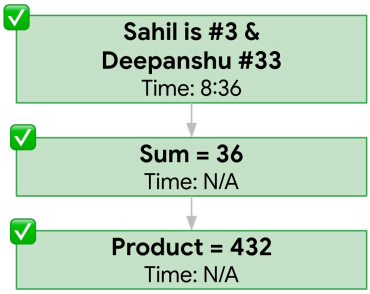
Gemini-2.5-Pro: Iterative Error Isolation



(a) **Iter 1:** Model was unable to even retrieve the correct timestamp, hence Temporal Localization and hence other unevaluated nodes are in yellow.

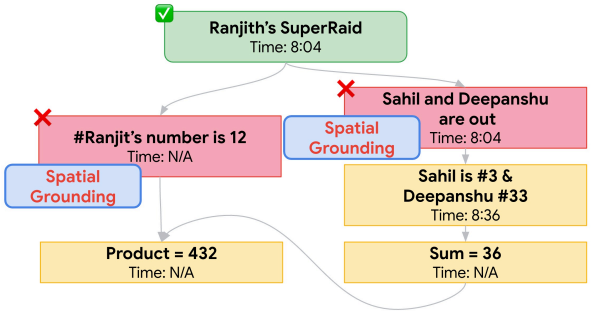


(b) **Iter 2:** Provided additional context of the timestamp of the raid. Finds Ranjit's number correctly but fails to locate others (Spatial Grounding).

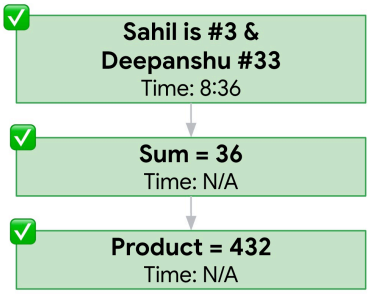


(c) **Iter 3:** After all previous contexts, model answers correctly.

GPT-5: Iterative Error Isolation

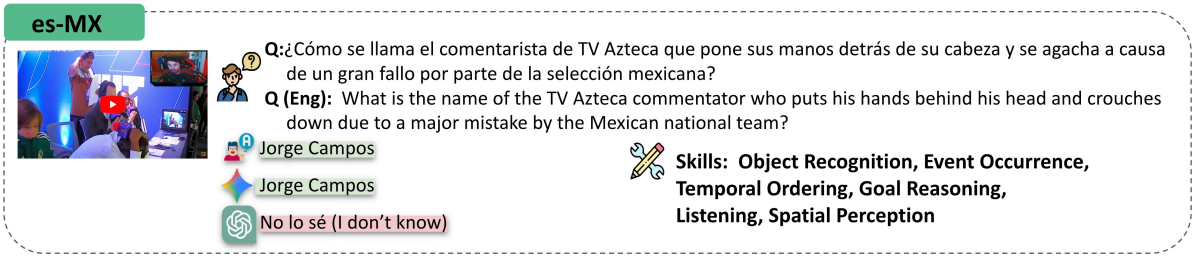


(a) **Iter 1:** Model was able to retrieve the correct timestamp. However, it didn't identify any of the players right Spatial Grounding and hence other unevaluated nodes are in yellow.



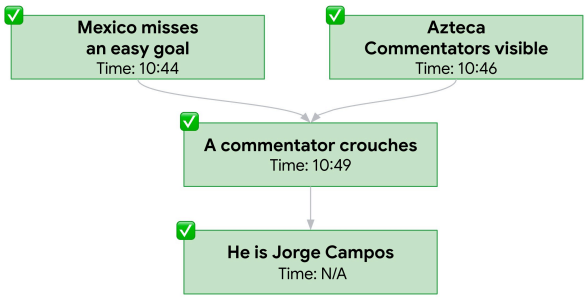
(b) **Iter 2:** Provided additional context of the raid and the player information from visited nodes. Model is able to complete all the remaining steps correctly.

Figure 8. Question from `ta-IN` locale of CURVE and *Error Isolation* graphs for Gemini-2.5-Pro and GPT-5.



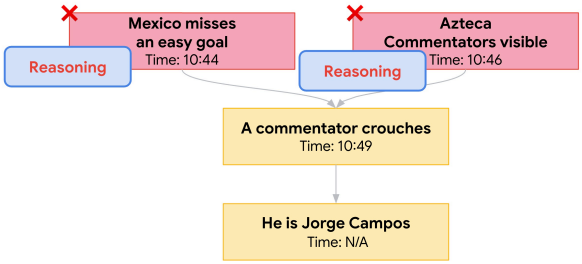
Context: A question from the `es-MX` locale. Question is translated here for better understanding. The reasoning trace is depicted as *Evidence Graph* in the figures below.

Gemini-2.5-Pro: Iterative Error Isolation

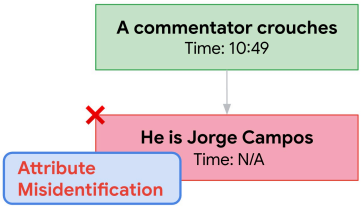


(a) **Iter 1:** Model was able to solve the question correctly in the first inference. Hence, all evidences are correctly mapped in the model reasoning traces.

GPT-5: Iterative Error Isolation

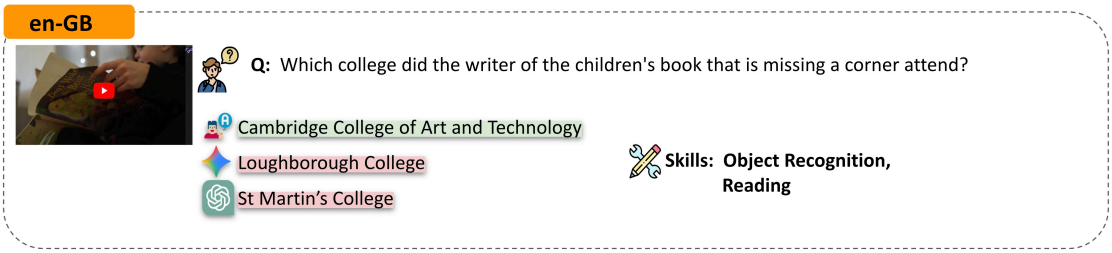


(a) **Iter 1:** Model did not even plan how to solve the problem (Reasoning). It directly decides that it does not know the answer. All other unevaluated nodes are in yellow.



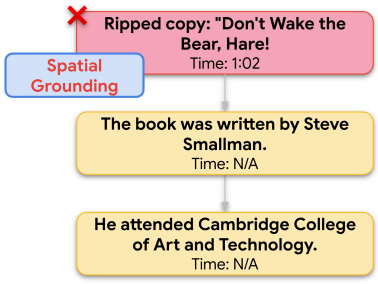
(b) **Iter 2:** The wrong evidences in **Iter 1** are additionally provided to the model. The model then correctly spatially grounds the crouching commentator but misidentifies him as Luis García. Since the model successfully grounded the target, the failure in the visual-task of naming is classified as a Attribute MisIdentification error rather than a lack of External Knowledge.

Figure 9. Question from `es-MX` locale of CURVE and *Error Isolation* graphs for Gemini-2.5-Pro and GPT-5. Gemini-2.5-Pro gets this question right, while GPT-5 makes an error.

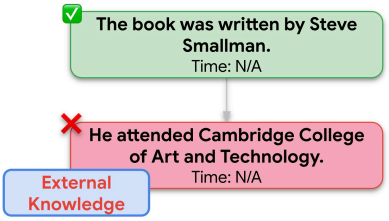


Context: A question from the en-GB locale. The reasoning trace is depicted as *Evidence Graph* in the figures below..

Gemini-2.5-Pro: Iterative Error Isolation

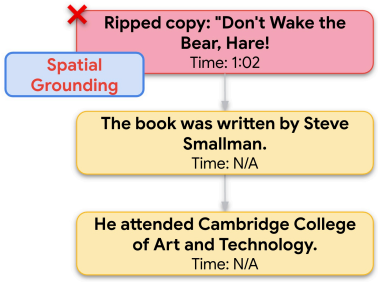


(a) **Iter 1:** Model was able to look at the correct timestamp. However, it didn't identify the required book with a ripped corner (Spatial Grounding). It instead focused on another book , and hence other unevaluated nodes are in yellow.

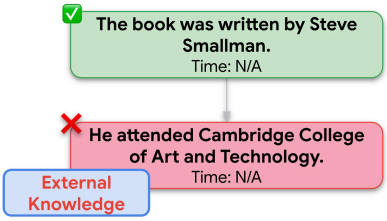


(b) **Iter 2:** Additional context of the book and the timestamp is provided. The model identifies the author correctly. However, it didn't identify the college he attended, an information that was not in the video (External Knowledge).

GPT-5: Iterative Error Isolation



(a) **Iter 1:** Model was able to look at the correct timestamp. However, it didn't identify the required book itself (Spatial Grounding). It instead focused on another book , and hence other unevaluated nodes are in yellow.



(b) **Iter 2:** Additional context of the book and the timestamp is provided. The model identifies the author correctly. However, it didn't identify the college he attended, an information that was not in the video (External Knowledge).

Figure 10. Question from en-GB locale of CURVE and *Error Isolation* graphs for Gemini-2.5-Pro and GPT-5. Both models make similar type of errors in this example.