

Evaluating accuracy in long-video analysis

Human-reviewed answer coverage and progress in understanding longer videos.

Version 1.3 · October 8, 2026 · Research note

Understanding beyond a moment

Long recordings carry ideas through time. A concept introduced near the beginning may become clearer through a later example, then acquire a qualification in the conclusion. A useful answer needs to connect those moments while preserving what each one means.

For Pelayar, long-video understanding is about that continuity: bringing relevant context into notes, explanations and questions about a recording. Duration changes the challenge because important information can be separated by extended discussion, topic changes and intervening examples.

Maintaining the thread

A long video is more than a collection of individual moments. Understanding it requires following the thread between them. The same term can appear repeatedly, a speaker can revisit an earlier claim, and an example can depend on a definition given much earlier.

We treat continuity as a design priority. An answer should keep the context needed for the user's request, preserve distinctions that matter and make connections clear. Simply producing a longer response does not establish deeper understanding.

The request itself can stay natural: "Help me understand this recording" or "Give me the main ideas." The work is to organize the relevant material into an answer the user can follow, without requiring them to specify every section in advance.

Bringing distant moments together

Consider a recording that introduces a concept, develops it through an example and returns to it later. An explanation should show how those parts fit together. A qualification near the end should remain attached to the idea it changes, rather than disappearing from the summary.

This also means respecting sequence. Events that happen close together are not necessarily related, and a later statement may revise an earlier one. Connections should follow the source instead of being inferred from proximity alone.

Connecting context across a recording



Bring the relevant context together in the answer

Illustrative relationship between moments in a video. This is not a measured coverage result.

From understanding to useful work

Video Analysis organizes material from a recording around the user's request. Agent Chat gives users a conversational way to ask questions, develop explanations and turn relevant information into working notes. These are different ways of working with the same underlying need: an answer that remains connected to its source.

The useful output depends on the task. Study notes need a clear structure and enough explanation to stand on their own. A focused question needs the relevant context without an unnecessary retelling of the entire recording. A comparison needs to preserve the differences between the ideas being compared.

Source references help users return to the recording when precision matters. The aim is to make a long video easier to work with while keeping its meaning intact.

What dependable continuity requires

Continuity has several dimensions. Context fidelity keeps a statement connected to its original meaning. Temporal order preserves how a discussion or event develops. Multimodal interpretation distinguishes what is spoken from what is shown. Relevance determines which of those details belong in the answer.

These dimensions need to work together. A concise response can omit a qualification; a detailed response can still connect the wrong moments. We therefore treat completeness, source support and sequence as distinct design concerns, rather than assuming that one stands in for the others.

Longer duration alone is not evidence of stronger understanding. The useful question is whether the answer preserves the relationships needed for the task as the recording unfolds.

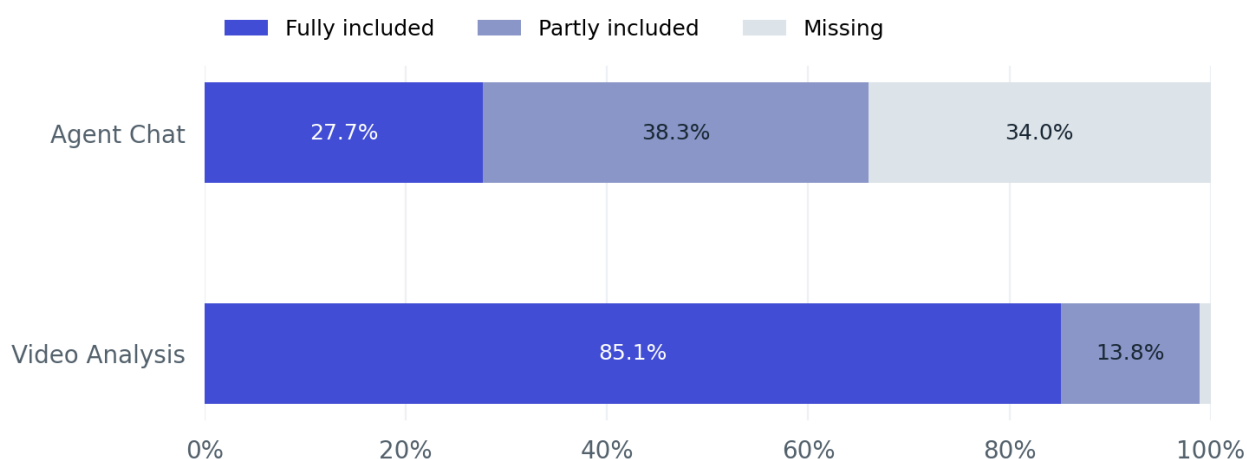
Measured answer coverage

A useful way to assess an answer is to check whether it includes the important points in a reviewed reference. Fully included means the point is covered correctly; partly included means some of it is covered; missing means it is absent. The primary coverage measure counts only fully included points, with partial and missing results shown separately.

In the reviewed pilot outputs, Video Analysis fully included 80 of 94 reference-point checks (85.1%). Agent Chat fully included 26 of 94 (27.7%). Both are assessed against the same reference checks on requests with a completed answer from each workflow.

This comparison shows how much of the checked reference reached the answer. It helps identify omissions; it is not a score for every factual statement the system might produce.

How much of the reference did each answer cover?



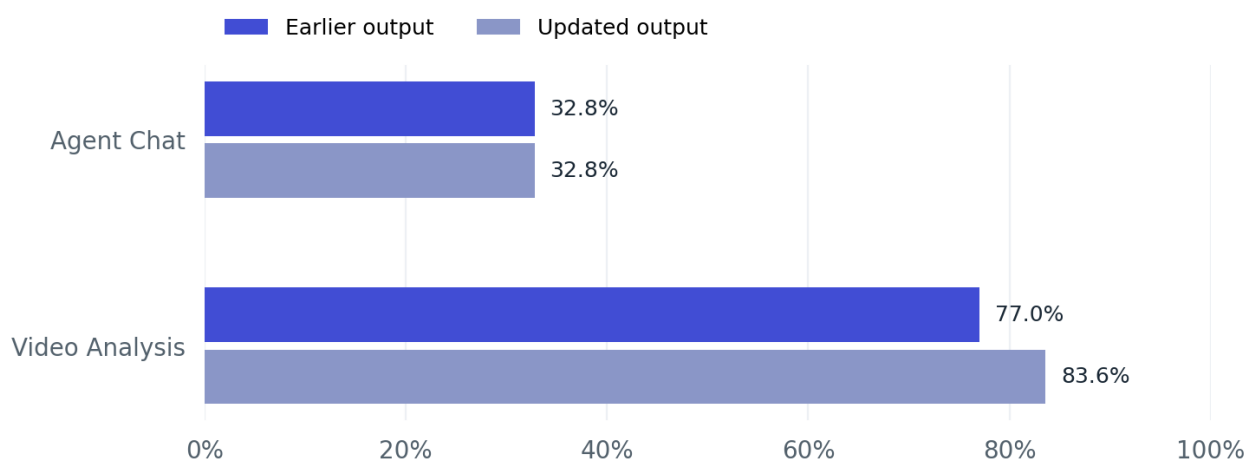
94 checked reference points per workflow, from the same completed, fully answerable requests. Fully included, partly included and missing are shown separately. Saved pilot outputs include revised answers.

Coverage across development

To understand progress, the comparison below keeps the requests and reference checks fixed before and after improvements. This avoids confusing a change in the sample with a change in coverage.

Video Analysis increased from 77.0% (47/61) to 83.6% (51/61). Agent Chat remained at 32.8% (20/61). These figures describe the reviewed development outputs on this matched set of requests.

Coverage before and after improvements



The same requests and 61 reference-point checks per workflow before and after improvements. This is a development comparison using these examples, not an independent held-out evaluation.

Research informing the approach

LongVideoBench frames long-video understanding around retrieving and reasoning over relevant details in extended video and subtitle inputs [1]. Video-MME emphasizes the range of information available through video frames, subtitles and audio [2]. Together, these works motivate attention to both context length and the type of evidence an answer needs.

Video-MME-v2 examines information aggregation, temporal dynamics and multimodal reasoning, including consistency across related questions [3]. Its framing is useful for thinking about continuity: finding an isolated detail and explaining how details relate are different challenges.

FActScore studies source support for individual claims in long-form text [4]. That perspective helps distinguish a useful explanation from one that merely sounds complete. The connection we draw from this research is a design principle: coverage of an idea and support for the statements about it should be considered separately.

Looking ahead

Our direction is to improve how Pelayar carries relevant context across longer recordings and turns that context into useful work. That includes clearer relationships between distant moments, stronger handling of changes in a discussion and better organization of explanations around the user's purpose.

The goal is straightforward: let people ask ordinary questions about substantial video material and receive answers that preserve the thread of what they watched.

About these results

A small development pilot with one reviewer and an assisted reference. Reference coverage measures inclusion of checked points, not overall factual accuracy, timestamp quality or performance as video duration increases.

The figures summarize saved development outputs, including revised answers. The continuity illustration is conceptual. The cited papers inform the approach; these results are not scores on their benchmarks.

References

- [1] Wu, H., Li, D., Chen, B., and Li, J. (2024). LongVideoBench: A Benchmark for Long-context Interleaved Video-Language Understanding. <https://arxiv.org/abs/2407.15754>
- [2] Fu, C., et al. (2024; revised 2025). Video-MME: The First-Ever Comprehensive Evaluation Benchmark of Multi-modal LLMs in Video Analysis. <https://arxiv.org/abs/2405.21075>
- [3] Fu, C., et al. (2026). Video-MME-v2: Towards the Next Stage in Benchmarks for Comprehensive Video Understanding. <https://arxiv.org/abs/2604.05015>
- [4] Min, S., et al. (2023). FActScore: Fine-grained Atomic Evaluation of Factual Precision in Long Form Text Generation. EMNLP. <https://aclanthology.org/2023.emnlp-main.741/>