HAMLET: HALLUCINATION ASSESSMENT IN LLMS FOR MULTI-PARTY, LONG-HORIZON, EVOLVING MEETING THREADS

Anonymous authors
Paper under double-blind review

ABSTRACT

Existing meeting benchmarks primarily evaluate models on a single meeting, treating each meeting as a self-contained record. This misses a key property of real-world meetings: facts and decisions evolve across a series of meetings for one project, where a proposal may be introduced, revised, and eventually adopted or remain unresolved. Hallucinations in this setting can mislead users and carry incorrect information into subsequent meetings, yet existing benchmarks provide limited analysis of such hallucination behaviour. We address these gaps with **HAMLET** (**H**allucination **A**ssessment in LLMs for **M**ulti-party, **L**ong-horizon, **E**volving meeting **T**hreads), which treats evolving discussion threads as the unit of evaluation. HAMLET extracts threads from AMI and ICSI meeting series and generates deterministic probes from their evolving states. Candidate answers are human-verified against the full series transcript, and a two-stage judge evaluates both reference-level correctness and unsupported content against the transcript. We further introduce HAMLET-Aug, which doubles thread length with synthetic on-topic discussion while preserving the thread state. Across 13 models, even frontier models find this setting challenging: the best model achieves 64.0% correctness. We find models can often track discussions to a resolution but tend to over-commit when no resolution exists, while many hallucinated details appear beyond the reference answer. On HAMLET-Aug, doubling thread length shifts hallucination towards additional content. These findings highlight the importance of measuring hallucination in evolving multi-meeting discussions.

1 INTRODUCTION

Meetings rarely stand alone. In many real-world settings, meetings run as a *series*, with the same team discussing the same matters over weeks. A useful meeting assistant therefore needs to track not only what was said in an individual meeting, but how the discussion evolves across meetings and avoid hallucinating when earlier statements are revised or left unresolved. Existing meeting benchmarks largely evaluate questions within individual meetings or merely treat questions as a retrieval problem: they test whether an answer can be found in the record, but not whether it reflects facts that may be proposed, revised, or decided over time (Thonet et al., 2025; Golany et al., 2024; Zhong et al., 2021). This leaves a fundamental gap in evaluating hallucination in cross-meeting discussions.

This shift from meetings as static documents to meetings as evolving processes introduces two coupled challenges. First, the assistant model must resolve *long-range dependencies*: evidence for a single question may be dispersed across tens of thousands of tokens and several meetings, yet still needs to be retrieved and combined correctly. Second, it must track *fact dynamics*: a fact can move from proposed to opposed, revised, and eventually decided, making the correct answer hiding in the discussion thread, rather than any individual span. For example, consider the question *"what did we decide about powering the remote?"* in the series we study. The correct answer is *watch-type batteries*. Yet the model may answer *kinetic charger*: a phrase that appears in the transcript, but only as a proposal in the first meeting, later revised before the thread reached a decision (Figure 1a). A retrieval-based check could therefore accept this answer as extractive, attributable, and citable, even though the retrieved phrase reflects an earlier state of the discussion rather than the final decision.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Existing meeting benchmarks also fall short as *measures* of hallucination. Some evaluations use coarse-grained factuality scores, such as 1–10 scores, making similar scores difficult to interpret and providing little insight into why a model failed (Hu et al., 2026). Some benchmarks also include unanswerable questions to test whether a model can recognize when the transcript does not include an answer rather than confidently hallucinating one, but usually these labels are assigned by rules. After auditing 100 sampled items from MeetingQA (Prasad et al., 2023) and MISeD (Golany et al., 2024), we find that most questions are in fact answerable (58% in MeetingQA, 34% in MISeD), suggesting that a substantial part of hallucination judgment in these sets may be label noise (Appendix B). Together, these limitations highlight the need to capture evolving discussions, construct questions systematically, and diagnose distinct hallucination modes.

Getting there raises two difficulties of its own. First, probe construction requires defensible gold answers, especially for unanswerable questions, and an objective way to select challenging probes. Second, probes must be diagnostic. They should reveal why models hallucinate and expose meaningful differences across models, rather than reducing performance to a single score.

We build **HAMLET** (Hallucination Assessment in LLMs for Multi-party, Long-horizon, Evolving meeting Threads) to address both challenges. It is built on the multi-party meeting series in AMI and ICSI (Carletta et al., 2005; Janin et al., 2003), and organised around cross-meeting discussion threads rather than individual meetings. This shift makes cross-meeting discussion the focus of evaluation, which existing meeting benchmarks do not capture. Each thread is represented as typed events (e.g., PROPOSE, SUPPORT, OPPOSE, REVISE, DECISION), from which deterministic question templates generate probes. Four probe types (**Retrieval**, **Absence**, **Contradiction**, and **Ambiguity**) target different failure modes (Figure 1b). An objective structural difficulty score controls probe selection, and every probe is double-annotated with evidence citations. We further provide **HAMLET-Aug**, a trajectory-preserving augmentation that lengthens discussion threads while preserving their underlying state transitions, enabling analysis of hallucination under extended contexts. Together, these designs make the evolving state of a discussion the object of hallucination evaluation, rather than evaluating each meeting in isolation.

Evaluating 13 models, we find substantial headroom: the best model answers 64.0% of probes correctly. More importantly, the benchmark reveals why they fail (section 4). Our contributions are:

- **Benchmark construction and release.** HAMLET turns evolving discussion threads into verifiable probes with controlled question generation and structural difficulty. We release human-verified probes over up-to-106k-token meeting series, together with annotations, gold answers, the evaluation protocol, and HAMLET-Aug for extended-context analysis [1] (§2).
- **Diagnostic evaluation.** Four probe types target distinct hallucination behaviours, evaluated with a two-stage protocol that checks both agreement with the reference answer and extra content in the answer against the transcript (§3).
- **Empirical findings.** Across 13 models, even frontier models show substantial difficulty on evolving multi-meeting discussions, with the best model only 64.0% correctness. We further analyse the failure modes revealed by HAMLET (§4).

2 THE HAMLET BENCHMARK

As shown in Figure 2, HAMLET is built from AMI and ICSI multi-party meeting series, where each series consists of meetings involving the same project. HAMLET extracts discussion threads from each series and constructs probes from their evolving states. The evaluated model receives the complete series transcript and a single probe question. We further augmented HAMLET by inserting state-consistent conversational content into the discussions, doubling their length to create **HAMLET-Aug** for controlled study of long-thread effects (§2.3).

2.1 THREADS, MOVES AND PROBE TYPES

Discussion thread extraction. Each meeting series concatenates 3–5 meetings (average 58.9k tokens; Appendix C). We extract discussion threads from each series. A thread is the discussion

[1] Data and code will be released upon publication.

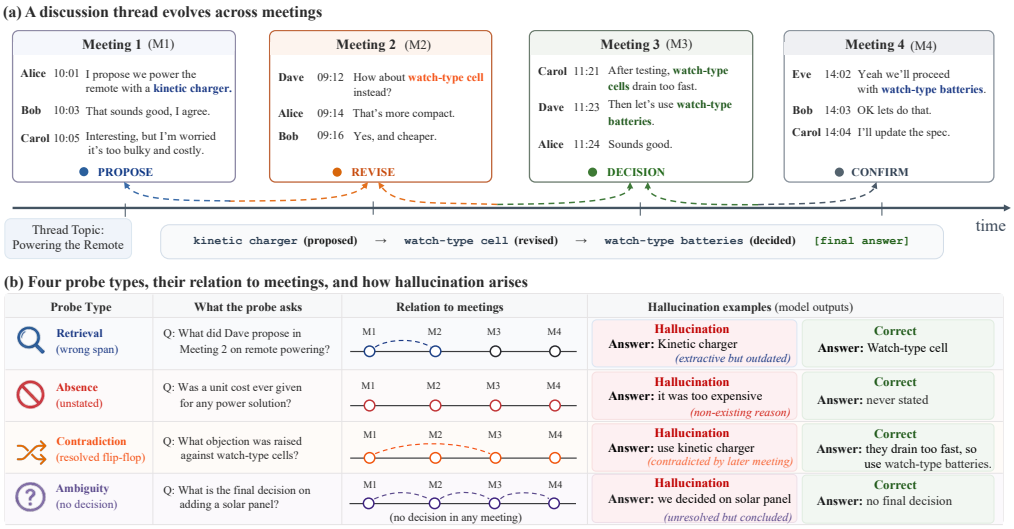


Figure 1: **HAMLET targets multi-party discussion threads that evolve across a meeting series.** (a) An example thread spanning four meetings, where *kinetic charger* is proposed in Meeting 1, revised to *watch-type cell* in Meeting 2, and decided as *watch-type batteries* in Meeting 3. (b) The four probe types in HAMLET, illustrated with their question, evidence, and example answers.

of a single matter, possibly across meetings. For example, in Figure 1, a thread may begin with a proposal in one meeting, be dropped, and then resume in a later meeting. Each contribution to a discussion thread is an *event*, anchored to an utterance and assigned a *move* type. The moves that change the thread state are, e.g., PROPOSE, SUPPORT, OPPOSE, REVISE, DECISION. Extraction is performed with GPT-5.4 (Appendix D), yielding 625 threads in HAMLET. This thread structure lets us track how a matter evolves across the series and infer its current state from the sequence of moves.

Probe types. To systematically probe hallucination in evolving meeting discussions, we design probes targeting distinct hallucination behaviours. We use four probe types. **Retrieval** tests whether a model can find and combine evidence across meetings and speakers. **Absence** tests whether it can reject a false presupposition. The question asks for a detail that the thread never establishes, so the correct answer should be "not determined." **Contradiction** tests whether a model can follow changes in a discussion. Earlier proposals may be revised or overturned, but the model must report correctly. **Ambiguity** tests whether a model can recognize an unresolved matter. When a thread has no decision, the model should report the open issue rather than conclude. We instantiate these probes with question templates. Each template is a question skeleton with slots filled from the thread. For example, *"What disagreement about {target} was left unresolved?"* An LLM then rephrases the question without changing its content. We use 16 templates in total, listed in Table 8, Appendix F.

2.2 Selection and Annotation

Difficulty-based selection. Filling the 16 templates across 625 threads yields 1,736 candidate probes, many of which are easy to answer because the underlying discussions are short, localized, and uncontested. We therefore select probes using an objective structural difficulty score computed solely from the thread and probe. There are five sub-scores, each normalized to $[0, 1]$, that measure the properties that make a probe harder: how many events a thread accumulates (*volume*), how widely it spreads across meetings and time (*dispersion*), how evenly its speakers and move types are represented (*distribution*), how far the probed event sits from the end of its meeting (*distance*), and how many superseded moves compete with the final state (*distractors*). Table 7 in Appendix E gives the formulas. Within each series$\times$template combination, we rank candidates by the difficulty score and retain the top two, yielding 175 probes balanced across series and templates.

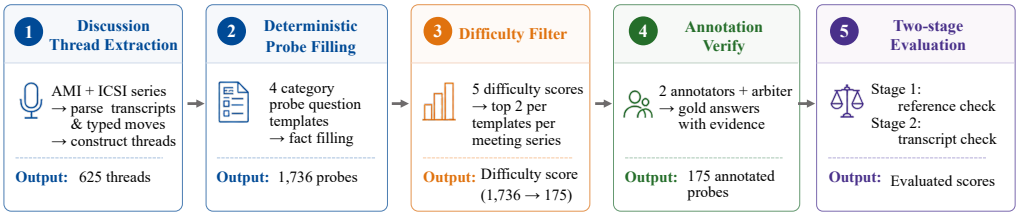


Figure 2: **Construction overview.** From thread extraction to question generation, probe selection, and human verification, followed by two-stage evaluation.

Double annotation with evidence citations. To reduce annotation burden while improving verification accuracy, we first use GPT-5.4 to generate a candidate answer for each probe, which annotators then independently verify against the full transcript series. Each annotator assigns an answerability label indicating whether the answer exists in the full transcript and provides utterance citations for the answer. Inner annotator agreement was 84.0% on the verdict and 89.7% on answerability (Gwet's AC1 0.786 / 0.878; Table 13). Disagreements are resolved by discussion with an additional annotator.

2.3 Trajectory-preserving augmentation

To further study the effect of longer discussions, we construct a trajectory-preserving augmentation of the longest discussion threads. We synthesize and insert filler events that do not change the thread state or trajectory. The transcript should therefore be extended while the correct answer remains unchanged. The same probes are reused on the augmented transcripts, enabling a controlled comparison with the original setting. We select the top 5% longest threads and insert events to extend each thread to 2× length. Outside the insertion windows, the transcripts are identical. This yields 104 probes with verified answers under the same annotation protocol as HAMLET (see Appendix G). We refer to the original set as **HAMLET** and the augmented set as **HAMLET-Aug**.

3 Evaluation Protocol

Free-form answers over meeting series can be long, making exact-match grading unsuitable and surface overlap insufficient. Evaluation therefore needs to assess both whether an answer covers the information in the reference and whether it introduces unsupported content. We therefore use a two-stage LLM judge (GPT-5.4, temperature 0; Figure 3): Stage 1 evaluates completeness and contradictions against the human-annotated reference, while Stage 2 checks additional content against the full transcript. We validate the judge's decisions against human labels.

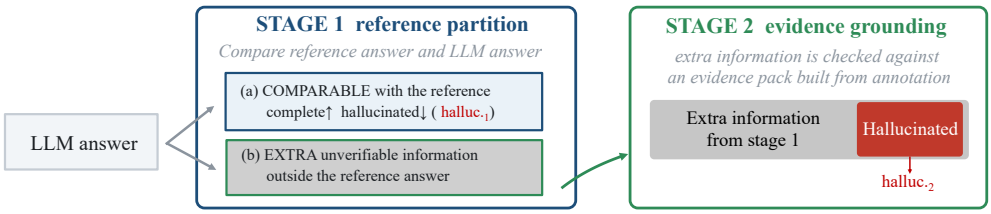


Figure 3: **Two-stage judging.** Stage 1 partitions a candidate against the human reference into *complete*, *hallucinated* and *extra*; stage 2 grounds the *extra* partition in the transcript evidence.

Stage 1: reference-based assessment. Stage 1 uses the human-verified reference answer as the evaluation target. The judge receives the question, reference answer, and model-generated answer, and independently labels *complete*, *hallucinated*, and *extra information* as binary dimensions, together with a rationale. *Complete* measures reference coverage (**compl.**), *hallucinated* measures contradictions to the reference (**halluc.**$_1$), and *extra information* identifies claims not covered by the reference (**extra**). The extra content is passed to Stage 2 for transcript-level verification.

Table 1: HAMLET results across 13 models. *compl.* ↑ measures reference answer coverage; *halluc.*$_1$ ↓ and *halluc.*$_2$ ↓ are stage-1 and stage-2 hallucination rates, respectively, with *extra* ↓ measuring content not exist in the reference. Overall-*halluc.* ↓ is the union of the two hallucination rates, and Overall-*correct* ↑ denotes complete answers with no hallucination at either stage. Best per column in bold.

		Stage-1			Stage-2	Overall	
Model	Family	compl. ↑	halluc.$_1$ ↓	extra ↓	halluc.$_2$ ↓	halluc. ↓	correct ↑
GPT-5.6-Sol	GPT-5.6	**.754**	.280	.777	**.051**	**.314**	**.640**
GPT-5.2	GPT-5.x	.686	.309	.886	.057	.337	.623
GPT-5.6-Terra	GPT-5.6	.709	.297	.800	.063	.349	.606
GPT-5.6-Luna	GPT-5.6	.640	.349	**.771**	.086	.394	.554
DeepSeek-V4-Flash	DeepSeek	.663	**.274**	.869	.120	.371	.549
GPT-5	GPT-5.x	.680	.349	.783	.131	.423	.531
GPT-5-mini	GPT-5.x	.577	.389	.834	.091	.440	.474
GPT-5.1	GPT-5.x	.691	.429	.914	.171	.503	.457
Qwen2.5-32B	Qwen	.411	.474	.914	.126	.520	.326
Qwen2.5-72B	Qwen	.394	.520	.897	.166	.594	.223
Qwen3-8B	Qwen	.331	.674	.903	.229	.749	.183
LongAlign-13B	Long-ctx	.177	.697	.926	.303	.794	.086
LongAlign-7B	Long-ctx	.177	.709	.909	.383	.806	.080

Stage 2: transcript evidence grounding. Stage 1 first identifies whether the LLM-generated answer contains extra information beyond the reference answer. Stage 2 then checks this against the relevant evidence in the transcript. We define **halluc.**$_2$ as the share of answers whose extra content contains information contradicted by the transcript. Internally, Stage 2 decomposes the extra content into claims and labels each as *supported*, *contradicted*, or *unverifiable*, with the human-cited transcript utterances. From the two stages, we also report two answer-level aggregates: an answer counts toward **Overall-halluc**. if either stage detects a contradiction, and toward **Overall-correct** only if it fully covers the reference and passes both stages.

Validation. We first manually verified the Stage 1 and Stage 2 judge results on a sample of 50 GPT-5.2 answers, obtaining 94% and 90% agreement with human labels, respectively. In addition, we compared the LLM-as-judge with traditional similarity metrics, including ROUGE (Lin, 2004) and BERTScore (Zhang et al., 2020). Among these metrics, ROUGE-L achieves the highest AUROC of 0.78 against the human-validated Stage 1 judge verdict (Appendix I), suggesting that lexical similarity is insufficient for reliably distinguishing hallucinated from non-hallucinated answers.

4 Experiments and Analysis

Setup. We evaluate 13 models spanning frontier closed-source models, smaller closed-source models, and open models designed for long-context processing (see Appendix H for the full model list and exclusion criteria). Each model receives the full series meeting transcript and generates answers at temperature 0 or using the same seed (see prompts in Appendix K). Our answering prompt is kept minimal, and we verify robustness to this choice by re-answering all probes using the answering prompts from three prior meeting benchmarks (Appendix H). For models with insufficient native context windows, such as Qwen and LongAlign models, we extend the context window to 128K tokens using RoPE scaling (Peng et al., 2024). Models with an explicit reasoning-effort control are evaluated at low, medium, high. Unless otherwise stated, we report results from the default setting, using medium reasoning effort when available.

4.1 Results on HAMLET

Overall results. Table 1 shows that HAMLET remains challenging across all models. The best overall correct rate is 64.0%, while the smaller open models with long-context capabilities remain near the bottom of the table. We also evaluate prior meeting benchmarks under our two-stage evaluation protocol (see Table 4 in Appendix B). HAMLET is substantially harder under the same evaluation.

Table 2: Per-category overall hallucination ↓ and correct rate ↑, defined as in Table 1. Hallucination is detected at either stage; correct answers are complete and hallucination-free at both stages.

	Retrieval		Absence		Contradiction		Ambiguity	
Model	halluc. ↓	corr. ↑	halluc. ↓	corr. ↑	halluc. ↓	corr. ↑	halluc. ↓	corr. ↑
GPT-5.6-Sol	**.321**	**.643**	.684	.263	.179	**.769**	**.242**	**.697**
GPT-5.2	**.321**	**.643**	.737	.263	.179	**.769**	.333	.606
GPT-5.6-Terra	.345	.607	.737	.263	**.154**	**.769**	.364	.606
GPT-5.6-Luna	**.321**	.583	.895	.105	.282	.718	.424	.545
DeepSeek-V4-Flash	.357	.536	.684	.316	.205	.718	.424	.515
GPT-5	.369	.560	.947	.053	.231	.744	.485	.485
GPT-5-mini	.381	.488	.895	.105	.205	.718	.606	.364
GPT-5.1	.524	.429	.842	.158	.333	.615	.455	.515
Qwen2.5-32B	.607	.238	**.474**	**.526**	.385	.462	.485	.273
Qwen2.5-72B	.548	.190	.684	.316	.615	.256	.636	.212
Qwen3-8B	.798	.155	.737	.263	.692	.205	.697	.182
LongAlign-13B	.917	.024	.684	.211	.744	.128	.606	.121
LongAlign-7B	.881	.036	.737	.158	.744	.103	.727	.121

The columns in Table 1 reveal where overall correctness is lost. *compl.* separates the model tiers: frontier systems cover the reference content on a majority of probes, whereas open models fall below 50%, indicating a substantial gap in retrieval and synthesis. Within the GPT family, however, hallucination largely determines the final ordering. GPT-5.1 achieves relatively high completeness but has the highest hallucination rate among the GPT systems. The *extra* column further shows that model answers frequently contain information absent from the reference, motivating transcript-level verification of this additional content. So stage 2 is necessary to capture hallucinations missed by the reference-based evaluation. For example, DeepSeek-V4-Flash has the lowest Stage 1 hallucination rate at 27.4%, but a Stage 2 rate of 12.0%, more than twice that of the best model. Overall correctness therefore reflects the joint effect of completeness and hallucination. Table 2 then examines how these failure patterns vary across probe types.

Finding 1: Models struggle to distinguish resolved decisions from unresolved information, especially under absence. Table 2 reports hallucination and correct rates across the four probe categories. Contradiction, where the thread changes direction but ends in a decision, is the easiest category for most models. Ambiguity uses similar threads and question forms but without an ending decision, yet its correct rate is substantially lower. The contrast suggests that models struggle to distinguish between information that has been resolved and information that remains unresolved: when a thread provides an explicit decision, models are more likely to produce a correct answer, but when no decision exists, they often commit to one anyway. Absence exposes this failure most directly. Its probes ask for details that the meetings never determined, making Absence the hardest category, with the highest hallucination rate and lowest completeness. Rather than recognizing that the requested information is unresolved, models tend to infer or fabricate a plausible answer.

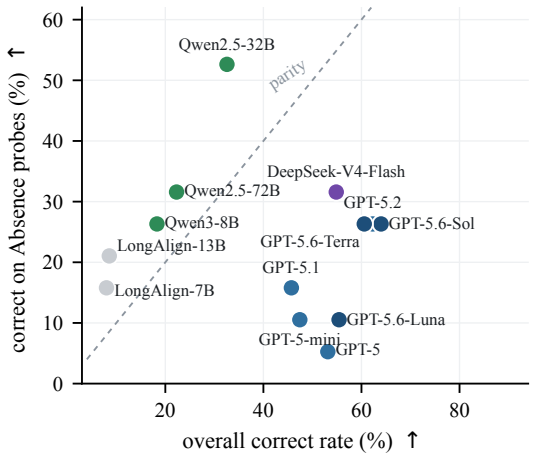


Figure 4: Overall correct rate vs. correct rate on Absence probes. The dashed diagonal indicates parity, where Absence accuracy equals overall accuracy.

Figure 4 shows Absence performance across models and reveals that more capable models tend to give an answer even when the meeting provides no evidence for one. A model's overall correct rate is statistically uncorrelated with its correct rate on Absence ($\rho = 0.02$, $p = 0.96$). The scatter also separates the families: GPT and DeepSeek configurations fall below the parity line, most losing more

than half of their overall correct rate on Absence, while Qwen models sit above it. Qwen2.5-32B, with half the overall correct rate of the strongest system (GPT-5.6-Sol), achieves twice its Absence correct rate. This suggests that helpfulness tuning biases models toward answering when evidence is insufficient. Greater capability does not overcome this bias. Thus, scaling the leaderboard does not necessarily scale honesty: models with higher overall performance may be less willing to abstain when the meeting leaves a question unresolved.

Finding 2: Reference-based judging misses fabricated information beyond the reference answer. Stage 2 grounds the parts of each answer that Stage 1 does not cover and detects fabrications. In the halluc.$_2$ column of Table 1, roughly one in twenty answers from the strongest model contains at least one claim contradicted by the transcript. The rate increases down the table, reaching 22.9% for the Qwen3-8B and 30–38% for the long-context baselines. Although the rate decreases with model capability, it reaches zero for no model. These errors usually arise from details that are not covered by the reference answer but are added by the models in their responses. Reference-only evaluation therefore misses this fabricated information.

4.2 HAMLET VERSUS HAMLET-AUG

As described in §2.3, we built HAMLET-Aug by inserting synthetic on-topic repeated discussions and re-asking 104 probes on the augmented transcripts. Comparing the two sets tests whether benchmark results remain consistent across transcripts and how longer discussion threads affect hallucination. Because the inserted discussion does not change any outside utterances, the gold answers should remain unchanged, and we verify them again using the same annotation process as in HAMLET.

Figure 5 decomposes each model's change across the two hallucination stage channels. The channels move in opposite directions: halluc.$_1$ falls for ten of the thirteen models, while halluc.$_2$ rises for ten. The decrease in halluc.$_1$ indicates that models are less likely to contradict the thread's probed state after the augmentation. The increase in halluc.$_2$ reveals a different effect. The inserted discussion does not resolve any question, but gives models more content from which to generate information beyond the reference answer. Models therefore produce more additional information and are more likely to be contradicted by the transcript. Thus, adding unresolved discussion does not uniformly hurt performance. It can make directly answerable information easier to recover while increasing hallucination in the additional content audited by Stage 2. We examine this shift further in §5.

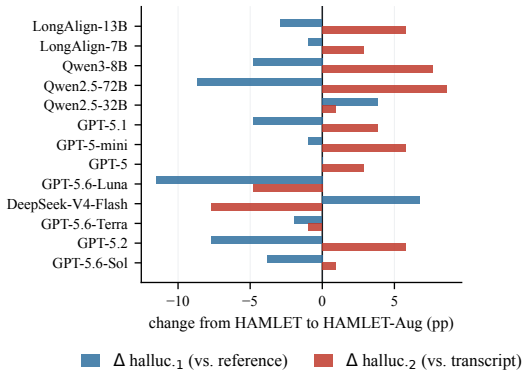
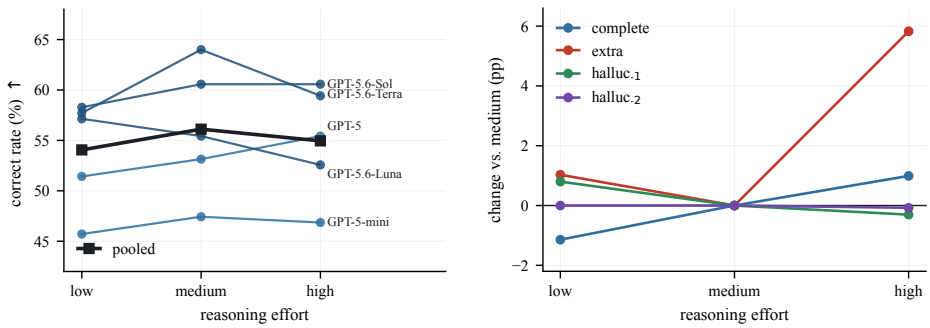


Figure 5: HAMLET-Aug comparison. Per-model changes from HAMLET to HAMLET-Aug on the 104 shared probes, decomposed into halluc.$_1$ (vs. reference) and halluc.$_2$ (vs. transcript), computed as the HAMLET-Aug rate minus the HAMLET rate (percentage points).

Finding 3: More reasoning effort adds answer content, not correctness; longer context shifts hallucination into extra content. Sweeping reasoning effort from low to high does not significantly improve overall correctness: the pooled correct rate is highest at medium effort (Figure 6a). The channel breakdown in Figure 6b shows that higher effort changes the composition of the answer rather than its correctness. From medium to high effort, *complete* increases, indicating more extensive coverage of the reference, and *extra* increases more, a much larger change in content beyond the reference. Meanwhile, both hallucination rates remain nearly unchanged. Thus, additional reasoning produces more elaborated answers, but most of the additional content falls outside what the reference answer covers and does not translate into higher correctness. This mirrors the limitation exposed by Stage 2: more answer content creates more opportunities for hallucinations to emerge, as models introduce additional claims that may not be supported by the discussion.



(a) Correct rate by effort level.

(b) Metrics change relative to medium effort.

Figure 6: Model performance across reasoning effort levels. (a) Correct rate across effort levels for the five models exposing the control. (b) Change relative to medium effort for each evaluation metric.

5 How Models Hallucinate: A Failure-Mode Study

When two models fail on the same probe, do they fail for the same reason? The findings above establish *where* the hallucination probe concentrates, particularly in unresolved threads and the unscored periphery. We now examine what these failures look like from two complementary directions. First, we fix the probes and vary the model: one probe from each category, answered by four models spanning the leaderboard, GPT-5.6-Sol, GPT-5-mini, DeepSeek-V4-Flash, and Qwen2.5-72B, and reviewed against the thread state (Appendix J). Second, we fix the model and vary the dataset: each answer on HAMLET is paired with the same model's answer to the same probe on HAMLET-Aug, allowing us to trace the hallucination rate difference.

Failure signatures of individual models. The dominant failure is shared: when the record leaves a question unresolved, models invent a determination that the discussion does not support. For example, asked why a participant backed a design choice that was never settled (Table 14, Appendix J), two models fabricate rationales and a third asserts that the idea was never raised. The remaining errors are model-specific. GPT-5.6-Sol misinterprets the status of the decision, placing a decision one meeting too early. DeepSeek-V4-Flash makes the most categorical unsupported claims, including definite negatives and invented disagreements, and often attributes content to the wrong speaker. GPT-5-mini stays within the correct thread but substitutes nearby content for the requested event. Qwen2.5-72B shifts the timeline as a whole, placing both the introduction and the adoption of a feature one meeting too early. Across all four models, the rejected answers are as specific and fluent as the accepted ones, and their errors are confined to the status and timing of decisions rather than to the topic under discussion. The answers therefore provide little surface evidence of these errors, and detecting them requires checking the answer against the state reached by the discussion.

Why halluc.$_1$ falls while halluc.$_2$ rises on HAMLET-Aug. HAMLET-Aug inserts discussion rounds that revisit an existing target without changing its decision, so the probed facts themselves are the same in both versions of the record. If halluc.$_1$ fell because the insertions make those facts easier to locate, the fall should be largest where the insertion touches the probe's evidence. Instead, when we split the probed answers by evidence locality (Figure 7), halluc.$_1$ falls only on probes whose evidence the insertion leaves untouched, and does not move where the insertion touches it. One possible explanation is that the reopened and still unsettled debate makes models more cautious about asserting a resolution anywhere in the record, while on touched probes this hesitancy is offset

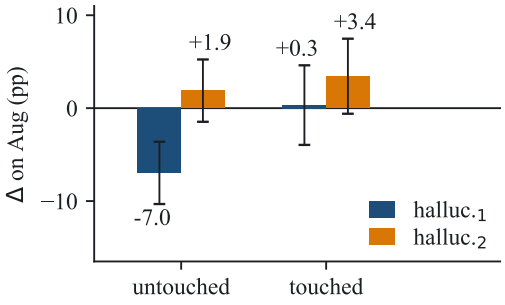


Figure 7: Change in the two hallucination rates from HAMLET to HAMLET-Aug over all models (error bars: 95% CIs), split by whether the synthetic insertion touches the probe's target.

by the back-and-forth discussion introduced around the probed evidence. Halluc.$_2$ is computed over answer content beyond what the question asks about, and it rises because models introduce more incorrect details in these parts on the augmented transcripts, including quantities, speaker attributions, and meeting indices, even when the relevant facts are identical in both versions (see Table 16). This increase is larger when the insertion overlaps with the probe's evidence, where the added discussion provides more similar and potentially confusable material. The same insertions therefore reduce unsupported resolutions in the core answer while increasing errors in its surrounding details. This distinction becomes visible only when answers are checked in our Stage 2 evaluation.

6 Related Work

Meeting question answering. Existing meeting benchmarks primarily evaluate question answering over individual meetings, including query-based summarisation (Zhong et al., 2021), extractive QA (Prasad et al., 2023), information-seeking dialogue (Golany et al., 2024), and long-context meeting QA (Thonet et al., 2025). More recent work such as MeetBench (Hu et al., 2026) evaluates meeting assistants over longer and more diverse tasks, but still largely treats each meeting or session as a self-contained unit. This is limiting for real-world projects, where relevant information may span multiple meetings and long temporal horizons, and where the meaning of an earlier statement can change as the discussion evolves. HAMLET instead evaluates evolving discussion threads across meeting series, with meeting series reaching up to 106k tokens.

Hallucination and faithfulness evaluation. General hallucination and faithfulness evaluation has been studied extensively across generation and source-grounded tasks (Ji et al., 2023; Maynez et al., 2020; Min et al., 2023; Es et al., 2024; Manakul et al., 2023). In meeting QA, however, hallucination is typically treated as a secondary aspect of answer quality rather than a diagnostic target. For example, MeetBench evaluates factuality on a 1–10 scale as one of several dimensions of meeting-assistant quality (Hu et al., 2026), while other meeting benchmarks primarily rely on reference answer matching (Prasad et al., 2023; Golany et al., 2024; Thonet et al., 2025). These approaches provide limited analysis of when and how models hallucinate, particularly when discussions span multiple long meetings and the status of information changes over time.

LLM judges. LLM-as-judge evaluation (Zheng et al., 2023; Liu et al., 2023) has become a common protocol for open-ended generation, with prior work assessing its reliability through agreement with human annotations (Tan et al., 2025). LLM judges have also been used to evaluate hallucination against human-annotated hallucinated content (Chuang et al., 2024; Tang et al., 2024). We follow this practice while restricting what the judge is asked to decide. Stage 1 compares the answer against a human-verified reference rather than requiring the judge to independently establish what is true. Content not covered by the reference proceeds to Stage 2, where it is checked piece by piece against retrieved transcript evidence. Both stages are validated against human annotations (Appendix I).

7 Conclusion

Meetings are evolving processes rather than self-contained records: what is proposed in one meeting may be revised, adopted, or left unresolved in the next. HAMLET turns this evolving state into a measurable hallucination evaluation task, with human-verified probes, controlled construction, and a two-stage evaluation that checks both reference-covered answers and additional claims against the transcript. Across 13 models, we find that models often recover decisions when a discussion reaches a resolution, but tend to over-commit when it does not. We also find that reference-based evaluation misses hallucinated details beyond the reference, while more reasoning or longer context does not necessarily make answers more reliable.

More broadly, these results suggest that long-horizon meeting understanding requires more than retrieving information from a long transcript: models must track how information changes and recognise when the discussion does not establish an answer. Future work can build on HAMLET and HAMLET-Aug to study longer and more diverse meeting trajectories, develop more reliable methods for evaluating unresolved information and unsupported additions, and investigate how models can better maintain the evolving state of a discussion over time.

AI USE STATEMENT

In this work, we used generative AI tools to generate synthetic data, reformat datasets, dataset meta info analysis, interpret the results and to implement an LLM-based evaluation component. We have not used generative AI tools to develop theoretical models or conceptual frameworks, formulate mathematical claims, propose or refine research hypotheses, design or provide feedback on the research methodology or experiments, assist with proofs. Additionally, we used generative AI tools to edit code and polish parts of the paper to improve readability. We have reviewed all AI-assisted work. We checked the consistency of data formatting, analysis results, experimental data and code, and compared the LLM judge with human judgements to assess their agreement. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work involves human annotation of model-generated answers and meeting transcripts for research evaluation. Annotators verified answers against source transcripts and identified unsupported or incorrect content. We use publicly available meeting datasets in accordance with their respective terms of use and do not collect new personal information or conduct interventions on human subjects. The synthetic discussion content introduced for HAMLET-Aug can be distinguished from the original meeting transcripts and is used only for controlled evaluation. We take care to preserve the underlying discussion state and verify the resulting benchmark instances before use. We take responsibility for the ethical use, documentation, and release of the resulting benchmark and evaluation materials.

REPRODUCIBILITY STATEMENT

We provide detailed descriptions of the benchmark construction, thread extraction and state representation, probe generation, human verification, two-stage evaluation protocol in the main paper and appendix. The experimental setup, model configurations, evaluation metrics, and analysis procedures are also documented to support reproduction of the reported results. The appendix further provides the judge validation, failure-mode analysis, and additional experimental details. We will release the benchmark data and evaluation materials, together with the code required to reproduce the reported analyses, subject to the terms of the underlying datasets.

REFERENCES

Jean Carletta, Simone Ashby, Sebastien Bourban, Mike Flynn, Mael Guillemot, Thomas Hain, Jaroslav Kadlec, Vasilis Karaiskos, Wessel Kraaij, Melissa Kronenthal, et al. The ami meeting corpus: A pre-announcement. In *International workshop on machine learning for multimodal interaction*, pp. 28–39. Springer, 2005.

Yung-Sung Chuang, Linlu Qiu, Cheng-Yu Hsieh, Ranjay Krishna, Yoon Kim, and James Glass. Lookback lens: Detecting and mitigating contextual hallucinations in large language models using only attention maps. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 1419–1436, 2024.

Shahul Es, Jithin James, Luis Espinosa Anke, and Steven Schockaert. Ragas: Automated evaluation of retrieval augmented generation. In *Proceedings of the 18th conference of the european chapter of the association for computational linguistics: system demonstrations*, pp. 150–158, 2024.

Lotem Golany, Filippo Galgani, Maya Mamo, Nimrod Parasol, Omer Vandsburger, Nadav Bar, and Ido Dagan. Efficient data generation for source-grounded information-seeking dialogs: A use case for meeting transcripts. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 1908–1925, 2024.

Kilem Li Gwet. Computing inter-rater reliability and its variance in the presence of high agreement. *British Journal of Mathematical and Statistical Psychology*, 61(1):29–48, 2008.

Yuelin Hu, Jun Xu, Bingcong Lu, Zhengxue Cheng, Hongwei Hu, Ronghua Wu, and Li Song. Meetbench-xl: Calibrated multi-dimensional evaluation and learned dual-policy agents for real-time meetings. *arXiv preprint arXiv:2602.03285*, 2026.

Adam Janin, Don Baron, Jane Edwards, Dan Ellis, David Gelbart, Nelson Morgan, Barbara Peskin, Thilo Pfau, Elizabeth Shriberg, Andreas Stolcke, et al. The icsi meeting corpus. In *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03).*, volume 1, pp. I–I. IEEE, 2003.

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.

Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, 2004.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-eval: Nlg evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pp. 2511–2522, 2023.

Potsawee Manakul, Adian Liusie, and Mark Gales. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pp. 9004–9017, 2023.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pp. 1906–1919, 2020.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pp. 12076–12100, 2023.

Bowen Peng, Jeffrey Quesnelle, Honglu Fan, and Enrico Shippole. Yarn: Efficient context window extension of large language models. In *International Conference on Learning Representations*, volume 2024, pp. 31932–31951, 2024.

Archiki Prasad, Trung Bui, Seunghyun Yoon, Hanieh Deilamsalehy, Franck Dernoncourt, and Mohit Bansal. Meetingqa: Extractive question-answering on meeting transcripts. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15000–15025, 2023.

Sijun Tan, Siyuan Zhuang, Kyle Montgomery, William Tang, Alejandro Cuadron, Chenguang Wang, Raluca Popa, and Ion Stoica. Judgebench: A benchmark for evaluating llm-based judges. In *International Conference on Learning Representations*, volume 2025, pp. 63277–63303, 2025.

Liyan Tang, Igor Shalyminov, Amy Wong, Jon Burnsky, Jake Vincent, Yu'an Yang, Siffi Singh, Song Feng, Hwanjun Song, Hang Su, et al. Tofueval: Evaluating hallucinations of llms on topic-focused dialogue summarization. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 4455–4480, 2024.

Thibaut Thonet, Laurent Besacier, and Jos Rozen. Elitr-bench: A meeting assistant benchmark for long-context language models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 407–428, 2025.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. BERTScore: Evaluating text generation with BERT. In *International Conference on Learning Representations (ICLR)*, 2020.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623, 2023.

Ming Zhong, Da Yin, Tao Yu, Ahmad Zaidi, Mutethia Mutuma, Rahul Jha, Ahmed Hassan, Asli Celikyilmaz, Yang Liu, Xipeng Qiu, et al. Qmsum: A new benchmark for query-based multi-domain meeting summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 5905–5921, 2021.

A NOTATION

Table 3 collects the symbols used in the main text.

Table 3: Notation used throughout the paper.

Symbol / term	Meaning
series	A group of meetings by the same team on the same project.
thread	Maximal set of utterances across a series concerning one decidable topic.
move	Typed label on a thread utterance: e.g. PROPOSE, SUPPORT, OPPOSE, REVISE, DECISION.
probe	(question, gold answer, probe type, template id, source thread).
complete	Stage-1 channel: candidate covers every element of the reference.
hallucinated	Stage-1 channel: candidate asserts something the reference contradicts.
extra	Stage-1 channel: candidate asserts content the reference does not mention.
halluc.$_1$	Share of answers with HALLUCINATED $= 1$ (stage 1).
halluc.$_2$	Share of answers whose extra content contradicts the transcript (stage 2).
overall halluc.	Answer-level union: halluc.$_1$ or halluc.$_2$ fires.
correct	COMPLETE with no hallucination at either stage; the headline metric.
supported / contradicted	Stage-2 labels of an extra claim against the transcript.
unverifiable	Stage-2 label: no transcript span decides the claim.
pp	Percentage points.

B REVIEW OF EXISTING BENCHMARKS

We evaluated four published benchmarks on meeting (MeetingQA (Prasad et al., 2023), MISeD (Golany et al., 2024), ELITR-Bench (Thonet et al., 2025), and MeetBench (Hu et al., 2026)) end to end with a single strong model (GPT-5.2), following each benchmark's official evaluation protocol. We then examined the results in two steps. First, for the two benchmarks with an explicit absence question subset (MeetingQA and MISeD), we manually audited the unanswerable labels against exactly the material available to the model. Second, we rescored the complete run for all four benchmarks, together with HAMLET, using our Stage-1 rubric, providing a common reference-based evaluation across datasets (Table 4).

MeetingQA (764 `is_impossible` items). We randomly sampled 50 items and manually audited them against the transcript input. Under the paper's own annotation rules, 58% of the unanswerable questions should have had an answer span. For example, the question *"So you have a meeting when?"* is answered directly by the next speaker: "we'll have a meeting, I think, it's in two days or something like that", yet the item is labelled unanswerable. GPT-5.2 answered 86% of the sampled items, and most of these answers are supported by the transcript under the corrected labels, further illustrating the impact of the original unanswerable-label errors.

MISeD (198 `isUnanswerable` items). On a sample of 50 items, manually audited against the full transcript, 34% of unanswerable items are in fact explicitly answerable. For example, for *"Did the participants mention any benchmarks for measuring the project's success?"* the gold answer asserts that the transcript mentions none, yet the project manager asks "you thought of some evaluation criteria?" and the marketing expert then lists them: *"fancy, innovative, easy to use, . . . "*. GPT-5.2 refused to answer 58% of the items whose unanswerable label was correct, and the refusal rate is 6% on incorrect labels. Thus, a hallucination rate computed from the raw unanswerable labels is difficult to interpret as a measure of model behaviour, because many of the underlying labels are inconsistent with the transcript.

ELITR-Bench and MeetBench. ELITR-Bench and MeetBench score answers against a reference rather than the transcript. ELITR-Bench uses a single 1–10 rubric and does not evaluate factual errors against the transcript. MeetBench also uses 1–10 scores for factual accuracy and other dimensions, but all scores are judged against the reference rather than the transcript.

Table 4 reports the same run scored with our Stage-1 rubric, placing the four benchmarks and HAMLET on a common evaluation scale. GPT-5.2 obtains a lower halluc.$_1$ on ELITR-Bench and MeetBench, while halluc.$_1$ is a bit higher on MeetingQA and MISeD with the original labels. However, the results on the latter two are difficult to interpret without accounting for the annotation issues identified above, particularly in their unanswerable subsets.

Table 4: GPT-5.2 (temperature 0) on the four existing meeting-QA benchmarks, all scored with the stage-1 rubric by the same judge (GPT-5.4). Greyed rows mix in the unanswerable subsets, whose labels the audit above found unreliable; the answerable rows are the meaningful comparison. The *extra* column is for information only and is not directly comparable across benchmarks because the length of the gold answer differs.

Benchmark	n	complete ↑	halluc.$_1$ ↓	extra
MeetBench (Hu et al., 2026)	93	.946	.118	.989
ELITR-Bench QA (Thonet et al., 2025)	130	.854	.138	.954
ELITR-Bench Conv (Thonet et al., 2025)	130	.839	.146	.939
MeetingQA, answerable (Prasad et al., 2023)	1712	.593	.064	.312
MeetingQA, all	2476	.453	.222	.387
MISeD, answerable (Golany et al., 2024)	430	.547	.240	.956
MISeD, all	628	.540	.315	.920

The conclusion we draw is not that these benchmarks are flawed for their intended purpose, but that labelling "there is no answer" over a long transcript is genuinely difficult, and can not be done by rule-based labelling, because asserting absence requires an exhaustive negative search. HAMLET instead derives its negative class from thread state and then verifies the derived labels with human annotators.

C Dataset Statistics

Table 5 summarises the scale of the test corpus: the seven held-out series comprise 28 meetings, and the per-series context (what a model receives for one probe) is reported in turns and in tokens.

Table 5: Test-corpus scale. Context size is per series, i.e. what a model sees for one probe.

Quantity	Value
Meetings	28
Turns	26k
Series	7
Turns per series (min–max)	1.38k–6.82k
Context tokens per series (min–max)	24.5k–106k (mean 58.9k)
Typed threads extracted	625
Typed events extracted	5.72k
Gold probes after difficulty filtering (HAMLET)	175
Reused probes on stretched transcripts (HAMLET-Aug)	104
Gold answer length in tokens (mean)	149

In addition to the test corpus, we release an unlabelled training split from AMI and ICSI meeting series. We apply the same extraction pipeline (Appendix D) and release the resulting typed threads, providing roughly six times as many threads as the test corpus (Table 6). The split contains no questions or gold answers and is intended for training and analysis.

D Thread Extraction

Threads are extracted per series with an LLM labeller (GPT-5.4) using 60-utterance windows with 12-utterance look-back to previous window. Because discussions can span meetings, the labeller maintains a *discussion registry* across meetings so revisited matters retain their original thread identifiers. Each thread has a *target* naming the matter under discussion, a short noun phrase such as

Table 6: Training-split scale. This split is released without questions or gold answers.

Quantity	Value
Meetings	218 (155 AMI / 63 ICSI)
Series	63 (40 with ≥ 2 meetings)
Turns	167k
Turns per series (min–max)	0.27k–38.3k
Context tokens per series (min–max)	4.2k–604k (mean 43.3k)
Typed threads extracted	3,810
Typed events extracted	40.0k

"remote control physical size". Each event records its meeting, utterance index, speaker, move label, and claims. There are nine move types: INTRODUCE, ASK, PROPOSE, SUPPORT, OPPOSE, REVISE, CLARIFY, DECISION, and ABANDON. Per-series statistics are stored with each fact sheet and will be released with the benchmark. Algorithm 1 summarises this procedure.

Algorithm 1 Discussion thread extraction

Require: meetings $M_1, \ldots, M_k$ of one series; window size $w = 60$; look-back $p = 12$; move set $\mathcal{A}$ = {INTRODUCE, ASK, PROPOSE, SUPPORT, OPPOSE, REVISE, CLARIFY, DECISION, ABANDON}
Ensure: registry R of threads; a thread is (identifier, target, event list); an event is (utterance index, speaker, target, move $\in \mathcal{A}$, claim)

$R \leftarrow \emptyset$
for $i = 1, \ldots, k$ **do**
  for each window W of w consecutive utterances of M_i **do**
    $C \leftarrow$ the p utterances before W
    $E \leftarrow$ LLM LABEL(R, C, W)
    for each event $e \in E$ **do**
      if a thread $t \in R$ already discusses the target of e **then**
        $\text{events}(t) \leftarrow \text{events}(t) \cup \{e\}$
      else
        $t' \leftarrow$ (fresh identifier, target of e, $\{e\}$)
        $R \leftarrow R \cup \{t'\}$
      end if
    end for
  end for
end for
return R

E DIFFICULTY SCORING

This section details the difficulty score of §2.2, and Table 7 defines the five sub-scores.

Constants. The normalisation constants in Table 7 are the raw values that map to a score of 0.5: $\text{med}(\#e) = 18$ events for *volume*. The entropy maxima are likewise fixed vocabulary sizes: $\log_2 x$ for x speaker roles and $\log_2 9$ for the 9 move types.

Aggregation. The first three sub-scores (volume, dispersion, distribution) describe the *thread*; the last two (distance, distractors) describe the *question*. A question probes an event of a thread, so it is scored from both angles. Its difficulty is the mean of the five sub-scores, dropping distance when the probe targets no evidence utterance (Absence probes).

Selection. Within each {meeting series}×{question template} cell, candidates are ranked by the difficulty score and the top two are kept, giving 175 probes, 22–27 per series. Ranking within cells rather than globally keeps the release set balanced over series and templates while still taking the hardest available instance of each question type.

Table 7: The five difficulty sub-scores. $\#M$ is the number of meetings in the series and $\#m$ the number the thread touches; $\#e$ is the thread's event count; $\#d = \#\text{PROPOSE}+\#\text{OPPOSE}+\#\text{REVISE}$ the number of internal decisive event moves; T the duration of the meeting, t the time of an event utterance in it, and $t_{\text{begin}}, t_{\text{end}}$ the times of the thread's first and last events; $\hat{H}$ is Shannon entropy normalized by its maximum; $\text{med}(\cdot)$ is the median of the feature over the thread.

Sub-score	Definition	Explanation
volume	$\frac{\#e}{\#e + \text{med}(\#e)}$	event count mapped to $[0, 1)$: 0.5 for a thread of median length, approaching 1 as the discussion grows;
dispersion	$\frac{1}{2}\left[\frac{\#m-1}{\#M-1} + \max_{meeting} \frac{t_{\text{end}} - t_{\text{begin}}}{T}\right]$	average of: across meetings, the share of the series' meetings the thread touches; within a meeting, the widest span of its events of a single meeting's duration; in [0, 1);
distribution	$\frac{1}{2}\left[\hat{H}(\text{speakers}) + \hat{H}(\text{moves})\right]$	average of the thread's events over speakers and over move types: in [0, 1), 0 when one participant and one move type dominates, 1 when uniform;
distance	$\frac{T - t}{T}$	the position of the probed event utterance in a meeting: in [0, 1), 0 at the meeting's end, near 1 at its start;
distractors	$\frac{\#d}{\#d + \text{med}(\#d)}$	count of superseded event moves of the probed event's target, mapped to $[0, 1)$;

F PROBE TEMPLATES

Table 8 lists the 16 released templates. Placeholders in braces are filled from thread fields; slots marked * are constructed at generation time and checked never to occur in the thread (Absence bait). After filling, an LLM rephrases the skeleton into a fluent participant-voiced question; the rephrasing may not add, drop or alter content.

G AUGMENTATION DETAILS

We construct HAMLET-Aug by locally stretching the longest discussion threads while preserving their original trajectory. For each selected thread, we first determine the target number of synthetic events as the full thread length. The synthetic events are inserted within one meeting, preferring a position before a DECISION, then before a REVISE or OPPOSE, and otherwise before a later event. A planner (GPT-5.4) generates each synthetic event containing a move and a claim. Then, role-conditioned speaker models realize these records as dialogue in the participants' voices. The generated turns are inserted at the selected position, with the timestamps of the original suffix shifted accordingly; all original utterances remain unchanged.

We validate the augmentation in three stages. First, deterministic checks verify the number of synthetic events, speaker and meeting identifiers, timestamp ordering, and the allowed move set ASK, PROPOSE, SUPPORT, OPPOSE, REVISE, CLARIFY. Synthetic content cannot contain a DECISION, ABANDON, or INTRODUCE move, or introduce a durable commitment, owner, requirement, or accepted final value. Second, an independent GPT-5.4 judge checks whether the augmentation preserves the original trajectory, avoids leakage from future events, faithfully realizes the planned events, preserves speaker and timing consistency, and reconnects coherently to the original next event. Third, we verify that the realized dialogue preserves the planned event order, count, speaker assignment, move, and claim, while leaving every original utterance unchanged.

We apply this procedure to 19 selected threads, corresponding to the 5% longest threads that contain probing targets, whose original lengths range from 40 to 190 events. The augmentation produces

Table 8: The 16 probe templates.

Cat.	ID	Question template
Retr.	R1	What did {speaker} propose about {target} in {meeting}?
Retr.	R2	Combining the points on {target}, what is the final specification?
Retr.	R3	Across all meetings, list every distinct proposal made about {target}, and mark which one was finally adopted.
Retr.	R4	How many distinct speakers contributed a proposal about {target} across the series?
Retr.	R5	In which meeting was {target} first introduced, and in which meeting was it finally decided?
Abs.	Ab1	What did the team decide about {aspect*} of {target}?
Abs.	Ab2	Why did {speaker} support {claim*} about {target}?
Abs.	Ab3	What was ultimately decided about {open question}?
Abs.	Ab4	What is the {type} of {target}?
Contr.	C1	What objection was raised against {proposed claim}, and how was it ultimately resolved?
Contr.	C2	What is the {type} of {target}?
Contr.	C3	During the discussion on {target}, {speaker} pushed back with the concern that {opposed claim}. How was the matter ultimately settled?
Contr.	C4	By the end, was everything about {target} fully settled, or were some parts of it still left open? If so, which parts?
Amb.	Am1	What disagreement about {target} was left unresolved?
Amb.	Am2	{open question}
Amb.	Am3	I'm writing up the minutes on {target}: what should I record as its final status?

1,359 synthetic events, yielding double length of the stretched threads. We then re-annotate the 104 reused probes under the HAMLET protocol to obtain the HAMLET-Aug gold labels.

HAMLET-Aug example. Table 9 shows an example stretched thread, discussion in the AMI series ES2004, whose target is the *remote-control anti-loss feature*. The original thread has 40 events spread over meetings ES2004b–d, 19 of them in ES2004c, so the target is 40 synthetic events. ES2004c contains no DECISION for this thread, so the insertion point falls before a REVISE, namely event ES2004c#103 at 30:10, where B revises the feature to voice recognition for finding the remote. The planner produces 40 events over the four speakers (12 ASK, 10 CLARIFY, 7 OPPOSE, 6 REVISE, 3 SUPPORT, 2 PROPOSE), each a move and a claim; its own exit-state summary reads: *"The inserted exchange keeps the anti-loss discussion open by questioning clip-on, bleep, TV-button, and speech-recognition variants without adopting any of them, adding no binding requirement or decision before the original thread continues."* The speaker models then realize each event in the participant's voice; for instance, the first planned event (B, ASK, *"Is the low-cost concern mainly about the speech-recognition part rather than the bleep itself?"*) becomes the turn *"Okay , um , is the low-cost issue mainly with the speech-recognition side of it rather than just having a bleep ?"*. Naturalization adds 62 further turns (17 hesitations, 24 self-restarts, 18 backchannels, 3 seam fillers) to the 40 event turns and 8 backchannels planned by the realizer, giving an inserted segment of 110 turns that spans 236 seconds; the meeting grows from 521 to 631 utterances, and the original suffix resumes with its timestamps shifted by that amount. Because every inserted move only re-asks, re-explains or re-contests options that were already on the table, the thread state at ES2004c#103 is the same in both transcripts, and the 104 reused probes keep their answers.

H Experimental Setup Details

Systems. GPT models (GPT-5.6-Sol, GPT-5.6-Terra, GPT-5.6-Luna, GPT-5.2, GPT-5.1, GPT-5, GPT-5-mini) are called through API; Qwen2.5-72B, Qwen2.5-32B, Qwen3-8B and LongAlign-7B/13B run with vLLM on 4×A100, extended to a 128k window with the rope scaling each model supports. Every system receives the full series transcript and a single question in one prompt, identical across systems, at temperature 0 where the API exposes it. DeepSeek-V4-Flash is called through API with thinking mode on at its default effort. Models with a reasoning-effort control (GPT-5.6-Sol/Terra/Luna, GPT-5, GPT-5-mini) are run at low, medium and high effort.

Original ES2004c prefix (unchanged). The last two thread events before the insertion point: A raises speech recognition or a bleep for finding the remote, subject to cost.

27:50 **A** [ES2004C#95, REVISE]: which mentions that our division has developed a new speech recognition s feature . We have to check into the for financial f feasibility whether we can incorporate it this at a low cost , but for like we had g um s response from the customers that they would like to have the feature of finding it . So it could be like , where is the remote , and the remote answers I am here . Some kind of thing or it gives a b bleep sound or some kind of sound and if this can be incorporated this would be more uh you can say trendy also ? And technologically innovative also .
28:10 **C**: Yeah .
28:28 **A** [ES2004C#96, CLARIFY]: So we we can check into the financial feasibility of incorporating this .

· · · 21 original turns on other topics (28:36–30:09) · · ·

Inserted synthetic segment (110 turns, 30:11–34:06). Tagged turns realize the 40 planned events; untagged turns are backchannels, hesitations and self-restarts added by naturalization.

30:11 **B** [ASK]: Okay , um , is the low-cost issue mainly with the speech-recognition side of it rather than just having a bleep ?
30:11 **A**: Okay .
30:15 **C**: Yeah .
30:17 **A** [CLARIFY]: Yeah , the cost concern was for putting in the fuller feature , not just any sound by itself .
30:23 **D**: Uh .
30:24 **D** [OPPOSE]: Mm , but even a simpler sound is still adding something people in the groups didn't directly ask for .
30:29 **C**: Um .
30:30 **C** [ASK]: Yeah , but did they actually reject finding a lost remote , or was it just the attachment idea they didn't mention ?
30:34 **C**: Yeah .
30:35 **B**: Mm-hmm .
30:36 **D**: It was the ,
30:37 **D** [CLARIFY]: It was the clip-on bit I was objecting to , not every possible way of locating the remote .
30:42 **B**: Okay so ,
30:43 **B** [REVISE]: Okay , so the attachment might need to stay separate from the sound-finding idea then .
30:47 **A**: Mm-hmm that ,
30:48 **A** [SUPPORT]: Mm-hmm , that fits , like , the earlier concern that carrying it on a person may not suit shared use .
30:54 **C**: Well
30:55 **C** [ASK]: Yeah , if it mostly stays in the room though , does a clip-on help with the sofa problem at all ?

· · · remaining synthetic turns omitted · · ·

Original ES2004c suffix (unchanged, shifted by $+236.4$ s). The thread resumes at its original next event, ES2004c#103 (originally at 30:10), and continues exactly as in the source transcript.

34:07 **B** [ES2004C#103, REVISE]: Yeah . Um , that means that there's no function for li the port , you know that it sits in , then pressing the button then having so we could incorporate voice recognition for the finding it . It's a bi i it's like a g it's a gadget , it's a
34:07 **C**: Yeah .

· · · remaining original turns omitted · · ·

Table 9: A real HAMLET-Aug insertion: discussion in AMI series ES2004 (*remote-control anti-loss feature*). The segment is inserted into ES2004c before a REVISE event. Gray rows are unchanged original utterances, tagged with their thread event where applicable.

Table 10: The full result of both judging stages on HAMLET (175 probes) and HAMLET-Aug (104 probes). Rows are ordered by HAMLET correct rate.

		HAMLET (175 probes)				**HAMLET-Aug (104 probes)**			
Model	Family	compl. ↑	halluc.$_1$ ↓	halluc.$_2$ ↓	extra ↓	compl. ↑	halluc.$_1$ ↓	halluc.$_2$ ↓	extra ↓
GPT-5.6-Sol	GPT-5.6	.754	.280	.051	.777	.702	.279	.087	.865
GPT-5.2	GPT-5.x	.686	.309	.057	.886	.663	.269	.115	.904
GPT-5.6-Terra	GPT-5.6	.709	.297	.063	.800	.702	.308	.067	.913
GPT-5.6-Luna	GPT-5.6	.640	.349	.086	.771	.721	.279	.029	.894
DeepSeek-V4-Flash	DeepSeek	.663	.274	.120	.869	.615	.356	.077	.865
GPT-5	GPT-5.x	.680	.349	.131	.783	.683	.365	.125	.875
GPT-5-mini	GPT-5.x	.577	.389	.091	.834	.596	.442	.154	.837
GPT-5.1	GPT-5.x	.691	.429	.171	.914	.673	.413	.212	.923
Qwen2.5-32B	Qwen	.411	.474	.126	.914	.375	.510	.154	.952
Qwen2.5-72B	Qwen	.394	.520	.166	.897	.462	.452	.240	.894
Qwen3-8B	Qwen	.331	.674	.229	.903	.413	.596	.288	.923
LongAlign-13B	Long-ctx	.177	.697	.303	.926	.192	.644	.356	.952
LongAlign-7B	Long-ctx	.177	.709	.383	.909	.183	.673	.337	.885

Table 10 reports the full result, with the per-channel scores of both judging stages on HAMLET and on HAMLET-Aug.

LLM-as-judge. Both stages use GPT-5.4 at temperature 0. Both stage prompts are released in Appendix K. Zero judge calls failed or returned unparseable output.

Robustness to the answering prompt. Our answering prompt is kept as minimal as possible (§K). To assess sensitivity to prompt choice, we generate all 175 HAMLET probes with GPT-5-mini (medium effort, the configuration used in Table 1) using the answering prompts from three prior meeting benchmarks, and score all runs with the same two-stage judge. As Table 11 shows, the effect is small: Overall-*correct* changes by at most 1.1 pp across the four prompts (paired t-test over the shared probes, $p \geq .76$), and no prompt produces a significant change in either hallucination rate.

Table 11: Prompt robustness: GPT-5-mini (medium) on HAMLET under four answering prompts, scored by the unchanged two-stage judge.

Answering prompt	compl. ↑	halluc.$_1$ ↓	extra ↓	halluc.$_2$ ↓	Overall-halluc. ↓	Overall-correct ↑
HAMLET (ours)	.577	.389	.834	.091	.440	.474
ELITR-Bench	.640	.406	.914	.177	.480	.463
MeetBench	.697	.354	.994	.177	.486	.469
MISeD	.617	.360	.817	.091	.429	.463

I Judge Validation Details

Human validation of LLM-as-judge. We sampled 50 GPT-5.2 answers, balanced by the judge's own faithfulness call (25 judged faithful and 25 judged unfaithful) and stratified by probe category. For each item, a human annotator reviewed the question, gold reference, model answer, and the judge's decision together with its cited inaccuracies, and marked whether the LLM-as-judge was correct. The judge agreed with the human on 94.0% of Stage 1 and 90% of Stage 2.

Traditional metric proxy. We also test whether traditional similarity metrics (Lin, 2004; Zhang et al., 2020) can replace the LLM-as-judge. Table 12 reports AUC and average precision for predicting Stage-1 verdicts. ROUGE-L performs best, but remains insufficient for reliable per-answer scoring. Traditional metrics can capture system-level ordering but not the per-answer judgements or Stage-2 grounding.

Annotation scheme. Each probe was independently labelled by two annotators, who saw the question, the GPT-5.4 candidate answer, and the full series transcript. For each item, they recorded question *answerability* (yes/no), a *verdict* on the candidate answer (correct/incorrect), and a free-text

Table 12: Traditional metrics as proxies for the "overall correct" verdict.

Metric	item-level AUC ↑	average precision ↑
ROUGE-L	.777	.789
ROUGE-1	.770	.802
ROUGE-2	.762	.773
BERTScore	.736	.763

reason citing supporting utterances. On average, annotating one question took around 20 minutes per annotator. An example of the annotation interface is shown in Figure 9. Disagreements were resolved through discussion with an additional annotator, and the arbitrated labels and corrected answer form the released gold. Table 13 reports agreement on the two categorical fields. Both fields are imbalanced between the two labels, we report raw agreement alongside Gwet's AC1 (Gwet, 2008) in Table 13.

Table 13: Inter-annotator agreement. Gwet's AC1 is the chance-corrected statistic appropriate under the high label prevalence of both fields (Gwet, 2008).

Set	Field	n	raw agreement	AC1
HAMLET	Verdict	175	.840	.786
HAMLET	Answerability	175	.897	.878
HAMLET-Aug	Verdict	104	.860	.819
HAMLET-Aug	Answerability	104	.860	.826

J Qualitative Analysis

Failure-mode audit. For each of the four case-study models, GPT-5.6-Sol, DeepSeek-V4-Flash, GPT-5-mini, and Qwen2.5-72B, we gather all HAMLET answers flagged by the two-stage judge and sampled 20 answers per model. Each sampled answer was read in full against the gold reference, the stage-1 and stage-2 verdicts and rationale, and assigned one primary error type.

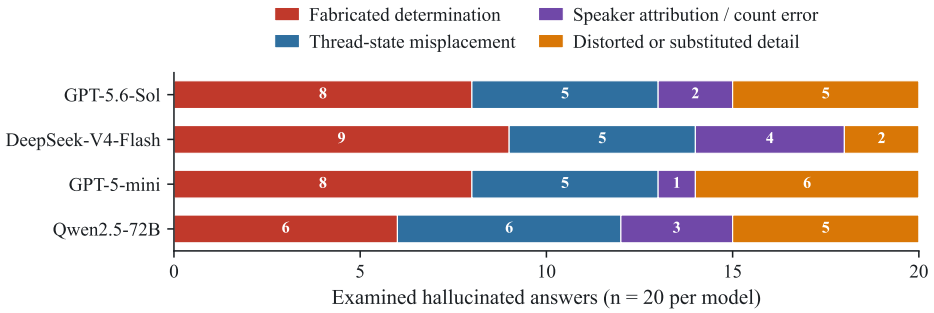


Figure 8: **What kind of error each system makes when it hallucinates.** Primary error type of 20 manually audited judge-flagged answers per system, sampled stratified by probe category.

The four error types we find in the answers are: *fabricated determination*, where the record does not establish a decision, or records agreement, but the answer supplies a decision, rationale, objection, or open disagreement; *thread-state misplacement*, where the discussion is real but its state is incorrect, such as assigning a decision to the wrong meeting, treating a superseded state as final, or flipping settled and open states; *speaker attribution / count error*, where the content is real but the "who" is wrong, such as assigning content to the wrong speaker or miscounting proposers; and *distorted or substituted detail*, where the "what" is wrong within a correctly located thread, such as inventing or altering lists, numbers, or options, or substituting adjacent material for the probed event.

Figure 8 shows the resulting distribution. *Fabricated determination* is the leading failure mode for all four systems. The second-ranked error varies by model: DeepSeek-V4-Flash shows more speaker-attribution errors and invented disputes; GPT-5-mini more often produces distorted or substituted details; Qwen2.5-72B more often misplaces the discussion state across meetings; and GPT-5.6-Sol shows a more even distribution, including flips between settled and unresolved states.

Case study. Table 14 fixes one probe per category and shows the answers of four systems: GPT-5.6-Sol, DeepSeek-V4-Flash, GPT-5-mini, and Qwen2.5-72B. Table 15 and Table 16 illustrate two mechanisms behind the halluc.$_2$ increase on HAMLET-Aug: the first fixes one probe and one system and compares the original and stretched transcripts; the second fixes one probe and shows three systems inventing a detail in the stretched transcript that none of them asserted on the original. §5 discusses both mechanisms at the model level, while §4.2 gives the aggregate picture. Here, we walk through the cases by probe type.

The errors differ by probe type. In Case 1, GPT-5.6-Sol keeps the meetings but misplaces the *decision* within them: it moves the decision one meeting earlier. Qwen2.5-72B instead shifts the meetings themselves. Thus, even when models retrieve the relevant part of the thread, they may still misidentify how the discussion evolved within it.

Case 2 exposes a different failure. The record does not determine an answer, yet three systems answer anyway, and they do so in mutually incompatible ways. GPT-5.6-Sol and GPT-5-mini invent different rationales for the supposed support, while DeepSeek-V4-Flash makes the opposite claim that the idea was never raised at all. Only the weakest of the four preserves the uncertainty in the record. This mirrors the aggregate pattern in Figure 4: stronger capability does not necessarily lead to abstention when the evidence is unresolved.

Two further patterns appear in the remaining cases. The first is *certainty inflation*. In Case 3, every failing system correctly identifies the discussion but changes its modality, turning an experiment described as optional ("if there's time") into a definite decision to try it. DeepSeek-V4-Flash is the only system that preserves the tentative state. The second is *displacement*. In Case 4, the question asks which disagreement remained open, but the two weaker systems answer with a different open issue from the same discussion. They substitute more salient nearby content for the specific event being probed.

Case 1 (Retrieval): placing a decision in the thread's history. *Which meeting first introduced the remote control speech recognition feature, and which meeting was it finally decided in?*	
Reference. First introduced in IS1009c by the user-interface designer; finally decided in IS1009d, where speech recognition was confirmed as the special feature for the product.	**GPT-5.6-Sol** The speech recognition feature was first introduced in IS1009c [. . .] and it was also selected as the remote's proposed special feature later in that same meeting.
	DeepSeek-V4-Flash First introduced in IS1009c by Francina, the User Interface Designer [. . .] finally confirmed in IS1009d, when the prototype was presented as having an inbuilt voice recognizer [. . .]
	GPT-5-mini The speech-recognition idea was first introduced in meeting IS1009c [. . .] It was finally decided/confirmed and built into the prototype in meeting IS1009d.
	Qwen2.5-72B First introduced in the second meeting (IS1009b) [. . .] It was finally decided in the third meeting (IS1009c) that speech recognition would be a special feature pending further research on its cost.
Case 2 (Absence): a support that was never explained. *Why did me003 support the idea that the candidate hidden variables for the belief-net middle layer should include a separate hidden node for whether the user is lost?*	
Reference. Not determined in meeting.	**GPT-5.6-Sol** me003 thought "lost" was an independent latent state: a user could be a tourist or on business and still be lost [. . .] A separate hidden node could therefore combine cues from discourse, context, and prosody [. . .]
	DeepSeek-V4-Flash [. . .] me003 supports adding a hidden node for "whether the user is lost" [. . .]
	GPT-5-mini Because being lost is an independent situational factor that strongly changes what kind of answer people want [. . .] me003 argued that it therefore should be represented as its own hidden node [. . .]
	Qwen2.5-72B In the meeting, me003 did not explicitly state a support for including a separate hidden node for whether the user is lost. However, the idea [. . .] could be inferred from the broader discussion [. . .]
Case 3 (Ambiguity): a possibility mistaken for a plan. *What did we decide to do about variance scaling of the two HTK input streams?*	
Reference. No final decision. Hari's suggestion to scale the variances so the streams get unequal weights was noted as easy to do without retraining, but left as an optional extra experiment "if there's time."	**GPT-5.6-Sol** We decided to try unequal weighting of the two HTK input streams by scaling their diagonal variances [. . .] time permitting [. . .]
	DeepSeek-V4-Flash No firm decision was made [. . .] it was left as an open possibility/extra experiment rather than something they committed to doing.
	GPT-5-mini We agreed to try Hari's variance-scaling trick [. . .] It was treated as a cheap/easy experiment to try [. . .]
	Qwen2.5-72B We decided to consider variance scaling [. . .] We agreed to try this as an experiment at the level of HTK recognition, and if it shows promise, we could further explore it.
Case 4 (Contradiction): which disagreement was the open one. *What disagreement about the candidate features for inferred user intentions was still left unresolved?*	
Reference. Whether to include an explicit feature for enterability of the destination—one practical feature, a split into "has door" vs. "public/accessible," or handled through object type. The distinction was not settled.	**GPT-5.6-Sol** They had not resolved whether to model physical enterability ("has a door") separately from practical or permitted accessibility ("is public") [. . .]
	DeepSeek-V4-Flash [. . .] left unsettled [. . .] whether related attributes like [. . .] "enterable/public-has-a-door" should be collapsed or kept separate [. . .]
	GPT-5-mini The group was still debating whether some observed cues [. . .] should point straight at the Mode (View/Enter/Approach) or only affect it indirectly via intermediate hidden nodes [. . .]
	Qwen2.5-72B There was a discussion about whether certain features, such as "discussed admission fee," should point directly to the mode [. . .] or to an intermediate node [. . .]

Table 14: Four probes, one per category, answered by four systems at the default settings of Table 1. Green blocks are answers the two-stage judge accepts; red blocks are answers it rejects, with the offending spans highlighted. Answers are abridged; bracketed ellipses are ours. The full answers and judge outputs are part of the released data.

HAMLET vs. HAMLET-Aug: the same answer, re-scored by a longer record. *Which meeting did we first introduce the transcription planning for the meeting data project in, and which meeting was it finally decided in?*

Reference. Transcription planning was introduced in Bmr001 and finally decided in Bmr003.

On HAMLET Transcription planning was first introduced in **Bmr001** and was finally decided in **Bmr003**, where the group settled on using linguistics students rather than a commercial service, with Jane likely coordinating. *(stage 2: claim unverifiable)*

On HAMLET-Aug Transcription planning was first introduced in **Bmr001**. The approach was ultimately decided in **Bmr003**, favoring linguistics students over commercial transcription services, with Jane potentially coordinating the work. *(stage 2: claim contradicted)*

Table 15: One probe and one model (GPT-5.6-Sol, medium) on the original and stretched transcripts. The main answer is consistent across both, and Stage 1 accepts both. The amber-highlighted claim is accepted against the original transcript in HAMLET but contradicted by the HAMLET-Aug transcript, revealing an error captured by Stage 2.

HAMLET vs. HAMLET-Aug: a detail invented only on the longer record. *What did we decide about which meetings to record next and who would take part in them?*

Transcript: "fe016: So there's probably — there's **three to four a week**, / me011: That's what we're aiming for."

GPT-5 (on HAMLET-Aug) [. . .] Frequency/scale: aim for a meeting or two per week to start [. . .]

GPT-5-mini (on HAMLET-Aug) [. . .] make it easy/spontaneous so they can record a few meetings per week (a target of ∼2–3/week was discussed) [. . .]

DeepSeek-V4-Flash (on HAMLET-Aug) [. . .] Adam said they should aim for one or two recordings a week, starting with the Speech group until the hardware and software were reliable [. . .]

Table 16: One probe and three models on HAMLET and HAMLET-Aug. The recording-rate target is stated identically in both transcript versions. On HAMLET, none of the three models mentions a rate and none is flagged by Stage 2. But on HAMLET-Aug, all three volunteer a rate, each wrong, and each answer is flagged.

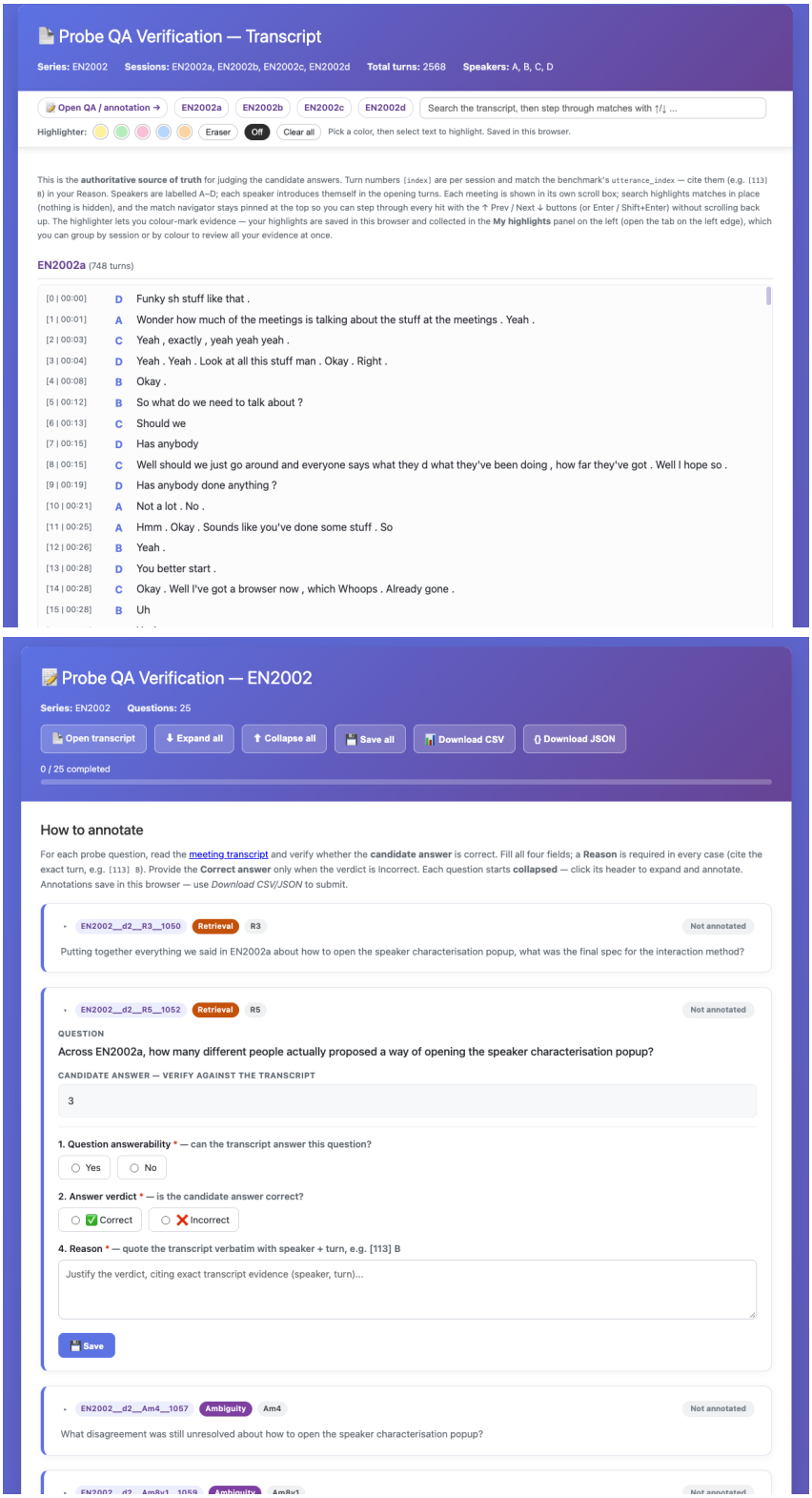


Figure 9: The annotation interface. Top: the transcript viewer, which renders a full meeting series with utterance indices for locating and checking cited evidence. Bottom: the verification panel, which shows one probe with its candidate answer and collects the answerability label, the verdict, and the cited reason.

K PROMPTS

This section provides the prompts used in this paper. The answering prompt is shared across all evaluated systems and provides each system with the full series transcript and a single question in one message (Section H). The two judge prompts implement the two-stage protocol described in Section 3 (Figure 3); each prompt consists of a system message followed by a user template, with `placeholders` filled at run time.

Answering Prompt (all evaluated systems)

```
You are a meeting assistant of the meeting series below. Answer the question based on the
provided meeting transcripts.
---
{transcript}
Question: {question}
Answer:
```

Stage-1 Judge Prompt (reference partition)

```
You compare a CANDIDATE ANSWER to a REFERENCE ANSWER (the ground truth) for a QUESTION, at
the level of the MAIN IDEA. Judge two things, each strictly true or false: COMPLETENESS
(recall -- does the candidate cover the reference's main idea?) and FAITHFULNESS /
GROUNDEDNESS (precision -- of the candidate's claims that can be compared with or inferred
from the reference, do none conflict with it?). Split the candidate's content into two
parts: (a) claims that can be compared with or inferred from the reference -- these are
the ones you judge for faithfulness; and (b) EXTRA information -- specific numbers or
claims that can neither be compared with nor inferred from the reference (unverifiable) --
these are only listed, never penalised. Parts (a) and (b) together make up the whole
candidate answer. Judge ONLY against the reference, not outside knowledge. Return STRICT
JSON, no prose.
---
### QUESTION
{question}
### REFERENCE ANSWER (ground truth)
{gold}
### CANDIDATE ANSWER (to evaluate)
{answer}
### TASK
Assess the candidate against the reference at the level of the main idea. First split the
candidate's content into (a) claims that can be compared with or inferred from the
reference and
(b) extra, unverifiable claims; then return STRICT JSON:
{
  "complete": true | false,
  "missing": ["<main idea of the reference the candidate fails to convey>", ...],
  "faithful": true | false,
  "conflict": ["<candidate claim comparable with the reference that conflicts with it>",
...],
  "extra": true | false,
  "extra_info": ["<specific number or claim that can neither be compared with nor
inferred from the reference>", ...],
  "rationale": "<=2 sentences",
  "confidence": "high" | "medium" | "low"
}
Rules:
- COMPLETENESS (recall). "complete": true if the candidate conveys the MAIN IDEA of the
reference needed to answer the question; false only if it omits a main necessary part of
that main idea.
- "missing": list the reference's main ideas (needed to answer the question) that the
candidate fails to convey; [] if none.
- FAITHFULNESS / GROUNDEDNESS (precision). Consider ONLY the candidate's claims that can
be compared with, or inferred from, the reference. "faithful": true if all such claims
are consistent with the reference (NONE conflicts with it); false if any of them conflicts
with the reference on the main idea.
- "conflict": list candidate claims that can be compared with the reference and that
conflict with it; [] if none.
- EXTRA / UNVERIFIABLE. "extra": true if the candidate states specific content (for
example a number or a named claim) that can neither be compared with nor inferred from the
reference; false otherwise.
- "extra_info": list those specific, unverifiable numbers or claims; [] if none. Do NOT
count them as inaccuracies -- they are simply unverifiable against the reference, not
wrong.
```

```
- Together, the claims you judge under FAITHFULNESS plus the claims you list under EXTRA
 should account for the candidate's content.
- Focus on the main idea; ignore style, length, and ordering.
- "confidence": how sure you are -- "high", "medium", or "low".
```

Stage-2 Judge Prompt (transcript grounding)

```
You verify claims against a meeting transcript. You are given an excerpt of the
TRANSCRIPT of a meeting, a QUESTION about that meeting, a candidate ANSWER to it, and part
of CLAIMS taken from that answer. For each claim decide whether the transcript excerpt
SUPPORTS it, CONTRADICTS it, or leaves it UNVERIFIABLE. Return STRICT JSON, no prose.
--------------------------------------------------------------------------------
### TRANSCRIPT EXCERPT
{evidence}
### QUESTION
{question}
### AI ANSWER
{answer}
### CLAIMS TO VERIFY
{claims}
### TASK
Judge each numbered claim against the transcript excerpt above and return STRICT JSON:
{
  "results": [
    {
      "id": <the claim's number>,
      "verdict": "supported" | "contradicted" | "unverifiable",
      "evidence": "<short quote from the excerpt, or \"\" if unverifiable>",
      "rationale": "<=1 sentence"
    }
  ]
}
Rules:
- "supported": the excerpt states the claim or directly implies it.
- "contradicted": the excerpt states something that cannot both be true with the claim.
- "unverifiable": the excerpt neither confirms nor denies it. The excerpt is only part of
the meeting, so a claim that is simply not covered is unverifiable, not contradicted.
- Quote the excerpt in "evidence" whenever you support or contradict a claim.
- Judge each claim on its own, and return exactly one result per claim, keeping the given
ids.
- Base every judgement on the excerpt alone, never on outside knowledge.
```