SIEVE: Slot-wise Evaluation of Information-Role Retention in Fixed-Length Meeting Summarization

Jumpei Nagasawa
Nestrium LLC
jumpei.nagasawa@nestrium.net

Seiya Shibayama
Laboro.AI Inc.
shibayamaseiya@outlook.jp

Abstract

Existing meeting-summary evaluations mainly ask whether a summary resembles a reference or is consistent with the source. They do not directly show which information roles needed for decision understanding are retained or dropped.

We propose SIEVE, a method that evaluates meeting summaries from seven information-role perspectives. For each role, SIEVE separately measures Presence, whether the role appears in the summary, and Fact Check, whether the retained content is free of contradictions with the source. This produces a role-level retention profile, allowing summarizers and length limits to be compared by what they preserve and lose rather than by a single aggregate score.

We apply SIEVE to 2,250 summaries from 50 MeetingBank municipal-meeting segments, generated by three models under 2-, 3-, and 6-sentence limits. The role-level profiles show that summaries more reliably retain topic, outcome, and action-detail information, while more often dropping background and concerns that shaped the discussion, as well as unresolved items left open afterward. Reference-similarity metrics widely used to evaluate MeetingBank summaries (ROUGE-L, BERTScore) correlate weakly or negatively with SIEVE Presence. Consistent with this, the dataset-provided reference summaries against which similarity is computed contain topic and action-detail information but few of the other roles. Reference similarity and the role retention that SIEVE measures are different perspectives on summary quality. SIEVE therefore supports auditing, selecting, and designing meeting summarizers based on which decision-relevant information survives compression.

1 Introduction

Meeting summarization is not mere compression. In decision-making meetings, what matters is not only what was on the agenda and what was decided, but also why it was put on the agenda, which concerns were contested, and what remained unresolved. If a short summary states "the budget was approved" or "the item was referred to committee" but drops this background, contestation, and unresolved tail, much of the meeting's meaning is lost. Moreover, in LLM agents and multi-step pipelines, meeting summaries are consumed as intermediate outputs by downstream processing; under a constrained budget, which information roles survive compression therefore affects not just human readability but the output quality of the entire LLM workflow.

Existing summary evaluation focuses mostly on similarity to a reference summary or on factual consistency. ROUGE-L and BERTScore measure how close a generated summary is to the reference; they do not directly evaluate how much of each source-side information role is retained. Factuality metrics check whether a summary's statements are consistent with the source, but which roles were dropped is outside their scope.

We propose **SIEVE (Slot-wise Information Evaluation with Verifiable Evidence)**, which decomposes a meeting segment into seven information roles — Topic, Outcome, Background, Operative Payload, Concrete Concern, Shared Concern, Unresolved Tail — and evaluates how much of each role is retained in a generated summary. Instead of a single score, SIEVE produces a per-role retention profile, allowing us to disentangle which aspects of a meeting short summaries preserve and which they drop.

We apply SIEVE to 2,250 summaries generated from 50 municipal-council meeting segments from MeetingBank. Empirically, we observe summarization failure patterns that are hard to see in a single score or in reference-similarity metrics: the retention gap between closure-bearing roles and deliberative context, the structural difficulty of recover-

Findings of the Association for Computational Linguistics: IJCNLP-AACL 2026, pages 3671–3699
November 6-10, 2026 ©2026 Association for Computational Linguistics

ing Unresolved Tail, and the interpretive boundary across evaluators on Background.

2 Related Work

Existing summary evaluation metrics. Existing summarization evaluation broadly falls into three families. (1) Reference-similarity metrics such as ROUGE, BERTScore, and embedding similarity measure the similarity between the generated summary and a reference summary (Lin, 2004; Zhang et al., 2020). (2) Factuality / faithfulness metrics such as FactCC, QAGS, FEQA, QAFactEval, SummaC, and AlignScore assess whether a summary's claims are supported by, or consistent with, the source (Kryscinski et al., 2020; Wang et al., 2020; Durmus et al., 2020; Fabbri et al., 2022; Laban et al., 2022; Zha et al., 2023). (3) Multi-dimensional evaluation such as G-Eval, UniEval, and FineSurE decomposes summary quality into general dimensions such as factuality, coherence, relevance, and conciseness (Fabbri et al., 2021; Zhong et al., 2022; Liu et al., 2023a; Song et al., 2024). Among these, FineSurE additionally evaluates keyfact-based completeness — whether the summary covers a list of keyfacts provided by humans or derived from the reference summary (Song et al., 2024). What these families share is the lack of a framework that evaluates, with the role as the unit, how much of each source-side information role is retained in the summary. Meeting summarization in particular is still evaluated mainly through reference similarity on benchmarks such as MeetingBank (Hu et al., 2023) and QMSum (Zhong et al., 2021), and, to our knowledge, no role-wise evaluation layer aligned with decision-making structure has been established. Even in meeting-specific studies, established automatic metrics have shown weak-to-moderate correlations with observable summarization errors, including cases of error masking (Kirstein et al., 2024). Note that we do not run factuality metrics as an experimental baseline against SIEVE; the claim here is a difference of measurement target: factuality metrics measure claim consistency, not role retention.

Content-unit evaluation. Content-unit approaches include Pyramid and the ACU protocol used in the RoSE benchmark (Nenkova and Passonneau, 2004; Liu et al., 2023b); they decompose a summary into content units and evaluate unit-wise. SIEVE also has a unit-wise structure, but differs in two respects: its unit is not a reference-derived atomic fact but one of seven meeting-specific information roles constituting the source-side decision structure, and it separates Presence (role retention) from slot-local Fact Check as independent dimensions.

3 Method: SIEVE

3.1 Overview

SIEVE is a diagnostic framework that separates the source-side description of roles from the summary-side judgment of role retention. For each segment, we first fix the source-side content of the seven information roles as an **event sheet**, and from this we construct a segment-specific slot-local rubric (Topic Gate / Quality Gate and Fact Check criteria). We then judge Presence and Fact Check for each generated summary. Canonical labels combine dual-judge agreement on concordant cells with human adjudication of all disagreement cells, grounded in the fixed SIEVE rubric; the two LLM judges are drawn from a model family disjoint from the generators.

3.2 Data

Data source. We use the train / validation / test splits of MeetingBank (Hu et al., 2023) — a city-council meeting summarization benchmark — at the original segment-transcript level. We impose a one-meeting-one-sample constraint that does not take multiple segments from the same meeting: items discussed in the same meeting share vocabulary and institutional context and are difficult to treat as independent samples.

Sample. We selected 50 segments by stratified sampling on decision type (25 decision-bearing, 25 non-decision), distributed across six public bodies (Alameda, Boston, Denver, King County, Long Beach, Seattle). Eligibility thresholds, body cap, and sampling details are in Appendix A.

3.3 Summary generation

Input. The input to the generator is the full segment transcript only; the dataset-provided reference summary, the event sheet, and metadata are not given at generation time.

Prompt and generators. The prompt is framed as a briefing to a policy analyst — a realistic use case in which a downstream reader needs to act on or follow up on a meeting and therefore needs not only what was decided but also what shaped the discussion and what remains open. It instructs

the model to summarize "what happened, what shaped the discussion, and what remained open afterward," with a specified sentence count (N = 2, 3, 6) and a plain-prose constraint (no bullets or headings); the full prompt is in Appendix A. The framing also avoids the under-specification of a naive summarization instruction. Because the prompt already requests what shaped the discussion and what remained open, the under-retention of deliberative roles is hard to read as an artifact of an experimental probe; it persists even under a realistic, role-aware summarization request. We used `deepseek-chat`, `deepseek-reasoner`, and `gemini-2.5-flash-lite` as generators.

Scale. The sentence budget has three levels: 2, 3, and 6 sentences. For each (segment $\times$ generator $\times$ budget) we generated 5 runs independently, for a total of $50 \times 3 \times 3 \times 5 = 2{,}250$ summaries and 15,750 slot cells.

3.4 Information-role slots

The information needed to understand a decision-making meeting corresponds to the following structural questions: what was discussed (T), how it was handled (O), what content was attached to the action (N), why this item came up in its current form (B), what individual / shared concerns were raised (L, S), and what remained unresolved (U). These are not reference-derived atomic content units but functional roles in the proceedings of a decision-making meeting.

As a concrete example of the L / S distinction: concerns voiced in individual speeches — three council members each objecting to different budget items, for instance — are captured under L (per-speech granularity), while a problem-side theme that recurs across multiple speakers — "lack of transparency in the budget process" repeatedly raised by different speakers, for instance — is captured under S (cross-speaker pattern). L and S are each a single slot per segment, judged with one Presence value; the distinction lies in what kind of source content is read into each slot, not in how many instances appear in the summary.

The seven slots are treated not as a comprehensive ontology of meetings but as a **minimal schema** for understanding decision-making. In each segment's event sheet, all seven slots are instantiated, including cases where a role appears as the absence of action, a deferral, or a non-finality (for example, in a non-decision segment, O is instantiated as "no vote / carried over to the next session"). This keeps the denominator for slot means fixed at all segments $\times$ 7 slots.

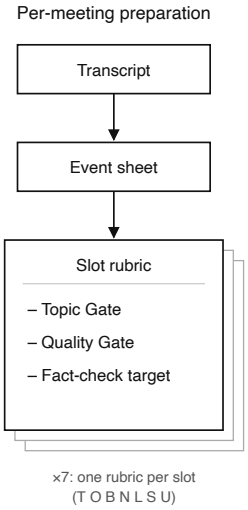


Figure 1: Per-meeting preparation. The event sheet and the seven slot-local rubrics are authored from the transcript only.

3.5 Event sheet

Rationale. To evaluate role retention, what each role concretely refers to in the source must be fixed. If this source-side interpretation is done anew at evaluation time, evaluators effectively re-summarize the source as part of evaluation, source-side interpretation itself wobbles across evaluators and runs, and the summary-side judgment is dragged along with it. SIEVE fixes the source-side role description once as an **event sheet**, and subsequent Presence / Fact Check judgments are made against the event sheet as a fixed reference. This is the central design choice of SIEVE.

The event sheet is a structured source-side reference (not a scoring rule) constructed from each segment's transcript (Figure 1); for each slot it records source-side content, transcript-level evidence, temporal information, source context for Fact Check, and related fields under a fixed authoring protocol. Field structure, treatment of temporal information, and authoring protocol are in Appendix C.

3.6 Presence: Topic Gate and Quality Gate

Two-gate structure. The Presence of each slot is determined by composing two gates (Figure 2). The **Topic Gate** asks whether the summary contains a statement corresponding to the slot's information role, and the **Quality Gate** asks whether —

Slot	Role	Definition and key design notes
T	Topic/Framing	substantive subject, policy direction, target of the agenda item — main framing axis
O	Outcome/Position	body's disposition (vote, approval, referral, etc.) on the acted-on item — main decision axis
N	Operative Payload	condition, number, requirement, scope, date attached to the action — action-near content paired with O
B	Agenda-Setting Background	prior event / condition / process / problem / opposition / mediation / revision / failure / deadline visibly connected to the current item; an explicit "why now" statement is not required
L	Concrete Concern	speech-level concrete concern (harm, obstacle, difficulty, loss, risk, tradeoff, constraint); a named place / group / area is not a requirement for Presence
S	Shared Concern Synthesis	recurring problem-side concern theme spanning multiple speakers (distinction from L: L is per-speech, S is a multi-speaker pattern)
U	Unresolved Tail	pending future handling, later review, unresolved implementation, explicit return path; its core is unfinishedness itself, not mere future orientation — an unrelated future item or a post-adjournment announcement does not satisfy U

Table 1: The seven information-role slots used by SIEVE.

given that the Topic Gate has passed — that statement carries enough specificity to convey the slot's content. Presence takes three values:

- 0.0: Topic Gate fail
- 0.5: Topic Gate pass + Quality Gate fail
- 1.0: Topic Gate pass + Quality Gate pass

0.5 is not an independent semantic label; it is a mechanical notation for the state "the role is expressed but lacks specificity."

Independence of Presence and Fact Check. Presence judges only **whether the slot's facet is expressed in the summary**; the source-correctness of the stated value is handled separately by Fact Check (Section 3.7). For example, if a summary stating a change in zoning classification cites the wrong code, Presence still passes and the error is recorded as a Fact Check error. For cells that pass the Topic Gate, Presence and Fact Check separately record role expression and source contradiction. Topic-Gate failures are recorded as non-retention and receive no Fact Check.

Each slot's gates are instantiated against the source-side slot description fixed in the event sheet. The table below indicates roughly what each gate asks per slot; complete operational definitions and examples are in Appendix B.1.

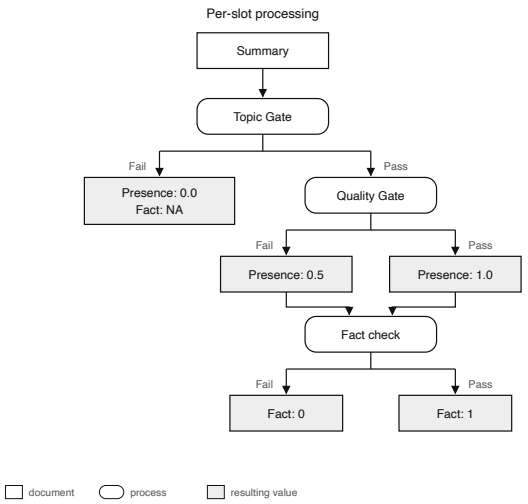


Figure 2: Per-slot processing of one summary. Presence is determined by the Topic Gate and the Quality Gate; Fact Check runs only for cells with Presence > 0 and yields 1, 0, or NA.

3.7 Fact Check

For each cell with Presence > 0, Fact Check asks whether any claim in the summary pertaining to that slot contradicts the source. Only contradictions count as errors; omissions and less-specific paraphrases do not, and claims about other slots' content are out of scope. The check fields per slot

Slot	Topic Gate (is the role expressed?)	Quality Gate (enough to convey slot content?)
T	identification of the substantive subject (procedural label alone is insufficient)	specification of policy direction + target
O	body's disposition on the acted-on item	acted-on item is identifiable
N	payload attached to the action / acted-on item	at least one of number / named requirement / scope boundary / date cutoff is identifiable
B	prior event / condition / process / problem connected to the current item	the connection between background and current item / form / handling is visible
L	concrete speaker-level concern / harm / risk / tradeoff	concern content is specific (no named place / group required)
S	problem-side concern theme spanning multiple speakers (L is per-speech, S is multi-speaker recurring)	the shared concern theme has identifiable content
U	pendingness marker (mere future orientation is insufficient)	identifiability of the pending item + visibility of the handling path

Table 2: Slot-local Topic Gate and Quality Gate criteria for Presence.

and typical contradiction patterns are in Appendix B.

3.8 Annotation and adjudication

Annotation framework. Canonical labels are obtained by taking dual-judge agreement on cells where the two LLM judges agree (with a 5–10% manual sample audit), and by full human adjudication of all disagreement cells (15.1% = 2,379 cells with Presence disagreement) by the first author. Cells where the judges agreed on Presence but disagreed on Fact Check were also human-adjudicated; in total, human adjudication covered 19.0% of all cells (2,987). The first author adjudicates referring only to the fixed SIEVE rubric, the slot-local rubric, the event sheet, and the raw transcript, without referring to the LLM judges' outputs (labels or rationales). LLM-as-judge is scalable but carries known biases (Zheng et al., 2023; Wang et al., 2024); we therefore use two LLM judges (GPT-5, Claude Sonnet 4.6) drawn from a model family disjoint from the generators (DeepSeek $\times$ 2, Gemini 2.5 Flash Lite) as a triage-and-proposal signal, not as independent final evaluators. The LLM judge configuration is detailed in Appendix A.

Calibration check. To check that the main labels are not strongly dependent on the first author's threshold, a second author blindly re-evaluated 200 Presence cells (a slot-stratified random subset of 100 plus a boundary-heavy diagnostic hard subset of 100). The second author was given only the fixed rubric, the slot-local rubric, and an event-sheet excerpt; the first author's final labels and the LLM judges' outputs were not provided. This is an audit by two annotators, not a large-scale multi-annotator IAA study, and we therefore treat it as a **calibration check and boundary check** rather than as a population-level reliability estimate (reported in Section 4.6).

Reference summary evaluation. The 50 reference summaries (350 cells) were also evaluated with the same event sheets and gate definitions as the generated summaries. The first author labeled all cells blind to the judge outputs; the exact Presence agreement between this first pass and the two LLM judges was 85.4% (GPT-5) and 86.9% (Claude Sonnet 4.6). The 46 cells (13.1%) on which the first pass disagreed with both judges on Presence, or with both on Fact Check, were finalized by the first author after re-reading the source text. The first author is a non-native English speaker and used machine translation as a reading aid; all judgments and evidence checks were made against the English originals.

Reproducibility artifacts. The full SIEVE rubric, event-sheet authoring protocol, rubric-construction / validation protocols, judge prompts, event sheets and slot-local rubrics for all 50 segments, and analysis code are released in the public repository (https://github.com/tanuking/sieve-paper-artifacts).

4 Results

All 95% confidence intervals reported in this section come from meeting-level bootstrap (10,000 iterations, with segment as the resampling unit to reflect the multi-level structure; Efron, 1979). **Mean Presence** denotes the per-summary average Presence across the seven slots, and **Total Presence** denotes their sum.

4.1 Role-specific retention

Applying SIEVE to the 2,250 summaries reveals a clear per-role compression pattern. Topic, Outcome, and Operative Payload form a high-retention closure-bearing group, while Background, Concrete Concern, and Shared Concern are lower-retention deliberative roles, and Unresolved Tail is a lower-retention unresolved role. The broad contrast mean(T, O, N) - mean(B, L, S, U) is 0.315 [0.269, 0.364]. Because B is threshold-sensitive (Section 4.6) and U is weakly correlated with B/L/S (Appendix E.6), we treat this as a retention contrast rather than a correlation cluster: the same point-estimate contrast holds for deliberative context alone (mean(T, O, N) - mean(B, L, S) = 0.290) and is even larger for U alone (mean(T, O, N) - U = 0.390). We interpret the result as this contrast rather than as a fine-grained ranking of the seven slots.

Slot	Mean Presence	95% CI	Fact Error
T	0.946	[0.923, 0.966]	0.038
O	0.911	[0.870, 0.944]	0.197
N	0.902	[0.871, 0.930]	0.123
L	0.661	[0.590, 0.726]	0.024
B	0.634	[0.554, 0.710]	0.050
S	0.593	[0.518, 0.663]	0.026
U	0.530	[0.453, 0.606]	0.092

Table 3: Slot-level Presence and Fact Error rates.

Fact Error is the rate of slot-local contradictions among cells with Presence > 0 (Fact Check is not applied to cells with Presence = 0). Action-near slots such as O (0.197) and N (0.123) have higher Fact Error rates, suggesting that retaining the role and correctly retaining the attached procedural / scalar detail are distinct failure modes. Conversely, the low Fact Error rates for background and concern slots should not be read as high factual precision: under a contradiction-only criterion, vague background or concern statements are less likely to trigger slot-local contradictions.

Slot-pair Presence correlations are dominated by L/S (r = 0.729), the strongest pair correlation in the matrix, while all pairs crossing closure-bearing slots (T/O/N) and the deliberative/unresolved side are weak, and U is weakly correlated with all other slots ($|r| \leq 0.142$). The slot-pair correlations therefore support a concern-centered deliberative structure (especially L/S), while U remains weakly correlated with the other slots (Appendix E.6). The same broad retention contrast holds across all three generators: T/O/N remain high, while the deliberative roles and U remain lower, although Gemini 2.5 Flash Lite is particularly low on U (0.360 vs DeepSeek 0.59–0.64; per-generator slot retention is in Appendix E.7).

4.2 Recoverability

To distinguish whether low-retention slots are merely unstable across runs or are hard to recover even with selection, we define **best-of-5** as the per-slot best Presence across the five runs of each (segment $\times$ generator $\times$ budget) condition. Because Presence takes values in $\{0, 0.5, 1.0\}$, we define a slot's **missing Presence mass** as 1 - (per-run mean Presence), the **recovered mass** as best-of-5 Presence - (per-run mean Presence), and the **recovery ratio** as recovered mass / missing mass (the operation is over fractional Presence mass, not over a count of binary cells).

T/O/N are mostly recovered by best-of-5 when they are dropped (recovery ratio ≥ 0.88), whereas U recovers only 0.513 [0.434, 0.594] of its missing Presence mass even at best-of-5. The contrast mean(T/O/N recovery) - U recovery is +0.398 [0.304, 0.491].

The low recovery of U suggests that unresolved-tail retention is not merely a run-selection problem; even same-prompt best-of-5 selection leaves much of the missing Presence mass unrecovered. Nor is it an artifact of segments having no unresolved tail to retain: excluding the 3 of 50 segments where a faithful summary cannot earn U credit (no pending handling path remains on the source side), mean

Slot	Miss.	Recov.	Recov. ratio	95% CI
T	0.054	0.052	0.959	[0.876, 1.000]
O	0.089	0.079	0.888	[0.826, 0.966]
N	0.098	0.086	0.886	[0.814, 0.960]
L	0.339	0.235	0.692	[0.593, 0.793]
S	0.407	0.261	0.640	[0.548, 0.734]
B	0.366	0.215	0.587	[0.495, 0.691]
U	0.470	0.241	0.513	[0.434, 0.594]

Table 4: Recoverability by slot under best-of-5 selection.

U Presence moves from 0.530 to 0.562, still the lowest of all slots. Run-stability diagnostics further indicate that Gemini exhibits larger across-run variability than the DeepSeek variants, picking up different slots in different runs (details in Appendix E.1).

4.3 Reference similarity does not substitute for SIEVE

ROUGE-L (Lin, 2004), BERTScore (Zhang et al., 2020; `roberta-large`), and sentence-embedding cosine (Reimers and Gurevych, 2019; `all-mpnet-base-v2`), all computed against each meeting's dataset-provided reference summary, did not act as proxies for SIEVE Presence. Across the 2,250 summaries, overall correlations are weak or negative, and reference similarity in particular has difficulty capturing the deliberative roles (B, L, S) and the unresolved tail (U).

Metric	Mean Presence r	Fact Error Rate r
ROUGE-L	-0.109	0.031
BERTScore	-0.138	-0.013
Embedding cosine	0.088	0.000

Table 5: Correlation between reference similarity metrics and SIEVE scores.

The magnitudes are small, so we do not interpret these metrics as strongly anti-correlated with SIEVE. The point is that the metrics show at most weak positive associations with Mean Presence and do not consistently track the role-retention profile. The decile diagnostics in Appendix E.2 suggest that this is not merely a null correlation: high-BERTScore summaries tend to retain topic / reference overlap while underweighting L/S/U, and this pattern holds within each sentence budget.

4.4 Role retention of the reference summaries themselves

When the 50 reference summaries are themselves evaluated with SIEVE, only T (0.810) and N (0.730) show high Presence; O/B/L/S/U remain at 0.020–0.100.

Slot	Presence	95% CI	Fact Error
T	0.810	[0.720, 0.890]	0.000
O	0.040	[0.000, 0.100]	0.000
B	0.100	[0.030, 0.180]	0.000
N	0.730	[0.610, 0.850]	0.000
L	0.020	[0.000, 0.060]	0.000
S	0.020	[0.000, 0.060]	0.000
U	0.020	[0.000, 0.060]	0.000

Table 6: Role retention of the dataset reference summaries.

n=50; meeting-level bootstrap 95% CIs. The Fact Error denominator is the set of cells with Presence > 0 (94 cells). Labels are the first author's human judgments.

4.5 Model and sentence-budget effects

Increasing the sentence budget improves Presence across all models; the budget 2→6 improvement in Mean Total Presence is +1.41 [1.17, 1.67] for DeepSeek reasoning, +1.41 [1.20, 1.63] for DeepSeek chat, and +1.15 [0.88, 1.42] for Gemini 2.5 Flash Lite. At the same budget, the DeepSeek variants outperform Gemini (at budget 6: +0.81 [0.55, 1.06] and +0.86 [0.64, 1.07] respectively), and DeepSeek reasoning and chat do not separate

(Δ = +0.05 [-0.11, 0.24], chat minus reasoning). Fine-grained model comparison is not the central claim of this paper; per-generator $\times$ per-budget numbers are in Appendix E.5.

4.6 Calibration check and Background boundary

To check that the main labels do not depend strongly on the first author's threshold, a second author blindly re-evaluated 200 Presence cells: a slot-stratified random subset of 100 plus a boundary-heavy diagnostic hard subset of 100. As this is an audit by two annotators rather than a large-scale multi-annotator IAA study, we treat it as a **calibration check and boundary check** rather than as a population-level reliability estimate. For the three-valued ordinal Presence scale we report quadratic weighted κ (Cohen, 1968).

Subset	n	Role	Exact	≤ 1 level	Severe	Quad. κ
Random	100	repr.	0.720	0.920	0.080	0.554
Hard	100	diagn.	0.620	0.970	0.030	0.279

Table 7: Second-author calibration audit results.

On the random subset, within-one-level agreement (difference ≤ 0.5) is 0.920 and severe disagreement ($0.0 \leftrightarrow 1.0$) occurs in only 8 of 100 cells (8.0%), concentrated on B (the full confusion matrix is in Appendix E.3). The hard subset is a diagnostic stress test by purposive sampling; its low κ does not indicate population-level reliability.

The audit largely supports the two ends of the main results — the high retention of T/O and the low retention of U — while showing that **Background is threshold-sensitive**. Across the audit, all 7 severe disagreements on B (6 on the random subset, 1 on the hard subset) fell in the direction main=0.0 $\rightarrow$ audit=1.0, with no opposing cases (details in Appendix E.3). Inspecting the content, the disagreement concentrated on a single interpretive axis: the main labels required B to mention a "precipitating prior condition that caused the item to be on the agenda in its current form" (the stricter of the two readings the published gate admits), while the audit labels tended to accept "general information that contextualizes the current discussion" as Background. This is an observation from two annotators and does not establish a population-level taxonomy of Background interpretation. Accordingly, the absolute level of B and the fine-grained slot ordering should be interpreted with care.

A robustness check stratified on segment type (decision-bearing vs non-decision) reproduced the main findings on both strata: closure-biased compression, the structural difficulty of U, and the weakness of reference similarity as a proxy for role retention (Appendix E.4).

5 Discussion

The weak or negative correlation between reference similarity and SIEVE Presence is consistent with a property of the references themselves: the reference summaries lack most of the information roles SIEVE measures. The reference summaries contain few roles beyond T and N (Section 4.4), and inspecting their text shows that 45 of the 50 (90%) contain no mention of in-session dispositions or proceedings such as passage, failure, vote counts, or same-day amendments; they read as agenda-item captions describing proposed actions. A caption that only describes the proposed action carries no information about how the item was handled or what remained open afterward, and the generated summaries in our experiment in fact retain more of the roles the references lack (B 0.585–0.683 vs 0.100, U 0.501–0.559 vs 0.020; Appendix E.4). This raises the possibility that optimizing reference similarity may under-reward information roles that are sparse in the references. Reference summaries should therefore be read not as an upper bound on performance but as a baseline that characterizes the dataset, and training or evaluation pipelines that treat reference similarity as gold may be systematically excluding the deliberative dimension of meeting records from their optimization target. This does not mean that ROUGE or BERTScore is invalid in general; they answer a different question than SIEVE.

SIEVE makes visible compression patterns that are hard to see in single-score evaluation. Short summaries retain closure-bearing roles — what was on the agenda and how it was handled — more stably than the deliberative roles that capture why the item is in its current form and which concerns shaped the discussion, plus the unresolved tail that captures what remained open. Because Background is threshold-sensitive (Section 4.6), however, the safest reading is not a fine-grained ordering of the seven slots but the contrast between closure-bearing roles and the lower-retention deliberative and unresolved roles. Importantly, this pattern emerges even though the generation prompt

explicitly asked for "what happened / what shaped the discussion / what remains open" — a realistic policy-briefing request that downstream readers would naturally make. The low retention of deliberative roles is therefore not an artifact of a naive compression instruction nor of an artificially constructed experimental probe; it arises even under a realistic, role-aware summarization request.

The closure-bearing retention contrast is observed in both sampled segment types. Mean Total Presence is nearly equal between decision-bearing and non-decision segments, and O stays at 0.886 on the non-decision stratum (Appendix E.4). Because the O gates count procedural dispositions such as referral, continuation, and retention, not only final passage, what O measures is not whether a decision was made but whether the handling of the item reaches the reader. That the retention contrast is preserved on non-decision segments as well suggests that the closure-bearing bias is not an artifact of a particular meeting type.

Unresolved Tail is the slot that most clearly shows how unresolved items are not naturally retained in short free-form summaries. T/O/N were mostly recovered by best-of-5 when they were dropped, while U recovered only about half of its missing Presence mass. This indicates that simply generating multiple runs under the same prompt and post-hoc selecting a better summary does not adequately guarantee retention of pending items and handling paths. This matters when meeting summaries are used as a handoff, an audit trail, a next-action management input, or context for a downstream LLM workflow — for such uses, it may be necessary not to ask a short narrative summary to carry unresolved items, but to extract open / pending / return-path items as a separate field, or to track pendingness in the source representation before generation. The present study does not compare prompt variants and therefore makes no general claim about the limits of prompt engineering; but at minimum, same-prompt best-of-5 selection is insufficient as a retention strategy for U.

More broadly, SIEVE frames compression quality as the question of which decision-relevant roles survive into downstream context. We do not directly measure downstream task degradation in this study, but the slot-wise profile provides a way to diagnose which information roles subsequent reasoning, retrieval, or handoff stages may lose access to.

6 Conclusion

We proposed SIEVE, a slot-wise framework for evaluating meeting summaries from seven information-role perspectives. Rather than asking only whether a summary resembles a reference or stays consistent with the source, SIEVE produces a role-level profile of what decision-relevant information is retained or dropped.

Applied to 2,250 summaries from municipal-meeting segments, SIEVE shows that short summaries preserve topic, outcome, and action-detail information more reliably than background, concerns, and unresolved items that explain how the discussion was shaped and what remains open. Reference-similarity metrics did not track this profile, and the reference summaries themselves turned out to contain little beyond topic and action-detail information. A summary can therefore look close to a reference while losing information needed for handoff, audit, or downstream reasoning.

These findings position SIEVE as a diagnostic tool for auditing, comparing, and designing meeting summarizers based on which information roles survive compression. Future work should test the slot schema in other meeting domains, broaden model comparisons, and refine evaluation designs for roles whose boundaries are sensitive to reasonable annotator interpretations.

Limitations

This study is limited to municipal and county council meeting segments; whether the same slot schema applies in other domains such as corporate meetings, medical meetings, or educational meetings has not been verified. The calibration audit is a re-evaluation of 200 Presence cells by two annotators; it is useful for indicating per-slot stability differences (in particular the threshold sensitivity of B), but is not a population-level reliability estimate. SIEVE requires an event sheet and a slot-local rubric and is not a cheap automatic evaluator — it is useful for audits of consequential meeting summaries and for methodological evaluation, but the authoring cost is a barrier to direct application at large-benchmark scale. Fact Check is a slot-local contradiction check and not a comprehensive factuality audit of the whole summary, so general factuality evaluation requires combination with other audit techniques. If a summary describes a pending item as resolved and contains no other qualifying pendingness marker, it fails the

U Topic Gate; the reversal is then recorded as U non-retention rather than a slot-local Fact Check error. We use a single role-aware policy-briefing prompt, three generators, and three sentence budgets (2/3/6), and do not test other prompt designs or model families; the results should therefore not be read as a general limit of prompt engineering. The study uses English municipal-meeting transcripts, and multilingual meeting settings remain untested. Decision-bearing versus non-decision labels were used for stratified sampling and robustness checks, but these decision-type labels were not independently validated by multiple annotators.

References

Anthropic. 2026. Introducing Claude Sonnet 4.6. https://www.anthropic.com/news/claude-sonnet-4-6. Accessed 2026-05-25.

Jacob Cohen. 1968. Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin*, 70(4):213–220.

DeepSeek. 2026. DeepSeek API documentation. https://api-docs.deepseek.com/. Accessed 2026-05-25.

Esin Durmus, He He, and Mona Diab. 2020. FEQA: A question answering evaluation framework for faithfulness assessment in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5055–5070.

Bradley Efron. 1979. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1):1–26.

Alexander R. Fabbri, Wojciech Kryscinski, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. SummEval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409.

Alexander R. Fabbri, Chien-Sheng Wu, Wenhao Liu, and Caiming Xiong. 2022. QAFactEval: Improved QA-based factual consistency evaluation for summarization. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2587–2601.

Google. 2026. Gemini 2.5 Flash-Lite model documentation. https://ai.google.dev/gemini-api/docs/models/gemini-2.5-flash-lite. Accessed 2026-05-25.

Yebowen Hu, Timothy Ganter, Hanieh Deilamsalehy, Franck Dernoncourt, Hassan Foroosh, and Fei Liu. 2023. MeetingBank: A benchmark dataset for meeting summarization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16409–16423.

Frederic Kirstein, Jan Philip Wahle, Terry Ruas, and Bela Gipp. 2024. What's under the hood: Investigating automatic metrics on meeting summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 6709–6723.

Wojciech Kryscinski, Bryan McCann, Caiming Xiong, and Richard Socher. 2020. Evaluating the factual consistency of abstractive text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 9332–9346.

Philippe Laban, Tobias Schnabel, Paul N. Bennett, and Marti A. Hearst. 2022. SummaC: Re-visiting NLI-based models for inconsistency detection in summarization. *Transactions of the Association for Computational Linguistics*, 10:163–177.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023a. G-Eval: NLG evaluation using GPT-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522.

Yixin Liu, Alexander R. Fabbri, Pengfei Liu, Yilun Zhao, Linyong Nan, Ruilin Han, Simeng Han, Shafiq Joty, Chien-Sheng Wu, Caiming Xiong, and Dragomir Radev. 2023b. Revisiting the gold standard: Grounding summarization evaluation with robust human evaluation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pages 4140–4170.

Ani Nenkova and Rebecca Passonneau. 2004. Evaluating content selection in summarization: The pyramid method. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 145–152.

OpenAI. 2026. GPT-5 model documentation. https://developers.openai.com/api/docs/models/gpt-5. Accessed 2026-05-25.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 3982–3992.

Hwanjun Song, Hang Su, Igor Shalyminov, Jason Cai, and Saab Mansour. 2024. FineSurE: Fine-grained

summarization evaluation using LLMs. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pages 906–922.

Alex Wang, Kyunghyun Cho, and Mike Lewis. 2020. Asking and answering questions to evaluate the factual consistency of summaries. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5008–5020.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, and Zhifang Sui. 2024. Large language models are not fair evaluators. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pages 9440–9450.

Yuheng Zha, Yichi Yang, Ruichen Li, and Zhiting Hu. 2023. AlignScore: Evaluating factual consistency with a unified alignment function. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pages 11328–11348.

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

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-Bench and chatbot arena. In *Advances in Neural Information Processing Systems 36*.

Ming Zhong, Yang Liu, Da Yin, Yuning Mao, Yizhu Jiao, Pengfei Liu, Chenguang Zhu, Heng Ji, and Jiawei Han. 2022. Towards a unified multi-dimensional evaluator for text generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2023–2038.

Ming Zhong, Da Yin, Tao Yu, Ahmad Zaidi, Mutethia Mutuma, Rahul Jha, Ahmed Hassan Awadallah, Asli Celikyilmaz, Yang Liu, Xipeng Qiu, and Dragomir Radev. 2021. 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*, pages 5905–5921.

A Generation, sampling, and judging details

Appendix A covers the implementation details of sampling, generation, and judging. Appendix B gives the complete rubric for slot-level Presence and Fact Check. Appendix C describes the event-sheet structure and a sample. Appendix D presents illustrative examples of Presence judgment. Appendix E reports supplementary quantitative results.

Sample selection. The eligibility thresholds (transcripts of 2,500–20,000 words and dataset-provided reference summaries of $\geq$15 words) were chosen so that the 2 / 3 / 6-sentence budgets impose meaningful compression pressure: below 2,500 words, the compression pressure is weak; above 20,000 words, the segment becomes an outlier and its context is unstable. When the same meeting contained multiple eligible segments, we used the longest transcript as a deterministic tie-breaker. From 882 eligible meetings, we selected 50 segments via a deterministic random ordering (fixed seed) within each decision-type stratum (25 decision-bearing, 25 non-decision), with a per-stratum body cap, raised only when a selection pass yielded no admissible candidate, to control over-sampling (the cap in effect at each admission is recorded in the sampling manifest). The final manifest of the 50 segments is included in the public repository.

Sample composition.

Public body	n
Alameda	10
Boston	7
Denver	11
King County	4
Long Beach	9
Seattle	9

Table 8: Distribution of sampled segments by public body.

Generation prompt template. All generators received the same prompt. `{N}` is the sentence budget and `{transcript}` is the segment transcript. The system message is fixed to `"You are a concise helpful assistant."` for all generators; `max_tokens = 8,192`.

```
You are summarizing one sampled public-meeting
segment for a policy
analyst. The analyst needs a concise, factual
briefing on what happened
in this segment, what shaped the discussion, and
what remains open
afterward.

Write exactly {N} sentences. Each sentence must
contain no more than
25 words. Use plain prose only — no bullet
points, numbered lists,
or headings. Each sentence should express one
main point.

Prioritize the developments that materially
changed, clarified, or
constrained the item during this segment. Avoid
trivial procedural
detail such as roll call, adjournment times, and
routine declarations,
but retain procedural outcomes when they are
substantively important.

=== TRANSCRIPT START ===
{transcript}
=== TRANSCRIPT END ===
```

Transcript length (words)	n
2,500 – 5,000	20
5,000 – 10,000	15
10,000 – 20,000	15

Table 9: Distribution of sampled segments by transcript length.

Length-constraint compliance. Sentence-count compliance (generated sentence count equal to the requested budget) is 99.0% (2,228/2,250), with per-condition compliance between 96.4% and 100%. The per-sentence 25-word instruction is violated in 0.61% of sentences (50/8,249, measured with a simple punctuation-based splitter), affecting 2.0% of summaries; non-compliant outputs were not filtered or regenerated. Mean summary length is 52.4 words (median 45).

Judge configuration. GPT-5 was invoked with `reasoning_effort = low`, Claude Sonnet 4.6 with `temperature = 0`, both with `max_tokens = 8,192`. The judge prompt is rendered with the fixed SIEVE rubric, the slot-local rubric for the relevant meeting, the relevant event-sheet excerpt, and the summary under evaluation; the judge proceeds in scoring order (Topic Gate $\rightarrow$ Quality Gate $\rightarrow$ Fact Check), with Presence locked after the Quality Gate. The full judge prompt is included in the public repository.

Model versions and access dates. All evaluated summaries were generated between 2026-05-14 and 2026-05-24. Generators: `deepseek-chat`,

`deepseek-reasoner` (DeepSeek API docs; accessed 2026-05-25), `gemini-2.5-flash-lite` (Google Gemini API; accessed 2026-05-25). The DeepSeek API names are version pointers rather than fixed model versions; according to the provider's changelog, during this period they served the non-thinking and thinking modes of DeepSeek-V4-Flash, respectively. LLM judges: GPT-5 (OpenAI API; accessed 2026-05-25), Claude Sonnet 4.6 (Anthropic API; accessed 2026-05-25).

B Slot-specific Presence and Fact Check definitions

B.1 defines the **Presence gates** that determine whether a slot is retained. B.2 lists the **Fact Check fields**, which are applied only to cells with Presence > 0 and only when the summary makes an explicit claim about the relevant field.

B.1 Presence gate operational definitions

Each gate is implemented as a named condition with a slot-specific canonical meaning and is instantiated against the source content of each segment. We describe the operational definitions for each of the seven slots below.

T (Topic/Framing).

- Topic Gate **Subject Identified**: the policy area, issue, or matter being discussed is named as a noun phrase. A procedural form alone (motion number, bill identifier, etc.) or a bare agenda label is insufficient.
- Quality Gate **Policy Direction + Target**: the summary expresses a policy direction toward the subject (a direction-bearing verb such as update, revise, clarify, expand, restrict, phase out, or a concrete rule effect such as defaulting / extending / narrowing) and the target to which the direction applies (entity, group, area, rule set, scope). A procedural disposition (approved, adopted, recommended) does not count as a direction.

O (Outcome/Position).

- Topic Gate **Disposition Verb Tied to Acted-on Item**: the summary records that the body took a disposition (vote, referral, approval, deferral, etc.) on the acted-on item. Discussion, questioning, support, or hearing alone is insufficient.
- Quality Gate **Acted-on Item Specific**: the acted-on item is specific enough that the reader can tell "an action on which item." A formal number or label suffices but is not required; a substantive description also qualifies.

B (Agenda-Setting Background).

- Topic Gate **Precipitating Context Visible**: the summary states, as an independent proposition, a prior event / condition / process / problem / opposition / mediation / revision / failure / deadline that is connected to the current item. An explicit "why now" statement is not required.
- Quality Gate **Connected to Current Item or Form**: the summary makes it visible that the background is connected to one of the current item, the proposed form, a revision, the discussion frame, or the handling approach.

N (Operative Payload).

- Topic Gate **Payload Attached to Action**: a payload (condition, requirement, scope, restriction, number, date) is expressed as attached to the action or the acted-on item. Single-sentence attachment and anaphora-linked adjacent sentences count. A free-standing background description that requires the reader to reconstruct the action–payload link is insufficient. Even if a recognizable payload is attached to the wrong actor / stage, Presence still passes (the replacement error is handled by Fact Check).
- Quality Gate (any-of; pass if any of the following is satisfied): **Number** (an identifiable numeric quantity such as a count, amount, percentage, duration, or threshold) / **Named Requirement** (an identifiable named requirement, restriction, condition, or carve-out) / **Scope Boundary** (an identifiable scope boundary: geographic, categorical, population-specific, etc.) / **Date Cutoff** (an identifiable date cutoff: effective date, expiration date, sunset clause, deadline).

L (Concrete Concern).

- Topic Gate **Concrete Concern Language**: the summary uses problem-side language such

as concern, harm, obstacle, difficulty, loss, risk, tradeoff, deficiency, shortage, gap, barrier, or failure. Problem-side vocabulary appearing only as the object or modifier of a solution-side verb is insufficient. Generic debate language is insufficient.

- Quality Gate **Specific Concern Content**: the concern content is specific enough that the reader can tell "a concern / harm / obstacle / risk / tradeoff about what." Generic references such as "concerns were raised" are insufficient. **A named place, named group, or bounded area is not required for Presence** (local anchoring, when stated, is a Fact Check field, but is not a basis for lowering Presence).

S (Shared Concern Synthesis).

- Topic Gate **Problem-Side Concern**: the concern theme is named in problem-side language such as problem, difficulty, deficiency, gap, fragmentation, dissatisfaction, or negative state. Problem-side vocabulary appearing only as the object or modifier of a solution-side verb is insufficient. Generic debate language is insufficient.

- Quality Gate **Specific Content**: the named concern theme has identifiable concern content (object of concern, contrast at issue, criticized mechanism, specific tension / tradeoff, problem content itself). A generic concern label alone is insufficient.

U (Unresolved Tail).

- Topic Gate **Pendingness Marker**: the summary contains a marker that explicitly signals an unfinished / pending / open state, such as "remained pending," "unresolved," "held over," "will proceed to a later full-body meeting / final action / final passage / final consideration," or "final action remained pending." A bare directive is insufficient.

- Quality Gate **Pending Item Identifiable + Handling Path Visible**: (i) what is pending (item, workstream, issue) is identifiable, and (ii) a concrete handling path is visible (future event path, process path, institutional custody path). Mere future orientation, an unrelated future item, or a post-adjournment announcement does not satisfy U.

B.2 Fact Check fields

We list the fields against which Fact Check operates per slot. These are not a required-mention checklist; source-comparison is performed only when the summary makes an explicit claim about the relevant field. The full roles and fail-condition details per field, and slot-specific rules, are included in the public repository.

T *Fact Check fields for Topic/Framing.*

- **Subject. Check:** Compare the substantive subject stated by the summary with the source-supported subject ; **Fail:** The summary explicitly states a policy area / issue / matter as the subject that differs from the source

- **Policy Direction. Check:** Compare the policy direction stated by the summary with the source-supported direction ; **Fail:** The summary states a direction that reverses or replaces the source-supported direction

- **Target. Check:** Compare the target stated by the summary with the source-supported target ; **Fail:** The summary states an entity / group / area / process / jurisdiction / scope as the target that differs from the source

O *Fact Check fields for Outcome/Position.*

- **Actor. Check:** Compare the acting body / actor stated by the summary with the source ; **Fail:** The summary attributes the disposition to an actor / body different from the source

- **Procedural Stage. Check:** Compare the procedural stage stated (or clearly implied) by the summary with the source ; **Fail:** The summary states or clearly implies a stage beyond the source-supported stage, or claims that an incomplete legal / electoral / administrative effect has been completed

- **Disposition Result. Check:** Compare the stated result with the source-supported result ; **Fail:** The summary states a result (passed / failed / deferred, etc.) different from the source

- **Acted-on Item. Check:** Compare the acted-on item stated by the summary with the source ; **Fail:** The summary states an ordinance / motion / amendment / appointment as the acted-on item that differs from the source

- **Vote Result. Check:** When the summary attaches a vote count / unanimity / passed-failed result, compare it with the source ; **Fail:** The summary states a vote result different from the source, or attaches a source-supported vote to the wrong action

B *Fact Check fields for Agenda-Setting Background.*

- **Trigger Condition. Check:** Compare the prior event / condition / process / problem / opposition / mediation / revision / failure / deadline stated by the summary with the source ; **Fail:** The summary states a triggering background different from the source

- **Agenda Connection. Check:** When the summary states the relation between the background and the current item, compare it with the source ; **Fail:** The summary reverses the relation, or states a reason / background relation different from the source

N *Fact Check fields for Operative Payload.*

- **Payload Carrier. Check:** Compare the carrier to which the summary attaches the payload (action / ordinance / motion / amendment / resolution / acted-on item) with the source ; **Fail:** The summary attaches the payload to a carrier different from the source

- **Payload Content. Check:** Compare the requirement / condition / restriction / permission / rule / operative content stated by the summary with the source ; **Fail:** The summary states operative payload content different from the source

- **Payload Scope. Check:** When the summary states the payload's applicable scope / category / population / geography / process / target, compare it with the source ; **Fail:** The summary states a scope / target different from the source

- **Payload Scalar or Baseline. Check:** When the summary states a number / date / duration / threshold / amount / deadline / baseline, compare it with the source ; **Fail:** The summary states a wrong value, attaches a correct value to the wrong payload role, or states a wrong baseline / comparison anchor

L *Fact Check fields for Concrete Concern.*

- **Obstacle Content. Check:** Compare the concern / harm / obstacle / difficulty / loss / risk / tradeoff / constraint / problem stated by the summary with the source ; **Fail:** The summary states concern content different from the source

- **Local Anchor. Check:** When the summary ties the concern to a named place / named group / bounded area / local anchor, compare it with the source (local anchor is not required for Presence) ; **Fail:** The summary states a local anchor different from the source

S *Fact Check fields for Shared Concern Synthesis.*

- **Concern Theme. Check:** Compare the recurring concern theme stated by the summary with the source ; **Fail:** The summary states a concern theme different from the source

- **Concern Content. Check:** When the summary states the concern's object / contrast / mechanism / tension / tradeoff / risk / consequence / affected group / scope, compare it with the source ; **Fail:** The summary states concern content different from the source

U *Fact Check fields for Unresolved Tail.*

- **Pending Status. Check:** Compare the unfinished / pending / resolved state stated by the summary with the source ; **Fail:** The summary states that a source-supported pending item is resolved / completed, or vice versa

- **Pending Item. Check:** Compare the item the summary says is pending with the source ; **Fail:** The summary states a pending item / workstream / ordinance / motion different from the source

- **Handling Path. Check:** When the summary states a future meeting / body / committee / process / report-back / review / drafting, compare it with the source ; **Fail:** The summary states a handling path / body / process / date / destination different from the source

B.3 Global rules

The following three global rules apply to all slots:

1. **Contradiction-only**: Fact Check is a contradiction check, not a completeness check (see Section 3.7).

2. **Slot-local**: Fact Check is limited to the selected slot and its Fact Check fields. Even if a summary claim about another slot's content is contradictory, it does not lower the Fact Check of the current slot to 0. We do not perform a summary-wide audit.

3. **Conflict patterns**: typical contradiction patterns fall into six types: (i) wrong scalar value, (ii) correct value attached to the wrong target, (iii) wrong temporal status, (iv) wrong decision / stage, (v) wrong actor / target / scope, (vi) wrong condition / cause / baseline. These are conflict patterns, not a checklist. These patterns were used as descriptive tags during adjudication, not as additional required fields.

C Event sheet structure and sample

C.1 Event sheet fields

For each of the seven slots, the event sheet records the following fields:

Field	Content
Label	Slot identifier (T / O / B / N / L / S / U)
Summary	A slot-local summary of the source-side content for the slot
Evidence	Transcript-level evidence (quotation or paraphrase)
Source vocabulary	The relevant vocabulary actually used in the source
Temporal status	Temporal positioning of the slot content (see below)
Source context for fact check	Source-side facts needed for Fact Check judgments
Answer locality	Concentration of the slot content within the transcript (local / distributed, etc.)
Marker presence	Presence of explicit lexical markers indicating the slot content

Table 10: Event sheet fields used for slot-local evidence.

C.2 Temporal status

Temporal status takes one of seven values: *already occurred* (past event), *currently existing* (current situation), *completed in session* (act completed within the segment), *pending future handling* (awaiting a later body), *future effect* (future effect to take place), *hypothetical or risk* (hypothetical or risk state), and *recurring concern* (a repeatedly discussed concern).

Temporal status is recorded as source context, not as an independent scoring field. It is materialized only when temporal positioning matters for slot judgment or Fact Check — for example, distinguishing a recommendation to a later body from a completed approval. It is particularly important for U, where pendingness and future handling are central.

Event sheets are constructed manually under a fixed protocol. The protocol document and the event sheets for all 50 segments are included in the public repository.

C.3 Sample event sheet: MTG31 Slot T

To illustrate the operational substance of an event sheet, we excerpt the full event sheet for slot T (Topic/Framing) of MTG31 (`LongBeachCC_03222016_16-0122`, Long Beach City Council, recycling RFP proceedings). The complete version including the remaining six slots (O / B / N / L / S / U), and event sheets for all 50 segments, are included in the public repository.

Segment metadata

Dataset-provided reference summary

> *Recommendation to adopt Specifications No. RFP PW 15-091 and award a contract to USA Waste of California, dba Waste Management, of Long Beach, for recyclable collection services, in an annual amount not to exceed $3,500,000, plus an annual Consumer Price Index adjustment; ... (Citywide)*

Slot T event sheet

```
label: T

summary: >
  The segment was substantively about Long
  Beach's citywide recyclable
  collection services procurement. The policy
  move under discussion
  shifted from the agenda's proposed award of the
  recycling contract
  to Waste Management to a narrower course:
  conduct an independent
  review of the recycling RFP process first, then
  let the Council
  decide whether to award the contract as
  recommended or restart
  the process.

evidence:
  - "Recommendation to award a contract to Waste
  Management for
    recyclable collection services in an annual
    amount not to exceed
    3.5 million citywide."
  - "Over the past few weeks, there's been a huge
  amount of
```

Field	Value
Meeting ID	LongBeachCC_03222016
Segment	item 16-0122 (recyclable collection services RFP)
Body	Long Beach City Council
Decision type	Other (non-decision; review-before-award)
Segment word count	10,306

Table 11: Metadata for the sample event-sheet segment.

```
    misinformation provided about the procurement
    process with
    regards to the recycling RFP."
  - "What we're doing is ... conduct the review
  ... Based on that
    review, the council will decide either to
    award the contract
    as recommended or we begin the process
    again."

source_vocabulary:
  - recyclable collection services
  - recycling RFP
  - procurement process
  - award a contract to Waste Management
  - citywide
  - conduct the review
  - award the contract as recommended
  - begin the process again

temporal_status:
  primary: currently_existing
  secondary:
    - pending_future_handling
    - hypothetical_or_risk

source_context_for_fact_check:
  subject: City of Long Beach citywide recyclable
  collection services
    procurement / recycling RFP
  policy_direction:
    - The agenda item originally proposed
    awarding the contract to
      Waste Management.
    - The adopted course was to conduct an
    independent review first,
      then decide whether to award as recommended
      or begin the
      process again.
  target: The recyclable collection services
  contract / recycling
    RFP process (not all municipal refuse or the
    entire solid-waste
    system).

answer_locality: Opening agenda and staff report;
Richardson/Price
  motion clarification; final vote sequence.

marker_presence:
  subject_identified: explicit
  policy_direction: explicit
  target: explicit
```

D Sample Presence judgments

As representative examples, we show Presence 0.0 / 0.5 / 1.0 judgments for T (Topic/Framing), B (Agenda-Setting Background), and U (Unresolved Tail). These are illustrative cases selected from evaluated cells and are included to explain rubric boundaries. For B, we also include one severe audit-disagreement case to illustrate the strict/permissive Background boundary discussed in Section 4.6.

D.1 T (Topic/Framing)

T = 0.0 (MTG7, DeepSeek reasoning, 2 sent., run 2)

> *Council deadlocked on delaying the R&D model, leaving previous implementation plans intact. Debate highlighted financial savings versus reduced sworn staffing and county pilot approval constraints.*

→ Slot context: a policy judgment on Long Beach's Rapid Medic Deployment EMS staffing model. The substantive subject (paramedic deployment / EMS staffing model) is not stated as a noun phrase; only procedural framing (deadlock, debate) appears. Topic Gate **Subject Identified** is not satisfied.

T = 0.5 (MTG2, DeepSeek reasoning, 2 sent., run 2)

> *Jay Osborne was confirmed as director of the King County Department of Human Resources via unanimous vote after an amendment.*

→ Slot context: confirmation of the new director of the King County HR Department. The subject (the director confirmation of the Department of Human Resources) is stated, but neither the policy

direction (standing up centralized HR) nor the target is expressed as a substantive description, so the Quality Gate fails.

T = 1.0 (MTG1, DeepSeek reasoning, 2 sent., run 1)

> *Council postponed CB430, approved rezoning CB509, and passed CB563 after debate over solar access for Earth Links garden. Affordable housing bills 626 and 625 advanced to second reading, leaving funding source and timing unresolved.*

→ Slot context: a multi-bill Denver City Council session. The substantive subjects are stated (the CB509 rezoning, the affordable-housing bills 626 and 625, solar access for the Earth Links garden), the policy moves are substantive rather than merely procedural (a rezoning; affordable-housing funding measures, with their funding source at issue), and the targets are identified (the named bills and the garden). The Quality Gate passes on these substantive elements; the procedural disposition verbs (postponed / approved / advanced) are not what satisfies it.

D.2 B (Agenda-Setting Background)

B = 0.0 (MTG1, DeepSeek reasoning, 2 sent., run 5)

Segment / meeting: 01_DenverCityCouncil_09122016 / MTG1. Generator / budget / run: deepseek_reasoning / budget_02 / run_05. Final B Presence: 0.0.

Generated summary text:

> *The council voted 7-5 to advance Council Bill 626, an alternative affordable housing fund, to second reading. After hearing from concerned community members, the council approved the Sun Valley rezoning to allow up to eight-story development.*

B source-side context: Several agenda items were shaped by source-side background conditions. Council Bill 430 was postponed because notice requirements had not been met. Council Bill 509 came forward because an old, narrow PUD tied to a former bakery no longer fit the separately addressed and separately taxed parcels. Council Bill 563 came forward to implement recently adopted Sun Valley and Decatur-Federal station-area plans and to move away from industrial zoning that was treated as a barrier to the neighborhood vision. Council Bill 626 arose because some council members viewed the original affordable-housing funding proposal as insufficient or too slow and wanted to explore more funding sources. Resolution 759 was shaped by prior debate over red-light-camera enforcement and yellow-light timing.

→ Slot context: the summary reports current action and debate content but does not state a separate prior condition, process, problem, or pressure that visibly shaped why the item came forward or took its current form. The Topic Gate for B therefore fails.

B = 0.5 (MTG7, DeepSeek reasoning, 2 sent., run 1)

Segment / meeting: 07_LongBeachCC_06102014 / MTG7. Generator / budget / run: deepseek_reasoning / budget_02 / run_01. Final B Presence: 0.5.

Generated summary text:

> *Council debated delaying RMD implementation but motions tied, leaving prior authorization intact. Public testimony opposed cuts, but no policy change was adopted.*

B source-side context: the discussion was shaped by several prior and current conditions: RMD had been discussed for years and included in the FY14 budget, implementation had been delayed, the Fire Department was projected to finish FY14 on budget because of one-time funds and salary savings, the FY15 proposed budget included RMD, and eliminating RMD would require about $1.4 million in other Fire Department reductions. The item also arose against County EMS pilot authorization and past budget-deficit reasoning for the model.

→ Slot context: the summary names background or prior conditions, so B is partly reached, but the connection to the current item, proposal form, or handling is too thin or incomplete for full Presence. Topic Gate passes, but Quality Gate fails.

B = 1.0 (MTG31, DeepSeek reasoning, 3 sent., run 3)

Segment / meeting: 31_LongBeachCC_03222016 / MTG31. Generator / budget / run: deepseek_reasoning / budget_03 / run_03. Final B Presence: 1.0.

Generated summary text:

> *City manager recommended rejecting all bids for a 30-day independent review af-*

ter auditor raised concerns. Council approved a compromise motion for a third-party review selected by the city attorney, avoiding contract extension. The review will examine written protests, verbal allegations, and RFP steps, reporting by June 14th.

B source-side context: the current action was shaped by a prior recycling RFP process, vendor protests and comments, questions raised at earlier council discussion, and a City Auditor memorandum expressing concern. The timeline pressure came from the existing Waste Management contract expiring on June 30, 2016, making a long audit, rebid, or extension operationally consequential.

→ Slot context: the summary states prior events, conditions, process history, and pressure-setting facts, and visibly connects them to the current item, debate frame, timing, or handling. Both Topic Gate and Quality Gate pass.

B audit disagreement: main/final B = 0.0, second-author audit B = 1.0 (IAA143)

Segment / meeting: 13_LongBeachCC_03102015 / MTG13. Generator / budget / run: deepseek_temp1_chat / budget_03 / run_03. Final B Presence: 0.0. Second-author audit B Presence: 1.0.

Generated summary text:

Council adopted mandatory spay/neuter for dogs with exemptions, effective October 1, 2015. A councilmember proposed grandfathering currently licensed unaltered dogs to mitigate revenue loss and budget impact. Opponents argued enforcement would burden low-income owners, while supporters cited reduced shelter euthanasia as the primary benefit.

B source-side context: the ordinance was shaped by a prior November 2014 council direction to return with an ordinance, a longer reform narrative dating back years, and continuing concerns that pet overpopulation, shelter intake, and euthanasia remained problems despite increased adoptions and prior voluntary efforts.

→ Slot context: this is a severe audit-disagreement example, included separately from the regular B = 1.0 case to illustrate the strict/permissive Background boundary. The main / final reading required a prior or existing condition that shaped why the item came forward or took its form; general topic context was insufficient, so B was labeled 0.0. The second-author audit accepted broader contextualizing information connected to the ordinance adoption — including the grandfather clause proposal, enforcement burden, and shelter euthanasia rationale — and labeled B as 1.0. This is exactly the type of interpretive divergence reported in Section 4.6.

D.3 U (Unresolved Tail)

U = 0.0 (MTG7, Gemini 2.5 Flash Lite, 2 sent., run 1)

Council debated EMS staffing changes and budget tradeoffs. Public comments addressed both fire response speed and city finances.

→ Slot context: in the source, the adoption of the Rapid Medic Deployment model is pending, and county pilot approval / monitoring / subsequent council action are stated as future handling paths. The summary contains no pendingness marker (remained pending, will return, final action, etc.), so the Topic Gate fails.

U = 0.5 (MTG10, DeepSeek chat, 3 sent., run 4)

The council also referenced future work on contract refinement and noted that staff would return with further information.

→ Slot context: future work on the Alameda Family Services contract. A pendingness marker (return) and a handling path (staff will return) are stated, but the pending item (what specifically is pending — a subset of contract finalization, a second-year extension, etc.) is not identifiable, so the Quality Gate fails.

U = 1.0 (MTG31, DeepSeek reasoning, 6 sent., run 3)

After the review, the council will decide whether to award the contract to Waste Management or reject all proposals and rebid.

→ Slot context: the contract award decision is pending until the independent review completes. Both the pending item (the award decision) and the handling path (a council decision after the review: award, or reject all proposals and rebid) are identifiable. The Quality Gate passes.

E Supplementary results

E.1 Run stability detail

Jaccard is the agreement of retained-slot sets (the set of slots with Presence = 1.0) across runs of the same condition; the **best-of-5 gap** (oracle gain) is the difference between the slot-wise best-of-5 total (the sum over slots of each slot's best Presence across the five runs; a slot-level oracle that no single summary need attain) and the per-run mean Total Presence under the same condition. At budget 6, the DeepSeek variants are run-stable on retained-slot sets, while Gemini 2.5 Flash Lite has lower Jaccard and a larger best-of-5 gap. Gemini's Jaccard is -0.141 [-0.182, -0.098] below DeepSeek reasoning and -0.136 [-0.178, -0.094] below DeepSeek chat, with a best-of-5 gap roughly 0.65 higher. Jaccard was computed on the strict (Presence = 1.0) retained-slot set, but the qualitative ordering across generators is preserved under the loose (Presence ≥ 0.5) variant.

Generator	Jaccard	95% CI	Best-of-5 gap	95% CI
DeepSeek reasoning	0.880	[0.855, 0.904]	0.692	[0.528, 0.864]
DeepSeek chat	0.875	[0.846, 0.902]	0.648	[0.508, 0.792]
Gemini 2.5 Flash Lite	0.739	[0.695, 0.783]	1.316	[1.080, 1.560]

Table 12: Run-stability diagnostics by generator.

E.2 Per-budget and decile diagnostics of reference similarity

Per-budget correlations. Conditioning on budget, ROUGE-L and BERTScore remain weakly negatively correlated with Mean Presence; embedding cosine turns slightly negative at budget 2 and slightly positive at budget 6, but $|r| < 0.06$ throughout. By slot, BERTScore $\times$ T_presence is positive at every budget (+0.20, +0.21, +0.11), while BERTScore $\times$ {L, S, U}_presence is negative: within each sentence budget, higher BERTScore is associated with higher T Presence and lower L/S/U Presence.

Pearson r vs. Mean Presence, n = 750 per budget.

BERTScore top vs bottom decile within budget. Extracting the top and bottom deciles within each budget, the top BERTScore decile is consistently shorter than the bottom decile, has higher surface overlap with the reference, and has lower Mean Presence; L/S/U Presence in particular is lower in the top decile at every sentence budget. These descriptive comparisons do not isolate the cause of the association.

n = 75 summaries per decile.

Reference overlap definition. "Ref. overlap" in the decile table is the token-level lexical overlap ratio between the generated summary and the reference summary: after lowercasing and simple whitespace tokenization, |tokens(cand) intersection tokens(ref)| / |tokens(cand)| (denominator = candidate). No stopword removal or stemming is applied.

Segment-level examples of similarity-retention dissociation. The generated summary of MTG31 (recycling procurement) is lexically distant from the reference summary but, under SIEVE, retains the auditor concerns, the independent review, the June 14 report, and the avoidance of contract extension, reaching Mean Presence 1.000. Conversely, on MTG41 (budget package), a title-like summary close to the reference (BERTScore 0.911) reaches only Mean Presence 0.286. Cases where reference similarity is high but role retention is low, and cases where reference similarity is low but role retention is high, both exist.

Budget	ROUGE-L	BERTScore	Embedding
2 sent.	-0.097	-0.110	-0.028
3 sent.	-0.063	-0.077	+0.016
6 sent.	-0.021	-0.087	+0.054

Table 13: Budget-specific correlations between reference similarity and SIEVE scores.

Budget	Decile	Length (w)	Ref. overlap	Mean Presence	T	L	S	U
2 sent.	top 10%	30.3	0.527	0.592	0.953	0.260	0.280	0.167
	bottom 10%	33.3	0.120	0.680	0.787	0.627	0.620	0.587
3 sent.	top 10%	40.5	0.516	0.670	0.980	0.373	0.347	0.427
	bottom 10%	48.0	0.150	0.783	0.847	0.807	0.793	0.713
6 sent.	top 10%	69.2	0.578	0.789	1.000	0.707	0.593	0.427
	bottom 10%	83.4	0.207	0.879	0.967	0.960	0.913	0.747

Table 14: BERTScore top- and bottom-decile diagnostics by sentence budget.

E.3 LLM triage and second-author audit details

LLM judge slot-level agreement. Slot-level Presence agreement between the two LLM judges (GPT-5, Claude Sonnet 4.6). B is the lowest on exact agreement and the highest on severe disagreement; O is the highest on exact and the lowest on severe.

This mirrors the second-author audit: Background is the slot where both automated triage and human calibration most clearly surface an operational boundary.

Human audit confusion matrices. Rows are the first author's final labels and columns are the second author's audit values. Severe disagreements ($0.0 \leftrightarrow 1.0$) total 11 (5.5%): 8 on the random subset and 3 on the hard subset. On the random subset the mass concentrates in the (1.0, 1.0) cell; on the hard subset the mass is spread broadly around the Quality Gate boundary (around 0.5).

Random subset (n = 100):

Hard subset (n = 100):

Main vs audit slot-mean comparison. Slot-level mean Presence on the random subset, compared between the main results (first author's final labels) and the second author's audit labels. B, which was lowest under the main results, moves up to the top region under the audit, while T and O remain the highest-retention slots in this subset and U remains among the lower-retention slots. Because each slot is limited to 14–15 cells in the random subset, this ranking is a diagnostic comparison rather than a population-level estimate.

N's rank change reflects both a lower audited mean ($0.929 \rightarrow 0.821$) and increases in several other slots; these small-sample rankings are descriptive.

Directionality of B severe disagreement. Of the 7 severe disagreements on B across the audit (6 in the random subset, 1 in the hard subset), all fell in the direction main=0.0 $\rightarrow$ audit=1.0, with no opposing cases. Under the null hypothesis of unbiased direction, this corresponds to binomial p = 0.016; but as one of the seven comes from the purposive hard subset and the observation is by n=2 annotators, we treat it as a descriptive directional figure rather than a population-level claim.

Slot	n	Exact	Adjacent (gap=0.5)	Severe (gap=1.0)
O	2,250	92.0%	6.0%	2.0%
T	2,250	86.3%	11.2%	2.6%
U	2,250	86.2%	8.6%	5.2%
N	2,250	85.6%	6.5%	7.9%
S	2,250	83.2%	6.2%	10.6%
L	2,250	82.7%	9.1%	8.3%
B	2,250	78.3%	5.4%	16.4%

Table 15: LLM-judge slot-level agreement before human adjudication.

First \ Second	0.0	0.5	1.0
0.0	5	14	8
0.5	0	2	1
1.0	0	5	65

Table 16: Presence confusion matrix for the random audit subset.

First \ Second	0.0	0.5	1.0
0.0	1	3	3
0.5	0	18	12
1.0	0	20	43

Table 17: Presence confusion matrix for the boundary-heavy audit subset.

Slot	Main mean	Audit mean	Main rank	Audit rank
T	1.000	1.000	1	1
O	0.929	0.964	2	2
N	0.929	0.821	2	4
L	0.600	0.767	4	6
S	0.571	0.821	5	4
U	0.571	0.643	5	7
B	0.433	0.900	7	3

Table 18: Slot-mean comparison between main labels and the second-author audit.

E.4 Per-segment-type results

Mean Total Presence is almost equal across the two strata (decision-bearing 5.198, non-decision 5.154), and the broad retention contrast (T/O/N high, deliberative roles and U lower) is reproduced on both. At the slot level the differences are small: B is higher on non-decision (0.683 vs 0.585) and U is higher on decision-bearing (0.559 vs 0.501).

U's best-of-5 recovery ratio is 0.522 on decision-bearing and 0.505 on non-decision segments (overall: 0.513). Within-stratum correlations between reference similarity and Mean Presence remain weak in both strata (all $|r| \leq 0.19$; decision-bearing / non-decision: ROUGE-L -0.018 / -0.188, BERTScore -0.084 / -0.187, embedding cosine +0.165 / +0.008).

Each stratum: n = 25 segments $\times$ 3 generators $\times$ 3 budgets $\times$ 5 runs = 1,125 summaries.

Segment type	n	Total	T	O	B	N	L	S	U	Fact Err
decision-bearing	1,125	5.198	0.940	0.936	0.585	0.906	0.668	0.604	0.559	0.099
non-decision	1,125	5.154	0.952	0.886	0.683	0.899	0.653	0.581	0.501	0.086

Table 19: SIEVE results by decision-bearing and non-decision segment type.

E.5 Model and budget effects detail

Mean Total Presence by generator $\times$ budget. All models improve with budget; at the same budget, the DeepSeek variants outperform Gemini.

Generator	2 sent.	3 sent.	6 sent.	Δ 2→6
DeepSeek reasoning	4.704	5.352	6.118	+1.414
DeepSeek chat	4.760	5.532	6.172	+1.412
Gemini 2.5 Flash Lite	4.166	4.470	5.312	+1.146

Table 20: Mean total Presence by generator and sentence budget.

E.6 Slot-pair Presence correlation

Slot-pair Presence correlations across the full set of 2,250 summaries. L/S has r = 0.729, the strongest of any pair in the matrix, consistent with sharing the same deliberative-concern dimension. B is weakly correlated with L (r = 0.263) and with S (r = 0.209), suggesting semantic proximity between agenda-setting background and the deliberative concerns. All pairs that cross closure-bearing slots (T/O/N) and the deliberative/unresolved side are weak, and U is weakly correlated with all other slots ($|r| \leq 0.142$).

Pearson r; n = 2,250 for every pair.

	T	O	B	N	L	S	U
T	—	0.045	0.117	0.168	0.075	0.096	0.032
O		—	0.033	0.151	0.034	0.048	0.075
B			—	0.118	**0.263**	**0.209**	0.048
N				—	0.028	0.021	0.072
L					—	**0.729**	0.142
S						—	0.079
U							—

Table 21: Slot-pair Presence correlations.

E.7 Per-generator slot retention

The same broad retention contrast (T/O/N high, deliberative roles and U lower) holds across all three generators. Gemini 2.5 Flash Lite is particularly low on U (0.360 vs DeepSeek reasoning 0.637, DeepSeek chat 0.593). The slot-level patterns of DeepSeek reasoning and chat are close, and the within-family variants do not separate at the slot level either.

Mean Presence; n = 750 cells per generator per slot (aggregated across all budgets).

Generator	T	O	B	N	L	S	U
DeepSeek reasoning	0.939	0.969	0.653	0.936	0.662	0.595	0.637
DeepSeek chat	0.960	0.944	0.701	0.934	0.707	0.650	0.593
Gemini 2.5 Flash Lite	0.939	0.819	0.547	0.837	0.613	0.533	0.360

Table 22: Per-generator slot retention by information role.