CIG: Measuring Conversational Information Gain in Deliberative Dialogues with Semantic Memory Dynamics

Anonymous ACL submission

Abstract

Measuring the quality of public deliberation requires evaluating not just civility or argument structure, but the informational progress a conversation makes. We introduce a framework for **Conversational Information Gain (CIG)** that evaluates each utterance in terms of how it advances collective understanding of the target topic. To operationalize CIG, we model an evolving **semantic memory** of the discussion: the system extracts atomic claims from utterances and incrementally consolidates them into a structured memory state. Using this memory, we score each utterance along three interpretable dimensions: **Novelty**, **Relevance**, and **Implication Scope**. We annotate 80 segments from two moderated deliberative settings (TV debates and community discussion) with these dimensions and show that memory-derived dynamics (e.g., the number of claim updates) correlate more strongly with human-perceived CIG than traditional heuristics such as utterance length or TF–IDF. We develop effective LLM-based CIG predictors paving the way for information-focused conversation quality analysis in dialogues and deliberative success. [1]

1 Introduction

Public deliberation—reasoned dialogue aimed at collective understanding and decision-making (Habermas, 1985) for public interest (Dewey and Rogers, 2012)—is fundamental to democratic societies. Yet, the quality of these exchanges, from community forums (Schroeder et al., 2024) to public debates (Montez and Brubaker, 2019), can vary widely from stagnation to productive collaboration.

Metrics for evaluating dialogue quality (Goddard and Gillespie, 2023) have largely focused on structural formality over substantive content. Approaches based on schemes like the Deliberative Quality Index (DQI; Steenbergen et al. (2003)) or computational proxies for civility (Anuchitanukul et al., 2022; Price et al., 2020) and argument structure (Wachsmuth et al., 2024) struggle to capture true informational progress. These surface-level signals can mislead: civil phrasing can conceal malicious intent (Kruk et al., 2024), and valuable insights might emerge from informal exchanges (Walton, 2008; Bohm and Weinberg, 2004). Such methods fail to distinguish constructive progress from bureaucratic talks (Montez and Brubaker, 2019; Walton, 2003).

Grounded in the concept of information impact as "A change, or the nature or magnitude of change, in the knowledge base of a subject domain of the recipient" (Meadow and Yuan, 1997), we define **Conversational Information Gain (CIG)** as the degree to which an utterance advances collective understanding toward the goal/topic. CIG is decomposed into three interpretable aspects—**Novelty**, **Relevance**, and **Implication Scope**—each capturing whether a contribution introduces new information, connects to the shared goal, and extends its implications to the public community beyond individual cases. To operationalize CIG, we require a representation of the evolving collective knowledge state against which each new utterance can be evaluated. Both annotators and models must know what has already been said. As shown in Figure 1, we implement this through a lightweight semantic memory that maintains a consolidated set of claims.

We first validate CIG through human annotation of 80 dialogue segments drawn from two moderated group-discussion settings—TV debates and community discussion—achieving moderate-to-high inter-annotator agreement for CIG and its aspects. We then validate automation by using an LLM under the annotators' information conditions (topic, short context, and a prior-memory summary), and show that the LLM's predictions

[1] Code and annotations will be released upon publication.

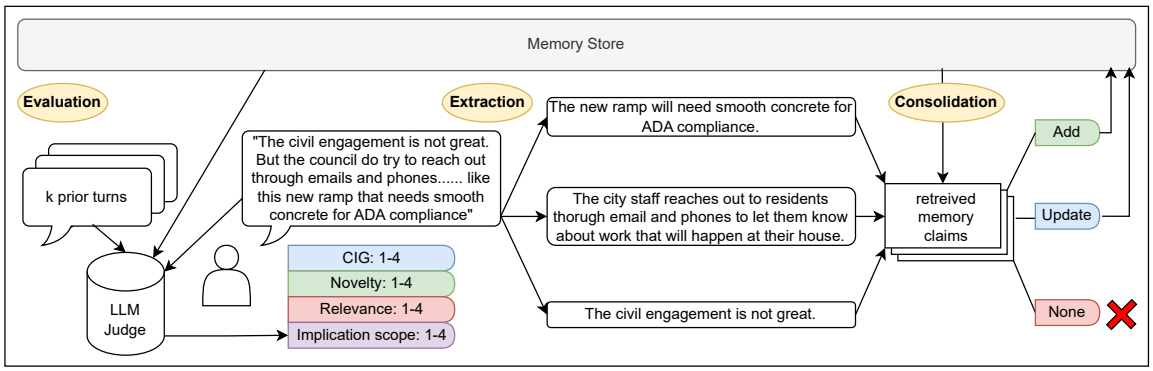

Figure 1: **Overview of the CIG pipeline.** Each utterance is **evaluated** with the Memory Store as knowledge context for Novelty, Relevance, Implication Scope, and overall CIG (1–4). The Memory Store is maintained through two modules: **Extraction**, which converts utterances into atomic claims; and **Consolidation**, which matches extracted claims against the retrieved memory, triggering ADD, UPDATE, or NONE operations.

closely track aggregated human judgments. Finally, by varying the model's prior context, we find that predictions based on retrieved memory summaries are highly correlated with those based on the full preceding transcript, indicating that memory summaries provide a compact yet faithful substitute for full-history context in automated CIG assessment.

We then analyze how memory dynamics relate to perceived informational progress, finding that simple memory-state signals (e.g., claim update counts) track human CIG ratings more consistently than common heuristics such as sentence entropy or TF–IDF. An unsupervised aggregation analysis reveals a "conjunctive bottleneck": an utterance's CIG is effectively limited by its weakest aspect. Finally, we provide a case study illustrating how CIG can be used to analyze downstream interaction dynamics in moderated discussions.

2 Related work

Measuring Informativeness in Conversation While definitions of informativeness vary across disciplines, they share a common conceptual core: the change a message induces in the recipient—whether in certainty (Hunt, 2003), utility (Glazer, 1993), or knowledge state (Meadow and Yuan, 1997). Particularly, dialogue evaluation has largely converged on two necessary conditions required to trigger this change: novelty (the presence of new signal) and relevance (the alignment of that signal to the goal context) (Ghosal et al., 2022; Maës et al., 2024a). However, these two dimensions alone are insufficient to capture the *magnitude* or *worthiness* of a contribution (Huang et al., 2020). While metrics like "impact" or "usefulness" attempt to proxy this third dimension, they are often vaguely defined or rely on subjective Likert scales that conflate personal preference with collective value (Lee et al., 2022; Qian et al., 2025). A complementary line of work extends classical information theory (Shannon, 1948), leveraging surprisal or information-density to trace information flow (Maës et al., 2022; Giulianelli et al., 2021; Tsipidi et al., 2024). However, such measures only partially align with human-perceived salience (Zarcone et al., 2016; Maës et al., 2022), limiting their precision in localizing information exchanges.

Agent Semantic Memory Agent memory modules provide a way to measure knowledge acquisition during conversation (Zhang et al., 2024). While originally developed to sustain coherence and personalization across long conversation sessions, these modules function by maintaining a selective, persistent state. In frameworks such as Mem0 (Chhikara et al., 2025), salient claims are extracted from utterances and an LLM applies update operations to a knowledge store. This approach improves temporal and multi-hop reasoning on long-dialogue benchmarks (Maharana et al., 2024) and reduces latency and token cost relative to full-history baselines.

3 Conversation Information Gain (CIG)

Following previous studies in group discussion analysis (Dowell et al., 2019; Maës et al., 2024b), we define **Conversational Information Gain (CIG)** as *how much a response advances the group's shared understanding of the topic or progress toward the goal, given both prior knowledge and the preceding dialogue*. We decompose

Aspect	Label	Definition / Anchor
Conversational Information Gain	1. No gain	Repeats or obstructs; no meaningful advance beyond the existing knowledge.
	2. Minimal gain	Small clarification or slight nuance that is noticeable but limited.
	3. Incremental	Adds new details/mechanisms within the same conceptual frame or ideas within the topic.
	4. Insightful	Reframes or introduce new ideas under the topic; shifts the conversation in a new valuable way.
Novelty	1. Not novel	Repetition/paraphrase of prior content or non/common-sense content.
	2. Minimally novel	Minor or mostly predictable detail added to an existing idea.
	3. Moderately novel	New evidence, concrete example, or supporting detail expanding an existing idea.
	4. Highly novel	New framework, principle, idea, or line of reasoning that opens a new direction.
Relevance	1. Not relevant	Off-topic; no connection to the conversation goal.
	2. Minimally relevant	Loose or indirect link; requires inference to connect.
	3. Moderately relevant	Substantially related but not central (e.g., side issue or counterpoint).
	4. Highly relevant	Directly and explicitly addresses the core topic or goal.
Implication Scope	1. Local	Manages the immediate moment; implication limited to participants/procedures.
	2. Bounded	Self-contained fact, feeling, or stance; no generalization beyond the case.
	3. Generalizing	Inductively generalizes a case or evidence to a broader audience.
	4. Universal	States an abstract principle, value, or norm with wide or universal applicability.

Table 1: Conversational Information Gain (CIG) rubric. Each aspect is rated on a 1–4 scale using Prior Knowledge and the preceding dialogue as the collective knowledge context.

CIG into three aspects—**Novelty**, **Relevance**, and **Implication Scope**—to connect general criteria for informativeness to the demands of public deliberation. Novelty and Relevance capture whether a contribution introduces new information and tied to the discussion goal, while Implication Scope captures how broadly the contribution's implications extend beyond the immediate case and people, reflecting the public orientation of deliberation. All constructs are rated on a four-level rubric (Table 1). Importantly, CIG reflects informational advancement rather than conversational quality per se—reiterations or coordination moves typically score low on CIG, even though they may still support the discussion indirectly.

To anchor high CIG levels, we draw on Chi (2009)'s typology of conceptual change which distinguishes *assimilation*—adding information within an existing mental model—from *accommodation*—restructuring the model itself. Accordingly, we separate our top two CIG levels: Level 3 (*Incremental*) reflects assimilation, where an utterance adds evidence, details, or mechanisms within the current framing; Level 4 (*Insightful*) reflects accommodation, where an utterance introduces a new framing or principle that qualitatively redirects the discussion. Detailed definitions for each aspect are in Table 1, and examples in Appendix Table 12.

Novelty *Assesses whether the information is new compared to the prior knowledge and preceding dialogue*. Traditional approaches often proxy novelty with n-gram overlap against prior text (Soboroff and Harman, 2005), but recent work shows that lexical novelty correlates poorly with human perception (Saakyan et al., 2025). Consequently, we rate novelty on a similar four-level scale as CIG, with levels reflecting the message's effect on the existing conceptual framing. The novelty score is intentionally independent of topical alignment and magnitude of impact.

Relevance *Measures how substantively a message relates to the main conversation topic or goal.* Since lexical-overlap proxies correlate weakly with human judgments in context-sensitive and implicit dialogue (Yeh et al., 2021; Dascal, 1977), we adopt the four-level (1–4) Topic–Conversation Relevance scale (Fan et al., 2024). Levels 1 (off-topic) and 4 (directly on-topic) form the endpoints; the key distinction is between the middle tiers. Level 2 (*minimally relevant*) covers content whose connection to the goal is indirect and requires a bridging inference (e.g., "school choice" for topic "housing affordability"), and Level 3 covers content that is clearly connected but not central, typically a recognized subtopic (e.g., "zoning restrictions").

Implication Scope *Measures the intended reach of a statement—who it is meant to matter to and how far its implications generalize.* Many dialogue-evaluation schemes include an "impact" dimension (Lee et al., 2022). For public deliberation, where perceived impact is subjective, we operationalize the concept—drawing on attributes like "generality" from Meadow and Yuan (1997)—by measuring the scope of entities a statement implicates. This reflects the public or community orientation of the discussion. In our four-level rubric, Level 1 denotes local or procedural contributions whose

significance is limited to the immediate participants (e.g., turn management). Level 2 captures bounded, case-specific content (e.g., a personal fact) not intended to generalize. Level 3 marks inductive generalization, where a specific fact is used to convey a broader pattern—such as a personal story evidencing social harm (Kessler et al., 2023). Level 4 denotes universal, principle-level claims aimed at wide public applicability. We emphasize that this hierarchy is descriptive rather than normative: effective deliberation requires a rhythm between grounded personal testimony (Level 2) and abstract principles (Level 4); thus, a higher score does not inherently imply a "better" contribution.

4 Entailment Based Semantic Memory

To automatically and incrementally score the CIG of conversational utterances, we require a dynamic representation of what the group has already established. To this end, we extend the Mem0 framework (Chhikara et al., 2025) to a multi-party setting, maintaining an evolving semantic memory that tracks how each speaker introduces and revises claims over time. As shown in Figure 1, the memory is managed through a two stage process of claim extraction and consolidation to provide the knowledge context for evaluating Novelty, Relevance, Implication Scope, and overall CIG.

Claim Extraction The first stage uses an LLM to decompose each utterance into a set of discrete, self-contained claims, converting context-dependent dialogue into a structured representation for semantic memory consolidation and storage (Appendix Table 13 details the prompt). To make each claim interpretable in isolation, the extractor uses local context to resolve ambiguities (e.g., coreference) and splits compound statements into atomic propositions, while normalizing away conversational fillers and hedging to recover the core propositional content (e.g., "I feel that maybe the transport cost is too high" $\rightarrow$ "The transport cost is too high"). Although this normalization removes epistemic markers, it enables more reliable NLI-based entailment checks; we preserve speaker-specific beliefs by storing each extracted claim together with its source speaker (and inferred addressee) in the memory state.

Multi-Party Memory Consolidation To manage redundancy and track knowledge evolution, each newly extracted claim A is compared against the memory store by first retrieving the top-k semantically similar candidates $\{B_i\}_{i=1}^{k}$ via vector search. An LLM-based NLI judge (bi-directionally) evaluates the claim and each candidate, mapping each pair to one of five relations: *equivalent*, *forward_entail*, *backward_entail*, *contradiction*, or *neutral*.

We then apply a deterministic policy to select an action $\alpha(A) \in \{\text{ADD}, \text{UPDATE}, \text{NONE}\}$. To strictly track speaker-specific belief trajectories without collapsing inter-speaker disagreements, we restrict operations to the same-speaker subset $\mathcal{S}(A) = \{B_i \mid \text{speaker}(B_i) = \text{speaker}(A)\}$. If $\mathcal{S}(A) = \emptyset$ or contains only *neutral* relations, we trigger ADD and insert A into the semantic memory; this ensures that parallel framings or contested claims from other participants are preserved as distinct entries. For $\mathcal{S}(A) \neq \emptyset$, *equivalence* or *backward_entailment* triggers NONE (taking priority over updates), while *contradiction* or *forward_entailment* triggers UPDATE on the single most similar item $B^\star = \arg\max_{B \in \mathcal{S}(A)} \text{sim}(A, B)$. Appendix Table 11 summarizes this mapping with examples, and Appendix Table 14 details the prompt.

	FORA	**INSQ**
Session-level (10 per setting, avg $\pm$ std)		
Utterances / Session	210.1 $\pm$ 56.5	205.8 $\pm$ 73.8
Words / Utterance	38.8 $\pm$ 11.9	36.4 $\pm$ 17.9
Speakers / Session	6.8 $\pm$ 1.0	8.8 $\pm$ 1.8
Annotated Segments (4 per session, avg $\pm$ std)		
Utterances / Segment	11.6 $\pm$ 4.4	11.3 $\pm$ 4.7
Speakers / Segment	4.5 $\pm$ 1.3	4.4 $\pm$ 1.0
Words / Utterance (non-skip.)	86.3 $\pm$ 52.5	88.6 $\pm$ 56.0
Gini (non-skip.)	0.42 $\pm$ 0.16	0.21 $\pm$ 0.15
Skipped Tokens (%)	17.8 $\pm$ 21.0	9.4 $\pm$ 9.0

Table 2: Session and segment level descriptive statistics for annotated dialogues from *FORA* and *INSQ*. **non-skip** are computed over non-skipped utterances.

5 Data and Annotation

We annotated transcripts from two deliberative discussion settings: ten TV debates from **INTELLIGENCE SQUARED (INSQ)** (Zhang et al., 2016) and ten community discussion from the **FORA corpus (FORA)** (Schroeder et al., 2024) about Durham city community future vision. To mitigate annotator bias and potential harms, we manually curated this set by excluding topics of a highly sensitive or recently polarizing nature. While both settings feature a moderator and multi-party dis-

	CIG	Novelty	Relevance	Scope
INSQ	0.589	0.506	0.669	0.597
FORA	0.567	0.583	0.566	0.510

Table 3: Krippendorff's α by corpus for CIG (Informativeness) and its three aspects.

cussion (see Table 2 for descriptive statistics), they differ in atmosphere and intent. INSQ debates are competitive events organized around a motion and two teams, featuring domain experts where the moderator primarily coordinates turn-taking. FORA assemblies are collaborative, brainstorming meetings where facilitators often participate as peers alongside general community members.

5.1 Preprocessing

For INSQ, we use the dataset's predefined structure and analyze only the discussion phase. FORA exhibits a comparable three-part format—introduction, discussion, and conclusion—but these boundaries are not annotated, so we use GPT-5 (OpenAI, 2025) to detect them and manually verify. In both settings, *conclusions* are discarded, and *introductions* are used to initialize the initial knowledge in the memory modules. We then segment the *discussion* phase of each episode into sub-topical units using GPT-5 with a shifting window and majority vote (see Appendix A.1 for segmentation details).

For each segment, we generate a prior memory summary by retrieving relevant existing memories via semantic similarity and summarizing them using GPT-5. We developed the summarization prompt using author-annotated session as seed data (see Appendix A.3 for development details). To focus the annotation on participants' contributions, moderator turns and truncated short utterances (< 5 words) (e.g. "I want to-.") are marked as *skipped*, accounting for 9–18% of tokens (see Table 2). From each episode, we select four segments for annotation, balancing factors like estimated reading time, skip ratio, and segmentation confidence (see Appendix A.2 for the selection criteria).

5.2 Annotation Protocol

We recruited 88 annotators on Prolific (£10.60/hour). All participants were required to be native English speakers with at least a bachelor's degree. We ultimately dropped 10 participants due to consistently low agreement with peers.[2] Each assignment (approximately 40 minutes) included a tutorial and a prescreen test, followed by four sampled segments from the same session presented in chronological order.

We deployed two annotation variants over the same material: one collected a single score for overall CIG per utterance (on a scale of 1–4); the other collected three separate ratings for *Novelty*, *Relevance*, and *Implication Scope*, each on a scale of 1–4. The annotation interface showed the topic at the top, with long-term memory (the prior memories summary) and short-term context (three preceding utterances) in a left panel, and the current target utterance for rating in a right panel (see Appendix A.5 for the interface snippet and design). Inter-annotator agreement, reported in Table 3, is moderate to high across all aspects and settings. Appendix Figure 5 shows the normalized label distributions.

6 Validating the CIG Aspects

To test whether overall CIG is explained by the three proposed aspects, we fit ordinal regression models on the 80 annotated segments and compare different predictor sets against a word-count baseline (Table 4). AIC-based model selection suggests that Novelty and Relevance account for most of the predictive signal. In contrast, Implication Scope—which we included to capture a deliberation-specific intuition that contributions with broader public reach might be perceived as more informative—provides little reliable benefit. In INSQ, the best fit is achieved by removing Scope, indicating that including it slightly degrades predictive power; in FORA, the full model improves only marginally over the Scope ablation.

Novelty and Relevance behave as expected and explain most of the variance in perceived CIG. In contrast, Implication Scope contributes surprisingly little. This contradicts our original deliberation-motivated hypothesis that "public-reach" statements (broader generalizations, principles) would systematically read as more informative than local or case-specific remarks. Our data suggest that perceived information gain is not monotonic in scope: low-scope utterances can be

[2] Our quality control process involved recruiting two annotators per session first. If their mean quadratic weighted kappa (QWK) score was <0.25, we recruited a third annotator. We then identified and removed the outlier annotator, ensuring at least two reliable annotations per segment.

Corpus	Model	AIC (↓)
INSQ	Base (word count)	472.58
	Base+novelty	424.39
	Base+relevance	444.29
	Base+implication scope	445.92
	Base+3 (all aspects)	416.60
	Ablation(-implication scope)	**416.01**
	Ablation(-relevance)	420.21
	Ablation(-novelty)	438.26
FORA	Base (word count)	643.25
	Base+novelty	599.28
	Base+relevance	580.12
	Base+implication scope	637.83
	Base+3 (all aspects)	**564.06**
	Ablation(-implication scope)	564.53
	Ablation(-relevance)	601.09
	Ablation(-novelty)	581.83

Table 4: Ordinal regression model comparison using AIC (lower is better). Base uses only utterance length. Base+3 includes all three aspects. Ablation models remove one aspect from the full model.

Corpus	Aspect	GPT-5	Human LOO
INSQ	CIG	0.457±0.020	0.656±0.131
	Novelty	0.587±0.024	0.637±0.245
	Relevance	0.452±0.018	0.431±0.202
	Imp. Scope	0.529±0.021	0.562±0.162
FORA	CIG	0.520±0.027	0.631±0.141
	Novelty	0.556±0.017	0.599±0.200
	Relevance	0.414±0.015	0.446±0.163
	Imp. Scope	0.479±0.008	0.559±0.226

Table 5: Mean absolute error for **GPT-5** when given only the memory summary—the same information provided to human annotators—compared to the **Human LOO** baseline.

highly informative when they introduce concrete evidence, while high-scope utterances can be uninformative when they restate values or abstractions without adding substance. Rather than functioning as a driver of CIG, Scope appears to index how participants reason (case-based versus principle-based) and may be more appropriate as a descriptive dimension for analyzing deliberative styles.

7 Prediction of CIG and Heuristic Aggregation

Predicting CIG with GPT-5 We next test whether segment-level CIG annotation can be automated with GPT-5 under the same information conditions as our human annotators. For each target segment, GPT-5 receives (i) the identical memory-based prior summary, (ii) the same preceding utterances, and (iii) the same tutorial exemplars for annotator as few-shot demonstrations, and outputs the same ratings for overall CIG and its aspects (Appendix Tables 15 and 16 detail the prompts). To contextualize error, we report a human leave-one-out (LOO) baseline, computed as each annotator's MAE against the mean of the remaining annotators and averaged across annotators, which estimates the expected deviation of an individual annotator from group consensus. As shown in Table 5, GPT-5's MAE under this matched-context setup is comparable to—and in several cases lower than—the human LOO baseline, suggesting that GPT-5 can reproduce aggregate human judgments at least as well as a typical annotator when constrained to the same prior context (additional models are reported in Appendix Table 18).

We next examine how alternative prior-knowledge inputs reproduce the GPT-5_SUMMARY predictions, as a check on whether the memory-derived summary retains the information that most influences GPT-5's ratings. Specifically, we compare four conditions: the full preceding transcript (`GPT-5_full`), the retrieved memory items shown verbatim without summarisation (`GPT-5_memory`), only the three most recent utterances (`GPT-5_short_prior`), and no prior context beyond the topic (`GPT-5_no_knowledge`). As shown in Table 6, `GPT-5_full` and `GPT-5_memory` yield the smallest mean deviations from GPT-5_SUMMARY, while restricting or removing prior context produces substantially larger errors. Overall, this suggests that the memory-based summaries provide a compact yet largely faithful approximation of the information in the full history that drives GPT-5's CIG estimates.

Comparing memory-dynamics signals to heuristic proxies We compare several informativeness-related heuristic proxies used in prior work—word-based, surprisal-based, entity-based, and TF–IDF features—against our memory-dynamics signals (per-utterance counts of extracted claims and claim updates) by computing Pearson correlations with the human-annotated CIG soft label (mean rating across annotators). As shown in Table 7, memory dynamics are the strongest correlates: relevance-gated updates (*Memory changes (Relv$^-$)*) achieve the highest correlation ($|r| = 0.727$), followed by *Memory changes (Any)* ($|r| = 0.720$) and *Extracted claim count* ($|r| = 0.713$), outperforming all other proxies. Token surprisal (sum) and utterance length (tokens) form a mid-tier baseline (both $|r| = 0.675$), while entity-based signals are sub-

Corpus	Aspect	GPT-5_full	GPT-5_memory	GPT-5_short_prior	GPT-5_no_knowledge
INSQ	CIG	0.265±0.021	**0.265**±0.007	0.350±0.013	0.394±0.024
	Novelty	0.353±0.009	**0.321**±0.028	0.649±0.038	0.764±0.044
	Relevance	0.159±0.026	0.131±0.013	**0.125**±0.007	0.151±0.025
	Imp. Scope	0.165±0.011	0.156±0.014	**0.149**±0.011	0.155±0.016
	Mean	0.236±0.017	**0.218**±0.016	0.318±0.017	0.366±0.027
FORA	CIG	**0.207**±0.008	0.220±0.010	0.293±0.020	0.336±0.008
	Novelty	0.279±0.006	**0.259**±0.019	0.383±0.031	0.492±0.025
	Relevance	0.151±0.015	**0.146**±0.012	0.148±0.006	0.167±0.005
	Imp. Scope	0.211±0.027	0.203±0.007	**0.195**±0.010	0.206±0.010
	Mean	0.212±0.014	**0.207**±0.012	0.255±0.017	0.300±0.012
Overall mean (both corpora)		0.224±0.015	**0.213**±0.014	0.286±0.017	0.333±0.020

Table 6: MAE of GPT-5 variants evaluated against **GPT-5_summary** reference, across both corpora and all aspects. Bold indicates the best performer across all variants.

Feature	$\lvert r \rvert$ w/ CIG
Memory changes (Relv$^-$)	0.728
Memory changes (Any)	0.721
Extracted claim count	0.714
Memory changes (Info$^-$)	0.712
Memory changes (Novo$^-$)	0.703
TF–IDF sum	0.701
Length (tokens)	0.675
TF–IDF max	0.640
Novel word count	0.601
Memory changes (Scope$^-$)	0.592
Entity count	0.456
Novel entity count	0.428
Novel entity ratio	0.306
Specificity (mean IDF)	0.235

Table 7: Pearson correlation between each proxy feature and the CIG average from **human** annotations. "$^-$" denotes memory-change counts restricted to utterances where the corresponding aspect score > 2.

stantially weaker (e.g., *Entity count* $|r| = 0.463$). Because annotators were provided memory-derived prior context, memory dynamics may be advantaged; we therefore replicate the same correlation analysis using GPT-5 predictions produced under full-transcript context (GPT-5_FULL) and observe similar overall pattern (Appendix Table 19). Appendix B details the proxies equations.

Claim-Level Predictions to Utterance Impression Motivated by broader interest in semantic salience (e.g., Biggs et al., 2012), we conduct an exploratory analysis to see how claim-level qualities roll up to utterance-level impressions of CIG (Wei et al., 2022). First, we use GPT-5 to predict scores (1-4) for every extracted claim on all three aspects (Novelty, Relevance, and Scope; see Appendix Table 17 for prompt details). We then use a two-step, unsupervised procedure to test how well these claim-level predictions can recover the human-annotated utterance-level CIG scores. This involves: (i) aggregating aspect-specific scores across multiple claims within an utterance (claim-aggregation on x-axis); (ii) aggregating the resulting aspect scores into a single CIG estimate (aspect-combination function on y-axis). We test a variety of operators for both steps.

The results, shown in Figure 2, reveal two consistent patterns. First, for the aspect-combination (y-axis), the 'min' operator performs best by a large margin across all claim-level aggregators (top row). This reveals a **"conjunctive bottleneck"**: an utterance's perceived informativeness is effectively gated by its weakest aspect. Second, for the claim-aggregation (x-axis), the best-performing operators are max pooling strategies like 'top-2 mean' (which achieves the single best MAE of 0.583) and 'top-quartile mean' (MAE = 0.627), when paired with the 'min' operator. Conversely, aspect-combination operators sensitive to high values, like 'softmax' and 'max', perform the worst.

8 Case Study: Moderator Dynamics

To demonstrate CIG's utility beyond static evaluation, we examine its capacity to diagnose interaction dynamics—specifically, the impact of moderation strategies on information flow. Facilitators play a crucial role in shaping deliberative quality, yet measuring the immediate informational "yield" of their interventions remains a challenge. To categorize these interventions, we employ the **WHoW taxonomy** (Chen et al., 2024), a domain-agnostic framework designed to analyze moderator behavior across varying contexts. WHoW classi-

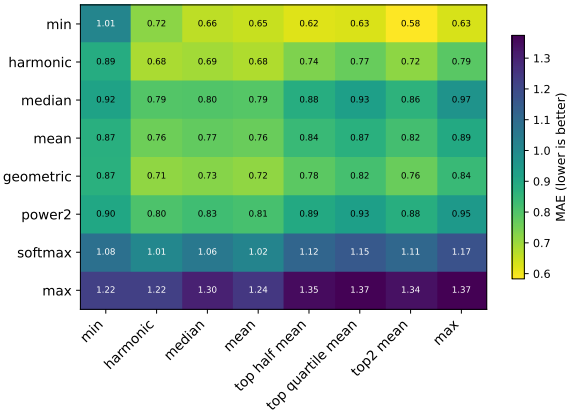

Figure 2: Heatmap of MAE for predicting human utterance-level CIG. The **y-axis** represents **Aspect Combination methods**. The **x-axis** represents **Claim Aggregation methods**.

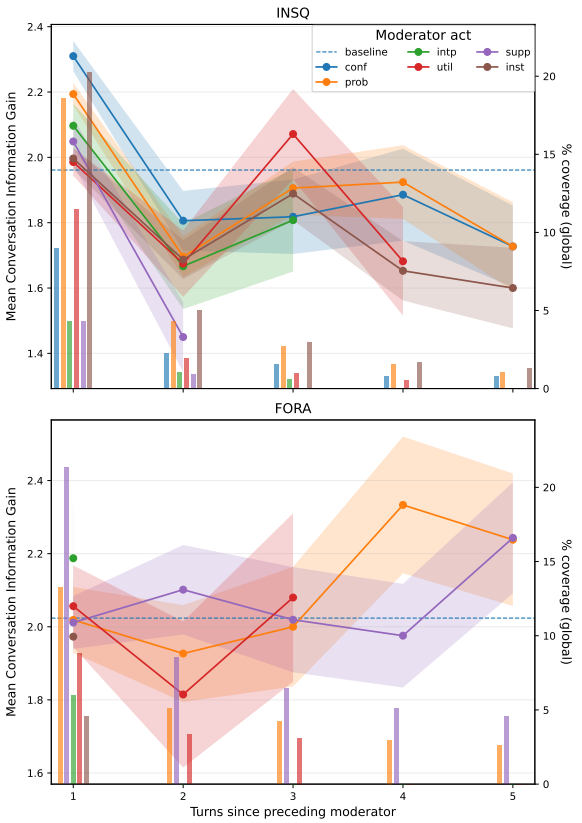

Figure 3: Mean speakers CIG (left y-axis) vs. turns since the last moderator act (x-axis), for both settings. Lines show the mean CIG for each moderator acts. Bars (right y-axis) show the global % coverage for each act. The dashed line is the corpus mean CIG.

fies moderator turns into six broad communicative acts: *Probing* (eliciting responses), *Confronting* (direct engagement between participants), *Interpretation* (interpretate previous content), *Supplementing* (contributing information), and *Utilities* (other functional acts). By mapping downstream CIG trajectories to these specific acts, we can isolate which strategies trigger informational gain and how these effects vary by context.

We apply CIG to assess how these interventions shape participants' CIG in our two corpora. We align each participant utterance to the most recent moderator turn, bucket outcomes by lag (1–5 turns), and stratify by the moderator's dialogue act. As Figure 3 shows, the two settings diverge significantly. In INSQ (top), moderators deploy a broader mix of acts—with confronting (`conf`) appearing most effective—and participant CIG typically peaks immediately after an intervention (lag 1). In FORA, peaks often emerge several turns later, suggesting a more organic buildup of informativeness among participants. Moderator moves in FORA are also less varied (notably, no `conf` acts are present) and include more supplement (`supp`) actions, indicating that facilitators in community settings often function as co-participants who contribute content directly.

9 Conclusions

We introduced CIG, a framework for measuring Conversational Information Gain in deliberative dialogue by focusing on informational progress rather than surface form. Grounded in the notion of information gain as a change in a knowledge state (Meadow and Yuan, 1997), our approach operationalizes CIG through a lightweight semantic memory that tracks the evolving set of claims introduced in a conversation. This memory is maintained via an LLM-based, NLI-guided mechanism, enabling each new utterance to be evaluated for Novelty, Relevance, Implication Scope, and overall CIG. We validated CIG on 80 segments from two diverse settings—INSQ debates and FORA community discussion—achieving moderate-to-high inter-annotator agreement.

Future work could integrate more recently advanced memory frameworks (Xu et al., 2025; Li et al., 2025) to strengthen consolidation and improve robustness. The NLI-based update mechanism may also support analyses of finer meaning-construction processes in dialogue (Poole-Dayan et al., 2025). More broadly, memory-based information gain signals could serve as interpretable supervision or even reinforcement learning reward signal for training or evaluating information-seeking and deliberation-support systems.

Limitations

Our study has several limitations. First, the CIG labels are collected under a constrained information condition (topic, a prior-knowledge summary, and a short preceding window), which may bias judgments toward what is captured by the summary rather than the full discourse; while we partially probe this with context ablations, the results may not generalize to settings where annotators have full access to long histories. Second, the semantic-memory pipeline relies on LLM extraction and NLI-style consolidation, so errors in claim parsing, retrieval, or entailment can propagate into memory dynamics and any downstream analyses. Third, due to efficiency constraints, each extracted claim is assigned only a single memory action, and at most one relation to one existing memory item is recorded, which can under-represent cases where a claim simultaneously relates to multiple prior claims. Finally, we study two moderated English-language corpora and exclude highly sensitive topics; broader coverage (languages, cultures, and adversarial/polarized domains) is needed to assess robustness and external validity.

Ethical considerations

We recruited crowd annotators via an online platform to rate the informativeness of short dialogue segments drawn from public deliberative settings. Participation was voluntary and based on informed consent; annotators were informed of their right to withdraw at any time without penalty and could opt out of rating specific segments after viewing the topic/context if they felt uncomfortable(Appendix D show anonymised version of the plain language statement for participation). Because public discussions can include sensitive or emotionally charged content, we limited topic selection and provided sufficient context to support informed skipping. We collected no direct personal identifiers beyond platform-managed participant IDs, and we release only de-identified annotations and processed transcripts. Compensation followed platform guidelines and was set to meet or exceed applicable minimum-wage expectations for estimated task duration. This study received approval from an institutional human research ethics review process (details withheld for anonymous review). We also acknowledged that the studies have used large language model, including Gemini-2.5-pro from google and GPt-5 from OpenAI, to assist writing refinement and research code debugging and data cleaning.

References

Atijit Anuchitanukul, Julia Ive, and Lucia Specia. 2022. Revisiting contextual toxicity detection in conversations. *ACM Journal of Data and Information Quality*, 15(1):1–22.

Adam T Biggs, Ryan D Kreager, Bradley S Gibson, Michael Villano, and Charles R Crowell. 2012. Semantic and affective salience: The role of meaning and preference in attentional capture and disengagement. *Journal of experimental psychology: human perception and performance*, 38(2):531.

David Bohm and Robert A Weinberg. 2004. *On dialogue*. Routledge.

Ming-Bin Chen, Lea Frermann, and Jey Han Lau. 2024. Whow: A cross-domain approach for analysing conversation moderation. *arXiv preprint arXiv:2410.15551*.

Prateek Chhikara, Dev Khant, Saket Aryan, Taranjeet Singh, and Deshraj Yadav. 2025. Mem0: Building production-ready ai agents with scalable long-term memory. *arXiv preprint arXiv:2504.19413*.

Michelene TH Chi. 2009. Three types of conceptual change: Belief revision, mental model transformation, and categorical shift. In *International handbook of research on conceptual change*, pages 89–110. Routledge.

Marcelo Dascal. 1977. Conversational relevance. *Journal of pragmatics*, 1(4):309–327.

John Dewey and Melvin L Rogers. 2012. *The public and its problems: An essay in political inquiry*. Penn State Press.

Nia MM Dowell, Tristan M Nixon, and Arthur C Graesser. 2019. Group communication analysis: A computational linguistics approach for detecting sociocognitive roles in multiparty interactions. *Behavior research methods*, 51(3):1007–1041.

Yaran Fan, Jamie Pool, Senja Filipi, and Ross Cutler. 2024. Topic-conversation relevance (tcr) dataset and benchmarks. *Advances in Neural Information Processing Systems*, 37:140159–140174.

Tirthankar Ghosal, Tanik Saikh, Tameesh Biswas, Asif Ekbal, and Pushpak Bhattacharyya. 2022. Novelty detection: A perspective from natural language processing. *Computational Linguistics*, 48(1):77–117.

Mario Giulianelli, Arabella Sinclair, and Raquel Fernández. 2021. Is information density uniform in task-oriented dialogues? In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Rashi Glazer. 1993. Measuring the value of information: The information-intensive organization. *IBM Systems Journal*, 32(1):99–110.

Alex Goddard and Alex Gillespie. 2023. Textual indicators of deliberative dialogue: A systematic review of methods for studying the quality of online dialogues. *Social Science Computer Review*, page 08944393231156629.

Jürgen Habermas. 1985. *The theory of communicative action: Volume 1: Reason and the rationalization of society*, volume 1. Beacon press.

Minlie Huang, Xiaoyan Zhu, and Jianfeng Gao. 2020. Challenges in building intelligent open-domain dialog systems. *ACM Transactions on Information Systems (TOIS)*, 38(3):1–32.

Darwin P Hunt. 2003. The concept of knowledge and how to measure it. *Journal of intellectual capital*, 4(1):100–113.

Daniel Kessler, Dimitra Dimitrakopoulou, and Deb Roy. 2023. Hearing personal experiences improves social evaluations compared to personal opinions, especially for polarized parties. *Especially for Polarized Parties (December 05, 2023)*.

Julia Kruk, Michela Marchini, Rijul Magu, Caleb Ziems, David Muchlinski, and Diyi Yang. 2024. Silent signals, loud impact: Llms for word-sense disambiguation of coded dog whistles. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12493–12509.

Mina Lee, Megha Srivastava, Amelia Hardy, John Thickstun, Esin Durmus, Ashwin Paranjape, Ines Gerard-Ursin, Xiang Lisa Li, Faisal Ladhak, Frieda Rong, et al. 2022. Evaluating human-language model interaction. *arXiv preprint arXiv:2212.09746*.

Zhiyu Li, Shichao Song, Chenyang Xi, Hanyu Wang, Chen Tang, Simin Niu, Ding Chen, Jiawei Yang, Chunyu Li, Qingchen Yu, et al. 2025. Memos: A memory os for ai system. *arXiv preprint arXiv:2507.03724*.

Eliot Maës, Philippe Blache, and Leonor Becerra-Bonache. 2022. Shared knowledge in natural conversations: can entropy metrics shed light on information transfers? In *26th Conference on Computational Natural Language Learning (CoNLL)*, pages 213–227.

Eliot Maës, Hossam Boudraa, Philippe Blache, and Leonor Becerra-Bonache. 2024a. Did you get it? a zero-shot approach to locate information transfers in conversations. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*.

Eliot Maës, Hossam Boudraa, Philippe Blache, and Leonor Becerra-Bonache. 2024b. Did you get it? a zero-shot approach to locate information transfers in conversations. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4877–4890, Torino, Italia. ELRA and ICCL.

Adyasha Maharana, Dong-Ho Lee, Sergey Tulyakov, Mohit Bansal, Francesco Barbieri, and Yuwei Fang. 2024. Evaluating very long-term conversational memory of llm agents. *arXiv preprint arXiv:2402.17753*.

Charles T Meadow and Weijing Yuan. 1997. Measuring the impact of information: defining the concepts. *Information processing & management*, 33(6):697–714.

Daniel John Montez and Pamela Jo Brubaker. 2019. Making debating great again: Us presidential candidates' use of aggressive communication for winning presidential debates. *Argumentation and Advocacy*, 55(4):282–302.

OpenAI. 2025. GPT-5 API. Large language model accessed via API.

Elinor Poole-Dayan, Deb Roy, and Jad Kabbara. 2025. An ai-powered framework for analyzing collective idea evolution in deliberative assemblies. *arXiv preprint arXiv:2509.12577*.

Ilan Price, Jordan Gifford-Moore, Jory Flemming, Saul Musker, Maayan Roichman, Guillaume Sylvain, Nithum Thain, Lucas Dixon, and Jeffrey Sorensen. 2020. Six attributes of unhealthy conversations. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 114–124.

Kun Qian, Maximillian Chen, Siyan Li, Arpit Sharma, and Zhou Yu. 2025. Bottom-up synthesis of knowledge-grounded task-oriented dialogues with iteratively self-refined prompts. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 827–844.

Arkadiy Saakyan, Najoung Kim, Smaranda Muresan, and Tuhin Chakrabarty. 2025. Death of the novel (ty): Beyond n-gram novelty as a metric for textual creativity. *arXiv preprint arXiv:2509.22641*.

Hope Schroeder, Deb Roy, and Jad Kabbara. 2024. Fora: A corpus and framework for the study of facilitated dialogue. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13985–14001.

Claude E Shannon. 1948. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423.

Ian Soboroff and Donna Harman. 2005. Novelty detection: the trec experience. In *Proceedings of human language technology conference and conference on empirical methods in natural language processing*, pages 105–112.

Marco R Steenbergen, André Bächtiger, Markus Spörndli, and Jürg Steiner. 2003. Measuring political deliberation: A discourse quality index. *Comparative European Politics*, 1:21–48.

Eleftheria Tsipidi, Franz Nowak, Ryan Cotterell, Ethan Wilcox, Mario Giulianelli, and Alex Warstadt. 2024. Surprise! uniform information density isn't the whole story: Predicting surprisal contours in long-form discourse. *arXiv preprint arXiv:2410.16062*.

Henning Wachsmuth, Gabriella Lapesa, Elena Cabrio, Anne Lauscher, Joonsuk Park, Eva Maria Vecchi, Serena Villata, and Timon Ziegenbein. 2024. Argument quality assessment in the age of instruction-following large language models. *arXiv preprint arXiv:2403.16084*.

Douglas Walton. 2003. A pragmatic theory of fallacy.

Douglas Walton. 2008. *Informal logic: A pragmatic approach*. Cambridge University Press.

Wenqing Wei, Sixia Li, and Shogo Okada. 2022. Investigating the relationship between dialogue and exchange-level impression. In *Proceedings of the 2022 International Conference on Multimodal Interaction*, pages 359–367.

Wujiang Xu, Kai Mei, Hang Gao, Juntao Tan, Zujie Liang, and Yongfeng Zhang. 2025. A-mem: Agentic memory for llm agents. *arXiv preprint arXiv:2502.12110*.

Yi-Ting Yeh, Maxine Eskenazi, and Shikib Mehri. 2021. A comprehensive assessment of dialog evaluation metrics. *arXiv preprint arXiv:2106.03706*.

Alessandra Zarcone, Marten Van Schijndel, Jorrig Vogels, and Vera Demberg. 2016. Salience and attention in surprisal-based accounts of language processing. *Frontiers in psychology*, 7:844.

Justine Zhang, Ravi Kumar, Sujith Ravi, and Cristian Danescu-Niculescu-Mizil. 2016. Conversational flow in oxford-style debates. *arXiv preprint arXiv:1604.03114*.

Zeyu Zhang, Xiaohe Bo, Chen Ma, Rui Li, Xu Chen, Quanyu Dai, Jieming Zhu, Zhenhua Dong, and Ji-Rong Wen. 2024. A survey on the memory mechanism of large language model based agents. *arXiv preprint arXiv:2404.13501*.

A Implementation details

A.1 Dialogue segmentation implementation

We segment each discussion into coherent topical units using an LLM-based boundary detector. Given the topic and prior context, the model proposes a list of segment intervals over utterance indices (please refer to Table 8 for the prompt), targeting segments of roughly 500–800 words and no more than 20 turns. To improve robustness, we run the segmenter $p{=}5$ times with diversified decoding and convert each run into a set of proposed breakpoints (segment starts). We then perform **weighted majority voting** over candidate breakpoints: for each proposed start index b, we add a vote of 1.0 to b and 0.5 to its immediate neighbors $(b{-}1, b{+}1)$, which makes the voting tolerant to small off-by-one differences across runs. Aggregating and normalizing votes yields a breakpoint confidence profile over utterances. Final breakpoints are selected as local peaks exceeding a threshold while enforcing minimum/maximum segment-length constraints, with a fallback to the highest-scoring breakpoint only when a cut is required to satisfy the maximum length. Finally, we merge adjacent segments when doing so reduces deviation from the desired word-count range (default 450–750 words) without exceeding an utterance cap, yielding a small set of well-formed segments per session.

Section	Prompt part (abridged)
Role & task	You are an expert dialogue analyst. Segment the interaction phase into coherent subtopic segments of **450–700 words** or $\leq$**20 speaker turns** each.
Inputs (template vars)	Topic/Goal: `{topic}` Prior dialogue summary: `{prior_summary}`
Instruction	Identify subtopics and boundaries for *all* utterances in the provided dialogue. Return **only** valid JSON matching the registered schema (no extra text). `segment_index` starts from 0.
Output format	`[{segment_index:int, utterances_interval:[int,int], segment_subtopic:str}, . . . ]`
Field definitions	`segment_index`: 0-based segment id. `utterances_interval`: `[start_idx, end_idx]` over utterance indices. `segment_subtopic`: short description of the segment's subtopic.

Table 8: Dialogue segmentation prompt and output schema. Curly-brace fields (e.g., `{topic}`) denote template variables filled at runtime.

A.2 Segment selection for annotation

To keep the human annotation workload feasible and affordable while still sampling informative portions of each episode, we select $k{=}4$ segments per conversation using a constrained, score-based procedure implemented in our task-generation script. For every candidate segment, we first filter the target utterances by marking moderator/audience turns as *skipped*, and also skipping truncated or fragmentary utterances (e.g., $\leq$5 words without terminal punctuation, or $\leq$3 words), since these are most likely low in informativeness. We then estimate the total reading time of the task by summing (i) the segment utterances, (ii) the generated prior-context summary shown to the annotator, and (iii) a short window of immediately preceding dialogue (up to $K{=}5$ turns) that fits within a fixed reading-time budget; the prior-history window is truncated if adding more context would exceed the budget.

Next, we compute a *segment quality* vector that balances (a) task feasibility and (b) expected annotation value. Feasibility terms include deviation from a target number of non-skipped utterances (to avoid segments that are too sparse or too dense), deviation from the reading-time budget, and the segmenter's confidence (mean boundary confidence). Expected value is approximated using the *context benefit* of the chosen summary type: for the same segment, we compare GPT-5's CIG/aspect scoring consistency under the selected context (e.g., `memory_summary`) versus a `no_summary` baseline, using the segment's stored `summary_scores`; segments where the summary yields a larger improvement are preferred because they are more sensitive to having an accurate prior-knowledge representation. Finally, we rank segments within each conversation using these features (via `rank_segments`) and take the top $k{=}4$ ranked segments, while retaining metadata such as skip ratios, participant count, and reading-time breakdown for reporting and audit.

A.3 Memory-Based Summarisation

Scoring CIG requires a *knowledge context* that reflects what has already been established in the discussion. Providing the full preceding transcript as context is often impractical: it is token-expensive for LLM-based scoring and cognitively burdensome for human annotators, who must judge informativeness relative to what the group already knows. While our memory module can retrieve semantically relevant prior claims, presenting the retrieved claim list directly to annotators can also be overwhelming and fragmentary, especially in long multi-party episodes. We therefore summarise the prior context into a short, readable prior-knowledge passage that (i) reduces prompt length for automated scoring and (ii) makes long episodes digestible for annotation without discarding the key claims and points of contention needed to judge Novelty, Relevance, Implication Scope, and overall CIG.

To develop and validate this summarisation procedure, we implemented multiple prior-context variants (direct transcript-based summary; prior-summary recursion; theme-aware decomposition; and a memory-based summary conditioned on retrieved semantic memories). We then ran a controlled comparison on one fully author-annotated seed episode (198 utterances): for each segment, we generated each context variant, asked GPT-5 to score overall CIG and the three aspects un-

Section	Prompt part (abridged)
Role & task	You are an expert dialogue analyst. Produce a coherent, highly readable summary of the prior context that is useful for interpreting the current segment.
Inputs (template vars)	Topic: `{topic}` Prior context: `{prior_dialogue}` *or* `{prior_summary}` *or* `{formatted_memories}` Current segment: `{current_dialogue}`
Length & style constraints	Plain prose; no bullet points. Target length: ∼250 words (bounded range enforced by the prompt). Summary must begin with: `"The prior conversation..."`
Faithfulness constraint	Use only information available in the provided prior context (no hallucination).
Output & formatting	Return **only** a JSON object (no extra text, no markdown): `{"summary":"<two-paragraph summary>"}`
Variants (prior-context representation)	**Direct summary:** condition on `{prior_dialogue}`. **Recursive summary:** update `{prior_summary}` using `{current_dialogue}`. **Theme-aware summary:** extract salient themes from `{current_dialogue}`, then summarise relevant parts of `{prior_dialogue}`. **Memory-based summary:** condition on retrieved memories `{formatted_memories}`.

Table 9: Summarisation prompt template (abridged) and its four variants, which differ only in how the prior context is represented (full transcript, recursive prior summary, theme-aware decomposition, or retrieved semantic memories). Curly-brace fields (e.g., `{topic}`) denote template variables filled at runtime.

Context variant	Aspect	Mean	SD	MSE↓	κ ↑
Human (reference)	imsc	2.292	1.179	–	–
	info	1.876	0.932	–	–
	novo	2.022	0.916	–	–
	relv	3.044	1.158	–	–
No summary	imsc	2.449	1.094	0.916	0.649
	info	2.256	1.156	1.044	0.556
	novo	2.310	1.182	1.178	0.492
	relv	3.157	1.244	1.018	0.649
Prior-summary only	imsc	2.464	1.130	0.952	0.647
	info	2.092	1.054	0.756	0.628
	novo	1.993	1.096	0.854	0.582
	relv	3.197	1.204	1.088	0.613
Direct summary	imsc	2.408	1.100	0.906	0.654
	info	2.088	1.024	0.686	0.651
	novo	1.974	1.074	0.800	0.600
	relv	3.237	1.162	0.953	0.650
Aspect-aware summary	imsc	2.398	1.108	0.850	0.677
	info	2.033	1.030	0.616	0.684
	novo	1.912	1.063	0.744	0.625
	relv	3.146	1.238	0.913	0.684
Memory-based summary	imsc	2.343	1.084	0.847	0.671
	info	2.010	1.016	0.595	0.690
	novo	1.916	1.072	0.785	0.607
	relv	3.146	1.241	0.956	0.671
Full transcript (LLM)	imsc	2.274	1.104	0.821	0.686
	info	1.985	1.025	0.548	0.716
	novo	1.912	1.046	0.730	0.624
	relv	3.102	1.228	0.971	0.660

Table 10: Summary-context ablation on one author-annotated seed session. "Mean"/"SD" report the rating distribution under each context variant; MSE and κ measure agreement against the author ratings (human reference row has no agreement scores). Lower MSE and higher κ indicate closer alignment.

der that context, and compared these predictions against the author labels. Table 10 shows that weak-context baselines (e.g., *no_summary* and *prior_summary*) yield substantially higher error and lower agreement—most notably for *info* and *novo*—whereas memory-based summaries consistently improve alignment (lower MSE, higher κ) while remaining comparable to transcript-dependent summaries. This result motivates our use of retrieved-memory summaries as the default prior context for segment-level CIG scoring.

A.4 Memory extraction and consolidation

To build a compact, queryable representation of what has been established so far in a multi-party discussion, we use a two-stage memory module: (i) *atomic claim extraction* from each target utterance, and (ii) *memory consolidation* that decides how each new claim should affect the evolving memory state. This design separates *content decomposition* (turn → propositions) from *state management* (propositions → persistent memory), allowing the downstream summarisation and CIG scoring components to operate over a stable set of self-contained claims rather than raw transcript text.

Atomic claim extraction. The extractor prompt (Table 13) instructs the model to act as an **Atomic-Fact Extractor** and output a list of *atomic*, *self-contained*, and *semantically distinct* propositions for a given target utterance. We require claims to be understandable without the original dialogue by resolving pronouns and deictic references using the provided context, but we constrain the extractor to *extract only from the target utterance* (context is used only for disambiguation). To avoid superficial surface variation, the prompt explicitly discourages paraphrase duplicates and speech-act descriptions (e.g., "asks/suggests"), and removes hedges and filler language unless a speech verb is needed for reported speech. The extractor returns only JSON in a fixed schema—a list of proposition objects with `speaker`, `target_speaker`, `claim`, and `turn_id` fields—and is capped at 30 claims per utterance to prioritize salience and keep the memory growth controlled.

Memory consolidation via speaker-aware NLI. Given the existing memory state and a set of newly extracted claims, the consolidator prompt (Table 14) updates memory using a deterministic, speaker-aware procedure grounded in bidirectional natural language inference (NLI). For each new claim A, we first retrieve candidate existing memories B that are semantically similar (prioritizing same-speaker items), then classify the logical relationship between A and B in both directions ($A \Rightarrow B$ and $B \Rightarrow A$). The bidirectional NLI outcomes are mapped to a compact relation set (*equivalent*, *forward_entail*, *backward_entail*, *contradiction*, *neutral*), and a strict target-selection ladder chooses *exactly one* eligible B (or none) to ensure consistent, auditable behavior.

ADD/UPDATE/NONE decisions and update semantics. The consolidator then maps each relation to one of three actions: ADD when a claim is novel (*neutral*) or comes from a different speaker (to preserve multi-party disagreement), NONE when a claim is redundant (*equivalent* or *backward_entail*), and UPDATE when a claim refines

NLI Relation	Action	Why	Example (A=new, B=existing)
equivalent	NONE	A $\Leftrightarrow$ B; duplicate, keep older	**A**: "I have a cat" vs **B**: "My pet is a cat"
forward_entail	UPDATE	A $\Rightarrow$ B; A refines/extends B (merge)	**A**: "I have a black cat." entails **B**: "I have a cat."
backward_entail	NONE	B $\Rightarrow$ A; A is weaker/redundant	**A**: "I have a cat." is entailed by **B**: "I have a black cat."
contradiction	UPDATE	A $\perp$ B; correct/replace B with A	**A**: "I have a black cat." vs **B**: "I have a white cat."
neutral	ADD	No strong link; new topic/info	**A**: "Decision deadline is next Monday." (no related B)

Table 11: Memory Consolidator rules: Mapping from NLI relation to memory action.

or corrects an existing one (*forward_entail* or *contradiction*). Updates are defined explicitly: contradictions replace the target memory, while forward entailment merges A and B into a more specific proposition. The module outputs a JSON list of `memory_updates` containing the action, the inferred logical relation, the source claim, and the selected target memory (or `null`), enabling downstream components to reconstruct the memory timeline and compute memory-dynamics features.

A.5 Annotation interface design

Figure 4 shows our web-based annotation interface, designed to make CIG judgments explicitly relative to shared prior context. The UI is organized into two synchronized panels. The left panel presents the *Prior Knowledge* for the current segment as a compact prior-discussion summary with 3 preceding utterances, while the right panel presents the *Target Utterances* one at a time, including the utterance ID, speaker name, and stance badge (e.g., *for*/*against*). For each utterance, annotators provide ordinal ratings for Novelty, Relevance, and Implication Scope using fixed four-level radio scales, and can navigate through utterances via *Previous/Next* controls with a progress indicator (e.g., "Utterance 1 of 6"). To reduce superficial scoring and encourage deliberate grounding in context, the right-hand rating panel is locked for the first 60 seconds whenever a new segment is loaded, forcing annotators to read the prior-knowledge summary before entering labels. In addition, the interface highlights keywords that appear in both the prior summary and the current utterance; clicking a highlighted keyword automatically scrolls the prior-knowledge panel to the corresponding mention, supporting rapid cross-referencing and helping annotators verify whether an utterance is truly novel or merely restating earlier claims.

B Informativeness-related proxy definitions

This appendix defines the informativeness-related proxy features used in our correlation analysis (Table 7). To avoid equation overflow in ACL two-column format, we use compact symbols in math and refer to implementation variable names (e.g., `token_count`) in text.

B.1 Length and volume

Length (tokens). Let T_i be the spaCy token sequence for utterance i (spaces excluded; punctuation retained). Utterance length (`token_count`) is:

$$n_i^{\text{tok}} = |T_i|. \tag{1}$$

Content-word volume. Let C_i be the multiset of *content* lemmas in utterance i, obtained by filtering spaCy tokens to alphabetic tokens that are not stopwords, numerals, or punctuation, and then lemmatising (fallback to lowercased surface form when lemma is unavailable). Content-token volume (`content_token_count`) is:

$$n_i^{\text{cont}} = |C_i|. \tag{2}$$

B.2 TF–IDF and lexical specificity (episode-level)

TF–IDF (episode-level). We fit a `TfidfVectorizer` *per episode* (utterance-as-document), using the content-lemma token list as the analyzer. For utterance i, let $W_i = \{w_{ij}\}$ denote its nonzero TF–IDF weights. We compute:

$$\begin{aligned} s_i^{\text{tfidf}} &= \sum_j w_{ij}, \\ m_i^{\text{tfidf}} &= \max_j w_{ij}, \\ \mu_i^{\text{tfidf}} &= \frac{1}{|W_i|} \sum_j w_{ij}. \end{aligned} \tag{3}$$

These correspond to `tfidf_sum`, `tfidf_max`, and `tfidf_mean`. Implementation: `norm=None`, `min_df=2`, `max_df=0.95`.

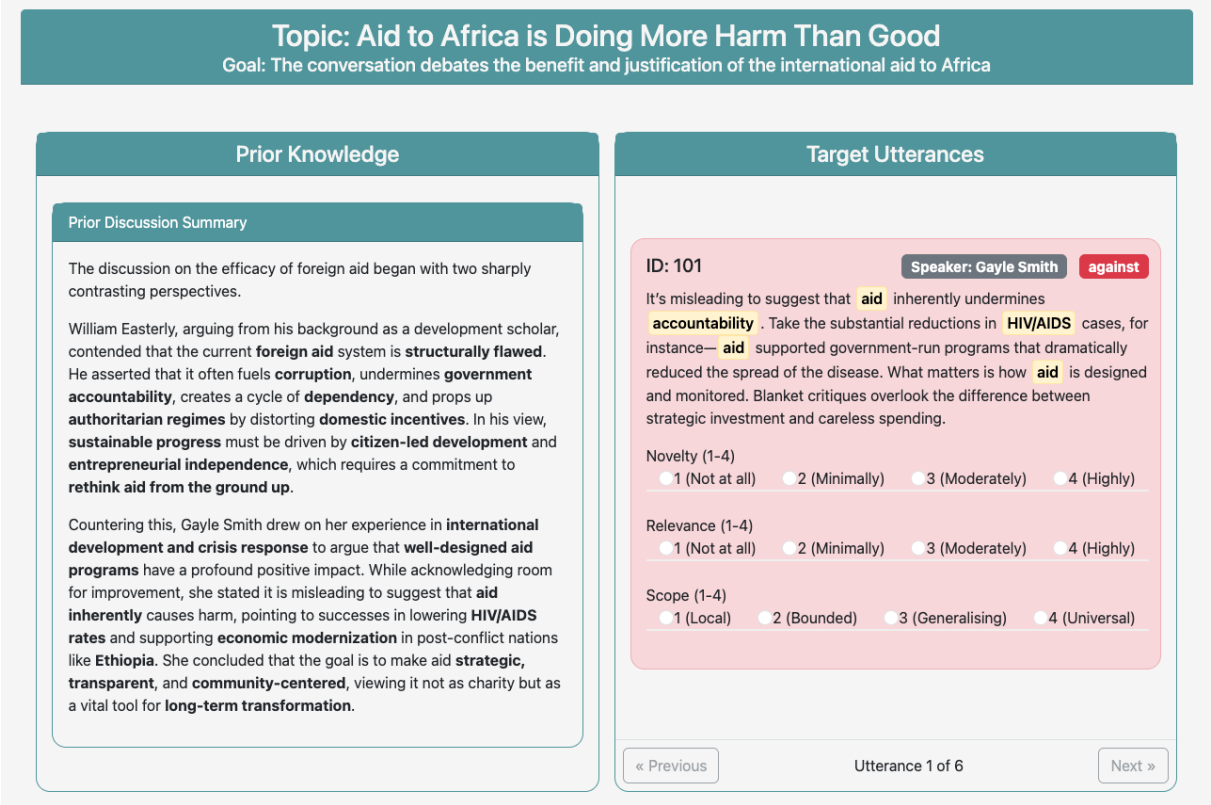

Figure 4: Annotation interface for segment-level CIG aspect rating. The left panel presents the prior-knowledge summary for the current segment, and the right panel displays target utterances with per-utterance ratings for Novelty, Relevance, and Implication Scope. The UI highlights overlapping keywords and supports click-to-scroll cross-referencing.

Lexical specificity (IDF). Let $\texttt{idf}_j$ be the episode-level IDF from the fitted vectorizer, and let J_i be indices of the terms present in utterance i (i.e., nonzero TF–IDF entries). We compute:

$$\mu_i^{\text{idf}} = \frac{1}{|J_i|} \sum_{j \in J_i} \texttt{idf}_j, \qquad \tilde{\mu}_i^{\text{idf}} = \text{median}(\{\texttt{idf}_j\}_{j \in J_i}). \tag{4}$$

These correspond to `specificity_mean_idf` and `specificity_median_idf`. If $J_i = \emptyset$, we set the specificity values to 0.

B.3 Lexical and entity novelty (cumulative)

Word novelty (cumulative). Let $V_{<i}$ be the set of content lemmas observed earlier in the same episode, and let L_i be the set of content lemmas in utterance i (derived from C_i). Define $N_i = L_i \setminus V_{<i}$ as the set of novel lemmas introduced at utterance i. We compute:

$$n_i^{\text{new}} = |N_i|, \qquad \rho_i^{\text{new}} = \frac{n_i^{\text{new}}}{\max(1, n_i^{\text{cont}})}. \tag{5}$$

These correspond to `novel_word_count` and `novel_word_density`. We also track cumulative debug counters such as `seen_vocab_size_so_far`$= |V_{<i}|$.

Entity novelty (cumulative). Let E_i be the multiset of named-entity surface forms in utterance i (lowercased), $U_i = \text{set}(E_i)$ the set of unique entity forms in the utterance, and $S_{<i}$ the set of entity forms observed earlier in the episode. We compute:

$$n_i^{\text{ent}} = |E_i|, \qquad n_i^{\text{ent-new}} = |U_i \setminus S_{<i}|, \qquad \rho_i^{\text{ent-new}} = \frac{n_i^{\text{ent-new}}}{\max(1, |U_i|)}. \tag{6}$$

These correspond to `entity_count`, `novel_entity_count`, and `novel_entity_ratio`. We use spaCy NER labels: `PERSON`, `ORG`, `GPE`, `LOC`, `NORP`, `EVENT`, `WORK_OF_ART`, `LAW`, `PRODUCT`, `FAC`, `LANGUAGE`. We additionally track token-normalized entity novelty:

$$\rho_i^{\text{ent-new-tok}} = \frac{n_i^{\text{ent-new}}}{\max(1, n_i^{\text{tok}})}, \tag{7}$$

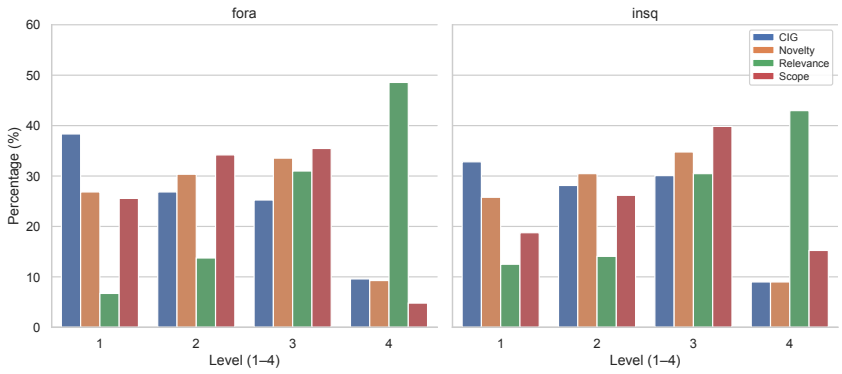

Figure 5: Normalized distributions of annotation levels (1–4) per aspect, and by corpus.

corresponding to `novel_entity_density_token`.

B.4 LM surprisal and predictability

LM surprisal / entropy. For a causal language model, token surprisal (bits) is $s_t = -\log_2 p(w_t \mid w_{<t})$. Let T be the number of evaluated tokens (excluding special tokens; determined via `attention_mask` and `start_idx`). We compute average and summed surprisal:

$$\bar{s}_i = \frac{1}{T}\sum_{t=1}^{T} s_t, \qquad S_i = \sum_{t=1}^{T} s_t = T \cdot \bar{s}_i. \tag{8}$$

These correspond to `sent_avg_h` and `sum_h`. We also compute a length-normalized variant:

$$\bar{s}_i^{\text{norm}} = \frac{\bar{s}_i}{\bar{h}(T)}, \tag{9}$$

where $\bar{h}(T)$ is the mean `sent_avg_h` among utterances of length T (`norm_sent_avg_h`).

Top-quartile token predictability. The pipeline also stores per-token log-probabilities $\ell_t = \log_2 p(w_t \mid w_{<t})$ in `tokens_h` (negative values). Let $\mathcal{Q}_i$ be the indices of the top 25% of tokens in utterance i ranked by ℓ_t (i.e., the most predictable tokens). We compute:

$$q_i = \frac{1}{|\mathcal{Q}_i|}\sum_{t\in\mathcal{Q}_i} \ell_t, \tag{10}$$

corresponding to `top_quatile_avg_ent`. Note that despite the variable name, this aggregates *log-probabilities* rather than surprisals; higher values (less negative) indicate more predictable tokens.

B.5 Memory dynamics

Memory dynamics (counts). Let $\mathcal{A}_i$ be the set of memory-update actions aligned to utterance i (from segment-level `memory_actions`). We define:

$$n_i^{\text{claim}} = |\mathcal{A}_i|, \qquad \Delta_i^{\text{mem}} = \sum_{a\in\mathcal{A}_i} \mathbb{I}[\texttt{event}(a) \neq \texttt{NONE}]\,. \tag{11}$$

These correspond to `claim_count` and `mem_delta`.

Aspect-gated memory changes. Let $\hat{y}_{i,d} \in \{1,2,3,4\}$ be the predicted aspect rating for utterance i and aspect d. For $d \in \{\texttt{info}, \texttt{novo}, \texttt{relv}, \texttt{imsc}\}$, we gate memory changes by requiring $\hat{y}_{i,d} > 2$:

$$\Delta_{i,d}^{\text{mem}} = \sum_{a\in\mathcal{A}_i} \mathbb{I}[\texttt{event}(a) \neq \texttt{NONE}]\; \mathbb{I}[\hat{y}_{i,d} > 2]\,. \tag{12}$$

The triad-gated variant additionally requires `novo`, `relv`, and `imsc` all exceed the midpoint:

$$\Delta_{i,\text{triad}}^{\text{mem}} = \sum_{a\in\mathcal{A}_i} \mathbb{I}[\texttt{event}(a) \neq \texttt{NONE}] \prod_{d\in D} \mathbb{I}[\hat{y}_{i,d} > 2]\,, \tag{13}$$

where $D = \{\texttt{novo}, \texttt{relv}, \texttt{imsc}\}$. These implement the Relv^{-}-style features reported in Table 7 by counting memory changes only when the corresponding aspect is rated above the midpoint. In the released code, gating labels are sourced from stored per-utterance aspect predictions (e.g., `claim_predictions["gpt-5"]`) and are only available for utterances included in rated segments (via `mem_rating_used`).

Feature	$\|r\|$ w/ CIG
Memory changes (Relv$^-$)	0.728
Memory changes (Any)	0.714
Extracted claim count	0.713
Memory changes (Info$^-$)	0.709
Memory changes (Novo$^-$)	0.704
TF–IDF sum	0.692
Length (tokens)	0.687
TF–IDF max	0.634
Novel word count	0.620
Memory changes (Scope$^-$)	0.613
Entity count	0.449
Novel entity count	0.422
Novel entity ratio	0.305
Specificity (mean IDF)	0.231

Table 19: Same as Table 7 but using the CIG soft label predicted by **GPT-5** as the target variable (reported for robustness in the appendix).

C Pipeline specification

Transcript segmentation is performed using GPT-5 with default inference settings (accessed in August, 2025). Claim extraction is carried out using GPT-5-mini, also with default inference settings (accessed in August, 2025).

For the retrieval layer, extracted claims are embedded using `Qwen/Qwen3-Embedding-0.6B`, a dense embedding model. The embedding vectors are stored in Qdrant, which serves as the vector database for retrieval.

Memory consolidation is performed using GPT-5-mini with default inference settings (accessed in August, 2025). Segment-level memory summarization is then conducted using GPT-5 with default inference settings (accessed in August, 2025). Finally, CIG ratings are produced using GPT-5 with default inference settings (accessed in August, 2025), with additional models evaluated in ablation experiments.

Local-model experiments are run on the Spartan cluster using `Qwen/Qwen3-4B-Instruct-2507` on a single NVIDIA A100 GPU.

D Plain language statement (anonymised)

Project overview. This research studies how community audiences perceive **informativeness** in public deliberative conversations (e.g., debates, community meetings, and other public forums).

What you will do. You will complete a short training/pilot task, then (optionally) additional annotation tasks. In each task, you will read a debate topic, brief background information, and short segments of conversation, and rate each target utterance on informativeness-related dimensions (e.g., Novelty, Relevance, and Scope of implications) using 1–4 scales.

Time commitment. Each task is designed to be completed within a bounded time window; a pilot phase is used to estimate average completion time and set payments accordingly.

Benefits. Your participation contributes to research on how people judge informative communication in deliberation, supporting the development of tools for more constructive and effective public conversations.

Payment. You will be compensated in line with platform guidelines and to meet or exceed applicable minimum-wage expectations given typical task duration.

Risks / discomfort. No significant risks are anticipated. However, some segments may contain sensitive, controversial, or emotionally charged content. You may stop at any time if you feel uncomfortable.

Voluntary participation / opt-out. Participation is voluntary. You may withdraw at any time without negative consequences. Each segment includes sufficient topic/context information beforehand, and you may opt out of rating any specific item if desired.

Privacy / data handling. No direct personal identifiers are collected by the research team beyond platform-managed participant IDs. Data will be de-identified prior to release; only anonymised annotations and derived research outputs will be shared publicly.

Results sharing. Aggregate findings and anonymised resources (e.g., code and de-identified annotations) may be released via academic publications and public repositories.

Contacts and complaints. For anonymous review, institutional identifiers and direct contact details are withheld. A non-identifying institutional ethics-review process approved the study; in the camera-ready version, the final manuscript will include the responsible office/contact channel for participant concerns.

Context. Topic: *Gun Reduces Crime*.
Prior summary: Pro speakers (Gary Kleck, Gayle Smith) argue that responsible firearm ownership deters criminals and enhances safety; opponent (R. Gil Kerlikowske) argues that widespread firearm availability increases risk to civilians and law enforcement and that tighter controls and prevention are more effective.
Immediate preceding turn (moderator): *"... you were talking about British police being unarmed... and yet Gary Kleck... said why are police armed... unless it is to deter assault... can you take that on?"*

Aspect	Level	Example utterance (given the context above)	Why this level
Overall CIG	1	R. Gil Kerlikowske: "That's just wrong—guns don't reduce crime."	Pure assertion; adds no new content/mechanism.
	2	R. Gil Kerlikowske: "To clarify, I'm talking about police carrying firearms, not civilian ownership."	Narrow clarification; limited informational advance.
	3	R. Gil Kerlikowske: "If officers are armed, some attackers may back off, which could reduce assaults on police even if other crimes don't change."	Adds a concrete mechanism within the existing deterrence frame.
	4	R. Gil Kerlikowske: "We should separate *deterrence* from *escalation*: arming police may deter some assaults but can also raise the lethality of encounters, so "reduces crime" depends on which outcomes we count."	Reframes the evaluation criterion; redirects the discussion.
Novelty	1	Gary Kleck: "Police carry guns to deter criminals—that's the point."	Restates a point already present in the prior summary.
	2	Gayle Smith: "In some places, officers patrol without firearms."	Small factual addition; limited conceptual novelty.
	3	R. Gil Kerlikowske: "Even the *possibility* of a gun being taken from an officer changes how police approach routine encounters."	Introduces a specific explanatory detail not established in context.
	4	Gary Kleck: "Deterrence is a *belief* effect: what matters is offenders' perceived risk of armed resistance, not simply the raw number of guns."	Introduces a new conceptual lens/principle.
Relevance	1	Gayle Smith: "Seattle is a hard place to drive in."	Off-topic with no connection to the motion.
	2	R. Gil Kerlikowske: "Public trust in police affects whether people cooperate with investigations."	Indirect link; requires bridging inference to connect to crime reduction via guns.
	3	R. Gil Kerlikowske: "Officer assaults are related to public safety, but they are not the full question of whether guns reduce overall crime."	Clearly connected but peripheral to the core claim.
	4	Gary Kleck: "If guns deter offenders, then wider lawful ownership can reduce crime—that's the motion we're debating."	Directly addresses the core topic/goal.
Implication Scope	1	R. Gil Kerlikowske: "Could you repeat the question?"	Procedural/local to the interaction.
	2	Gayle Smith: "I have a 2 years old boy."	Personal information; bounded to an individual case.
	3	R. Gil Kerlikowske: "My neighborhood has been unsafe in the last 5 years, and I think it is not just that, but wider region."	Generalizes beyond a single case to a broader pattern.
	4	Gary Kleck: "Public policy should preserve a general right of self-defense while minimizing harms to the broader community."	Principle-level claim with wide public applicability.

Table 12: Exemplar utterances illustrating Levels 1–4 for Overall CIG and its three aspects under a fixed deliberation context. Speaker names are restricted to those appearing in the prior summary. Examples are illustrative and not drawn verbatim from the corpus.

Section	Prompt part (abridged)
Role	You are an **Atomic-Fact Extractor**.
Goal	Given a conversational context and a target utterance, extract a list of **atomic claims**.
Core principles	**Self-Contained:** each claim is understandable without the original dialogue. **Atomic:** each claim is the smallest proposition that can be independently true or false. **Semantically Distinct:** do not output paraphrase duplicates; each claim must be unique.
Extraction rules	(1) Extract explicit claims literally stated in the target utterance. (2) Extract salient implicit claims a listener would confidently infer. (3) Focus on content, not speech acts (avoid describing asking/suggesting).
What to avoid	Do not describe the act of speaking (avoid "Asks...", "States...", "Suggests..."). Do not include hedge/filler (e.g., "I think", "maybe", "kind of"). Exception: a speech verb is allowed only for reported speech (e.g., "Stephen said that ...").
Output schema	Return **only** valid JSON: `{"memories":[{"speaker":"...", "target_speaker":"...", "claim":"...", "turn_id":"..."}, ...]}` If none: `{"memories":[]}`
Quality checks	Use the **Context** only to resolve ambiguity (pronouns/deictics); extract only from the **Target utterance**. One JSON object per proposition; do not extract more than 30 claims (keep the most important). If speaker appears as "Name (role)", keep only the full name. Return JSON only (no markdown, no extra text).

Example (from prompt)	**Context:**
	1. Speaker 1: I hope my kids own guns.
	2. Speaker 2: I am thinking the opposite.
	Target utterance:
	3. Speaker 3: When I look at the statistics about how that adds to the risk of suicide, the risk of being misused, the risk of it being stolen, used in a domestic quarrel, I think it's just too much of a risk.
	Output:
	`{"memories":[`
	`{"speaker":"Speaker 3","target_speaker":"Everyone","claim":"Having a gun increases the risk of suicide.","turn_id":"3"},`
	`{"speaker":"Speaker 3","target_speaker":"Everyone","claim":"Having a gun increases the risk of misuse.","turn_id":"3"},`
	`{"speaker":"Speaker 3","target_speaker":"Everyone","claim":"Having a gun increases the risk of theft.","turn_id":"3"},`
	`{"speaker":"Speaker 3","target_speaker":"Everyone","claim":"Having a gun increases the risk of use in domestic quarrels.","turn_id":"3"},`
	`{"speaker":"Speaker 3","target_speaker":"Everyone","claim":"Owning a gun is too risky.","turn_id":"3"}`
	`]}`

Other examples	Example A, B, C, E, F, . . .

Table 13: **Abridged** single-task prompt for multi-party atomic claim extraction. The table summarizes the role, constraints, output schema, and one representative in-context example; remaining examples are omitted for space.

Section	Prompt part (abridged)
Role	You are a **Multi-Party Memory Consolidator**. For each newly extracted claim, decide whether to **ADD**, **UPDATE**, or **NONE** using NLI.
Input	`existing_memories`: JSON array of stored proposition objects. `newly_extracted_claims`: JSON array of new proposition objects.
Core procedure	(1) Retrieve candidates: search for the most semantically relevant existing memory, prioritizing same-speaker items. (2) NLI: classify relation between A (new) and B (existing) in both directions ($A \Rightarrow B$, $B \Rightarrow A$). (3) Decide action: apply speaker-aware decision rules. (4) Emit one JSON update object for A.
NLI relation set	Map bidirectional NLI to one of: *equivalent* ($A \Leftrightarrow B$), *forward_entail* ($A \Rightarrow B$ only), *backward_entail* ($B \Rightarrow A$ only), *contradiction*, *neutral*.
Target selection (priority)	Choose *exactly one* target B (or none) by the following ladder: 1) same speaker & *equivalent* 2) same speaker & *backward_entail* 3) same speaker & (*contradiction* or *forward_entail*) 4) different speaker & any non-neutral relation Else: no eligible $B \rightarrow$ treat as neutral (ADD, target=null). Ties within a rung: pick the highest confidence (or highest similarity).
Action mapping	Same speaker: *equivalent*, *backward_entail* $\rightarrow$ NONE; *forward_entail*, *contradiction* $\rightarrow$ UPDATE; *neutral* $\rightarrow$ ADD. Different speaker: always ADD.
UPDATE semantics	If *contradiction*: replace B with A (source is A). If *forward_entail*: merge A with B into a more specific claim (source is merged claim).
Output schema	`{"memory_updates":[{"action":"ADD|UPDATE|NONE", "logical_relation":"...", "source":{...}, "target":null|{id,...}}]}` Return `{"memory_updates":[]}` if nothing changes.
Quality checks	Rewrite claims to be context-independent; resolve pronouns; remove hedges/filler. `target_speaker` denotes a person/group (not an object). `target` may only include: `id, speaker, target_speaker, claim, turn_id`.

Example (from prompt)	**Existing memory:**
	`{ "id":"mem_012", "speaker":"Sam", "target_speaker":"Everyone", "claim":"The financial cost of death penalty appeals is very high.", "turn_id":"8"}`
	New claim A:
	`{ "speaker":"Sam", "target_speaker":"Everyone", "claim":"The financial cost of death penalty appeals exceeds the cost of life imprisonment.", "turn_id":"14"}`
	Output update:
	`{ "action":"UPDATE", "logical_relation":"forward_entail", "source":{... turn 14 ...}, "target":{... mem_012 ...}}`

Table 14: Abridged prompt for multi-party memory consolidation, including the NLI relation set, deterministic target-selection ladder, and action mapping that produces ADD/UPDATE/NONE updates. One illustrative example is shown; additional examples in the original prompt are omitted for brevity.

Section	Prompt part (abridged)
Role & task	You are an expert dialogue analyst acting from the perspective of a community audience. Rate each **TARGET** utterance on **Informativeness** (1–4).
Input (template vars)	Topic/Goal: `{topic}` Prior knowledge + preceding dialogue: `{context}` Target utterances: `{target}` (evaluate indices `{start}`–`{end}`, total `{total}`).
Baseline assumption	Treat the `Prior Knowledge` section and preceding dialogue as the **Shared Knowledge Baseline**; repeating/paraphrasing baseline content → low scores.
Dimension definition	**Conversational Information Gain (Informativeness)**: how much the utterance advances shared understanding or progress on the topic, given baseline. Scale anchors: 1=no gain; 2=minimal; 3=incremental; 4=insightful.
Output & schema	Return **only** valid JSON (no extra text): `[{"utterance_index":int, "informativeness":int, "context_type":INFO}, . . . ]` *(*`context_type` *is hard-coded for identification.)*
Example (from prompt)	**Topic:** Gun Reduces Crime. **Prior knowledge:** brief debate summary (Kleck/Smith vs Kerlikowske). **Target utterances:** indexed 1–4. **Illustrative output:** `[{"utterance_index":1, "informativeness":3}, {"utterance_index":2, "informativeness":1}, {"utterance_index":3, "informativeness":1}, {"utterance_index":4, "informativeness":4}]`

Table 15: **Abridged** info-only rating prompt used to score segment utterances on Informativeness (CIG). For space, we *shorten* (i) the system preamble and reminder lines, (ii) the full prose definitions for the 1–4 scale, and (iii) the example's prior-knowledge summary and dialogue excerpt; the table preserves the exact `{topic}`/`{context}`/`{target}` inputs, rating dimensions, and the required JSON output schema.

Section	Prompt part (abridged)
Role & task	You are an expert dialogue analyst acting from the perspective of a community audience. Rate each **TARGET** utterance on **Novelty**, **Relevance**, and **Implication Scope** (each 1–4).
Input (template vars)	Topic/Goal: `{topic}` Prior knowledge + preceding dialogue: `{context}` Target utterances: `{target}` (evaluate indices `{start}`–`{end}`, total `{total}`).
Baseline assumption	Treat the `Prior Knowledge` section and preceding dialogue as the **Shared Knowledge Baseline**; repetitions → low **Novelty**.
Dimension definitions	**Novelty**: newness relative to baseline. **Relevance**: substantive connection to topic/goal. **Implication Scope**: scope of intended reach/generality.
Output & schema	Return **only** valid JSON (no extra text): `[{"utterance_index":int, "novelty":int, "relevance":int, "implication_scope":int, "context_type":MIX}, . . . ]` *(*`context_type` *is hard-coded for identification.)*

Example (from prompt)	**Topic:** Gun Reduces Crime.
	Prior knowledge: brief debate summary (Kleck/Smith vs Kerlikowske).
	Target utterances: indexed 1–4.
	Illustrative output:
	`[{"utterance_index":1,"novelty":3,"relevance":3,"implication_scope":3},`
	`{"utterance_index":2,"novelty":1,"relevance":4,"implication_scope":4},`
	`{"utterance_index":3,"novelty":3,"relevance":1,"implication_scope":1},`
	`{"utterance_index":4,"novelty":4,"relevance":3,"implication_scope":3}]`

Table 16: **Abridged** mixed-aspects rating prompt used to score segment utterances on Novelty, Relevance, and Implication Scope. For space, we *shorten* (i) the system preamble and reminder lines, (ii) the full prose definitions and all anchor examples for each 1–4 scale, and (iii) the example's prior-knowledge summary and dialogue excerpt; the table preserves the exact `{topic}`/`{context}`/`{target}` inputs, the rated dimensions, and the required JSON output schema.

Section	Prompt part (abridged)
Role & task	You are an expert dialogue analyst acting from the perspective of a community audience. Rate each **TARGET claim** on **Informativeness (CIG)**, **Novelty**, **Relevance**, and **Implication Scope** (each 1–4).
Input (template vars)	Topic/Goal: `{topic}` (Optional) Dialogue context: `{dialogue_context}` Existing memories (shared baseline): `{existing_memories}` Target claims: `{claims}`
Baseline assumption	Treat `Existing memories` as the **Shared Knowledge Baseline**; repetitions/paraphrases of the baseline → low **Novelty** (and typically low **Informativeness**).
Dimension definitions	**Informativeness (CIG)**: forward-looking impact on shared understanding/progress. **Novelty**: newness relative to baseline/context only. **Relevance**: substantive connection to the topic/goal (procedural/meta → low). **Implication Scope**: intended reach/generality of the claim (local → universal).
Output & schema	Return **only** valid JSON (no extra text): `[{"id":int, "informativeness":int, "novelty":int, "relevance":int, "implication_scope":int}, . . . ]`
Example (from prompt)	**Topic:** Should our state retain the death penalty? **Existing memories:** e.g., "provides justice for victims' families"; "appeals are very costly". **Target claims:** e.g., "risk of executing an innocent person"; "appeals cost exceeds life imprisonment"; plus off-topic claims. **Illustrative output:** `[{"id":13, "informativeness":4, "novelty":4, "relevance":4, "implication_scope":4}, {"id":14, "informativeness":3, "novelty":3, "relevance":4, "implication_scope":4}, . . . ]`

Table 17: **Abridged** claim-level rating prompt used to score extracted claims on Informativeness (CIG), Novelty, Relevance, and Implication Scope given an `existing_memories` baseline. For space, we *shorten* (i) the system preamble and reminder lines, (ii) the full prose definitions and all anchor examples for each 1–4 scale, and (iii) the example's dialogue context and claim list; the table preserves the exact template inputs (`{topic}`, `{dialogue_context}`, `{existing_memories}`, `{claims}`), the rated dimensions, and the required JSON output schema.

		Fora					Insq				
Model	**Context**	**CIG**	**Nov.**	**Rel.**	**Scope**	**Mean**	**CIG**	**Nov.**	**Rel.**	**Scope**	**Mean**
GPT-5	Full	**0.512**	0.556	0.429	0.483	0.495	**0.411**	0.577	0.444	0.541	**0.493**
	Memory	0.531	0.574	0.419	0.496	0.505	0.459	0.567	**0.427**	0.526	0.495
	No Knowledge	0.594	0.700	0.422	0.488	0.551	0.470	0.700	0.436	0.539	0.536
	Short Prior	0.572	0.620	0.417	0.495	0.526	0.459	0.637	0.440	0.536	0.518
	Summary	0.520	0.556	**0.414**	0.479	**0.492**	0.457	0.587	0.452	0.529	0.506
GPT-5-Mini	Full	0.551	0.574	0.456	**0.438**	0.505	0.487	0.726	0.529	0.484	0.557
	Memory	0.591	0.557	0.467	0.453	0.517	0.501	0.697	0.535	0.467	0.550
	No Knowledge	0.646	0.542	0.448	0.463	0.524	0.557	0.599	0.520	0.472	0.537
	Short Prior	0.622	0.539	0.460	0.455	0.519	0.577	**0.530**	0.551	0.473	0.533
	Summary	0.569	0.617	0.475	0.472	0.533	0.546	0.806	0.547	0.489	0.597
Gemini-2.5-Flash	Full	0.563	0.616	0.622	0.474	0.569	0.543	0.653	0.615	0.496	0.577
	Memory	0.560	0.543	0.575	0.476	0.539	0.554	0.600	0.502	0.467	0.531
	No Knowledge	0.654	0.632	0.661	0.492	0.610	0.677	0.621	0.648	0.465	0.603
	Short Prior	0.588	0.570	0.533	0.471	0.540	0.689	0.652	0.710	**0.453**	0.626
	Summary	0.589	0.585	0.579	0.452	0.551	0.611	0.628	0.579	0.458	0.569
Qwen3-4B	Full	0.622	0.571	0.467	0.593	0.563	0.597	0.656	0.589	0.597	0.610
	Memory	0.643	0.523	0.425	0.599	0.547	0.726	0.617	0.509	0.569	0.605
	No Knowledge	0.619	0.723	0.469	0.655	0.617	0.536	0.812	0.569	0.626	0.636
	Short Prior	0.535	0.520	0.462	0.575	0.523	0.601	0.673	0.494	0.586	0.589
	Summary	0.579	**0.519**	0.472	0.588	0.539	0.658	0.618	0.528	0.593	0.599

Table 18: MAE results for different models, contexts, and aspects across Fora and Insq corpora. The last column in each corpus group shows the mean MAE across the four aspects. The lowest MAE in each column is highlighted in bold.