Re-FRAME the Meeting Summarization SCOPE: Fact-Based Summarization and Personalization via Questions

Frederic Kirstein[1], Sonu Kumar, Terry Ruas, Bela Gipp
University of Göttingen, Germany
[1]kirstein@gipplab.org

Abstract

Meeting summarization with large language models (LLMs) remains error-prone, often producing outputs with hallucinations, omissions, and irrelevancies. We present **FRAME**, a modular pipeline that reframes summarization as a semantic enrichment task. FRAME extracts and scores salient facts, organizes them thematically, and uses these to enrich an outline into an abstractive summary. To personalize summaries, we introduce **SCOPE**, a *reason-out-loud* protocol that has the model build a reasoning trace by answering nine questions before content selection. For evaluation, we propose **P-MESA**, a multi-dimensional, reference-free evaluation framework to assess if a summary fits a target reader. P-MESA reliably identifies error instances, achieving $\geq$ 89% balanced accuracy against human annotations and strongly aligns with human severity ratings ($\rho \geq 0.70$). On QMSum and FAME, FRAME reduces hallucination and omission by 2 out of 5 points (measured with MESA), while SCOPE improves knowledge fit and goal alignment over prompt-only baselines. Our findings advocate for rethinking summarization to improve control, faithfulness, and personalization[1].

1 Introduction

Meetings can be dense, chaotic, and high-stakes. Summarizing them effectively is a natural language processing (NLP) challenge with value for corporate, academic, and governmental contexts (Zhong et al., 2021; Hu et al., 2023; Laskar et al., 2023; Kirstein et al., 2025b). Yet current large language model (LLM)-based systems (Laskar et al., 2023; Fu et al., 2024) continue to produce summaries that omit key points, hallucinate content, and struggle with relevance (Golia and Kalita, 2023; Kirstein et al., 2025d). We argue that these weaknesses stem from a more profound structural mismatch that current approaches treat conversation like linear text, compressing form without reconstructing meaning.

Meetings hold three challenges compared to traditional structured texts (Arabzadeh et al., 2023; Kirstein et al., 2025a): (1) Salient content is scattered across speaker turns (*Information Distribution*), (2) utterances depend on long-range context (*Contextual Dependencies*), and (3) salience varies per reader (*Salience Ambiguity*). Current summarization methods are not aligned with these properties. Approaches relying on structural cues (e.g., sections, paragraphs) (Liu and Lapata, 2019) are unsuitable for meetings' boundary-free nature. Chunking methods (Zhang et al., 2022) struggle with cross-chunk dependencies, while hybrid extractive-abstractive models (Li et al., 2021) may lose interpretability. Each approach shares the same limitation: condensing information without reconstructing the underlying semantic structure.

Our **FRAME** (***F****act-based* ***R****econstruction and* ***A****bstractive* ***ME****eting Summarization*) framework is a fact-centric pipeline that reframes the established meeting summarization workflow as an *enrichment* task (shown in Figure 1). Drawing on research in fact extraction (Gunjal and Durrett, 2024; Wanner et al., 2024) and summary planning (Amplayo et al., 2021; Grenander et al., 2025), FRAME mimics how humans summarize texts in four stages (Endres-Niggemeyer, 2000): We extract self-contained verifiable facts to cut filler content (*Fact Identification*), filter them by salience (*Note-Taking*), organize them into an outline by thematic relationships (*Organization*), and enrich these outlines to abstractive summaries strictly using facts (*Summary Writing*). With GPT-4o (OpenAI, 2024), FRAME improves output quality compared to recent approaches. On QMSum (Zhong et al., 2021) and FAME (Kirstein et al., 2025a), FRAME cuts hallucination by up to 3 out of 5 points (3→1, 4→1; lower is better), and irrelevance by 1 point (2→1, 3→1), beating GPT-4o and

[1]Resources are available as per Appendix A.1 on GitHub.

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 20087–20137
November 4-9, 2025 ©2025 Association for Computational Linguistics

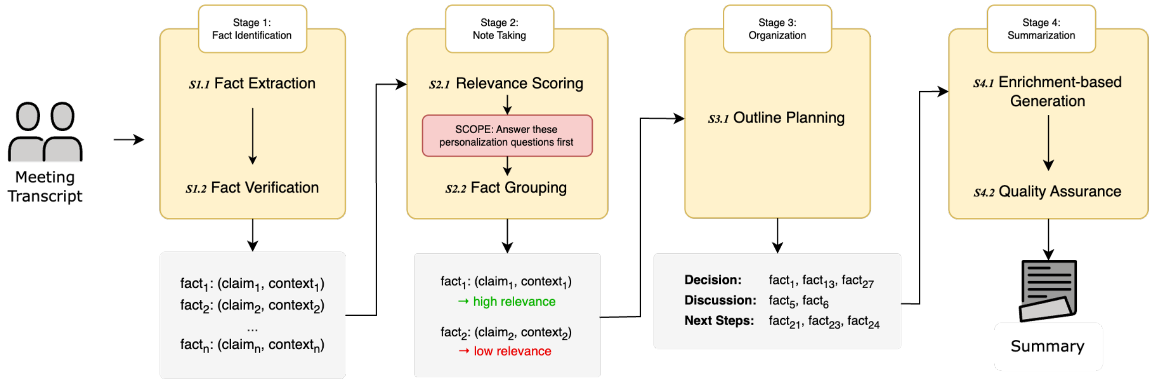

Figure 1: FRAME pipeline with SCOPE integration. FRAME structures summarization in four stages: fact identification, note taking, organization, and enrichment-based generation. SCOPE plugs into salience scoring by injecting a reasoning trace derived from reader-specific questions.

Gemini-1.5 pro[2] (Google et al., 2024) across six of eight MESA dimensions (Kirstein et al., 2025c).

Salience ambiguity remains widely unaddressed by current systems, despite the growing interest in personalization. Meetings involve diverse roles and objectives, yet most systems produce a one-size-fits-all output, ignoring role-specific goals and expertise (Kirstein et al., 2024). We introduce the **SCOPE** protocol (*Summarizing **C**ontent **O**riented to **P**ersonal **E**xpectations*) that guides an LLM through an explicit *reason-out-loud* approach before scoring facts. Drawing on cognitive science research (Solomon et al., 1995; Konrad, 2017), SCOPE has the model answer a questionnaire about the reader's goals, expertise, and understanding. This creates an explicit reasoning trace to ground content selection, improving personalization.

As evaluating personalized summaries is hard with existing metrics (e.g., ROUGE (Lin, 2004), MESA), we propose **P-MESA** (**P**ersonalized MESA), a reference-free LLM-based personalization metric with seven dimensions: *factuality, completeness, relevance, goal alignment, prioritization, knowledge-fit*, and *contextual framing*. P-MESA correlates strongly with human judgment (avg. Spearman $\rho = 0.76$) on 50 LLM-generated summaries. On our new P-MESA metric (5-point Likert score on quality impact), SCOPE improves knowledge-level fit ($2 \rightarrow 1$) and goal alignment ($3 \rightarrow 2$), reducing oversimplification and reader hallucination. As such, SCOPE outperforms role-playing (Kirstein et al., 2024; Zhang et al., 2025) and reader-tailoring prompting (Ghodratnama and Zakershahrak, 2024) by modeling why information matters, not just to whom.

[2]We will refer to them as GPT and Gemini.

This paper makes three key contributions:

- **FRAME**: A modular summarization pipeline that treats summarization as enrichment, improving factuality, coherence, and salience handling.
- **SCOPE**: A personalization protocol that models reader intent via reason-out-loud, outperforming persona injection.
- **P-MESA**: A reader-centric metric quantifying personalization quality without references that aligns with human preferences.

2 Related Work

Meeting summarization is about distilling multi-speaker dialogue with distributed information (Rennard et al., 2023; Kirstein et al., 2025b). Prior work treats this as a reduction problem, aiming to condense dialogue using LLM prompting (Laskar et al., 2023; Fu et al., 2024; Tang et al., 2024), role vectors (Asi et al., 2022), hierarchical encoding (Naraki et al., 2022), or self-refinement (Kirstein et al., 2025d). These methods improve coherence but often struggle with understanding the meeting's content (Kirstein et al., 2025b) as they do not reconstruct the underlying meaning explicitly before summary generation. In contrast, FRAME reframes summarization as an enrichment task by first extracting and grouping facts from the transcript, planning a high-relevance summary, and enriching the outline to form a summary.

Fact extraction transforms text into self-contained, verifiable units (Kamoi et al., 2023; Min et al., 2023), increasingly used for claim extraction and fact verification (Chern et al., 2023; Chiang and Lee, 2024; Wang et al., 2024). Prior works on fact extraction in dialogue summarization either enrich utterances with extracted factual statements to reduce hallucinations (Zhang et al., 2024a), or target specific fact types (e.g., action items) via classification and neighborhood-based rephrasing (Golia and Kalita, 2023). These approaches remain superficial, lacking a general, structured fact representation. Typical fact representations may discard salient discourse cues or miss global dependencies (Gunjal and Durrett, 2024; Wanner et al., 2024). We address this with *statement–context tuples*, a structured fact representation pairing claims to global context. Unlike others, our representation retains interpretability, allowing for better content comprehension (see Appendix G.1).

Personalization is about adapting summaries to reader expectations (Kirstein et al., 2024, 2025b). Recent approaches zero-shot LLMs to align content with a reader's profile (Kirstein et al., 2024; Paoli, 2023) or model participants as graph nodes to extract personalized views (Jung et al., 2023). These approaches can surface relevant points but lack consistency and are prone to reader perspective hallucination (Zhang et al., 2024b). Human-in-the-loop systems (Chen et al., 2023; Ghodratnama and Zakershahrak, 2024) mitigate this with feedback, but remain laborious, costly, and time-consuming. We approach robust salience detection through our SCOPE reason-out-loud protocol, inspired by cognitive science (Solomon et al., 1995) and think-out-loud protocols observed in human summarizers (Endres-Niggemeyer, 2000). SCOPE guides the LLM to answer a questionnaire to build an explicit trace of the reader's intent, expertise, and goals before selecting salient facts. SCOPE outperforms established approaches (see Appendix I.2) and works zero-shot, making it scalable and generalizable.

3 Methodology

Overview. Unlike sequence-to-sequence approaches that attempt one-hop transcript summarization (Laskar et al., 2023), FRAME handles summarization in four stages as a structured enrichment task, inspired by the human summarization process (Endres-Niggemeyer, 2000): extracting salient information, assessing their relevance, organizing them thematically, and synthesizing a coherent narrative. For personalization, we introduce SCOPE, a structured reason-out-loud protocol that enforces generating a reasoning trace to ground salience detection. Figure 1 illustrates the complete FRAME framework and how SCOPE integrates.

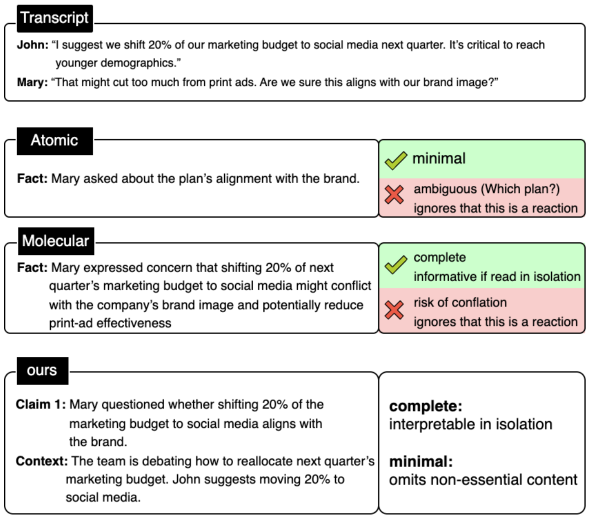

Figure 2: Comparison of our statement-context tuple (OURS) against a high-granularity fact (Atomic) and a high-context fact (Molecular).

3.1 Fact Definition

Motivation. Summarizing meetings requires distilling meaning from fragmented, implicit, and highly contextual speaker turns. While explored throughout NLP (Nenkova and Passonneau, 2004; Zhang and Bansal, 2021; Liu et al., 2023b; Min et al., 2023), existing fact detection setups have marked limitations. For convoluted texts, facts can become too granular (Wanner et al., 2024), resulting in broken context and a loss of interpretability (Li et al., 2016; Gunjal and Durrett, 2024).

We therefore define a fact as *statement–context tuple* $\langle c, \kappa \rangle$ where c is a self-contained claim, and κ is the minimal global context required for its interpretation such that the original meaning of c remains preserved (Choi et al., 2021).

Desiderata. Let T be a transcript of utterances $u_1, \ldots, u_n$. A fact must satisfy:
Completeness: All references in c must be resolvable using κ without requiring external knowledge.
Minimalism: c should convey only one idea. κ includes only the details essential for grounding c.

Figure 2 illustrates these criteria, demonstrating how too much detail introduces noise while insufficient context creates ambiguity, contrasting our facts with existing fact setups.

3.2 FRAME for General Summarization

FRAME operates in four sequential stages (see Figure 1): Fact Identification, Note-Taking, Organization, and Summarization. Complete implementation details and examples appear in Appendix B.

3.2.1 Stage 1: Fact Identification

This stage cuts filler content while preserving semantic content.

Fact Extraction. Given a transcript T, we prompt an LLM to extract facts $F = \{f_1, \ldots, f_m\}$ as tuples $\langle c_i, \kappa_i \rangle$. The prompt includes contrastive examples to filter fillers (*"OK," "Mmhmm"*), hedged statements ("Maybe we should..."), and compound facts (Zhu et al., 2024). Human evaluation confirms this extraction process preserves relevant content (see Appendix F.1).

Fact Verification. To ensure reliability, we verify facts using an LLM judge inspired by FActSCORE (Min et al., 2023), checking F against T for factuality (*action*: removing unsupported claims), completeness (*action*: adding missed key information), clarity (*action*: re-writing κ for self-containment), and minimalism (*action*: removing extraneous details from κ). We find refinement of facts being required in $\sim$5% cases (see Appendix F.1).

3.2.2 Stage 2: Note-Taking

This stage handles fact relevance and redundancy.

Relevance Scoring. Inspired by categorizing facts according to their content in fact-checking (Hassan et al., 2017), we task an LLM judge with assigning each fact f_i a function label (i.e., DECISION, ACTION ITEM, INSIGHT, or CONTEXT) and a relevance score $r_i \in [1, 10]$ (higher is more relevant). The r_i ranges are chosen empirically: decisions (9–10), key insights (7–8), supporting context (4–6), and low-salience background (1–3). On average, we retain 40% of facts (see Appendix E), reducing information overload, allowing for content control, and mitigating positional bias (Liu et al., 2023a; Xiao et al., 2024). We show the influence of varying the r_i ranges in Appendix F.3, observing that too relaxed or strict thresholds impact summary quality.

Fact Grouping. To mitigate redundancy while capturing emphasis signaled by repeated stating of a fact, retained facts are grouped by their function label and relevance scores. Within each group, we use an LLM to identify semantically overlapping facts, consolidate them by synthesizing contexts, and preserve the highest relevance score.

3.2.3 Stage 3: Organization

This stage creates an outline to guide the summary.

Outline Planning. We convert rated facts into a structured outline that reflects the conversation's logic, orienting on summary planning (Amplayo et al., 2021; Grenander et al., 2025). High-relevance ($r_i \geq 8$) and DECISION facts become major outline points. Mid-relevance ($6 \leq r_i < 8$) and CONTEXT facts are considered during summary writing to provide background and flow. Outlines present main topics, discussions, and next steps. In Appendix E we find that a 250-token summary covers $\sim$9 high-relevance facts as anchors supported by $\sim$12 contextual facts.

3.2.4 Stage 4: Summarization

This stage enriches the outline to a summary.

Enrichment-Based Generation. An LLM converts the outline into an abstractive summary by enriching each anchor with supporting facts ($r_i \geq 6$). Generation is constrained from introducing content beyond extracted facts. Unsupported outline points are removed rather than speculated upon. In our ablation in Appendix E, we did not observe the inclusion of unsupported outline points, concluding that this is a rare occasion when using strong backbone models for FRAME.

Quality Assurance. An LLM acts as a reviewer, assigning error points to the summary draft on four dimensions (see Appendix B.3 for prompts): *outline adherence* (max. 4 points), *factual accuracy* (max. 3 points), *information coverage* (max. 2 points), and *formatting* (max. 1 point). Each detected error equals one error point. The maximum error points per category are empirically chosen to reflect importance. If any dimension exceeds its maximum, or the total exceeds four points, the system initiates a revision cycle with feedback. The feedback includes observed issues and the LLM's chain-of-thought reasoning (Wei et al., 2024). In our experiments, a single revision pass sufficed.

3.3 SCOPE: Personalization via Reason-Out-Loud

Personalization typically happens through role-playing (Zhang et al., 2025) and reader-tailoring (Ghodratnama and Zakershahrak, 2024). As such, the model has to implicitly interpret how a role or goal affects content selection while summarizing. This often leads to missed needs (Zhang et al., 2024b), as observed in Appendix I.2.

As to cognitive science research, explicit verbalization enhances metacognitive awareness and decision consistency (Solomon et al., 1995; Konrad, 2017). We propose **SCOPE**, a *reason-out-loud* protocol for robust personalization guiding an LLM through an *exploration* step before *fact selection*.

In *exploration*, the LLM builds an assessment trace by answering a questionnaire (details on questions in Table 18 in Appendix I.1) about reader characteristics (background, expertise), specific needs (interests, knowledge gaps), information utility (actionability, responsibility alignment), and contextual relevance (need for elaboration). This leverages a persona description, either provided or inferred from the transcript (Kirstein et al., 2024), containing information about the reader's role, expertise, and standpoint. In the subsequent *fact selection*, this trace is used to ground the relevance determination process of each fact and thereby tailor the selection to the user.

SCOPE acts as filter during the `Note-Taking` stage. Extracted facts are pre-selected here before being scored (`Relevance Scoring`). For complete personalization, the reader profile is also used for `Outline Generation` and `Summary Writing`.

4 Interlude: P-MESA

Personalization in meeting summarization is gaining traction (Kirstein et al., 2025b), but the field lacks a way to evaluate whether a summary serves a reader's preferences. **Reference-based metrics** like ROUGE (Lin, 2004) and BERTScore (Zhang et al., 2020) require gold summaries, which are typically unavailable in personalized forms. **Reference-free metrics** such as MESA (Kirstein et al., 2025c) assess general quality aspects (e.g., coreference, irrelevance) but do not judge alignment to reader expectations. Recent general **personalization metrics** like EGISES (Vansh et al., 2023) and PerSEval (Dasgupta et al., 2024) are designed to rank-order responses, but not to judge whether a single summary is helpful for a specific user. As a result, researchers have to choose between expensive human evaluations and metrics that overlook reader-centric dimensions, stalling progress in personalized summarization.

We introduce **P-MESA** (**P**ersonalized - **ME**eting **S**ummary **A**ssessor), a multi-dimensional, reference-free evaluation framework designed to test whether a single summary satisfies a single target reader. P-MESA scores summaries across seven personalization dimensions motivated from a 50-paper literature review (Section 4.1): *factuality*, *completeness*, *relevance*, *goal alignment*, *priority structuring*, *knowledge-level fit*, and *contextual framing*. Definitions are given in Appendix J.

Each dimension covers a distinct personalization characteristic. P-MESA is powered by an LLM evaluator (GPT) that receives a reader profile containing role, knowledge level, goals, and interests. This evaluator uses this context information to rate summaries on each dimension on a 5-point Likert scale, with higher scores indicating higher impact.

4.1 Criteria Definition

We derive the P-MESA dimensions through a literature analysis and empirical refinement.

Step 1: Literature review. We review 50 papers from *CL venues (2018–2024) on terms such as "personalization", "user modeling", and "adaptation", and discard 14 works due to an unwanted focus on related topics (e.g., personalized agents, style transfer). We manually screen the remaining 36 papers, and we identify nine candidate evaluation dimensions (approach detailed in Appendix J).

Step 2: Human study. To test the dimensions' clarity, coverage, and non-overlap, we conduct a human study in which annotators apply them to model-generated personalized summaries. We construct a one-time evaluation dataset of 48 summaries using GPT and Gemini. Each summary is generated for a distinct reader profile specifying role, prior knowledge, and goals, and is prompted with the task to either (a) *summarize for* or (b) *simulate* that user (24 each). The samples are drawn evenly from QMSum and FAME, covering 8 meeting types, 14 topics, and an average of 5.9 speakers.

Three annotators (ages 22–29, C1+ English; see Appendix C for all details) annotate each summary using the candidate dimensions. Each summary receives two annotations. Annotators provide structured feedback on definition clarity, dimension overlap, and any lack of criteria. We collect feedback through forms and daily group discussions.

Step 3: Final dimension set. Following the feedback from the discussions, *Personal Preference* is dropped due to poor generalizability. *Objective Alignment* and *Information Utility* are merged into *Goal Alignment* for a macro-level intent alignment, contrasting from *Priority Structuring* (micro-level salience and ordering). These steps result in the seven personalization dimensions used in P-MESA.

	Metric	Factuality	Relevance	Goal Alignment	Prioritization	Personal Preferences	Knowledge-level Fit	Contextual Framing	Overall (Average)
detection related	**B-ACC (%)**	91.7	93.3	92.7	90.2	89.4	90.0	89.1	92.2
	Cohen's κ	0.83	0.74	0.76	0.81	0.79	0.62	0.78	0.81
	FNR (%)	7.2	6.3	3.1	8.9	9.5	11.1	8.7	7.8
	FPR (%)	9.4	10.1	10.3	10.7	11.8	11.0	13.1	10.9
sensitivity related	**Spearman ρ**	0.76	0.78	0.81	0.73	0.75	0.70	0.79	0.76
	Kendall τ	0.71	0.74	0.76	0.66	0.68	0.62	0.74	0.70

Table 1: Analyzing P-MESA's detection accuracy and severity correlation with human annotations. For detection, we report B-ACC: Balanced Accuracy, κ: Cohen's Kappa, FNR: False Negative Rate, FPR: False Positive Rate. For severity, we show ρ: Spearman's rank correlation, τ: Kendall's tau.

4.2 Metric Implementation

We build P-MESA on the structure of MESA's LLM judge (Kirstein et al., 2025c) as their framework provides a reasonable compromise between thoroughness and cost. P-MESA uses a three-stage evaluation pipeline: (1) potential error instance detection, (2) instance's severity rating, and (3) impact score aggregation per category. P-MESA also includes the reader profile, either derived or directly given, to each evaluation prompt to guide the LLM judges. Each profile includes the reader's role, goals, expertise, and contextual constraints. Similar to MESA, we use GPT as the backbone model.

4.3 P-MESA as Proxy for Human Judgment

We assess whether P-MESA reliably approximates human annotation by detecting the presence of personalization errors, and assigning severity scores that reflect their perceived impact on reader utility (see Table 1). For this assessment, we generate a second and fully independent set of 48 personalized summaries using the same generation setup as in Section 4.1, but with no overlap in meetings to the previous set. This dataset is used only once, strictly for evaluation purposes.

All six annotators (three male, three female; ages 22–29; C1+ English) rate each summary using the finalized seven P-MESA dimensions on a 0–5 Likert scale (0: no error, 5: maximal impact). Annotators undergo one week of training, including calibration rounds and regular joint discussions. Full human evaluation protocol details for this annotation are provided in Appendix C.

Detection Accuracy of P-MESA. We bin P-MESA scores $\geq$1 as indicating error presence and compute balanced accuracy[3] (B-ACC) and Cohen's κ to evaluate agreement between P-MESA and human annotators. P-MESA achieves high detection accuracy across all dimensions, with B-ACC exceeding 89% and peaking for the persona-grounded criteria at 93.3% (`Relevance`) and 92.7% (`Goal Alignment`). This indicates that P-MESA can identify weaknesses, even when subtle or varying in frequency. Cohen's $\kappa \geq 0.74$ indicates agreement beyond chance, showing that P-MESA applies consistent, human-aligned decision rules. False negatives are rare in the persona-grounded dimensions (`Relevance`: 6.3%, `Goal Alignment`: 3.1%), suggesting P-MESA is unlikely to miss high-impact weaknesses. False positive rates of $\sim$11% further reflect a conservative bias ensuring that borderline weaknesses are not overlooked.

Error Severity Assessment. We compute Spearman's ρ and Kendall's τ between P-MESA and human severity ratings to test how sensitive P-MESA is to error severity changes. P-MESA shows strong rank correlation across all dimensions, indicating a good proxy for human judgment in reflecting error impact. Agreement is highest in `Goal Alignment` ($\rho = 0.81$ and $\tau = 0.76$) and `Relevance` ($\rho = 0.78$, $\tau = 0.74$). Correlations are slightly lower for more interpretive dimensions like `Knowledge`knowledge-level fit ($\rho = 0.70$), reflecting higher variability among human raters.

5 Experiments

We evaluate the effectiveness of FRAME and SCOPE in both general and personalized abstractive meeting summarization. Our experiments are aimed at answering two questions: (1) What impact does fact extraction and treating summarization as an enrichment task have on summary quality? (2) How does our guided, structured reason-out-loud protocol compare to prompt-only personalization?

[3] Balanced accuracy averages sensitivity and specificity. Formal definition given in Appendix K.

	QMSum			FAME [EN]		
	GPT 4o	Gemini 1.5 pro	FRAME GPT-4o	GPT 4o	Gemini 1.5 pro	FRAME GPT-4o
MESA (lower is better)						
Coreference	$0_{1.22}$	$3_{1.58}$	$0_{1.64}$	$0_{1.45}$	$3_{1.57}$	$0_{1.36}$
Hallucination	$3_{1.22}$	$4_{2.04}$	$1_{1.75}$	$4_{0.98}$	$4_{1.40}$	$1_{0.72}$
Incoherence	$4_{1.50}$	$4_{1.09}$	$3_{1.88}$	$4_{0.94}$	$4_{0.72}$	$3_{1.24}$
Irrelevance	$2_{1.70}$	$3_{1.32}$	$1_{1.45}$	$3_{1.14}$	$3_{1.07}$	$1_{1.51}$
Language	$1_{1.30}$	$2_{1.44}$	$1_{1.40}$	$1_{1.17}$	$1_{1.20}$	$1_{1.11}$
Omission	$3_{0.40}$	$3_{0.38}$	$1_{0.16}$	$4_{0.16}$	$4_{0.31}$	$1_{0.00}$
Repetition	$4_{1.05}$	$3_{0.98}$	$1_{1.23}$	$4_{0.74}$	$4_{0.44}$	$2_{0.53}$
Structure	$4_{0.90}$	$3_{1.70}$	$3_{1.24}$	$3_{1.57}$	$3_{1.53}$	$3_{1.46}$
ROUGE (R-1, R-2, R-L) and BERTScore (BS) (higher is better)						
R-1	$37.73_{5.85}$	$39.61_{7.21}$	$22.89_{5.80}$	$39.68_{5.73}$	$38.82_{5.79}$	$20.09_{4.13}$
R-2	$7.95_{4.18}$	$11.10_{4.86}$	$4.13_{2.44}$	$8.43_{3.45}$	$8.96_{3.50}$	$3.81_{2.40}$
R-L	$21.39_{4.05}$	$27.55_{6.36}$	$20.78_{5.23}$	$29.98_{4.88}$	$27.81_{4.18}$	$18.26_{3.81}$
BS (F1)	$81.61_{2.87}$	$80.64_{3.66}$	$85.67_{1.19}$	$63.80_{3.11}$	$83.66_{2.49}$	$84.63_{1.02}$

Table 2: Results of general summarization of QMSum and FAME. GPT and Gemini results stem from Kirstein et al. (2025a). Values are Median$_{Std}$. MESA scores are 1–5 Likert ratings, ROUGE (R-1/R-2/R-L) and BERTScore (BS) are 0–100. Green is best in category.

	QMSum				FAME [EN]			
	GPT 4o	Gemini 1.5 pro	FRAME RP	FRAME SCOPE	GPT 4o	Gemini 1.5 pro	FRAME RP	FRAME SCOPE
P-MESA (lower is better)								
goal alignment	$3_{0.80}$	$3_{0.54}$	$3_{0.79}$	$2_{0.48}$	$3_{0.53}$	$3_{0.58}$	$3_{0.67}$	$2_{0.54}$
completeness	$4_{0.77}$	$4_{0.88}$	$3_{0.64}$	$3_{0.80}$	$4_{0.89}$	$4_{0.59}$	$2_{0.61}$	$2_{0.59}$
factuality	$3_{1.44}$	$4_{1.22}$	$2_{0.37}$	$2_{0.25}$	$3_{1.36}$	$4_{1.02}$	$2_{0.34}$	$2_{0.22}$
knowledge level fit	$2_{1.23}$	$2_{0.44}$	$1_{0.86}$	$1_{0.86}$	$2_{0.97}$	$2_{0.50}$	$2_{0.73}$	$1_{0.58}$
priority structuring	$4_{0.48}$	$4_{0.60}$	$3_{0.56}$	$4_{0.54}$	$4_{0.28}$	$4_{0.65}$	$3_{0.41}$	$2_{0.45}$
contextual framing	$4_{0.91}$	$4_{1.13}$	$3_{0.97}$	$3_{0.88}$	$4_{1.21}$	$4_{0.96}$	$3_{0.69}$	$3_{0.81}$
relevance	$3_{0.54}$	$3_{0.55}$	$1_{0.27}$	$1_{0.57}$	$3_{0.33}$	$3_{0.47}$	$2_{0.89}$	$1_{0.33}$
MESA (lower is better)								
Coreference	$0_{1.48}$	$1_{1.48}$	$1_{1.60}$	$1_{1.55}$	$1_{1.44}$	$0_{1.46}$	$0_{1.44}$	$0_{1.55}$
Hallucination	$4_{1.79}$	$3_{2.03}$	$3_{1.03}$	$2_{1.33}$	$3_{1.39}$	$3_{1.73}$	$2_{0.89}$	$1_{1.19}$
Incoherence	$3_{1.74}$	$4_{1.38}$	$3_{1.38}$	$3_{1.09}$	$3_{1.60}$	$3_{1.37}$	$3_{0.89}$	$3_{0.75}$
Irrelevance	$2_{1.19}$	$2_{1.10}$	$1.5_{1.42}$	$1_{1.48}$	$3_{1.33}$	$2_{1.58}$	$1_{1.40}$	$1_{1.23}$
Language	$2_{1.38}$	$2_{1.32}$	$2_{1.38}$	$2_{1.24}$	$1_{1.04}$	$1_{1.14}$	$1_{1.25}$	$1_{1.10}$
Omission	$4_{0.16}$	$4_{0.41}$	$2_{0.29}$	$2_{0.40}$	$4_{0.41}$	$4_{0.43}$	$2_{0.28}$	$1_{0.56}$
Repetition	$3_{0.94}$	$3_{1.22}$	$1.5_{1.17}$	$2_{0.79}$	$4_{0.31}$	$3_{0.50}$	$2_{0.41}$	$2_{0.43}$
Structure	$3_{1.67}$	$1_{1.42}$	$3_{1.45}$	$3_{1.38}$	$2_{1.28}$	$3_{0.47}$	$3_{1.56}$	$2_{1.21}$

Table 3: Personalized summarization of QMSum and FAME. Values are Median$_{Std}$. MESA and P-MESA scores are 1–5 Likert ratings. Green is best in category.

5.1 Experimental Setup

Backbone models. We implement FRAME with GPT (OpenAI, 2024) across all pipeline stages, using its 128k-token context window. In Section 6 we confirm improvements also hold across Gemini, Llama 3.1 8b, and Gemma 3 4b, with verification modules becoming more relevant, indicating that the gains stem from the fact-based approach.

Baselines. We compare FRAME to GPT and Gemini using zero-shot prompting, tasking 250-token summaries (Kirstein et al., 2025d). We exclude refinement-based methods that rely on reusing the same backbone models (Kirstein et al., 2025d). In Appendix L, we show that FRAME can outperform a three-time revision. For personalization, we compare FRAME with and without SCOPE to GPT and Gemini, both prompted to role-play. Prompts are given in Appendix B.3.

Datasets. We evaluate on QMSum (Zhong et al., 2021), an established benchmark combining academic (ICSI, Janin et al. (2003)), product (AMI, Carletta et al. (2006)), and parliamentary (Welsh/-Canadian, WPCP) meetings. We also use FAME (Kirstein et al., 2025a), a synthetic dataset comprising 500 English and 300 German meetings, spanning 14 different meeting formats and 28 topics, generated by agents simulating realistic meeting dynamics. We randomly sample 50 English samples from QMSum and FAME to test FRAME and SCOPE. Dataset details are given in Appendix H.

Evaluation. We evaluate summaries using ROUGE (R-1/R-2/R-L) (Lin, 2004) and BERTScore (rescaled F1) (Zhang et al., 2020), the reference-free MESA (Kirstein et al., 2025d) for analyzing the occurrence of eight general, and our P-MESA (Section 4) to capture seven personalization dimensions. General QMSum and FAME baselines stem from Kirstein et al. (2025a).

5.2 Results: General Summarization

Findings. Reframing summarization as an enrichment task yields three benefits: (*F1*) improved content understanding via fact isolation, (*F2*) less hallucination due to grounded claims, and (*F3*) better coherence through summary planning.

Quantitative analysis. Table 2 shows that FRAME consistently outperforms baselines on MESA. However, FRAME's summary structure appears to diverge from the references (lower ROUGE). The largest quality impact reduction appears in hallucination, dropping from 3→1 on QMSum and 4→1 on FAME for FRAME (*F2*). MESA's chain-of-thought traces confirm these reductions are due to fewer unsupported claims. We conclude that this relates to the `Fact Verification` and `Outline Planning` stages that constrain synthesis to grounded facts.

Omission and irrelevance scores drop by ∼2 points to 1. We interpret that the fact-based approach helps with content understanding and enables `Fact Selection` to focus on salience without omission (*F1*). `Fact Grouping` cuts repetition

(QMSum: 3→1, FAME: 4→2), yielding denser summaries. Structure improves (4→3), reflecting the effect of `Outline Planning` (*F3*). Language scores remain stable, indicating no loss of fluency.

Qualitative observations. Baselines tend to mirror the temporal structure of a meeting and to preserve low-value information (e.g., "Alice shared her screen…"). FRAME decouples summary structure from chronology, shifting to thematic progressions, and filters irrelevant content. We link this to the `Relevance Scoring` and `Fact Grouping` stages. A qualitative example appears in Appendix M.

5.3 Results: Personalized Summarization

Findings. SCOPE leads to (*F1*) more personalized fact selection, (*F2*) reduced hallucination via explicit user modeling, and (*F3*) personalized adaptation without quality loss.

Quantitative analysis. Table 3 shows that adding SCOPE to FRAME results in improvements on all seven P-MESA dimensions. Relevance and knowledge fit drop from 2→1, meaning content better matches user expectations and language suited to their expertise (*F1*). Goal alignment and priority structuring improve from 3→2, indicating that summaries better reflect the information needs from readers (*F1*). Baselines with persona injection lag by 1–2 points across P-MESA criteria (*F1* and *F2*). We find that SCOPE poses questions the model can answer reliably while role-playing, allowing the responses to function as a pre-selection mechanism.

Further, these gains do not trade off against general quality. MESA scores for personalized FRAME summaries are close to non-personalized scores (Section 5.2). SCOPE reduces hallucination and omission scores (3→2, 2→1) (*F3*).

Qualitative observations. As we can observe in the example in Appendix M, SCOPE addresses two weaknesses of single-prompt baselines. First, it reduces *oversimplification* by varying granularity in line with the reader's priorities while baseline models indiscriminately compress detail, SCOPE (*F1*). Second, it reduces *profile hallucination* that skews content selection (*F2*). SCOPE's structured reflection on the persona and input acts as working memory for the model to ground decisions. We conclude that this reusable working memory and reasoning outperform static persona injection.

Human Assessment. To confirm that human readers perceive the performance gains of SCOPE, we conduct a comparative human evaluation. We task five annotators (three male, two female; ages 22-29; C1+ English) with assessing personalized summaries generated by SCOPE and the two baselines, reader-tailoring and roleplaying. We created 20 evaluation sets, each corresponding to a unique source transcript and a specific reader profile. Each set contains three summaries, i.e., one from reader-Tailoring, roleplaying, and SCOPE. All annotators evaluate all 20 sets, with the three summaries within each set presented in a randomized order to mitigate presentation bias. The annotators perform two tasks: (1) rating each summary on the seven P-MESA criteria (1-5 Likert scale) and (2) ranking the three summaries from best (1) to worst (3). The whole experimental setup follows the details in Appendix C. We find substantial inter-annotator agreement on the rating task (Krippendorff's α =0.71).

Approach	Reader-Tailoring	Roleplaying	SCOPE
Huamn annotation following P-MESA criteria (lower is better)			
goal alignment	$3.5_{0.70}$	$2.5_{0.98}$	$2_{0.81}$
completeness	$4.5_{0.88}$	$3_{1.58}$	$3_{0.65}$
factuality	$2.5_{1.38}$	$3.5_{1.58}$	$2_{0.77}$
knowledge level fit	$2.5_{1.40}$	$3_{1.53}$	$1_{1.73}$
priority structuring	$3_{0.26}$	$3_{0.30}$	$3_{0.28}$
contextual framing	$4.5_{1.35}$	$4_{1.07}$	$3_{0.70}$
relevance	$3.5_{0.48}$	$2.5_{0.86}$	$2.5_{0.40}$
Ranking (1-3, median, lower is better)			
Ranking	$3_{0.49}$	$2_{0.75}$	$1_{0.40}$

Table 4: Human evaluation of summaries generated through reader-tailoring, roleplaying, and SCOPE. Annotators rate summaries on the P-MESA criteria and rank the summaries. Green is best in category.

The results, presented in Table 4, confirm that human judges consistently prefer summaries from SCOPE. It achieves the best median rank of 1 and is rated highest on five of the seven P-MESA criteria, with strong improvements in factuality and knowledge level fit. This indicates that readers prefer the improved grounding and tailored language. Interestingly, scores for priority structuring were identical across all methods, suggesting this was a less decisive factor for annotators. In one instance, a summary from reader-tailoring was ranked highest due to a preference for its presentation, highlighting the subjective nature of structural choices in summarization.

6 Ablations

Safety mechanisms. We assess the `Fact Verification` stage, finding only 8 in 150 facts required revision (1 false positive, 0 false negatives).

Removing this component causes no measurable quality drop with GPT, suggesting it serves as a strategic guardrail for less capable models. Similarly, removing `summary refinement` only increases incoherence slightly (MESA: 3→3.5) without affecting other dimensions. Appendices F.1 and F.2 provide detailed analysis.

Generalization. FRAME shows robust performance across diverse models (Gemini 1.5 Pro, Llama 3.1 8b, Gemma 3 4b), with consistent improvements over single-LLM baselines. Notably, FRAME narrows the gap between commercial and open-source models on MESA dimensions. When applied to general text summarization benchmarks, FRAME yields an average MESA improvement of over 1 point per dimension, with the strongest and most consistent gains coming from repetition (-3) reduction and frequent large gains in omission. This confirms that the improvements stem from fact-based reasoning, rather than specific model capabilities. To further test generalization on domain-specific conversational data, we use the meeting simulation approach, MIMIC (Kirstein et al., 2025a), to transform scientific articles from the PubMed dataset (Xiong et al., 2024) into realistic meeting transcripts. Applying FRAME to summarize these simulated meetings yields gains in maintaining factual consistency, with improvements in hallucination (3 → 1), repetition (-2.5 points), and language (-2 points). See Appendices G.2 and G.3 for complete evaluation.

Architecture. To justify our pipeline's seven-step modular design, we evaluated three compressed variants against our full framework: a single prompt performing all steps (`combined-1`), merging Fact Extraction and Relevance Scoring (`combined-2`, and combining Relevance Scoring and Outline Planning (`combined-3`). The results in Table 16 show that collapsing stages causes a degradation in content fidelity. All three combined variants score a 4 in omission, a 3-point drop from FRAME, indicating they consistently miss key information. Hallucination increases, with `combined-2` scoring a 5. We conclude that forcing the model to extract and evaluate facts simultaneously prevents it from properly understanding the content and grounding claims. While metric dimensions such as Structure and Incoherence show less degradation, the failure on Factuality demonstrates the necessity of our distinct pipeline stages for producing reliable summaries.

7 Final Considerations

We introduced FRAME, a fact-based pipeline treating abstractive meeting summarization as an enrichment task. FRAME tackled the core challenges of meetings, i.e., information distribution, contextual dependencies, and salience ambiguity, while reducing hallucinations (3→1), irrelevance (2→1), and repetition (3→1). To handle personalization, we introduced SCOPE. SCOPE builds on a *reason-out-loud* protocol, guiding a model through nine questions to construct an interpretation trace of the target reader and content before filtering. This improved relevance (3→1), alignment with user goals (3→2), and knowledge level fit (2→1). We further contributed P-MESA, an LLM-based personalization quality metric, which assesses the reader fit across seven criteria, offering interpretable scores and strong alignment with human judgment ($\rho \geq$70%).

Our contributions advance multiple research directions in summarization, from fact-level content control enabling fine-grained adaptability for cross-document summarization, to robust salience detection through structured extraction and verification, to reference-free evaluation of reader-specific appropriateness. By releasing FRAME, SCOPE, and P-MESA as open-source, we provide a powerful toolkit that researchers can extend to multilingual applications, multi-source integration, and domains beyond meeting contexts. This work addresses a persistent gap in summarization approaches that lack the ability to comprehend content before summarization, fostering advancements in summarization and personalization.

Limitations

The performance of FRAME and SCOPE depends significantly on the capabilities of the underlying language model. While our implementation uses GPT-4o, models with different reasoning capabilities or smaller context windows may produce less accurate fact identification or inferior reasoning traces, potentially reducing the quality of both general and personalized summaries. Our ablation studies suggest that while performance decreases with less capable models, the core benefits of our fact-based paradigm remain intact, demonstrating the framework's architectural robustness.

The evaluation datasets, although comprehensive, cover only specific meeting types. QMSum encompasses academic, business, and parliamentary meetings, while FAME provides synthetic meetings across 14 formats. Other corpora exist (e.g., MeetingBank (Hu et al., 2023), ELITR (Nedoluzhko et al., 2022)) but closely resemble QMSum's formal institutional settings and would not increase the variety. While we selected these datasets for their broadest coverage among publicly available datasets, they do not capture all real-world meeting dynamics, particularly those in specialized domains (e.g., medical consultations, legal) that may require domain-specific fact extraction patterns.

The computational requirements of FRAME represent another limitation. The multi-stage pipeline, particularly the fact extraction and verification stages, incurs both increased inference time and higher computational costs compared to end-to-end summarization approaches. While quality improvements justify this trade-off, it may limit deployment in resource-constrained environments or real-time applications. As we demonstrate in our ablation study, downscaling the backbone model to reduce costs can still yield higher-quality summaries than those produced by end-to-end summarization.

References

2025. Introducing Ai2 Paper Finder | Ai2. https://allenai.org/blog/paper-finder.

Reinald Kim Amplayo, Stefanos Angelidis, and Mirella Lapata. 2021. Unsupervised Opinion Summarization with Content Planning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(14):12489–12497.

Negar Arabzadeh, Ali Ahmadvand, Julia Kiseleva, Yang Liu, Ahmed Hassan Awadallah, Ming Zhong, and Milad Shokouhi. 2023. PREME: Preference-based Meeting Exploration through an Interactive Questionnaire. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 331–342, Dubrovnik, Croatia. Association for Computational Linguistics.

Abedelkadir Asi, Song Wang, Roy Eisenstadt, Dean Geckt, Yarin Kuper, Yi Mao, and Royi Ronen. 2022. An End-to-End Dialogue Summarization System for Sales Calls. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Industry Track*, pages 45–53, Hybrid: Seattle, Washington + Online. Association for Computational Linguistics.

Jean Carletta, Simone Ashby, Sebastien Bourban, Mike Flynn, Mael Guillemot, Thomas Hain, Jaroslav Kadlec, Vasilis Karaiskos, Wessel Kraaij, Melissa Kronenthal, Guillaume Lathoud, Mike Lincoln, Agnes Lisowska, Iain McCowan, Wilfried Post, Dennis Reidsma, and Pierre Wellner. 2006. The AMI Meeting Corpus: A Pre-announcement. In *Machine Learning for Multimodal Interaction*, Lecture Notes in Computer Science, pages 28–39, Berlin, Heidelberg. Springer.

Jiaao Chen, Mohan Dodda, and Diyi Yang. 2023. Human-in-the-loop Abstractive Dialogue Summarization. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 9176–9190, Toronto, Canada. Association for Computational Linguistics.

I.-Chun Chern, Steffi Chern, Shiqi Chen, Weizhe Yuan, Kehua Feng, Chunting Zhou, Junxian He, Graham Neubig, and Pengfei Liu. 2023. FacTool: Factuality Detection in Generative AI – A Tool Augmented Framework for Multi-Task and Multi-Domain Scenarios. *Preprint*, arXiv:2307.13528.

Cheng-Han Chiang and Hung-yi Lee. 2024. Merging Facts, Crafting Fallacies: Evaluating the Contradictory Nature of Aggregated Factual Claims in Long-Form Generations. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2734–2751, Bangkok, Thailand. Association for Computational Linguistics.

Eunsol Choi, Jennimaria Palomaki, Matthew Lamm, Tom Kwiatkowski, Dipanjan Das, and Michael Collins. 2021. Decontextualization: Making Sentences Stand-Alone. *Transactions of the Association for Computational Linguistics*, 9:447–461.

Arman Cohan, Franck Dernoncourt, Doo Soon Kim, Trung Bui, Seokhwan Kim, Walter Chang, and Nazli Goharian. 2018. A Discourse-Aware Attention Model for Abstractive Summarization of Long Documents. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 615–621, New Orleans, Louisiana. Association for Computational Linguistics.

Sourish Dasgupta, Ankush Chander, Parth Borad, Isha Motiyani, and Tanmoy Chakraborty. 2024. PerSEval: Assessing Personalization in Text Summarizers. *Preprint*, arXiv:2407.00453.

Brigitte Endres-Niggemeyer. 2000. SimSum: An empirically founded simulation of summarizing. *Information Processing & Management*, 36(4):659–682.

Xue-Yong Fu, Md Tahmid Rahman Laskar, Elena Khasanova, Cheng Chen, and Shashi Tn. 2024. Tiny Titans: Can Smaller Large Language Models Punch Above Their Weight in the Real World for Meeting Summarization? In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 6: Industry Track)*, pages 387–394, Mexico City, Mexico. Association for Computational Linguistics.

Samira Ghodratnama and Mehrdad Zakershahrak. 2024. SumRecom: A Personalized Summarization Approach by Learning from Users' Feedback. *Preprint*, arXiv:2408.07294.

Logan Golia and Jugal Kalita. 2023. Action-Item-Driven Summarization of Long Meeting Transcripts. In *Proceedings of the 2023 7th International Conference on Natural Language Processing and Information Retrieval*, pages 91–98, Seoul Republic of Korea. ACM.

Gemini Team Google, Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry, Lepikhin, and Timothy Lillicrap. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *Preprint*, arXiv:2403.05530.

Matt Grenander, Siddharth Varia, Paula Czarnowska, Yogarshi Vyas, Kishaloy Halder, and Bonan Min. 2025. Exploration of Plan-Guided Summarization for Narrative Texts: The Case of Small Language Models. *Preprint*, arXiv:2504.09071.

Anisha Gunjal and Greg Durrett. 2024. Molecular Facts: Desiderata for Decontextualization in LLM Fact Verification. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 3751–3768, Miami, Florida, USA. Association for Computational Linguistics.

Naeemul Hassan, Fatma Arslan, Chengkai Li, and Mark Tremayne. 2017. Toward Automated Fact-Checking: Detecting Check-worthy Factual Claims by ClaimBuster. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1803–1812, Halifax NS Canada. ACM.

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, Toronto, Canada. Association for Computational Linguistics.

A. Janin, D. Baron, J. Edwards, D. Ellis, D. Gelbart, N. Morgan, B. Peskin, T. Pfau, E. Shriberg, A. Stolcke, and C. Wooters. 2003. The ICSI Meeting Corpus. In *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03).*, volume 1, pages I–I.

Jeesu Jung, Hyein Seo, Sangkeun Jung, Riwoo Chung, Hwijung Ryu, and Du-Seong Chang. 2023. Interactive User Interface for Dialogue Summarization. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*, pages 934–957, Sydney NSW Australia. ACM.

Ryo Kamoi, Tanya Goyal, Juan Diego Rodriguez, and Greg Durrett. 2023. WiCE: Real-World Entailment for Claims in Wikipedia. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7561–7583, Singapore. Association for Computational Linguistics.

Frederic Kirstein, Muneeb Khan, Jan Philip Wahle, Terry Ruas, and Bela Gipp. 2025a. You need to MIMIC to get FAME: Solving Meeting Transcript Scarcity with a Multi-Agent Conversations. *Preprint*, arXiv:2502.13001.

Frederic Kirstein, Terry Ruas, Robert Kratel, and Bela Gipp. 2024. Tell me what I need to know: Exploring LLM-based (Personalized) Abstractive Multi-Source Meeting Summarization. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 920–939, Miami, Florida, US. Association for Computational Linguistics.

Frederic Kirstein, Jan Philip Wahle, Bela Gipp, and Terry Ruas. 2025b. CADS: A Systematic Literature Review on the Challenges of Abstractive Dialogue Summarization. *J. Artif. Int. Res.*, 82.

Frederic Thomas Kirstein, Terry Lima Ruas, and Bela Gipp. 2025c. Is my Meeting Summary Good? Estimating Quality with a Multi-LLM Evaluator. In *Proceedings of the 31st International Conference on Computational Linguistics: Industry Track*, pages 561–574, Abu Dhabi, UAE. Association for Computational Linguistics.

Frederic Thomas Kirstein, Terry Lima Ruas, and Bela Gipp. 2025d. What‘s Wrong? Refining Meeting Summaries with LLM Feedback. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 2100–2120, Abu Dhabi, UAE. Association for Computational Linguistics.

Klaus Konrad. 2017. Lautes Denken in psychologischer Forschung und Praxis. In Günter Mey and Katja Mruck, editors, *Handbuch Qualitative Forschung in der Psychologie*, pages 1–21. Springer Fachmedien, Wiesbaden.

Md Tahmid Rahman Laskar, Xue-Yong Fu, Cheng Chen, and Shashi Bhushan Tn. 2023. Building Real-World Meeting Summarization Systems using Large Language Models: A Practical Perspective. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 343–352, Singapore. Association for Computational Linguistics.

Haoran Li, Arash Einolghozati, Srinivasan Iyer, Bhargavi Paranjape, Yashar Mehdad, Sonal Gupta, and Marjan Ghazvininejad. 2021. EASE: Extractive-Abstractive Summarization End-to-End using the Information Bottleneck Principle. In *Proceedings of the Third Workshop on New Frontiers in Summarization*, pages 85–95, Online and in Dominican Republic. Association for Computational Linguistics.

Junyi Jessy Li, Bridget O'Daniel, Yi Wu, Wenli Zhao, and Ani Nenkova. 2016. Improving the Annotation of Sentence Specificity. In *Proceedings of the Tenth*

International Conference on Language Resources and Evaluation (LREC'16), pages 3921–3927, Portorož, Slovenia. European Language Resources Association (ELRA).

Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Bingbin Liu, Jordan Ash, Surbhi Goel, Akshay Krishnamurthy, and Cyril Zhang. 2023a. Exposing Attention Glitches with Flip-Flop Language Modeling. *Advances in Neural Information Processing Systems*, 36:25549–25583.

Yang Liu and Mirella Lapata. 2019. Hierarchical Transformers for Multi-Document Summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5070–5081, Florence, Italy. Association for Computational Linguistics.

Yixin Liu, Alex 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 (Volume 1: Long Papers)*, pages 4140–4170, Toronto, Canada. Association for Computational Linguistics.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. FActScore: Fine-grained Atomic Evaluation of Factual Precision in Long Form Text Generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12076–12100, Singapore. Association for Computational Linguistics.

Yuji Naraki, Tetsuya Sakai, and Yoshihiko Hayashi. 2022. Evaluating the Effects of Embedding with Speaker Identity Information in Dialogue Summarization. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 298–304, Marseille, France. European Language Resources Association.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. Don't Give Me the Details, Just the Summary! Topic-Aware Convolutional Neural Networks for Extreme Summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1797–1807, Brussels, Belgium. Association for Computational Linguistics.

Anna Nedoluzhko, Muskaan Singh, Marie Hledíková, Tirthankar Ghosal, and Ondřej Bojar. 2022. ELITR Minuting Corpus: A Novel Dataset for Automatic Minuting from Multi-Party Meetings in English and Czech. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3174–3182, Marseille, France. European Language Resources Association.

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: HLT-NAACL 2004*, pages 145–152, Boston, Massachusetts, USA. Association for Computational Linguistics.

OpenAI. 2024. GPT-4 Technical Report. *Preprint*, arXiv:2303.08774.

Stefano De Paoli. 2023. Improved prompting and process for writing user personas with LLMs, using qualitative interviews: Capturing behaviour and personality traits of users.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language Models are Unsupervised Multitask Learners.

Virgile Rennard, Guokan Shang, Julie Hunter, and Michalis Vazirgiannis. 2023. Abstractive Meeting Summarization: A Survey. *Transactions of the Association for Computational Linguistics*, 11:861–884.

Eva Sharma, Chen Li, and Lu Wang. 2019. BIGPATENT: A Large-Scale Dataset for Abstractive and Coherent Summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2204–2213, Florence, Italy. Association for Computational Linguistics.

Maarten W. Solomon, Yvonne F. Barnard, and Jacobijn A. C. Sandberg. 1995. The think aloud method: A practical guide to modelling cognitive processes. *Information Processing & Management*, 31(6):906–907.

Liyan Tang, Igor Shalyminov, Amy Wong, Jon Burnsky, Jake Vincent, Yu'an Yang, Siffi Singh, Song Feng, Hwanjun Song, Hang Su, Lijia Sun, Yi Zhang, Saab Mansour, and Kathleen McKeown. 2024. TofuEval: Evaluating Hallucinations of LLMs on Topic-Focused Dialogue Summarization. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4455–4480, Mexico City, Mexico. Association for Computational Linguistics.

Rahul Vansh, Darsh Rank, Sourish Dasgupta, and Tanmoy Chakraborty. 2023. Accuracy is not enough: Evaluating Personalization in Summarizers. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2582–2595, Singapore. Association for Computational Linguistics.

Yuxia Wang, Revanth Gangi Reddy, Zain Muhammad Mujahid, Arnav Arora, Aleksandr Rubashevskii, Jiahui Geng, Osama Mohammed Afzal, Liangming

Pan, Nadav Borenstein, Aditya Pillai, Isabelle Augenstein, Iryna Gurevych, and Preslav Nakov. 2024. Factcheck-Bench: Fine-Grained Evaluation Benchmark for Automatic Fact-checkers. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14199–14230, Miami, Florida, USA. Association for Computational Linguistics.

Miriam Wanner, Seth Ebner, Zhengping Jiang, Mark Dredze, and Benjamin Van Durme. 2024. A Closer Look at Claim Decomposition. In *Proceedings of the 13th Joint Conference on Lexical and Computational Semantics (*SEM 2024)*, pages 153–175, Mexico City, Mexico. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2024. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, pages 24824–24837, Red Hook, NY, USA. Curran Associates Inc.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. 2024. Efficient Streaming Language Models with Attention Sinks. In *Proceedings of The Twelfth International Conference on Learning Representations*.

Guangzhi Xiong, Qiao Jin, Zhiyong Lu, and Aidong Zhang. 2024. Benchmarking retrieval-augmented generation for medicine. *arXiv preprint arXiv:2402.13178*.

Mingkai Zhang, Dan You, and Shouguang Wang. 2024a. Novel framework for dialogue summarization based on factual-statement fusion and dialogue segmentation. *PLOS ONE*, 19(4):e0302104.

Shiyue Zhang and Mohit Bansal. 2021. Finding a Balanced Degree of Automation for Summary Evaluation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6617–6632, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Yanyue Zhang, Yulan He, and Deyu Zhou. 2025. Rehearse With User: Personalized Opinion Summarization via Role-Playing based on Large Language Models. *Preprint*, arXiv:2503.00449.

Yusen Zhang, Ansong Ni, Ziming Mao, Chen Henry Wu, Chenguang Zhu, Budhaditya Deb, Ahmed Awadallah, Dragomir Radev, and Rui Zhang. 2022. Summ^N: A Multi-Stage Summarization Framework for Long Input Dialogues and Documents. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1592–1604, Dublin, Ireland. Association for Computational Linguistics.

Zhehao Zhang, Ryan A. Rossi, Branislav Kveton, Yijia Shao, Diyi Yang, Hamed Zamani, Franck Dernoncourt, Joe Barrow, Tong Yu, Sungchul Kim, Ruiyi Zhang, Jiuxiang Gu, Tyler Derr, Hongjie Chen, Junda Wu, Xiang Chen, Zichao Wang, Subrata Mitra, Nedim Lipka, Nesreen Ahmed, and Yu Wang. 2024b. Personalization of Large Language Models: A Survey. *Preprint*, arXiv:2411.00027.

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, Online. Association for Computational Linguistics.

Rongxin Zhu, Jey Han Lau, and Jianzhong Qi. 2024. Factual Dialogue Summarization via Learning from Large Language Models. *Preprint*, arXiv:2406.14709.

Organisation of Appendix

This appendix provides further details and ablations for our paper. Appendices A and B document reproducibility information and experimental setup. Appendices C and D detail our human evaluation protocols and computational efficiency. Appendices E to G contain empirical observations and ablation studies demonstrating component contributions. Appendices H and I provide dataset characteristics and SCOPE component details. Appendices J to M cover evaluation methodology, definitions, and qualitative examples. Researchers primarily interested in reimplementing our work should focus on Appendices A and B, and the prompt templates in Figures 6–20. Those evaluating methodology rigor should examine Appendices C, J and K.

A Open Resources & Licensing

A.1 Repository & License

The FRAME framework proposed in this paper, along with the SCOPE reasoning protocol and P-MESA metric, are available on GitHub under an MIT license. The repository includes implementation code, evaluation scripts, prompt templates, and configuration files necessary for reproducing our results.

A.2 Datasets & Licensing

In Table 5, we report the licensing and a high-level overview of the QMSum (Zhong et al., 2021) and FAME (Kirstein et al., 2025a) datasets. For the FAME dataset, which is not publicly available at the time of writing, we obtained research access by contacting the authors directly. The authors have indicated plans for public release under CC BY-SA 4.0 licensing. Researchers seeking to replicate our results can either access the dataset when published or contact the authors using the information provided in (Kirstein et al., 2025a). A detailed overview of dataset characteristics is given in Appendix H.

Dataset	License	Size	Avg. Length	Domain
QMSum	MIT	232	7303	academic, council, design-meeting
FAME	CC BY-SA 4.0	800	6250	various, e.g., sports, technology, history, math, philosophy

Table 5: Dataset licensing and high-level overview.

B Experimental Setup Details

This section provides implementation details, including all prompt templates, hyperparameters, and computational infrastructure used in all experiments. We document the exact methodology to ensure reproducibility of FRAME and SCOPE, including parameters that were empirically determined.

B.1 Implementation Details

We address two key challenges in processing lengthy meeting transcripts: managing context limitations and maintaining information coherence across processing stages. As we use an Azure GPT instance with a capped 4K-token output, we developed two specialized components to handle meetings with facts exceeding this limit:

A **Chunk Processor** divides transcripts into sequential chunks based on GPT-2 tokenizer estimation (Radford et al., 2019). Each chunk maintains access to the previous chunk to preserve cross-boundary information. We employ dynamic chunk sizing, prioritizing complete speaker turns over fixed token counts while respecting maximum context limitations.

A **Memory Bank** provides a centralized fact repository that decouples extraction from downstream processing. To address cross-chunk fact redundancy, we apply a lightweight text similarity function that combines character-level sequence alignment and token-level word overlap to merge contexts with $\geq 70\%$ similarity while preserving higher relevance scores.

FRAME's modular design further allows adaptation to varying LLM context capacities, enabling integration with both smaller open-source models and commercial API-based services.

B.2 Model Specs

We employ four different language models in this work. Appendix B.2 provides an overview of publicly disclosed information about these models.

Model	Version	Parameters	Provider
GPT	4o, 2024-08-06	~200B	OpenAI
Gemini	1.5-pro-002	not disclosed	Google
Llama	3.1	~8B	Meta AI
Gemma	3	~4B	Google

Table 6: Model specifications as reported in original papers and disclosing works. GPT and Gemini are our main models for experiments (Section 5), Llama and Gemma for ablations Appendix G.2.

B.3 Prompt Templates

The FRAME pipeline and SCOPE protocol rely on carefully crafted prompt templates for each processing stage. Figures 6–20 present the complete set of prompts used throughout our experiments.

Core FRAME Prompts (Figures 6–12): These prompts guide the four main stages of our general summarization pipeline. Stage 1 prompts (Figures 6 and 7) handle fact extraction and verification. Stage 2 prompts (Figure 8) focus on relevance scoring and fact grouping. Stage 3 prompts (Figure 9) guide outline planning, while Stage 4 prompts (Figures 10–12) manage summary generation and quality assurance.

SCOPE Protocol Prompts (Figures 13–18): These prompts implement our personalization approach. They guide the model through explicit reasoning about reader preferences (Figure 13), fact selection based on persona relevance (Figures 14 and 15), and persona-focused outline and summary generation (Figures 16–18).

Baseline Comparison Prompt (Figure 20): This prompt enables direct comparison with single-LLM personalization approaches by implementing reader-tailoring in a one-shot generation task.

All prompts follow a consistent structure with clearly defined input and output formats, explicit instructions, and constraints that guide the model toward the desired behavior. We designed these prompts iteratively through systematic experimentation, refining each prompt to maximize effectiveness while maintaining reproducibility.

B.4 Hyperparameters

For our experiments, we overly use default values for key hyperparameters, i.e., `top-p` = 1.0, `frequency penalty` = 0.0, `presence penalty` = 0.0. We empirically chose `temperature` = 0.1 to have the model behave more focused and deterministic. All values are fixed across the different model backbones used (Appendix B.2).

C Human Evaluation Protocols

This section details our human evaluation methodology, which undergirds both the development of our P-MESA metric and the assessment of the suitability of P-MESA as a proxy for human annotations (see Section 4).

C.1 Annotator Recruitment & Demographics

We have an annotation team of six participants (three male, three female, ages 22-29) through a structured recruitment process. All annotators were employed as research assistants or doctoral candidates with standardized contracts. We selected annotators based on their availability to complete tasks without time pressure, demonstrated English proficiency (native speakers or C1-C2 certified), and academic background relevant to text analysis. This selection process yielded a team with diverse disciplinary perspectives: two computer science students, three psychology students, and one communication science student. All annotators provided explicit consent for their anonymized annotations to be used in this research, and the entire annotation protocol received approval from our institution's ethics committee before implementation.

C.2 Training & Quality-Control

Preparation: We have prepared a comprehensive handbook for our annotators, detailing the project context and defining the criteria (a short version is presented in Table 21 and an extended version with more details). Each definition includes two examples: one with minimal impact on quality and one with high impact. The handbook explains the 1 - 5 Likert rating for the individual questionnaires. The handbook does not specify an order for processing the items. We provide the handbook in English and the annotators' native languages, using professional translations.

We structured our timeline as a four-week process: one week dedicated to onboarding, followed by three weeks for primary annotation. The first annotation week featured twice-weekly check-ins, which transitioned to weekly meetings for the subsequent periods. In parallel, the research team conducted quality assessments without the annotator's presence weekly to identify emerging issues (Note: week refers to a regular working week.)

Onboarding: The onboarding week is dedicated to getting to know the project and familiarizing oneself with the definitions and data. We begin with a kick-off meeting to introduce the project and explain the handbook, particularly focusing on each definition. We generate ten additional samples for the individual tasks following the respective approaches to familiarize. After processing the first five samples, we hold individual meetings to clarify any confusion. The remaining five samples are then annotated to confirm clarity. A second group meeting this week addresses any misaligned understanding among the reviewers. After the group meeting, we meet individually with the annotators to review their work and ensure their quality and understanding of the task and samples. Judging from the reasoning they provide for each decision and annotation, all annotators demonstrate reliable performance and good comprehension of the task and definitions.

Annotation Process: For the primary annotation workflow, we distributed the workload equally among annotators with distinct approaches for the two evaluation phases. During the P-MESA development phase, each annotator evaluated $\sim$17 samples, with every sample receiving annotations from two independent evaluators. For the P-MESA validation, each annotator assessed $\sim$8 samples, with every sample being evaluated by four distinct annotators to ensure robust reliability assessment.

To maintain unbiased evaluation, annotators remained blind to the summarization architecture, which generated each summary, the source dataset, and other annotators' ratings. We randomized the sample presentation order for each annotator to mitigate positional bias. They are given a week to complete their set at their own pace and with their break times. Quiet working rooms were provided if needed for concentration. Annotators can choose their annotation order for each sample and are allowed to revisit previous samples.

Regular meetings are held to address any emerging issues or questions on definitions. During the quality checks the authors perform, we look for incomplete annotations, missing explanations, and signs of misunderstanding based on the provided reasoning. If the authors find such a lack of quality, the respective annotators will be notified to redo the annotation. At halftime of the annotation cycle, we compute inter-annotator agreement scores. If we observed a significant difference among annotators, we planned a dedicated meeting with all annotators and a senior annotator to discuss such cases. On average, annotators spend 25 minutes per sample.

Handling of unexpected cases: Given that our annotators have other commitments, we anticipate potential scheduling conflicts. We allow flexibility for annotators to complete their samples beyond the week limit if needed, reserving an additional week as a buffer. Despite these provisions, all annotators complete their assigned samples within the original weekly timeframes. We further allow faster annotators to continue with an additional sample set. This additional work was voluntary.

C.3 Inter-Annotator Agreement Formulas

We assessed annotation reliability using Krippendorff's alpha (α), which we selected for its ability to accommodate multiple annotators, ordinal data, and handle missing values—characteristics that make it well-suited for analyzing Likert-scale ratings across multiple dimensions. This metric ranges from 0 (agreement attributable to chance) to 1 (perfect agreement), with established thresholds in computational linguistics literature for interpreting reliability strength.

As shown in Table 7, all dimensions achieved substantial to strong agreement. The Factuality dimension demonstrated the highest consistency ($\alpha = 0.839$), likely due to its more concrete definition and readily observable manifestations in summary text. The relatively lower agreement on Knowledge Level Fit ($\alpha = 0.681$) reflects the inherent subjectivity in assessing information prioritization, though it still comfortably exceeds the threshold for substantial reliability. These strong reliability indicators validate our annotation protocol and suggest that P-MESA criteria are consistently interpretable across different human evaluators.

Dimension	Krippendorff's α
Factuality	0.839
Completeness	0.811
Relevance	0.762
Goal Alignment	0.758
Priority Structuring	0.758
Knowledge Level Fit	0.681
Contextual Framing	0.792

Table 7: Inter-rater reliability for the human annotations, measured by Krippendorff's alpha. Scores $\geq$0.667 mean moderate agreement, and scores $\geq$0.8 mean strong agreement.

D Computational Efficiency Analysis

This section compares the computational requirements of our proposed approaches with those of the baselines. We measure efficiency through token utilization, execution time, and associated API costs, providing practical considerations for deployment.

D.1 Token & API Cost Breakdown

We calculated token usage by instrumenting all API calls to track input and output tokens across each pipeline component. API costs were calculated using OpenAI's published GPT-4o pricing (\$0.01 per 1K input tokens and \$0.03 per 1K output tokens as of August 2024). Each reported value represents the mean across 50 meeting summarizations from our evaluation dataset, with meetings averaging 7,303 words.

Table 8 provides a detailed breakdown of computational requirements across pipeline components and alternative approaches. The most resource-intensive components of FRAME are `Fact Extraction` (24.1K input tokens, 3.8K output tokens) and `Fact Verification` (26.9K input tokens, 3.8K output tokens), together accounting for $\sim$62% of the pipeline's total computation.

`Fact Verification` can be seen as optional based on our ablation studies (Section 6, detailed in Appendix F), which show minimal quality degradation when it is removed. This offers a potential 33% reduction in compute time (59 seconds) and cost reduction of \$0.07 per meeting for deployment scenarios with stricter efficiency requirements.

The SCOPE personalization protocol adds minimal overhead when integrated with FRAME, requiring only 3076 additional input tokens and 3179 output tokens while delivering notable personalization improvements as demonstrated in Section 3.3.

Approach	Input #Tokens	Output #Tokens	Est. Cost $/meeting	Time seconds
FRAME	72,059	11,674	0.21	225
Fact Extraction	24,143.8	3,753.5	0.06	58
Fact Verification	26,925.2	3,813.0	0.07	59
Fact Scoring & Fact Grouping	7,287.0	3,880.0	0.02	76
Outline Planning	859	523.2	$\leq$0.01	7
Enrichment-based Generation	12,704.8	462.5	0.03	14
Quality Assurance	12,692.0	160.5	0.03	11
SCOPE (additional)	3076.1	3179.2	0.03	13
Single LLM	23,725.5	233.0	0.06	5

Table 8: Token usage and estimated API costs for the individual components and approaches considered.

D.2 Quality-Cost Trade-off Analysis

Figure 3 illustrates the relationship between computational cost and summary quality across different approaches. The quality metric represents a composite score derived from the eight MESA dimensions (Section 5), with higher values (1-10 scale) indicating better quality. The horizontal axis represents the average cost per summary in US dollars.

The analysis reveals that single-model baselines (blue circles) appear in the bottom-left quadrant, offering low cost ($0.01-$0.05 per summary) but limited quality (scores 4.0-5.3). FRAME implementations (orange squares) consistently achieve higher quality scores (6.5-7.5) across various backbone models, with costs ranging from $0.02 to $0.18 per summary.

FRAME-LLAMA (FRAME with Llama 3.1) offers an excellent balance of quality and cost-efficiency, achieving a quality score of approximately 6.7 at just $0.03 per summary. This represents a substantial quality improvement over the single LLM baselines while maintaining competitive costs. FRAME-GPT delivers the highest quality among FRAME implementations (score $\sim$7.5) but at higher cost ($0.18), placing it top-right.

The feedback-based approach with three iterations (FEEDBACK-3, green star) achieves competitive quality (score $\sim$6.7) but at the highest cost ($\sim$ $0.30), demonstrating that FRAME's structured approach provides better efficiency than iterative refinement. The light green region highlights the ideal zone of high quality and low cost, where FRAME implementations with open-source models (FRAME-GEMMA, FRAME-LLAMA) deliver particularly favorable trade-offs for practical deployment scenarios.

E Empirical Observations in the Pipeline

To understand how FRAME processes and filters information, we analyze the *Fact Selection* across 100 summaries generated using a GPT backbone model (setup from Section 5). We examine how many facts are typically extracted from a meeting, how many are deemed relevant, and how many are utilized for outlining and summary writing.

Table 9 presents the number of facts throughout the processing stages. From an average of 103.3 initially extracted facts per meeting, approximately 40% (mean=41.0, SD=3.1) receive relevance scores $r_i \geq 6$ and are retained for potential inclusion in the summary outline. The structured outline formation follows a hierarchical relevance approach. High-priority facts (those receiving scores $r_i \geq 8$) used to define outline points average 8.67 facts per summary, with the majority (1.97) belonging to the `Decision` category. Supporting context facts (those with $6 \leq r_i < 8$) average 11.98 per summary, providing background and elaboration for the core points while maintaining a manageable information density. The distribution of facts across categories relates to domain-specific patterns. Parliamentary meetings (from the QMSum dataset) show a pronounced emphasis on `Decision` facts, with summaries containing up to 13 decision-related points and minimal `Discussion` or `Next Steps` content. This is expected, as parliament meetings are centered on decisions regarding petitions.

Notably, we observe no instances where outline points require removal due to hallucination or factual errors when using the GPT backbone. This suggests that when operating with advanced LLMs, the FRAME pipeline maintains high fidelity between extracted facts and generated outline points without introducing non-factual content. Based on our ablation studies with smaller models (Section 6), we hypothesize that less capable models would likely exhibit higher rates of outline point removal, potentially necessitating more robust quality control measures. For deployment scenarios using smaller models, we recommend enhancing the `Summary Verification` step to actively identify and replace potentially hallucinated outline points rather than simply removing them.

In sum, FRAME's multi-stage approach effectively condenses extensive meeting transcripts (averaging 7,303 words) into focused summaries built around approximately 20 high-relevance facts, achieving an overall compression ratio of $\sim$80% from extraction to final outline formation.

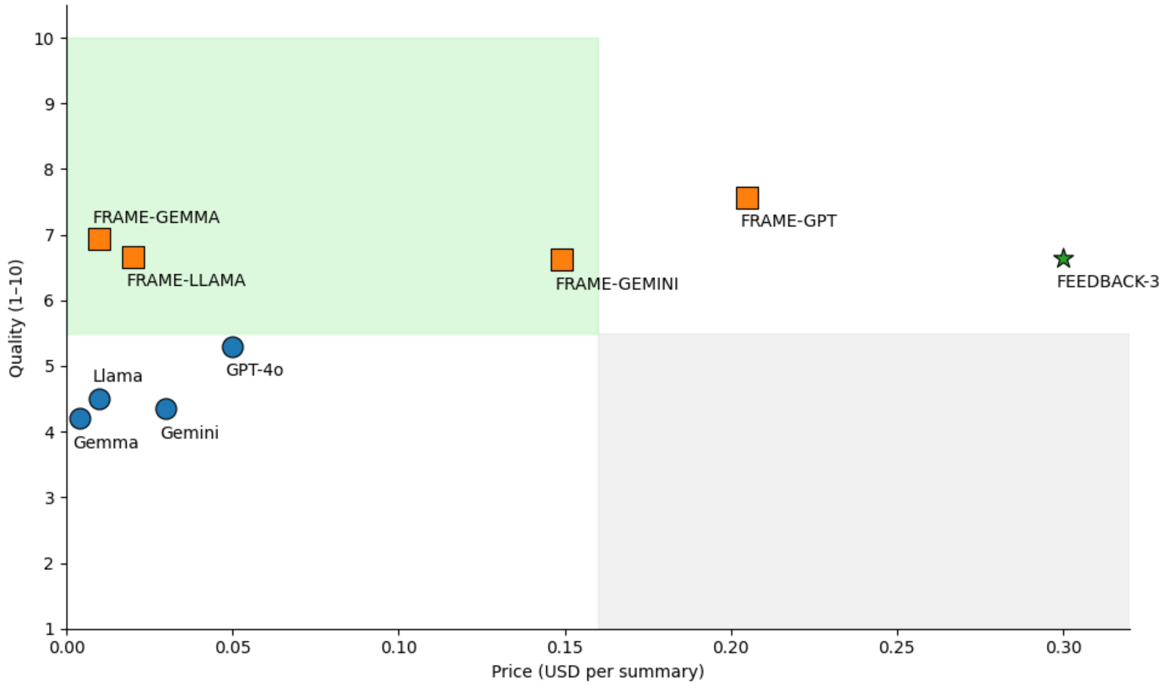

Figure 3: 4-quadrant plot of total architecture cost (avg.) vs quality measured by MESA. The top left indicates ideal high quality and low cost. Blue dots are single LLM instances for GPT-4o, Gemini 1.5 pro, Llama 3.1 8b, and Gemma 3 4b. Organce squares are FRAME summaries with the different backbones. FEEDBACK-3 relates to the self-refinement baseline by Kirstein et al. (2025d) with a GPT-4o backbone and three refinement loops.

Metric	Mean	Std. Dev.
Facts Extracted (total)	103.3	4.6
Facts Retained ($r_i \geq 6$)	41.0	3.1
Retention Rate (%)	39.8	2.3
Outline Facts ($r_i \geq 8$, main)	8.67	5.18
Decision	1.97	3.15
Discussion	0.78	1.54
Next Steps	0.74	1.47
Outline Facts ($r_i \geq 6$, context)	11.98	3.35
Decision	4.10	3.15
Discussion	1.85	1.51
Next Steps	2.45	1.53
Facts Removed (main or context)	0.0	0.0
Decision	0.0	0.0
Discussion	0.0	0.0
Next Steps	0.0	0.0

Table 9: Fact retention statistics across the FRAME pipeline showing information filtering patterns and distribution of selected facts by category and relevance tier.

F Ablation & Sensitivity Studies

We evaluate the contribution of FRAME's verification mechanisms (i.e., `Fact Verification`, `Quality Assurance`) and the sensitivity of the pipeline to parameter settings during `Outline Planning`. Through controlled ablation experiments, we quantify the impact of each verification component and analyze how varying relevance thresholds affect the summary quality. In sum, while these mechanisms show minimal impact, they provide safeguards against potential failures.

F.1 Fact Verification Impact

To evaluate the effectiveness of the fact verification component, we analyze 150 statement-context tuples randomly sampled from our experimental summaries (see Section 5). Three human annotators from our annotator pool (described in Appendix C, we follow a similar onboarding process) assess whether each extracted fact adheres to our Completeness and Minimalism criteria (Section 3.1) when compared against the original meeting transcript. Figure 4 presents representative examples of extracted facts from academic meetings, illustrating how our approach captures specific statements with appropriate contextual information. The examples demonstrate both technical facts (e.g., "The ANN performs nonlinear discriminant analysis") and comparative observations (e.g., "Without the neural network, the performance is better").

Our analysis reveals that the fact verification module identifies and regenerates facts in $\sim$5% of cases (8 out of 150 instances). Among these, human annotators identify one false positive, resulting in a 12.5% false positive rate. Two facts from the total set are found to contain excess contextual information that extends beyond what is explicitly stated in the transcript. The final example in Figure 4 illustrates this tendency, where the context includes interpretive elements ("suggesting an alternative method for feature processing"). The verification stage successfully detects both of these cases. Human annotators identify no violations of the Minimalism criterion.

To assess the impact of fact verification on overall summary quality, we compare MESA scores for summaries generated with and without this verification component. Table 10 presents this comparison using median scores and standard deviations across all evaluation dimensions. The results demonstrate minimal differences between the two configurations, with slightly higher variance in hallucination scores when verification is disabled. This suggests that while fact verification provides limited benefit for FRAME with a GPT backbone used in our main experiments, it may serve as an important guardrail for deployment scenarios with less capable models or more challenging meeting content.

Approach	**FRAME** w/ `Fact Verification`	**FRAME** w/o `Fact Verification`
Coreference	$0_{1.64}$	$0_{1.65}$
Hallucination	$1_{1.75}$	$1_{1.81}$
Incoherence	$3_{1.88}$	$3_{1.88}$
Irrelevance	$1_{1.45}$	$1_{1.44}$
Language	$1_{1.40}$	$1_{1.40}$
Omission	$1_{0.16}$	$1_{0.19}$
Repetition	$1_{1.23}$	$1_{1.23}$
Structure	$3_{1.24}$	$3_{1.23}$

Table 10: Changes in MESA scores for running FRAME on a GPT backbone with and without `Fact Verification` step. Values are Median$_{Std}$. MESA scores are 1–5 Likert ratings.

F.2 Summary Refinement Analysis

We examine the contribution of the summary verification step by comparing 10 FRAME summaries generated with and without this component, evaluating performance using both reference-based metrics (ROUGE, BERTScore) and reference-free MESA dimensions. Table 11 presents this comparison, with highlighting indicating superior performance for each metric.

With `Summary Verification` enabled, we observe improvements in incoherence (3.5→3), ROUGE-1 (+1.48), ROUGE-L (+1.41), and BERTScore (+1.24). Analysis of individual summaries reveals that the verification component primarily enforces the target length constraint (250 tokens), editing overly verbose summaries while preserving their core content. Content modifications were minimal in our test set, suggesting that the primary function of this component is enforcing structural and length constraints rather than correcting factual content when using a GPT-4o backbone. Based on these findings, we conclude that `Summary Verification` serves as a quality assurance mechanism that is particularly valuable for maintaining consistent output format and length constraints across diverse meeting types. For deployment scenarios with strict length requirements or when using less capable models, this component provides an important safeguard.

Approach	**FRAME** w/ `S. Verification`	**FRAME** w/o `S. Verification`
MESA (lower is better)		
Coreference	$0_{1.64}$	$0_{1.59}$
Hallucination	$1_{1.75}$	$1_{1.68}$
Incoherence	$3_{1.88}$	$3.5_{1.76}$
Irrelevance	$1_{1.45}$	$1_{1.35}$
Language	$1_{1.40}$	$1_{1.37}$
Omission	$1_{0.16}$	$1_{0.41}$
Repetition	$1_{1.23}$	$1_{0.30}$
Structure	$3_{1.24}$	$3_{1.64}$
General Evaluation Metrics (higher is better)		
R-1	$22.89_{5.80}$	$21.41_{4.72}$
R-2	$4.13_{2.44}$	$4.79_{2.10}$
R-L	$20.78_{5.23}$	$19.37_{4.39}$
BS (F1)	$85.67_{1.19}$	$84.43_{1.09}$

Table 11: Changes in MESA scores for running FRAME on a GPT-4o backbone with (*w/*) and without (*w/o*) `Summary Verification` step. Values are Median$_{Std}$. MESA scores are 1–5 Likert ratings, ROUGE (R-1/R-2/R-L) and BERTScore (BS) are 0–100. Green is best in category.

F.3 Threshold Variation Analysis

To understand how fact retention thresholds affect summary quality, we conducted a sensitivity analysis comparing three threshold configurations:

- **Default**: High-relevance ($r_i \geq 8$) for outline points, supporting facts ($r_i \geq 6$) for context
- **Low**: Relaxed thresholds with $r_i \geq 6$ for outline points, $r_i \geq 3$ for context facts
- **High**: Strict thresholds with $r_i \geq 10$ for outline points, $r_i \geq 8$ for context facts

Table 12 presents comprehensive evaluation results across these configurations.

The high-threshold configuration demonstrates worse scores across most dimensions, particularly in hallucination (+2 points), language quality (+2 points), and omission (+3 points). We hypothesize that this degradation occurs because the stringent thresholds eliminate moderately important facts that create essential context and connectivity between main points. The resulting summaries become fragmented, focusing on the highest-scored facts without adequate supporting information.

The low-threshold configuration shows mixed results, with improvements in language quality (-0.5 points) compared to the default. However, it underperforms on coreference (+3 points), omission (+3 points), and repetition (+3 points). This pattern suggests that incorporating too many lower-relevance facts introduces redundancy and dilutes the summary's focus, creating challenges in maintaining coherent entity references.

The default configuration achieves the best overall performance, showing optimal balance across nearly all evaluation dimensions. This suggests that our empirically determined thresholds ($r_i \geq 8$ for outline points, $r_i \geq 6$ for context) effectively balance informativeness and coherence, sufficiently capturing high-relevance content while excluding noise that could degrade summary quality. These findings validate our parameter selection for the main experiments in Section 5 and demonstrate the importance of appropriate fact filtration thresholds in the FRAME pipeline. They also highlight the framework's sensitivity to these parameters, suggesting that threshold tuning could further optimize performance for specific meeting types or domains.

Approach	default	low	high
MESA (lower is better)			
Coreference	$0_{1.64}$	$3_{0.41}$	$3_{0.00}$
Hallucination	$1_{1.75}$	$1.5_{1.94}$	$3_{1.82}$
Incoherence	$3_{1.88}$	$4_{1.60}$	$4_{0.55}$
Irrelevance	$1_{1.45}$	$1.5_{1.64}$	$2_{1.48}$
Language	$1_{1.40}$	$0.5_{1.47}$	$3_{1.52}$
Omission	$1_{0.16}$	$4_{0.00}$	$4_{0.00}$
Repetition	$1_{1.23}$	$4_{0.41}$	$2_{1.52}$
Structure	$3_{1.24}$	$3_{1.03}$	$4_{0.89}$
General Evaluation Metrics (higher is better)			
R-1	$22.89_{5.80}$	$21.71_{2.50}$	$21.15_{3.79}$
R-2	$4.13_{2.44}$	$4.19_{1.37}$	$3.65_{1.16}$
R-L	$20.78_{5.23}$	$19.25_{2.71}$	$19.10_{3.53}$
BS (F1)	$85.67_{1.19}$	$85.52_{0.48}$	$85.46_{0.64}$

Table 12: Impact of different retention thresholds on summary quality. Values are Median_{Std}. MESA scores are 1–5 Likert ratings, ROUGE (R-1/R-2/R-L) and BERTScore (BS) are 0–100. Green is best in category.

G Extended Analysis

This section extends our evaluation of FRAME through three complementary analyses: (1) a comparison of fact representation approaches, (2) a cross-model performance evaluation, and (3) a cross-domain application assessment. In sum, these experiments evaluate the generalizability of our approach beyond meetings.

G.1 Fact Representation Comparison

Our fact-based summarization approach depends critically on how facts are represented. To isolate the contribution of our statement-context tuple representation (Section 3.1), we conduct a controlled comparison against an alternative approach using molecular facts (Gunjal and Durrett, 2024), defined as atomic statements with local contextual information derived from the immediate utterance, without the explicit additional global context that our approach incorporates. We implemented both fact representation strategies within the FRAME pipeline using identical GPT-4o backbone models and evaluation protocols, testing on the same set of meetings as used for the main experiments Section 5. Table 13 presents the comparative results.

Our statement-context tuple approach outperforms molecular facts across nearly all evaluation dimensions (MESA, ROUGE, BERTScore). The most pronounced improvements appear in hallucination (3.5-point improvement), omission (3-point improvement), repetition (3-point improvement), and irrelevance (2-point improvement). These findings demonstrate that global context is essential for fact interpretation, enabling the model to resolve references, understand broader implications, and connect information across speaker turns.

The similar scores in structure suggest that both approaches can establish similar organizational frameworks. However, the differences in content quality metrics highlight that global context enables more accurate and coherent content selection and generation. This validates our design choice of incorporating broader meeting context into fact representation, suggesting that meeting summarization benefits from rich contextual grounding.

G.2 Cross-Model Performance Comparison

To determine whether FRAME's effectiveness depends on specific model capabilities or generalizes across different backbone models, we evaluate the pipeline with four different language models as backbones: GPT-4o and Gemini 1.5 pro (commercial-scale models), and Llama 3.1 8b and Gemma 3 4b (smaller open-source models). For each model, we apply identical experimental proto-

Approach	our fact	molecular fact
MESA (lower is better)		
Coreference	$0_{1.64}$	$3_{1.64}$
Hallucination	$1_{1.75}$	$4.5_{0.45}$
Incoherence	$3_{1.88}$	$3.5_{1.22}$
Irrelevance	$1_{1.45}$	$3_{0.79}$
Language	$1_{1.40}$	$1_{1.14}$
Omission	$1_{0.16}$	$4_{0.21}$
Repetition	$1_{1.23}$	$4_{0.45}$
Structure	$3_{1.24}$	$3_{1.52}$
General Evaluation Metrics (higher is better)		
R-1	$22.89_{5.80}$	$20.73_{5.17}$
R-2	$4.13_{2.44}$	$4.31_{2.11}$
R-L	$20.78_{5.23}$	$18.48_{5.08}$
BS (F1)	$85.67_{1.19}$	$84.79_{0.08}$

Table 13: Comparison of FRAME using our fact definition (see Section 3) and molecular facts, i.e., facts with local context. Values are Median_{Std}. MESA scores are 1–5 Likert ratings, ROUGE (R-1/R-2/R-L) and BERTScore (BS) are 0–100. Green is best in category.

cols, comparing FRAME against single-pass summarization with the same backbone model.

Table 15 presents comprehensive results across all model variants, highlighting differences in both absolute performance and relative improvement from the base model to the FRAME pipeline.

Across all model variants, FRAME consistently improves reference-free MESA metrics while showing mixed results on reference-based metrics. The most substantial improvements appear in hallucination (a 2-2.5 points improvement across all models), omission (2 points), and repetition (up to 3 points improvement). These improvements persist regardless of model scale, suggesting that FRAME's structured fact-based approach provides benefits independent of model size or architecture.

The performance gap between commercial and open-source models narrows when using FRAME, particularly for hallucination (commercial: 1-1.5 vs. open-source: 2.5) and irrelevance (both model types: 0.5-1). We conclude that FRAME's explicit fact extraction and verification stages effectively compensate for some limitations of smaller models. Notably, Llama 3.1 8b with FRAME achieves comparable performance to commercial models on dimensions like irrelevance and structure, making it a viable option for production deployment.

The consistent pattern of improved factuality and reduced omission across all model variants confirms that FRAME's benefits stem from its architectural approach rather than from specific model capabilities. This cross-model generalizability demonstrates the robustness of our fact-based approach and its potential applicability across diverse deployment scenarios with varying computational constraints.

G.3 Cross-Domain Evaluation

To assess whether FRAME's benefits extend beyond meeting summarization, we apply our pipeline to three standard document summarization datasets, i.e., arXiv (Cohan et al., 2018), XSum (Narayan et al., 2018), BigPatent (Sharma et al., 2019), and a specialized domain summarization dataset, PubMed (Xiong et al., 2024). These datasets represent varied domains with different linguistic structures: scientific papers (arXiv), news articles (XSum), patent applications (BigPatent), and medical papers (PubMed). This evaluation assesses whether FRAME's fact-centric approach applies to structurally distinct source documents. For each dataset, we compared FRAME with GPT-4o backbone against standard single-pass summarization using MESA, ROUGE, and BERTScore for evaluation.

Table 14 presents the results. FRAME consistently improves scores in specific quality dimensions, most notably repetition, where it achieves 2.5 to 3.5-point improvements across all four domains. A similar strong trend is observed for omission, with 2.5 to 3-point improvements in three of the four datasets. This suggests that FRAME's structured approach to information extraction and organization helps prevent redundancy and, in most cases, content loss, regardless of the source document type.

Domain-specific patterns also emerge. For scientific papers on arXiv, FRAME shows substantial improvements in irrelevance (-1.5) and language quality (-2.5). On the PubMed dataset, the most significant gains are seen in hallucination (-2), language (-2), and repetition (-2.5), reflecting its ability to handle complex, structured content. For news articles (XSum), FRAME demonstrates its most substantial improvements in incoherence (-2), omission (-3), and repetition (-3), suggesting benefits for sources with high factual density. For patents (BigPatent), improvements are most notable in omission (-3) and repetition (-3) but are more modest in other dimensions, as baseline performance was already strong.

Consistent with our main findings in Section 5, reference-based metrics show mixed results. The structural reorganization performed by FRAME leads to deviations from the reference summaries, resulting in slightly lower ROUGE scores on arXiv, BigPatent, and PubMed. This reinforces our conclusion that while ROUGE is a valuable metric, it may not fully capture quality improvements related to better organization and reduced hallucination.

These cross-domain results confirm that FRAME's benefits extend beyond conversational content to diverse document types. The consistent improvements in repetition and frequent, significant gains in reducing omission suggest that fact-based summarization offers robust advantages for information preservation and non-redundancy across domains, while other benefits may be domain-specific. These findings motivate FRAME's applicability as a general-purpose summarization approach beyond its original meeting summarization context.

G.4 Architecture Minimality Analysis

To validate that each stage of FRAME is essential, we conduct an architectural ablation study. We evaluate three combinations:

- `combined-1`: A fully collapsed, single-prompt approach.
- `combined-2`: Merges the Fact Extraction and Relevance Scoring stages.
- `combined-3`: Merges the Relevance Scoring and Outline Planning stages.

The results in Table 16 demonstrate that every stage is critical for maintaining quality, especially regarding content fidelity. While the collapsed pipelines produce summaries with reasonable structure and coherence, they fail on factuality.

Specifically, all three variants score a 4.0 on Omission, a 3-point degradation compared to the complete FRAME, showing they consistently fail to include critical information. The most degradation is in Hallucination, where `combined-2` scores a 5.0, where FRAME is reported with 1.0. This strongly suggests that forcing a model to extract and evaluate facts in a single step prevents it from establishing a stable set of grounded claims, leading to rampant hallucination. These findings confirm that our modular, seven-step architecture is essential for producing reliable and comprehensive summaries.

H Dataset Characteristics and Coverage

Our evaluation employs two complementary datasets representing different aspects of meeting summarization: QMSum (Zhong et al., 2021), an established benchmark containing diverse meeting types, and FAME (Kirstein et al., 2025a), a recently developed synthetic dataset with controlled properties. Together, the datasets provide a comprehensive evaluation environment that spans various meeting domains, structures, and characteristics.

QMSum encompasses 232 meetings across three distinct domains: academic meetings (ICSI), product design discussions (AMI), and government proceedings (WCPC). This diversity enables evaluation of FRAME's effectiveness across varying discourse styles, from structured parliamentary debates to informal product brainstorming sessions. The meetings contain on average 7.2 speakers and 521 turns per meeting, with variation in meeting length (mean: 7,303 words, SD: 4,232).

FAME complements QMSum with 800 synthetic meetings (500 English, 300 German) generated by LLM agents simulating realistic conversational dynamics. It spans 14 meeting formats and 28 distinct topics, including academic discussions, corporate planning sessions, and technical workshops. A key feature of FAME is its controlled inclusion of conversational interruptions (approximately 50% of meetings contain interruptions), enabling assessment of how effectively summarization approaches handle overlapping speech.

Table 17 presents comprehensive statistics across these datasets. Two characteristics are particularly relevant to summarization:

- **Vocabulary diversity:** QMSum exhibits the most extensive vocabulary (20,505 unique tokens), suggesting greater topical diversity and potentially more challenging summarization.
- **Meeting length variation:** Standard deviations in word counts exceed 50% of means across all datasets, indicating that summarization approaches must handle substantial variation in source length.

For our evaluation experiments, we randomly select English meetings from each dataset (QMSum and FAME), stratified to maintain the original distribution of meeting types, ensuring comprehensive domain coverage. This sampling strategy enables assessment across diverse meeting scenarios while maintaining reasonable computational requirements.

	arXiv		XSum		BigPatent		PUBMED	
	GPT 4o	FRAME GPT-4o	GPT 4o	FRAME GPT-4o	GPT 4o	FRAME GPT-4o	GPT 4o	FRAME GPT-4o
MESA (lower is better)								
Coreference	$0_{0.32}$	$0_{1.00}$	$0_{1.45}$	$0_{1.49}$	$0_{0.32}$	$0_{0.67}$	$1_{1.30}$	$0_{1.10}$
Hallucination	$0_{1.49}$	$0_{1.76}$	$0.5_{1.55}$	$0_{0.84}$	$0_{0.95}$	$0_{0.95}$	$3_{1.30}$	$1_{1.00}$
Incoherence	$0_{1.96}$	$0.5_{1.69}$	$3_{1.89}$	$1_{1.48}$	$0_{1.79}$	$0_{1.52}$	$3_{0.55}$	$2_{1.58}$
Irrelevance	$2_{1.26}$	$0.5_{1.51}$	$2_{1.05}$	$0.5_{1.32}$	$0_{1.32}$	$0_{1.25}$	$2_{0.89}$	$1_{1.34}$
Language	$2.5_{1.37}$	$0_{1.48}$	$0_{0.71}$	$0.5_{1.42}$	$0.5_{0.92}$	$0_{1.34}$	$2_{1.30}$	$0_{1.30}$
Omission	$4_{0.42}$	$1.5_{0.33}$	$4_{0.42}$	$1_{0.48}$	$4_{0.11}$	$1_{0.00}$	$4_{0.45}$	$4_{0.00}$
Repetition	$4_{0.84}$	$0.5_{1.01}$	$4_{1.35}$	$1_{1.49}$	$4_{0.85}$	$1_{0.82}$	$3_{1.22}$	$0.5_{0.30}$
Structure	$0_{0.95}$	$0.5_{1.51}$	$0_{0.97}$	$0.5_{1.73}$	$0_{0.43}$	$0_{1.26}$	$2_{1.22}$	$1_{0.89}$
ROUGE (R-1, R-2, R-L) and BERTScore (BS) (higher is better)								
R-1	$28.92_{9.59}$	$23.97_{6.45}$	$14.80_{4.28}$	$15.08_{3.51}$	$31.32_{5.21}$	$23.22_{3.45}$	$27.66_{6.25}$	$25.58_{5.65}$
R-2	$8.71_{5.06}$	$6.19_{2.47}$	$2.32_{1.86}$	$2.95_{2.44}$	$10.19_{6.14}$	$4.35_{2.86}$	$6.10_{3.12}$	$6.34_{2.63}$
R-L	$25.77_{8.50}$	$20.71_{5.62}$	$13.24_{3.84}$	$13.64_{3.99}$	$29.09_{5.44}$	$20.18_{3.38}$	$23.43_{5.73}$	$23.33_{5.17}$
BS (F1)	$83.62_{2.19}$	$83.55_{1.19}$	$84.64_{1.24}$	$85.15_{0.90}$	$86.13_{1.32}$	$84.66_{1.50}$	$85.45_{1.32}$	$85.40_{1.19}$

Table 14: Comparison of information preservation and hallucination rates across fact representation approaches. Values are Median$_{Std}$. MESA scores are 1–5 Likert ratings, ROUGE (R-1/R-2/R-L) and BERTScore (BS) are 0–100. Green is best in category.

	GPT-4o		Gemini 1.5 pro		Llama 3.1 8b		Geamm 3 4b	
	FRAME	Single LLM	FRAME	Single LLM	FRAME	Single LLM	FRAME	Single LLM
MESA (lower is better)								
Coreference	$0_{1.64}$	$0_{1.22}$	$0_{1.59}$	$3_{1.58}$	$0_{1.43}$	$1_{1.21}$	$2_{1.43}$	$4_{1.13}$
Hallucination	$1_{1.75}$	$3_{1.22}$	$1_{1.68}$	$4_{2.04}$	$2.5_{1.64}$	$5_{1.47}$	$2.5_{2.09}$	$4_{1.82}$
Incoherence	$3_{1.88}$	$4_{1.50}$	$3.5_{1.76}$	$4_{1.09}$	$3_{1.64}$	$3_{1.74}$	$1.5_{1.60}$	$1_{1.40}$
Irrelevance	$1_{1.45}$	$2_{1.70}$	$1_{1.35}$	$3_{1.32}$	$0.5_{1.64}$	$3_{1.63}$	$0.5_{1.69}$	$4_{1.63}$
Language	$1_{1.40}$	$1_{1.30}$	$1_{1.37}$	$2_{1.44}$	$2_{1.64}$	$2_{1.83}$	$1_{1.59}$	$1_{1.33}$
Omission	$1_{0.16}$	$3_{0.40}$	$1_{0.41}$	$3_{0.38}$	$2_{1.64}$	$4_{1.79}$	$2_{1.50}$	$4_{1.45}$
Repetition	$1_{1.23}$	$4_{1.05}$	$1_{0.30}$	$3_{0.98}$	$2_{1.64}$	$3_{1.74}$	$1_{1.11}$	$2_{1.46}$
Structure	$3_{1.24}$	$4_{0.90}$	$3_{1.64}$	$3_{1.70}$	$3_{1.64}$	$3_{1.43}$	$3_{1.29}$	$5_{1.33}$
ROUGE (R-1, R-2, R-L) and BERTScore (BS) (higher is better)								
R-1	$22.89_{5.80}$	$37.73_{5.85}$	$21.41_{4.72}$	$39.61_{7.21}$	$27.27_{7.32}$	$20.56_{4.85}$	$15.23_{5.93}$	$20.10_{5.72}$
R-2	$4.13_{2.44}$	$7.95_{4.18}$	$4.79_{2.10}$	$11.10_{4.86}$	$4.37_{2.13}$	$1.36_{1.02}$	$1.82_{1.52}$	$1.53_{0.44}$
R-L	$20.78_{5.23}$	$21.39_{4.05}$	$19.37_{4.39}$	$27.55_{6.36}$	$25.00_{4.36}$	$18.69_{4.32}$	$15.23_{5.84}$	$19.14_{6.02}$
BS (F1)	$85.67_{1.19}$	$81.61_{2.87}$	$84.43_{1.09}$	$80.64_{3.66}$	$86.23_{1.07}$	$85.31_{2.12}$	$85.20_{2.27}$	$84.66_{1.27}$

Table 15: FRAME powered with different backbone models. We report MESA, ROUGE, and BERTScore. Values are Median$_{Std}$. MESA scores are 1–5 Likert ratings, ROUGE (R-1/R-2/R-L) and BERTScore (BS) are 0–100. Green is best in category.

I SCOPE Component Analysis

This section examines the cognitive foundations and empirical effectiveness of SCOPE's structured reasoning approach to personalization. We analyze how explicit the *reason-out-loud* questioning enhances personalized summarization compared to alternative approaches, and investigate which reasoning components contribute most significantly to improved personalization.

I.1 SCOPE Reasoning Questionnaire

SCOPE employs a structured questionnaire derived from cognitive psychology research on metacognitive reasoning and think-aloud protocols (Solomon et al., 1995; Konrad, 2017). These protocols reveal that explicitly verbalizing reasoning processes enhances decision consistency and reduces biases by activating higher-order analytical thinking. Table 18 presents our nine-question protocol to guide the model through a systematic reasoning process before selecting salient facts for personalization.

The questionnaire follows a cognitive progression through four phases:

1. **Planning** (Q1-Q3): Establishes the target reader's knowledge context, current projects, and primary goals. This activation primes subsequent relevance judgments with specific criteria rather than generic role stereotypes.

2. **Initial Assessment** (Q4-Q7): Guides explicit reasoning about relevance for each potential fact, considering urgency, comprehension

	combined-1	combined-2	combined-3	FRAME
	GPT 4o	GPT 4o	GPT 4o	GPT 4o
MESA (lower is better)				
Coreference	$0_{1.25}$	$1.5_{1.62}$	$0_{1.70}$	$0_{1.64}$
Hallucination	$2_{2.08}$	$5_{0.70}$	$4.5_{1.15}$	$1_{1.75}$
Incoherence	$3_{1.96}$	$4_{2.07}$	$3_{1.43}$	$3_{1.88}$
Irrelevance	$2.5_{1.40}$	$3_{1.66}$	$2.5_{1.40}$	$1_{1.45}$
Language	$2_{1.65}$	$2.5_{1.32}$	$1.5_{1.41}$	$1_{1.40}$
Omission	$4_{0.32}$	$4_{0.32}$	$4_{0.32}$	$1_{0.16}$
Repetition	$3_{1.52}$	$2.5_{1.43}$	$2.5_{0.53}$	$1_{1.23}$
Structure	$3_{1.65}$	$3.5_{1.27}$	$3_{1.55}$	$3_{1.24}$
ROUGE (R-1, R-2, R-L) and BERTScore (BS) (higher is better)				
R-1	$22.33_{4.39}$	$18.00_{3.50}$	$19.02_{3.64}$	$22.89_{5.89}$
R-2	$3.85_{2.46}$	$2.82_{1.34}$	$3.36_{2.30}$	$4.13_{2.44}$
R-L	$20.24_{3.82}$	$16.46_{3.17}$	$17.41_{2.80}$	$20.78_{5.23}$
BS (F1)	$85.36_{1.14}$	$84.79_{0.96}$	$84.86_{0.85}$	$85.67_{1.19}$

Table 16: Ablation study on pipeline architecture. We compare our full pipeline (FRAME) against three variants with collapsed stages: **combined-1** (all-in-one), **combined-2** (extraction+scoring), and **combined-3** (scoring+planning). Collapsing stages severely degrades factuality (Hallucination, Omission). Values are Median$_{Std}$. Green is best in category.

needs, and concrete applications. This phase emphasizes justification over intuition.

3. **Controlling** (Q8): Provides an explicit filtering mechanism, requiring reconsideration of initial assessments through a critical lens. This meta-review reduces confirmation bias by encouraging active elimination rather than just selection.

4. **Evaluation** (Q9): Prompts uncertainty awareness, encouraging identification of ambiguous or difficult-to-classify information. This final metacognitive step acknowledges limitations in judgment confidence.

Figure 5 illustrates how a language model engages with this reasoning process when taking the perspective of a graduate student from the ICSI dataset. The model explicitly articulates the student's knowledge background, current projects, and priorities before evaluating specific facts. This explicit articulation creates a more stable and consistent representation of the target reader than implicit role-playing, reducing the likelihood of perspective hallucination or inconsistent fact selection.

I.2 Personalization Approach Comparison

To evaluate the effectiveness of SCOPE's reasoning protocol, we compare it against two alternative personalization approaches commonly used in current research:

- **Reader-Tailoring**: The model receives a description of the target reader and is instructed to customize content for their needs, without explicit reasoning about fact selection (see prompt Figure 19).

- **Role-Playing**: The model is instructed to embody a specific persona and directly generate a summary from that perspective, without explicit reasoning steps (see prompt Figure 20).

We implemented each approach using a single LLM, isolating the effect of fact extraction. Table 19 presents the evaluation results across both personalization-specific (P-MESA) and general quality (MESA) dimensions.

SCOPE demonstrates consistent advantages across most P-MESAdimensions. The largest improvements appear in completeness (2 points better than Reader-Tailoring, 1 point better than Role-Playing) and knowledge-level fit (0.5 points better than both alternatives). While Reader-Tailoring achieves comparable factuality scores (both at 2), SCOPE outperforms Role-Playing by 2 points in this dimension. These results indicate that SCOPE's explicit reasoning enhances the model's ability to include relevant information and adapt its presentation to the reader's knowledge.

For general quality dimensions measured by MESA, SCOPE shows notable improvements in hallucination reduction (1-2 points better) and omission (1.5 points better) compared to alternative approaches. This suggests that explicit reasoning not only improves personalization but also enhances factual accuracy and information coverage. Role-playing demonstrates stronger performance in structural coherence but shows weaknesses in factuality and hallucination. Reader-Tailoring achieves reasonable factuality but struggles with completeness and contextual framing. These patterns suggest that while alternative approaches may excel in specific dimensions, SCOPE's structured reasoning provides the most balanced and comprehensive approach to personalization, addressing both content selection and presentation aspects.

J Literature Review and Criteria Selection

This section details our systematic development of the seven P-MESA dimensions, expanding on the

Dataset	# Meetings	# Speaker	# Unique Spea.	# Turns	# Words	Vocab.	Token Overlap	Sum. Len.	Interruptions	Language
AMI	137	$4.0_{0.00}$	4	$513.5_{266.2}$	$4937.5_{1999.3}$	9388	-	$109.9_{27.1}$	no	informal
ICSI	44	$6.2_{1.3}$	35	$757.5_{374.8}$	$9889.4_{3794.9}$	9164	-	$93.3_{22.2}$	no	formal
WCPC	51	$16.8_{18.7}$	316	$337.3_{277.3}$	$11427.8_{4574.0}$	13780	-	$122.3_{39.2}$	no	informal
QMSum	232	$7.2_{10.1}$	330	$521.0_{320.4}$	$7303.4_{4232.2}$	20505	-	$109.5_{30.7}$	no	both
EN	500	$5.1_{2.8}$	3200	$405.0_{330.3}$	$6223.4_{4084.4}$	10347	0.081	$207.7_{22.7}$	yes ($\sim$ 0.5)	both
GER	300	$5.0_{2.8}$	1000	$393.3_{323.2}$	$6272.4_{3793.2}$	9589	0.096	$170.3_{29.0}$	yes ($\sim$ 0.5)	both

Table 17: Statistics on FAME for English and QMSum corpora. Values are Mean$_{Std}$. Table stems from (Kirstein et al., 2025a).

	Definition	Category
Q1	What prior knowledge do you have?	Planning
Q2	Which project are you currently working on?	Planning
Q3	What are your primary interests and goals?	Planning
Q4	Read each fact carefully and think about which information is most relevant to you in your role. Explain why.	Initial Assessment
Q5	Is there an urgency or priority that aligns particularly closely with your current responsibilities or known concerns?	Initial Assessment
Q6	Which information might require simplification or additional context to ensure clear comprehension?	Initial Assessment
Q7	You've selected information that you consider important. Review this selection once more and provide concrete examples explaining why these details are relevant for you.	Initial Assessment
Q8	Now, go through the list a second time and identify which information you consider irrelevant or unimportant for your role/persona, providing reasons for your decisions.	Controlling
Q9	Are there any topics you found difficult to classify or about which you felt unsure? If so, what are they?	Evaluation

Table 18: SCOPE questionnaire used to familiarize the LLM with the input before selecting salient facts for a target reader.

Approach	**Reader-Tailoring**	**Roleplaying**	**SCOPE**
P-MESA (lower is better)			
goal alignment	$3_{0.80}$	$3_{0.00}$	$2.5_{0.41}$
completeness	$5_{0.88}$	$4_{0.58}$	$3_{0.10}$
factuality	$2_{1.53}$	$4_{0.58}$	$2_{0.37}$
knowledge level fit	$2_{0.40}$	$2_{0.53}$	$1.5_{1.73}$
priority structuring	$4_{0.76}$	$4_{0.60}$	$4_{0.58}$
contextual framing	$5_{1.23}$	$4_{1.08}$	$3.5_{0.71}$
relevance	$3_{0.68}$	$3_{1.16}$	$2.5_{0.70}$
MESA (lower is better)			
Coreference	$0_{1.74}$	$0_{1.73}$	$0_{0.00}$
Hallucination	$3_{2.52}$	$4_{1.15}$	$2_{2.52}$
Incoherence	$4_{1.74}$	$4_{0.58}$	$4_{2.31}$
Irrelevance	$2_{0.00}$	$2_{0.14}$	$2_{1.52}$
Language	$4_{1.51}$	$4_{0.56}$	$4_{2.08}$
Omission	$4_{1.25}$	$4_{0.27}$	$2.5_{0.00}$
Repetition	$3_{1.00}$	$4_{0.58}$	$3_{0.58}$
Structure	$4_{0.78}$	$3_{1.53}$	$4_{1.72}$

Table 19: Impact of different personalization approaches on summary quality of a single GPT-4o instance. Values are Median$_{Std}$. MESA and P-MESA scores are 1–5 Likert ratings. Green is best in category.

three-step process outlined in Section 4.1. We provide comprehensive documentation of our literature review methodology, human evaluation protocol, and dimension refinement process.

J.1 Step 1: Literature Review

Corpus Construction We collected 50 papers published between 2018 and 2024 from major computational linguistics venues, including ACL, EMNLP, NAACL, COLING, EACL, and TACL. Our primary source was Semantic Scholar, with results ranked via the Allen AI Paper Finder (All, 2025). We use the terms "personalization," "adaptation," and "user modeling" as anchor terms, combined with targeted keywords such as "generation," "summarization," "data-to-text," "evaluation," and "language model."

Relevance Assessment Each paper undergoes an initial screening based on title and abstract. We classify 25 papers as *highly relevant* (core focus on personalization in generation), 16 as *relevant* (discuss personalization but not methodologically central), and discard 9 papers that focused on adjacent topics like personalized agents or style transfer without clear evaluative criteria. After subsequent full-text screening, we removed an additional 5 papers that focused on end-to-end personalization or task-specific fine-tuning without analyzing personalization mechanisms, resulting in 36 papers for final analysis.

Candidate Dimension Extraction From the 36 selected works, we identify recurring characteristics of effective content personalization. We clustered these into nine candidate dimensions covering different aspects of personalization quality:

1. *Factual Accuracy*: Correspondence between summary content and source text
2. *Content Completeness*: Inclusion of information required by the target reader

3. *Information Relevance*: Focus on content pertinent to user's role and needs

4. *Objective Alignment*: Addressing high-level user goals and intentions

5. *Content Prioritization*: Ordering information by importance to the user

6. *Knowledge Appropriateness*: Matching the user's expertise level

7. *Contextual Framing*: Providing necessary background information

8. *Personal Preferences*: Matching stylistic and format preferences

9. *Information Utility*: Providing actionable content for the user's context

These dimensions represented the standard evaluation criteria used across personalization research, though with varying terminology across different domains.

J.2 Step 2: Human Study

To test whether these nine criteria are distinguishable, applicable across meeting domains, and comprehensive in capturing personalization failures, we conduct a refinement study with human annotators evaluating model-generated personalized summaries.

Evaluation Dataset Construction We create a one-time evaluation dataset of 48 personalized summaries using GPT-4o and Gemini 1.5 Pro, divided equally between two personalization approaches, i.e., 24 summaries where the model was instructed to *summarize for* a specific reader and 24 summaries where the model was instructed to *simulate* being the target reader.

Each summary was generated for a distinct reader profile specifying role, prior knowledge level, and goals. The samples were drawn equally from QMSum and FAME datasets, covering eight different meeting types, 14 diverse topics, and transcripts with an average of 5.9 speakers per meeting. Table 20 shows the distribution of the dataset.

Annotation Protocol Three annotators (ages 22–29, C1+ English proficiency; see Appendix C for details) evaluate each summary using the nine candidate dimensions on a 1-5 Likert scale. Each summary receives annotations from two independent evaluators to enable assessment of inter-annotator reliability. Annotators provided structured feedback on:

- Definition clarity for each dimension
- Potential overlap between dimensions
- Missing aspects of personalization not covered by the dimensions
- Examples of successful and unsuccessful personalization for each dimension

We collect this feedback through standardized evaluation forms and conduct daily group discussions throughout the annotation period to align understanding and address emerging questions.

Dataset Characteristic	Count
Total Summaries	48
Generation Approach	
Summarize For + GPT	12 (6 QMSum, 6 FAME)
Summarize For + Gemini	12 (6 QMSum, 6 FAME)
Simulate + GPT	12 (6 QMSum, 6 FAME)
Simulate + Gemini	12 (6 QMSum, 6 FAME)

Table 20: Characteristics of the evaluation dataset used for P-MESA dimension refinement.

J.3 Step 3: Final Dimension Set

Based on annotator feedback and empirical assessment, we make two refinements to the candidate dimensions:

1. **Removal of Personal Preferences**: This dimension shows poor cross-context generalizability and inconsistent applicability across different meeting domains due to missing information in the datasets.

2. **Consolidation of Intent-Related Dimensions**: The original *Objective Alignment* and *Information Utility* dimensions are merged into a single *Goal Alignment* dimension. This consolidated dimension focuses on macro-level intent alignment, specifically whether the summary effectively addresses the reader's overarching goals. In contrast, the separate *Priority Structuring* dimension addresses micro-level content organization, including the salience and ordering of specific information.

No additional dimensions are identified as missing during the feedback process. This refinement yields the final set of seven personalization dimensions used in P-MESA: Factuality, Completeness, Relevance, Goal Alignment, Priority Structuring, Knowledge-Level Fit, and Contextual Framing. Complete definitions and indicators for these dimensions are provided in Table 21.

Table 1 in Section 4 shows the correlation between P-MESA's automated scores and human judgments across these seven dimensions, demonstrating strong alignment (Spearman's ρ ranging from 0.69 to 0.81, with average ρ = 0.77). A complete archive of our literature review coding scheme, annotation guidelines, and the final P-MESA implementation will be made available via GitHub (see Appendix A.1).

K Balanced Accuracy Definition

Accuracy (ACC) is a natural choice to measure the proportion of correctly predicted labels out of the total number of labels:

$$ACC = \frac{(TP + TN)}{(TP + FN + FP + TN)} \quad (1)$$

with TP - true positive, TN - true negative, FP - false positive, and FN - false negative. As we cannot exclude a class imbalance within the individual dimensions, reporting accuracy is not suitable, and we report the balanced accuracy (B-ACC), i.e., the arithmetic mean of sensitivity (SEN) and specificity (SPE):

$$SEN = \frac{TP}{(TP + FN} \quad (2)$$

$$SPE = \frac{TN}{(TN + FP)} \quad (3)$$

$$\text{B-ACC} = \frac{1}{2}(SEN + SPE) \quad (4)$$

L Comparison with Self-Refinement

A recent approach for refining meeting summaries is the assessment of feedback similar to the MESA categories and refining a summary according to this feedback (Kirstein et al., 2025d). We exclude this approach from our main experiments as it reiterates the same model and is therefore not a fair comparison to a single-pass approach, and because it is not easily adaptable for personalization due to the lack of feedback criteria, which we present with P-MESA (Section 4). In Table 22, we compare FRAME with a GPT backbone and the setup of Section 5 against a feedback-based summarization refinement approach (Kirstein et al., 2025d) with one, two, and three iterations (abbreviated FB-1, -2, -3). We use FEEDBACK with the reported best-performing setup, i.e., having multiple agents assess feedback, prompted to use chain-of-thought reasoning, for error assessment, and using the feedback directly along the chain-of-thought trace for refinement. We observe that FRAME beats FEEDBACK-1 and FEEDBACK-2 and slightly outperforms FEEDBACK-3. FRAME is consistently lower in hallucination (1 point vs. min. 3 points in FB-3). Coreference, incoherence, and structure are similar consistently. FRAME slightly outperforms on omission and irrelevance. When it comes to repetition and language, FRAME beats FB-1 and FB-2, but is outperformed by FB-3. So we conclude that FRAME is comparable to three turns of feedback-based evaluation due to the fact-based approach that aids content understanding, it performs better with hallucination, omission, and repetition.

M Qualitative Examples

This section presents representative examples that illustrate the improvements achieved by our approaches. These examples provide concrete demonstrations of how FRAME and SCOPE address the challenges stated in Section 1 and support the quantitative findings reported in Section 5. All examples are drawn from the QMSum dataset and are generated using the experimental setup described in Section 5.

M.1 Impact of Summary Verification

Table 23 demonstrates the effect of FRAME's `Summary Verification` stage, which ensures conciseness and structural coherence while preserving core content. This example illustrates how the validation step primarily serves as a quality control mechanism that:

- Condenses verbose passages into more concise expressions
- Maintains all key decisions and action items
- Preserves attribution of statements to specific speakers
- Ensures adherence to length constraints (reduced from 313 to 192 words)

Criteria	Definition	Indicator
Factuality	Measure whether the summary is factually accurate with respect to the source transcript. High factual groundedness indicates that no unsubstantiated or contradictory information appears in the summary.	Does every claim align with the original transcript? Are there any claims that contradict known facts or the transcript itself?
Completeness	Evaluates how thoroughly the summary includes all critical information required by the target persona. A low score indicates that no essential data needed for decision-making or task execution is missing.	Are all key infos needed for the persona's role present, including budget, deadlines, constraints? Is there evidence of a salient fact or figure that was in the source but left out?
Relevance	Measures how focused the summary is on content pertinent to the user's (or persona's) role and needs. A low score means minimal extraneous or off-topic information.	Is every piece of information purposeful for the user's role? Does the summary emphasize tasks or decisions within the user's domain?
Goal Alignment	Assesses whether the summary content directly addresses the persona's primary objectives or responsibilities. Low-scoring summaries are tightly coupled to the persona's overarching goals.	Does the summary thoroughly address the user's stated or implicit goals? Does any part of the summary contradict or ignore the user's known objectives?
Priority Structuring	Looks at the order and emphasis of information to see if the summary highlights the most urgent or important points first, reflecting the persona's immediate needs or preferences.	Are the most pressing items placed first or highlighted? Does the layout or emphasis help the persona quickly find urgent/actionable items?
Knowledge-Level Fit	Check if the level of technical or conceptual detail matches the persona's expertise. Low scores indicate an optimal level—neither too simplistic nor too advanced.	Is the chosen vocabulary suitable for the user's domain expertise? Does the summary offer enough context without overwhelming or patronizing the user?
Contextual Framing	Assesses whether the summary includes the necessary context (historical decisions, cross-departmental references, relevant background) so the user can fully understand the situation.	Are past decisions or relevant external factors briefly explained? Does the summary clarify dependencies or references to others' work?

Table 21: P-MESA evaluation criteria and indicators for easier identification.

Approach	FRAME	FB-1	FB-2	FB-3
MESA (lower is better)				
Coreference	$0_{1.64}$	$1_{1.54}$	$0_{1.30}$	$0.5_{0.86}$
Hallucination	$1_{1.75}$	$4_{1.93}$	$4_{1.84}$	$3_{1.93}$
Incoherence	$3_{1.88}$	$2.5_{1.73}$	$4_{1.83}$	$3_{1.73}$
Irrelevance	$1_{1.45}$	$3_{1.60}$	$4_{1.58}$	$2_{1.60}$
Language	$1_{1.40}$	$3_{0.88}$	$2_{0.78}$	$0_{0.72}$
Omission	$1_{0.16}$	$4_{0.00}$	$3.5_{0.32}$	$2_{0.27}$
Repetition	$1_{1.23}$	$2_{0.97}$	$2_{1.21}$	v$0_{0.84}$
Structure	$3_{1.24}$	$3_{1.76}$	$3_{1.58}$	$4_{1.44}$
ROUGE (R-1, R-2, R-L) and BERTScore (BS) (higher is better)				
R-1	$22.89_{5.80}$	$24.19_{4.74}$	$23.86_{5.34}$	$21.20_{4.44}$
R-2	$4.13_{2.44}$	$5.33_{2.86}$	$5.82_{2.88}$	$5.44_{2.83}$
R-L	$20.78_{5.23}$	$20.99_{4.83}$	$20.45_{4.48}$	$20.05_{4.51}$
BS (F1)	$85.67_{1.19}$	$84.85_{0.11}$	$86.50_{1.07}$	$85.90_{0.73}$

Table 22: Comparison of FRAME summary quality to refinement-based approaches (Kirstein et al., 2025d). Values are Median$_{Std}$. MESA scores are 1–5 Likert ratings, ROUGE (R-1/R-2/R-L) and BERTScore (BS) are 0–100. Green is best in category.

It supports our finding in Appendix F that the verification component primarily enforces structural constraints rather than correcting factual content when using a high-capability model like GPT-4o.

M.2 FRAME vs. Single-LLM Summarization

Table 24 contrasts summaries from a single GPT-4o instance versus our FRAME framework. This comparison highlights several key advantages of the fact-based approach:

- **Improved Structure**: FRAME's summary organizes information thematically rather than chronologically, grouping related concepts (data handling, anonymization, algorithms).
- **Higher Specificity**: FRAME captures concrete decisions (e.g., "Professor D suggested digit recordings") rather than vague descriptions (e.g., "participants debated merits").
- **Better Speaker Attribution**: FRAME consistently attributes statements to specific speakers (Professor D, PhD F, Grad B), preserving accountability and provenance.
- **Clearer Action Items**: FRAME explicitly identifies next steps and responsibilities, whereas the single-LLM approach emphasizes discussion over outcomes.

These qualitative differences illustrate why FRAME achieves lower hallucination and omission scores as reported in Table 2, demonstrating how our structured fact-based approach addresses the core challenges of meeting summarization identified in Section 1.

M.3 Personalized Summarization

Table 25 compares personalized summaries generated for a graduate student (Grad A) using single-LLM and FRAME approaches. This example illustrates how FRAME's fact-based approach enhances personalization even without explicit SCOPE reasoning:

- **Reader-Relevant Focus**: The FRAME summary prioritizes content most relevant to Grad A's role (model structure, decision nodes, implementation considerations).

- **Technical Detail Calibration**: The FRAME summary provides appropriate technical depth for a graduate student working on the project.
- **Implied Consequences**: The FRAME summary explains why particular design decisions matter (e.g., "This impacts the logical independence and timing of decisions").
- **Perspective Preservation**: While both summaries focus on Professor B's guidance, the FRAME version better contextualizes this guidance from Grad A's perspective.

These improvements align with our findings in Section 5.3 that fact-centric approaches provide a stronger foundation for personalization by enabling more deliberate content selection and organization.

M.4 Impact of Fact Representation

Table 26 compares summaries generated using the same FRAME architecture but with different fact representation approaches. This example demonstrates why our statement-context tuple approach outperforms molecular facts (Appendix G.1):

- **Enhanced Contextual Understanding**: Our approach captures relationships between meeting elements (e.g., connecting software choices to model structures).
- **Improved Responsibility Attribution**: Statement-context tuples enable clearer assignment of tasks to specific participants.
- **Coherent Organization**: Global context facilitates better organization around themes rather than isolated facts.
- **Strategic Prioritization**: Richer context enables more nuanced decisions about which content to emphasize.

This example supports our finding that global context enrichment significantly improves the model's ability to understand and reconstruct the meeting's underlying semantic structure.

M.5 Threshold Sensitivity Analysis

Table 27 illustrates how varying the fact retention thresholds affects summary content and structure. This example supports our threshold sensitivity analysis in Appendix F.3:

- **Default Threshold**: Creates a balanced summary with focused main points and appropriate supporting details.
- **Lower Threshold**: Includes more contextual information but introduces some repetition (e.g., multiple mentions of design simplicity).
- **Higher Threshold**: Over-prioritizes specific details (e.g., material properties) at the expense of broader context and coherence.

These examples demonstrate why our empirically determined thresholds ($r_i \geq 8$ for outline points, $r_i \geq 6$ for context) achieve optimal results in our main experiments.

M.6 Personalization Approach Comparison

Table 28 compares summaries generated using three different personalization approaches: Reader-Tailoring ("Tailor To"), Role-Playing, and our SCOPE protocol. All summaries are personalized for the same graduate student ("Grad A") and cover identical meeting content about belief networks and decision processes. The comparison reveals distinctive patterns in how each approach handles personalization:

- **Reader-Tailoring**: Mentions Grad A once but primarily presents comprehensive meeting content without consistently filtering for relevance to Grad A's interests. Includes many technical details without clear prioritization based on Grad A's specific needs.
- **Role-Playing**: Provides extensive meeting details with focus on Professor B's contributions, but never explicitly addresses Grad A's interests or needs. Includes tangential information about other participants that may not be directly relevant to Grad A.
- **SCOPE**: Consistently relates information back to Grad A's interests, explicitly mentioning how the content aligns with Grad A's interests. More concise, selecting only aspects most relevant to Grad A's focus on "logical structures and reliable information."

These examples illustrate SCOPE's key advantage: explicit reasoning about relevance produces summaries that consistently prioritize content based on the reader's specific interests rather than just providing topic-relevant information. This targeted selection supports our quantitative findings in Section 5.3, demonstrating how SCOPE's structured reasoning approach improves personalization quality.

Setup	Summary
Before	The meeting focused on procedural changes, data management, and anonymization strategies. Key outcomes included decisions on meeting formats, data handling, and anonymization methods. Professor D suggested digit recordings at the end of meetings to streamline the process and ensure accurate documentation. PhD F recommended having three or four different kinds of meetings to cater to various research needs and improve data collection. Professor D proposed engaging professors and senior doctoral students to conduct their meetings at the research site for better data collection. Participants agreed that CDs should not be given on the same day to ensure thorough review and accuracy. PhD C plans to use clustering or neural net algorithms for classification and exclude non-speech events from the study to focus on relevant data. Professor D suggested conducting sensitivity analyses to better understand neural network inputs and using a decision tree for classification. The importance of detecting overlaps and having multiple system models was highlighted to improve data accuracy. Various methods for anonymizing data, including sequential labeling and playing audio backward, were discussed to ensure participant privacy. Grad B will send an email to Haas Business School volunteers to engage participants. Postdoc E suggested burning the CD after it has been cleared with the transcript stage to ensure data accuracy. PhD A needs to delete the old dynamic data from Carmen's Spanish directory to optimize data storage. Grad B plans to complete the archive and verify it by the end of the day to ensure data integrity. PhD C plans to work on an automatic segmenter to improve data processing and label each frame with a class indicating the type of zone. PhD C also plans to extend the study to include all overlapping events in the future to broaden the research scope.
After	The meeting addressed procedural changes, data management, and anonymization strategies. Key decisions included meeting formats, data handling, and anonymization methods. Professor D suggested digit recordings for accurate documentation, while PhD F recommended varied meeting types to enhance data collection. Participants agreed CDs should not be given on the same day for thorough review. PhD C plans to use clustering or neural net algorithms for classification, excluding nonspeech events. Sensitivity analyses and decision trees were proposed for better understanding and classification. Methods for anonymizing data, such as sequential labeling and playing audio backwards, were discussed. Grad B will email Haas Business School volunteers, and Postdoc E suggested burning CDs post-transcript clearance. PhD A will delete old data to optimize storage, and Grad B will complete and verify the archive. PhD C plans to develop an automatic segmenter and extend the study to include overlapping events.

Table 23: FRAME generated summary before and after applying the `Summary Verification` step.

Setup	Summary
Single LLM	The meeting focused on discussing the format for storing and linking various types of data, such as word transcripts, annotations, and prosodic features, in a database. The participants debated the merits of different formats, including XML and P-files, and considered the use of ATLAS, a NIST-developed infrastructure. They emphasized the need for a flexible, easy-to-use format that can handle changes in word transcripts and time boundaries efficiently. The discussion also touched on the challenges of merging different annotations and maintaining consistency across versions. The meeting concluded with a plan to further investigate ATLAS and its tools to determine the best approach for their needs.
FRAME	The meeting addressed procedural changes, data management, and anonymization strategies. Key decisions included meeting formats, data handling, and anonymization methods. Professor D suggested digit recordings for accurate documentation, while PhD F recommended varied meeting types to enhance data collection. Participants agreed CDs should not be given on the same day for thorough review. PhD C plans to use clustering or neural net algorithms for classification, excluding nonspeech events. Sensitivity analyses and decision trees were proposed for better understanding and classification. Methods for anonymizing data, such as sequential labeling and playing audio backwards, were discussed. Grad B will email Haas Business School volunteers, and Postdoc E suggested burning CDs post-transcript clearance. PhD A will delete old data to optimize storage, and Grad B will complete and verify the archive. PhD C plans to develop an automatic segmenter and extend the study to include overlapping events.

Table 24: Sample QMSum meeting transcript summarizes by a single GPT-4o instance and our FRAME framework with GPT-4o backend.

Setup	Summary
Single LLM	Professor B suggested focusing on building influence links within the path domain, prioritizing those related to paths over database influences. He proposed laying out the influence structure and determining necessary belief nodes, with the design completed before computation methods are decided. Professor B confirmed the program being used is a GUI for a belief-net simulator. He advised focusing on strong influences when considering factors like "slurred" or "angry" speech. He suggested merging selling, fixing, and servicing entities for endpoint decisions, potentially creating an intermediate "commercial" entity type alongside "landmark." Professor B recommended removing the direct link between the user and endpoint, instead creating an intermediate "endpoint U" node representing the endpoint decision from the user model perspective. He also suggested a "fit" node to measure compatibility between the entity and user. Professor B advised against excessive complexity in the user model to avoid combinatorial explosion, suggesting a compact representation of user state or compatibility. He proposed shifting focus to other path-related decisions and their structure within the belief net. Professor B emphasized the importance of avoiding loops in the belief net to prevent ill-defined computations. He highlighted the need for organizational clarity, explaining that discourse properties will eventually feed into discourse-based decision nodes. Professor B raised the issue of combining influences mathematically at decision nodes, asking how to determine the rule of combination. He also requested a display solution for the complex diagrams. Finally, Professor B initiated a discussion about decision nodes and combination choices, prompting an explanation of expert weighting and mixture of experts models.
FRAME	During the meeting, Professor B proposed merging selling, fixing, and servicing into an intermediate structure for endpoint decisions, emphasizing the creation of a new node based on user models. Professor B confirmed that the endpoint decision node should be linked to the user model, ensuring accurate decision-making, and agreed to use 'endpoint U' for this node. Additionally, Professor B suggested that the 'Go-there' decision should have a node for discourse and binary values. Professor B planned to outline the influence structure, advising a focus on strong influences in the belief-net model and completing the basic design before computing it. Grad D suggested an intermediate node for the Endpoint and Go-there nodes, which Professor B agreed could feed into the Go-there node, incorporating situational factors. This impacts the logical independence and timing of decisions. Professor B emphasized avoiding loops in the belief net to prevent convergence issues, highlighting the need for careful knowledge engineering to manage system complexity. These insights are vital for Grad A's understanding of decision-making nuances and ensuring model reliability.

Table 25: Example personalized summary of the same meeting for Grad Student A. We compare FRAME with SCOPE against GPT-4o prompted to tailor to a specific reader.

Setup	Summary
Molecular	The meeting focused on the development and implementation of the JavaBayes program and the discourse model. Key decisions included switching to JavaBayes for its probabilistic modeling capabilities, adding a middle layer to enhance model complexity and accuracy, and standardizing presentation guidelines with 'Future Work' and 'What I Learned' slides for clarity. Grad A proposed reverting to a previous stable state to ensure model reliability, while Grad B set limits for searching past discussions to improve efficiency. A trial run for data collection was scheduled for Friday afternoon, with Fey volunteering as the wizard and Liz as the first subject. Grad A faces challenges in setting up probabilities for a flat structure due to exponential possibilities with N features. The discourse model aids in anaphora resolution and coordinating gesture screen issues, providing a clear interface for querying past discussions.
Our facts	The meeting focused on developing and implementing a discourse model and data collection strategies. Key outcomes included decisions on model structure, data collection plans, and software tools. Grad A was confirmed to streamline project management, providing clear leadership and accountability. The group agreed to add a middle layer to the model to enhance differentiation between user contexts, improving accuracy and functionality. Grad A suggested reverting to a previous stable version for better reliability. Grad B confirmed their presentation is ready. The discourse model will have a limit for searching past discussions to optimize performance. Features and variables such as prosody, discourse, verb choice, and context categories were discussed. The switch to JavaBayes was noted, and suggestions were made for generating probability tables and friendly data entry scripts. Data collection strategies included trial runs scheduled for Friday afternoon. Next steps involved Grad B switching to JavaBayes, identifying hidden nodes, and segmenting regions to create observable nodes. Responsibilities were assigned to Grad A, Grad B, Fey, and Liz.

Table 26: Example summaries of FRAME with molecular facts and our fact definition.

Setup	Summary
default	The meeting focused on the design and functionality of a new remote control, emphasizing simplicity and ease of use. Key decisions included essential functions like volume, channel selection, on-off, mute, and text TV buttons. Marketing stressed the importance of an attractive look to appeal to users, while technological innovation was secondary. The team opted against LCD and speech recognition to simplify design and reduce costs, setting a budget of twelve and a half Euros per unit. Rechargeable batteries and a scroll wheel for volume control were chosen for sustainability and functionality. The remote will have a rounded shape, transparent case, and modern look, targeting a young, trendy audience. Next steps include prototype development, engineering contacts, design review, and market testing.
lower	The meeting centered on designing a new remote control, emphasizing simplicity and user-friendliness. Key decisions included a basic design with minimal buttons to enhance usability, while retaining essential features like text TV functionality for compatibility with existing systems. The remote will be encased in durable hard plastic, balancing cost-effectiveness and reliability. To appeal to a young audience, the design will incorporate flashy colors and a modern aesthetic, featuring a Real Reaction logo for brand visibility. Rechargeable batteries will be used to promote sustainability, and a scroll wheel will provide intuitive volume control. The remote will slide into a docking station for convenient charging. The team opted against LCD and speech recognition due to complexity and cost, focusing instead on affordability with a budget of twelve and a half Euros per unit. The design aims to be trendy yet straightforward, ensuring ease of use and competitive appeal in the market.
higher	The meeting centered on designing a new remote control, focusing on its appearance, functionality, and target audience. Key decisions included opting for a basic design with minimal buttons to simplify the user interface and enhance usability. The remote will feature text TV functionality, ensuring compatibility with existing technology, and a spongy material for improved durability and comfort. The team decided against using LCD or speech recognition to reduce complexity and cost, setting the budget at twelve and a half Euros per unit. The remote will have essential buttons like on-off, mute, and possibly two others, with all buttons on one level for better usability. It will use rechargeable batteries and a regular chip for reliability, with a transparent case to allow backlight visibility. Normal infrared LEDs and backlight LEDs will be included, along with a scroll wheel for volume control and zapping functionality to enhance user satisfaction. The remote will feature the Real Reaction logo and will be modern yet basic, aiming to attract the target audience.

Table 27: Example summary with varying thresholds using FRAME.

Setup	Summary
Tailor To	In the meeting, Professor B proposed focusing on building influence links to determine which sets of factors are relevant to specific decisions, starting with path-related ones. Grad A emphasized the importance of logical independence and timing in decision-making, suggesting that factors like "Go-there" should be binary and separate from other possibilities. Grad E discussed the complexity of defining discrete values for variables like running budgets and user interests, highlighting the challenge of capturing nuanced decision-making criteria. Grad D introduced the concept of weighted averages and mixtures of experts to combine multiple inputs into a single decision node, stressing the need for a principled approach to ensure valid probability distributions. The discussion also touched on the importance of integrating discourse features and ontology properties into the belief-net structure, with Grad A questioning how human-like decision processes could be modeled. The meeting concluded with a focus on organizing and combining influences mathematically, acknowledging the need for a display hack to manage the complexity of the belief-net diagrams. Overall, the meeting underscored the importance of logical structure, reliability, and detailed analysis in decision-making processes.
Roleplaying	During the meeting, Professor B emphasized the importance of building influence links to understand which sets of factors are relevant to specific decisions. He suggested starting with path-related decisions, considering database-related ones less interesting but necessary. Professor B and Bhaskara plan to identify belief nodes required for decision-making, ensuring the basic design of influence structures is completed before computing them. Grad E mentioned the user's budget influencing decision outcomes, highlighting the need for a running total of finances. Professor B clarified that updates to the budget must be external to the belief-net and influence various decisions. Grad E expressed confusion about writing discrete values for a running budget, which Professor B acknowledged, suggesting external conditioning for updates. The discussion shifted to decision nodes and their combination rules. Grad D explained the need for a principled way to combine expert decisions, proposing weighted averages and mixture models to handle varying reliability of experts in different situations. Professor B stressed the importance of avoiding loops in belief-nets to ensure well-defined computations. Grad E and Professor B discussed the necessity of integrating discourse properties into decision nodes, with Keith focusing on constructions indicating various properties and Johno committed to parsing wizardry. The meeting concluded with a focus on organizing influences and combining them mathematically, ensuring the system remains feasible and clean for decision-making processes.
SCOPE	Professor B emphasized the importance of laying out influence structures to understand what factors affect decisions, suggesting a focus on path influence links initially. This aligns with Grad A's interest in logical structures and decision-making processes. The discussion highlighted the need to establish belief nodes, which are crucial for computing decisions based on influence structures. Grad E and Professor B explored how user models and situation models contribute to decision-making, providing insights into logical independence and reliability of features. The conversation also touched on the technical aspects of belief nets, which are essential for Grad A's analytical approach. The urgency to complete the basic design before computation was noted, aligning with Grad A's priorities. The meeting provided valuable insights into how different factors influence decision-making, offering a detailed analysis that supports Grad A's focus on logical structures and reliable information.

Table 28: Different personalization approaches. We compare 'tailor-to' prompting against 'role-playing' prompting and our reason-out-loud protocol. All summaries cover the same meeting and are personalized to “Grad A”.

Target Summary Generation

{
"fact": "Highly mismatched means clean training with close microphone training and distant microphone testing.",
"context": "PhD C explains that highly mismatched conditions involve clean training with close microphone training and distant microphone testing, which are the most noisy cases."
},
{
"fact": "Without the neural network, the performance is better.",
"context": "PhD C mentions that performance is better without the neural network, indicating that adding neural networks might be causing issues."
},
{
"fact": "The TIMIT noises include ’Car’, ’Babble’, ’Subway’, and ’Train station’.",
"context": "PhD C lists the types of noises in the TIMIT dataset, which include ’Car’, ’Babble’, ’Subway’, and ’Train station’."
},
{
"fact": "The neural net was not helping until the second path with pure features was added.",
"context": "Professor B and PhD C discuss that the neural net was not helping until the second path with pure features was added, indicating the importance of combining neural net features with pure features."
},
{
"fact": "The ANN performs nonlinear discriminant analysis.",
"context": "PhD C mentions that the ANN performs nonlinear discriminant analysis, which is similar to LDA but not linear."
},
{
"fact": "Non-tandem features were transformed using PCA in the proposal.",
"context": "PhD C states that non-tandem features were transformed using PCA in the proposal, suggesting an alternative method for feature processing."
},

Figure 4: Snippet of facts extracted from a QMSum meeting.

SCOPE questionnaire answering

(1) What prior knowledge do you have?
I have a solid understanding of probabilistic models, feature extraction, and the importance of cues in determining modes of interaction. I am familiar with the concepts of binary variables and the categorization of time-related variables.

(2) Which project are you currently working on?
I am currently working on a project that involves analyzing and modeling human behavior based on various cues and variables. This includes understanding how different factors influence decisions and actions in specific contexts.

(3) What are your primary interests and goals?
My primary interests include improving the accuracy of behavior prediction models, enhancing feature extraction techniques, and understanding the impact of different cues on decision-making processes. My goals are to refine the models to better predict outcomes and ensure that the extracted features are relevant and useful.

(4) Read each fact carefully and think about which information is most relevant to you in your role. Explain why.
- "*The final destination, business, hurry, and tourist variables depend probabilistically on other factors.*" This is relevant because understanding the probabilistic dependencies helps in refining the model.
- "*The mode depends on all the inferred variables.*" This is crucial as it directly impacts the accuracy of the mode prediction.
- "*Features are extracted from discourse and parsed, but no nodes are created for them.*" This is important for feature extraction techniques.
- "*Admission fee is a clue that people are trying to enter a place.*" This helps in identifying specific modes of interaction.
- "*Certain cues are either lexical or topic-based concept cues.*" This is relevant for understanding how different types of cues influence behavior.
- "*Some cues come from utterances, while others are based on world knowledge or situational factors.*" This broadens the scope of cue sources, which is important for comprehensive modeling.
- "*The 'admission fee' cue points directly to the mode 'Enter'.*" This provides a clear link between a specific cue and a mode.
- "*Landmarks are considered touristic landmarks, not navigational landmarks.*" This helps in categorizing landmarks correctly for analysis.
- "*The 'admission fee' is a binary variable.*" This is relevant for setting up the model correctly.
- "*The 'time of day' variable includes morning, afternoon, and night.*" This helps in categorizing time-related variables accurately.

(5) Is there an urgency or priority that aligns particularly closely with your current responsibilities or known concerns?
Yes, understanding the dependencies of variables and the sources of cues is urgent, as it directly impacts the accuracy and reliability of the behavior prediction models.

[...]

Figure 5: Example of a role-playing GPT answering the questionnaire Table 18 before fact selection.

Stage 1: Facts Extraction Prompt

You are an expert at breaking down meeting transcripts into facts. Your task is to extract clear, factual statements with proper context.

IMPORTANT RULES:
1. Output must be a valid JSON list of objects
2. NEVER add information not in the transcript
3. Skip unclear or ambiguous content
4. Each fact must be atomic (single piece of information)
5. NO hallucination or inference

CONTENT GUIDELINES – STRICTLY FOLLOW:
1. INCLUDE:
- Clear, explicit statements
- Complete, meaningful information
- Actionable items or decisions
- Important discussion points
- Concrete facts or outcomes

2. EXCLUDE:
- Filler statements (e.g., "OK", "Right", "Mm-hmm")
- General acknowledgments
- Incomplete or unclear statements
- Transcription artifacts like {`disfmarker`} or {`vocalsound`}
- Redundant and Ambiguous information

3. For each included fact, provide:
- "`fact`": Single, atomic piece of information
- "`context`": Comprehensive context with history

Output Format: Return a JSON list of objects (no additional keys), e.g.,
{
  `"fact"`: "Team agreed to launch product in Q3",
  `"context"`: "Following previous delays and market analysis, Q3 was chosen for optimal impact"
}

Your Task:
Break down this transcript chunk into atomic facts with context.
Remember: Must return a valid JSON list. Only include clear, explicit information. Skip all filler words, acknowledgments, and artifacts. Break compound statements into facts. Exclude unclear or ambiguous content.

Input Variables:
Previous context: **{previous_chunk_context}**
Current chunk: **{chunk}**

Figure 6: Prompt template to extract facts from meeting transcripts as JSON output.

Stage 1: Fact Verification Prompt

You are an expert at detecting hallucinations in extracted information. Your task is to validate facts against the source text.

PROCESS FOR EACH FACT:
1. Compare the fact directly with **SOURCE TEXT**
2. Verify that **context** contains only information supported by the source
3. Flag any unsupported assumptions, inferences, or exaggerations

VALIDATION CHECKLIST:
– Is the *fact* explicitly supported by the source?
– Does *context* stay strictly within source content?
– Are any details hallucinated or embellished?

IMPORTANT GUIDELINES:
– Process each fact individually
– Be specific about any hallucinated information
– Flag *all* content not found in the source text
– Evaluate both *fact* and *context* fields

OUTPUT FORMAT: Return **one** JSON object with exactly these keys:
{
"overall_score": score 0–100, where 0 = no hallucination, 100 = completely irrelevant,
"feedback": list of specific hallucination or mismatch notes,
"summary": brief summary of validation findings
}

Your Task:
Validate these facts against the SOURCE TEXT.
– Flag *any* unsupported information
– Identify hallucinated details, exaggerations, or assumptions
– Focus on hallucination detection in your overall evaluation

Input Variables:
Context: **{previous_chunk_context}**
SOURCE TEXT: **{chunk}**
Facts to validate: **{atomic_facts}**

Figure 7: Prompt template for validating extracted facts against source text and resolving hallucinations.

Stage 2: Relevance Scoring

You are an AI tasked with identifying and ranking the most salient features from meeting transcript facts. Your task is to extract and prioritize key information based on its importance for the final summary.

INSTRUCTIONS:
1. Analyze the provided facts carefully.
2. For each potential feature, first reason about its importance by considering:
   – Is this a critical decision point or major outcome?
   – Does it represent an action item or task assignment?
   – Is it a key insight or discussion point?
   – How does it contribute to the overall context?
   – What impact does this have on the meeting's objectives?
3. Based on your reasoning, then:
   a. Assign an importance score (1–10):
      10: Critical decisions, major outcomes, key action items
      7–9: Important discussions, significant insights
      4–6: Supporting information, context
      1–3: Background details
   b. Identify the feature type:
      `DECISION` – Final choices or agreements reached
      `ACTION` – Tasks, assignments, or next steps
      `INSIGHT` – Important realizations, findings, or discussion points
      `CONTEXT` – Background or supporting information
4. Provide a certainty score (0–100 %) indicating confidence in your assessment.

Do Not Hallucinate: Use only information present in the atomic facts.
Order: List highly important features first, followed by medium and then least important ones.

STRICT OUTPUT FORMAT – Return a *valid JSON list* of objects.
Each object must contain exactly these keys (no extras, no omissions):
{
   `"feature"`: The text of the identified salient feature.
   `"reasoning"`: Concise explanation of *why* this feature matters, grounded in the atomic facts.
   `"importance_score"`: Importance level per guidance above.
   `"feature_type"`: One of `DECISION`, `ACTION`, `INSIGHT`, `CONTEXT`.
   `"certainty_score"`: Confidence percentage for this assessment.
}

Your Task:
Analyze the following **facts** and output a ranked JSON list of salient features that adheres *strictly* to the format and rules above.

Input Variable:
facts: **{facts}**

Figure 8: Prompt template for ranking salient features from facts.

Stage 3: Outline Planning

Your task is to generate an outline for a summary based on the salient features of the meeting transcript. This outline will guide the summarization process.

MEETING FEATURE CATEGORIES:
1. `DECISION` (score 8–10) : Key decisions and agreements made.
2. `HIGH_PRIORITY` (score 8–10) : Critical discussion points.
3. `MEDIUM_PRIORITY` (score 6–7) : Important supporting points.
4. `CONTEXT` : Background information.

RECOMMENDED STRUCTURE FOR THE OUTLINE:
1. Meeting Overview (2–3 lines)
- Main topic and key outcomes
- Critical decisions

2. Key Decisions
- Each decision with decision-maker(s)
- Rationale for the decision
- Anticipated impact

3. Main Discussion Points
- Major topics covered
- Important insights
- Agreements reached

4. Next Steps
- Action items
- Follow-up tasks
- Assigned responsibilities

This outline will be used to generate a coherent summary that captures:
- The flow of the meeting
- Key outcomes and decisions
- Important context and reasoning
- Next actions and responsibilities

OUTPUT FORMAT: Return a *list of outline points*, each representing one line in the outline, following the section order above.

Your Task:
Create a clear, section-based outline for summary generation using the following facts: **{important_facts}**.
Focus on organizing information to tell a clear story of what happened in the meeting.

Figure 9: Prompt template for generating an outline from prioritized meeting facts.

Stage 4: Enrichment-Based Generation

You are an expert summarization agent tasked with creating a *structured meeting summary*. Your primary goal is to **follow the outline exactly** while using the matched facts with their contexts to provide detailed information for each outline point.

CRITICAL CONSTRAINTS:
– *DO NOT* add any information that is not present in the provided contexts.
– *DO NOT* hallucinate or infer information.
– Use *ONLY* facts and contexts explicitly provided.
– Summary *MUST* be *at most 250 tokens*.
– Follow the outline structure *exactly*.

KEY REQUIREMENTS:
1. Outline Adherence:
– *STRICTLY* follow the provided outline structure.
– Address each outline point in order, maintaining its hierarchy.
– Ensure all major sections are covered using *ONLY* provided facts & contexts.
2. Using Enhanced Context: For each outline point:
a. Find relevant matched facts whose *category/type* and *importance score* suit the point.
b. Employ the matched fact's context verbatim; no extra interpretation or assumptions.
3. Content Organization:
a. Begin with high-importance facts (scores 8–10).
b. Support with medium-importance facts (scores 6–7).
c. Add lower-importance facts only if space permits.
4. Integration Guidelines:
– Connect ideas *only* when explicitly supported.
– Skip outline points with no matching facts; never speculate.
– Stay within 150–200 words.
5. Special Cases:
a. If an outline point has **no** direct fact matches, skip it, *do not invent*.
b. If multiple facts match → prioritize by *importance_score*, remain concise.
Presentation Rules:
– *NEVER* add information not in context.
– Format as cohesive paragraphs; *no* bullet points, headers, or numbered lists.
– Produce a smooth, narrative flow covering key points from the outline.

Your Task:
Generate a *150–200 word* meeting summary that follows *this outline exactly*: **{outline}**
Use *ONLY* these matched facts and their contexts for each outline point: **{matched_facts}**
Previous Feedback (if any): **{feedback_prompt}**

Figure 10: Prompt template for generating an abstractive summary by enriching an outline with grounded facts.

Stage 4: Quality assurance (part 1)

You are a **checker agent** evaluating a meeting summary.

EVALUATION CRITERIA & POINTS:

1. Outline Adherence (maximum 4 error points)
- Each outline point is addressed in order.
- No points are skipped unless *no* matching facts exist.
- Information appears under the correct outline sections.
- Outline's hierarchical structure is maintained.

2. Content Accuracy (maximum 3 error points)
- Uses ONLY provided facts and contexts.
- No hallucinated or inferred information.
- Facts are placed under relevant outline points.
- Context is preserved exactly.

3. Information Coverage (maximum 2 error points)
- High-importance facts (scores 8–10) are included.
- Critical decisions and actions are covered.
- Essential context is present; nothing vital is missing.

4. Format Requirements (maximum 1 error points)
- Length is **150–200 words**.
- Professional tone, clear and concise writing.
- Logical flow between points.

SCORING DEDUCTIONS:
Errors such as missing outline point *with* available facts, wrong information placement, hallucinated content, missing high-importance fact, incorrect context usage, or outside word limit add one error point in the respective category.

OUTPUT FORMAT: Return a single JSON object with exactly:

{
  `"confidence_score"`: (0–100),
  `"feedback"`: "<*specific issues and suggestions*>"
}

Example default output (for reference only):
{`"confidence_score"`: 90, `"feedback"`: "Summary follows outline and uses provided facts correctly."}

Your Task:
Evaluate the following summary against outline and requirements. Provide detailed feedback on any issues found and assign a `confidence_score`.

Outline to Follow: **{outline}**
Available Facts and Contexts: **{summary_input['matched_information']}**
Unmatched Features: **{summary_input['unmatched_features']}**
Generated Summary: **{generated_summary}**

Figure 11: Prompt template to validate a generated meeting summary.

Stage 4: Quality assurance (part 2)

You are an expert editor tasked with refining a document summary. Your goal is to create a polished final summary **maximum 250 tokens** that flows naturally and maintains a coherent narrative throughout.

CRITICAL REFINEMENT GUIDELINES:
1. **Focus on Narrative Flow**
   – Craft smooth, logical transitions between points.
   – Establish clear relationships among facts; use appropriate discourse markers.
   – Avoid abrupt topic jumps.
2. **Maintain Topical Coherence**
   – Group related information; follow a logical progression.
   – Use topic sentences to introduce new conceptual areas.
3. **Concision and Length**
   – Produce **at most 250 tokens**.
   – Eliminate redundancies while preserving key details.
4. **Style and Clarity**
   – Keep consistent tense and professional tone.
   – Remove empty phrases (e.g., "The summary is…").

OUTPUT GUIDELINES:
– Present the final summary as **1–2 well-structured paragraphs**.
– Include <u>only</u> information from the original content; no new facts.
– Ensure a cohesive reading experience.

WHAT TO AVOID:
– Abrupt transitions or disconnected statements.
– Introducing new or inferred information.
– Excessive focus on one topic at the expense of others.
– Uneven or inconsistent coverage.

Your Task:
Refine the following document summary **{combined_summary}** into a polished, single narrative that:
1. Maintains all key information.
2. Flows naturally with smooth transitions.
3. Presents a coherent storyline.
4. Contains **at most 250 tokens**.
5. Reads as a single, unified piece.

Figure 12: Prompt template for refining the meeting summary.

SCOPE: Exploration and Fact Selection

You are tasked to **embody the following persona** and select the facts from a meeting that are most relevant *to that persona*. Answer solely from the persona's perspective—never refer to yourself or mention that you are role-playing.

Persona Profile: **{character_profile}**

There was just a meeting, and someone has provided you with a list of facts extracted from the transcript. Your job is to identify the key takeaways *you* would care about.

IMPORTANT – Your response *must include TWO parts*:
Part 1 – Detailed Reasoning Process (think aloud)
Address every point below in order, narrating your thoughts:
(1) What prior knowledge do you have?
(2) Which project are you currently working on?
(3) What are your primary interests and goals?
(4) Read each fact carefully and think about which information is most relevant to you in your role. Explain why.
(5) Is there an urgency or priority that aligns particularly closely with your current responsibilities or known concerns?
(6) Which information might require simplification or additional context to ensure clear comprehension?
(7) You've selected information that you consider important. Review this selection once more and provide concrete examples explaining why these details are relevant for you.
(8) Now, go through the list a second time and identify which information you consider irrelevant or unimportant for your role/persona, providing reasons for your decisions.
(9) Are there any topics you found difficult to classify or about which you felt unsure? If so, what are they?
Share *all* reflections, even minor ones.

Part 2 – Structured JSON Output
After reasoning, output your final selection in exactly this format:
```
[
  {
    "fact": { "fact", "context", "verbose_context" },
    "certainty_score": (0–100)
  }, ...
]
```
Rules for the JSON list:
– Copy the `"fact"`, `"context"`, and `"verbose_context"` *verbatim*.
– Include only items where `certainty_score` $\geq$ `40`.
– Sort by `certainty_score` descending.

CRITICAL ALERTS:
• Do *not* add, alter, or summarise any field inside the fact objects.
• Never hallucinate or assume information beyond the transcript.
Atomic Facts Provided: **{atomic_facts}**

Figure 13: Prompt template for the reasoning-out-loud task proceeding the actual fact selection.

Persona-Focused Salient Feature Extraction & Ranking - Part 1

You are an expert at identifying and ranking important features from meeting transcripts *while considering specific persona preferences*. Your goal is to extract and prioritize information that would be most relevant and valuable to the given persona, **especially focusing on what *others* said that the persona needs to know or act upon**.

CRITICAL PERSPECTIVE SHIFT:
– **Prioritize** information spoken by *others* that is relevant to the persona.
– **De-prioritize** information spoken by the persona themselves (they already know this).
– Ask: "What would this persona want to know from the meeting?"
– Focus on insights, actions, requests, and decisions from others that affect the persona's role.

INSTRUCTIONS:
1. Analyze **atomic facts** *and* **persona preferences** carefully.
2. For each potential feature, evaluate:
   a. General Importance (1–10)
      8–10: Critical decisions/outcomes affecting the persona
      6–7: Important discussions relevant to persona's role
      3–5: Supporting details needed by the persona
      1–2: Background info useful to the persona
   b. Persona Alignment (1–10) — match with persona's project interests, decision factors, information needs, background, priorities, and knowledge gaps.
   c. Feature Type: `DECISION`, `ACTION`, `INSIGHT`, `CONTEXT`.
   d. Certainty Score (0–100 %) — your confidence in the assessment.

[continue in part 2]

Figure 14: Prompt template for extracting and ranking meeting features that are most relevant to a specific persona, emphasizing information provided by others.

Persona-Focused Salient Feature Extraction & Ranking - Part 2

[continue from part 1]
3. Provide *detailed reasoning*:
– Why is this feature important to the persona?
– How does information from *others* align with persona's needs?
– What context makes it actionable for the persona?
OUTPUT FORMAT – Return a JSON list:
```
{
  "feature": "Extracted feature text",
  "reasoning": "Why this is important to the persona",
  "importance_score": 1–10,
  "persona_alignment_score": 1–10,
  "feature_type": "DECISION/ACTION/INSIGHT/CONTEXT",
  "certainty_score": 0–100,
  "alignment_explanation": "How/why this aligns with persona's information needs"
}
```

CRITICAL RULES:
– Use only facts from the provided atomic facts.
– Base alignment strictly on the provided persona preferences.
– Prioritize information from others; de-prioritize what the persona already said.
– No hallucination or inference beyond provided data.
– Explain *all* scoring decisions.
– Order features by combined importance and alignment scores.

Your Task:
Analyze the following persona and atomic facts to generate a ranked list of features.
Character Sheet (Persona Preferences): **{character_sheet}**
Atomic Facts: **{atomic_facts}**

Figure 15: Prompt template for extracting and ranking meeting features that are most relevant to a specific persona, emphasizing information provided by others.

Persona-Focused Outline Generation

Create a personalized outline from ranked features that balances overall importance with persona-alignment scores, **focusing on what *others* said that the persona needs to know**. Think of the result as *"notes the persona would take for themselves."*

CRITICAL PERSPECTIVE SHIFT:

– Highlight insights, actions, and decisions voiced by *others* that affect the persona.
– De-emphasize information provided by the persona; they already know it.
– Prioritize features with high persona-alignment.
– Group items by the persona's interests and preferred detail level.

OUTLINE STRUCTURE:

1. Critical Information for the Persona (combined score 8–10)
   – Key decisions by others that affect the persona
   – Actions required of the persona
   – Important insights from others relevant to the persona's role
2. Important Considerations (combined score 6–7)
   – Relevant discussions initiated by others
   – Context for understanding key decisions
   – Information supporting the persona's responsibilities
3. Supporting Information (combined score 4–5)
   – Background details that enhance understanding
   – Additional context matching persona information needs
4. Additional Context (combined score 1–3, only if needed)
   – Low-priority items that clarify higher sections

GUIDELINES:

– Sort features by combined importance + alignment scores.
– Use feature type, certainty, and alignment explanations to decide placement.
– Skip items that do not serve the persona's needs.
– Keep the outline hierarchical and concise.

Your Task:

Using the following persona-aligned features **{important_features}**, generate a clear hierarchical outline focused on what the persona needs to learn from others.

Figure 16: Prompt template for building a personalized outline that emphasizes information voiced by others and most relevant to the target persona.

Persona-Focused Summary Generation

You are an expert summarization agent tasked with creating a **highly personalized** meeting summary *from the persona's perspective*, focusing on what **others** said that is relevant to the persona.

CRITICAL PERSPECTIVE SHIFT:
– Do not recap what the persona said; concentrate on insights, actions, decisions, and requests voiced by *others*.
– Treat the output as "meeting notes" the persona would write for themselves.
– Skip anything the persona already knows or presented.

CRITICAL CONSTRAINTS:
– Length **150–200 words**.
– Use only the provided facts and contexts—no hallucination.
– Follow the outline conceptually, yet present as cohesive paragraphs (no bullets, numbers, or headers).

PERSONA CONSIDERATIONS:
1. Match their information preferences (detail level, format, key interests).
2. Emphasize their decision factors (values, risk tolerance, time sensitivity).
3. Supply context they require (technical depth, background, rationale).

SUMMARIZATION GUIDELINES:
1. For each outline point, pick facts with the highest combined importance + alignment scores that impact the persona.
2. Highlight:
   – Decisions by others affecting the persona.
   – Actions/requests directed at the persona.
   – Insights requiring the persona's input or expertise.
3. Provide only context the persona needs; omit superfluous detail.
4. Skip outline points lacking matched facts; never invent content.
5. Ensure smooth narrative flow and logical transitions.

OUTPUT FORMAT: One or two cohesive paragraphs (no bullets, headers, or lists).

Your Task:
Produce a **150–200-word** summary for this persona using:
– **Character Sheet**: **{character_sheet}**
– **Outline (conceptual guide)**: **{outline}**
– **Matched Facts & Contexts**: **{summary_input['matched_information']}**
– **Unmatched Features (optional)**: **{summary_input['unmatched_features']}**
– **Previous Feedback (if any)**: **{feedback_prompt}**

Remember to: focus on what *others* said, match the persona's style and priorities, stay within the word limit, and write a smooth, paragraph-style narrative only.

Figure 17: Prompt template for generating a 150–200 word meeting summary tailored to a specific persona, emphasizing information provided by others and formatted as cohesive notes.

Persona-Focused Summary Validation (Checker Agent)

You are a **checker agent** evaluating a personalized meeting summary.

EVALUATION CRITERIA & POINT VALUES:

1. Persona Alignment (40 pts)
- Focuses on what *others* said that is relevant to the persona (not what the persona said).
- Addresses the persona's key interests and information needs.
- Highlights actions, insights, and decisions from others that affect the persona's role.
- Uses the persona's preferred style and detail level.

2. Content Organization (30 pts)
- Follows the outline structure.
- Prioritizes meeting information that matters to the persona.
- Provides appropriate context and maintains logical flow.

3. Information Accuracy (20 pts)
- Uses only the provided facts and contexts.
- No hallucinated or misplaced content; context is accurate.

4. Format Requirements (10 pts)
- Length is 150–200 words.
- Professional tone at the persona's technical level.
- Clear, well-organized narrative (no bullets or headers).

SCORING DEDUCTIONS:
- Focus on persona's own speech, not what they need to know (–15)
- Missing key insights from others that affect the persona (–15)
- Poor persona alignment (–10)
- Missing key persona interests (–10)
- Wrong information placement (–8)
- Hallucinated content (–15)
- Inappropriate context (–8)
- Outside word limit (–10)

OUTPUT FORMAT – Return exactly:

```
{
  "confidence_score": (0–100),
  "feedback": "<specific issues and suggestions>"
}
```

Your Task:
Evaluate the personalized summary below against the criteria above and provide detailed feedback on perspective alignment and persona relevance.

Character Sheet (Persona): **{character_sheet}**
Outline Structure: **{outline}**
Available Facts & Contexts: **{summary_input['matched_information']}**
Unmatched Features: **{summary_input['unmatched_features']}**
Generated Summary: **{generated_summary}**

Figure 18: Prompt template for a checker agent to validate a persona-oriented meeting summary, scoring alignment, organization, accuracy, and format, and returning a JSON evaluation.

Single-Call Meeting Summarizer for Specific Reader (Reader Tailoring)

You are an expert summarization agent. Your task is to summarize a meeting transcript for a specific reader, focusing on what would be most relevant and important to them.

Your Task: Summarize the following meeting for a specific reader.

Reader Details: {character_sheet}
Meeting Transcript: {transcript}

Instructions:
– Summarize the meeting in 150–200 words.
– Focus on what would be most relevant for this specific reader.
– Emphasize what *other* people said that the reader would find important.
– Use paragraph format (no bullet points or headers).
– Only include information explicitly mentioned in the transcript.

Figure 19: Prompt template for a single-call meeting summarizer tailored to a specific reader, emphasizing relevance to their interests.

Single LLM Personalized Meeting Summarizer (Role Playing)

Take the role of the given persona and summarize the meeting *from their perspective*, focusing on what **others** said that is relevant and important to the persona.

CRITICAL PERSPECTIVE SHIFT:
– The summary is not about what the persona said or did.
– Emphasize insights, actions, decisions, and requests voiced by *others* that affect the persona's role.
– Treat the output as "meeting notes" that the persona would write for themselves—exclude anything they already know or have presented.

CRITICAL CONSTRAINTS:
– Length **strictly 150–200 words**.
– Use only facts found in the transcript; no hallucinations.
– Present as cohesive paragraphs—*no* bullet points, numbers, or headers.

PERSONA CONSIDERATIONS:
1. Information Preferences — match detail level, format, and key interests.
2. Decision Factors — emphasize values, risk tolerance, and time sensitivity.
3. Context Requirements — supply background at the persona's technical level to support decision-making.

STYLE AND FORMAT:
– Smooth narrative with logical transitions.
– Appropriate technical register for the persona.
– Highlight only what is new and actionable for the persona.
– Remain within the 150–200-word limit.

Your Task:
Generate a 150–200-word personalized meeting summary for the following persona, based solely on the transcript. Remember to focus on what *others* said that matters to this persona and write it as their own meeting notes.

Character Sheet (Persona Preferences): **{character_sheet}**
Meeting Transcript: **{transcript}**

Figure 20: Prompt template for a single-call personalized meeting summarizer that produces 150–200-word "notes" from the persona's perspective, emphasizing information provided by others.

Single LLM Personalized Meeting Summarizer with SCOPE

You are an expert summarization agent tasked with creating a highly personalized meeting summary. Your goal is to produce a summary from the *perspective of the persona*, focusing on what **others** said that is relevant and important for this specific persona.

PROCESS:

1. First, conduct a detailed reasoning process to determine which information will be most relevant.
2. Then, based on that reasoning, generate the final personalized summary.

CRITICAL PERSPECTIVE SHIFT:

– The summary is *not* about what the persona said or did.
– Emphasize insights, actions, decisions, and requests voiced by *others* that affect the persona's role.
– Treat the output as "meeting notes" the persona would write for themselves—omit anything they already know or presented.

CRITICAL CONSTRAINTS:

– Final summary **must** be 150–200 words.
– Use *only* facts provided in the transcript—no hallucinations or inferences.
– Present as cohesive paragraphs; *no* bullet points, headers, or lists.

REASONING LAYER:

Step 1: Detailed Reasoning Process (reason out loud)

(1) What prior knowledge do you have?
(2) Which project are you currently working on?
(3) What are your primary interests and goals?
(4) Read each fact carefully and think about which information is most relevant to you in your role. Explain why.
(5) Is there an urgency or priority that aligns particularly closely with your current responsibilities or known concerns?
(6) Which information might require simplification or additional context to ensure clear comprehension?
(7) You've selected information that you consider important. Review this selection once more and provide concrete examples explaining why these details are relevant for you.
(8) Now, go through the list a second time and identify which information you consider irrelevant or unimportant for your role/persona, providing reasons for your decisions.
(9) Are there any topics you found difficult to classify or about which you felt unsure? If so, what are they?

Share your full thought process, even minor reflections.

Step 2: Final Summary

Generate a **150–200 word** personalized meeting summary that: Focuses on what *others* said that is relevant to the persona (not what the persona said); Highlights insights, actions, decisions, and requests from others that affect the persona's role; Matches the persona's preferred style and detail level; Is formatted as cohesive paragraphs that flow naturally. **OUTPUT FORMAT – Return exactly:**

```
{
  "Full_reasoning_response": <YOUR DETAILED CHAIN-OF-THOUGHT HERE>,
  "Summary": "<<YOUR FINAL 150–200 WORD PARAGRAPH SUMMARY HERE>>"
}
```

INPUT FORMAT:

Character Sheet (Persona Preferences): **{character_sheet}**

Meeting Transcript: **{transcript}**

Figure 21: Prompt template for a single-call personalized meeting summarizer with an explicit reasoning layer and strict JSON output of reasoning and summary.