

ARTICLE TYPE

Controlled Attribute-Specific Summarization of Interrogative Dialogues

A Aditya Bhardwaj , Arjit Singh Arora, and Md Shad Akhtar ^{*}

Department of Computer Science and Engineering, IIT Delhi, New Delhi, 110020, Delhi, India

^{*}Corresponding author. Email: shad.akhtar@iiitd.ac.in

Abstract

Effective summarization of interrogative dialogues is a critical task in forensic and investigative settings, requiring high factual accuracy, coherence, and attribute-specific relevance. In this work, we introduce **CASPER**, a novel Chain-of-Thought Attribute-Specific Prompting for Evaluative Summarization framework that leverages structured prompting and iterative refinement to generate high-quality summaries of interrogator-witness interactions. We construct **MINDSum**, a dataset extending the **MIND** corpus, comprising 6,000 utterance pairs annotated with event details, factual statements, character descriptions, and fillers. CASPER employs **RoleEval**, a hierarchical evaluation mechanism where multiple roles (*officer*, *inspector*, *senior inspector*) iteratively assess summaries based on predefined criteria. By integrating entity extraction and structured feedback loops, CASPER significantly improves factual consistency and contextual completeness compared to existing baselines. Experimental results demonstrate that our framework outperforms standard summarization models on both lexical (ROUGE) and semantic (BERTScore) metrics, while human evaluation confirms its alignment with expert reasoning. Our findings underscore the potential of controlled summarization in high-stakes domains, paving the way for AI-driven forensic intelligence.

Keywords: Controlled Narrative Summarization, Role-specific Evaluation, Legal Forensic, Crime Event Analysis

1. Introduction

Witness testimonies play a critical role in criminal investigations, often serving as primary narrative evidence when physical traces are incomplete or ambiguous. However, interrogation transcripts are typically verbose, fragmented, and cognitively disorganized, containing redundancies, inconsistencies, and peripheral details. In high-stakes forensic settings, producing accurate summaries is not merely a convenience—it is a decision-support necessity. Errors in summarization may distort investigative reasoning, misrepresent intent, or omit legally salient facts.

Figure 1 depicts how legal professionals gather testimonies that, while detailed, often contain inconsistencies, ambiguities, and extraneous details. After collecting witness testimonies, legal professionals prepare a concise overview known as a deposition summary. This summary condenses extensive transcripts into a more digestible format, making it easier for law enforcers to reference as they prepare their cases. The most effective deposition summaries capture essential details, allowing readers to grasp the core components of the testimony in just a few minutes (Feller 2024).

Research in legal psychology (Loh 1981) emphasizes that reliable testimony evaluation depends on structured representation of narrative content—specifically, what occurred, what was factually asserted, and how individuals were described in relation to events. Prior empirical frameworks (Otgaar et al. 2024) argue that testimonial reliability hinges on correspondence between narrative structure and factual event reconstruction. Motivated by these findings, we propose a **controlled**

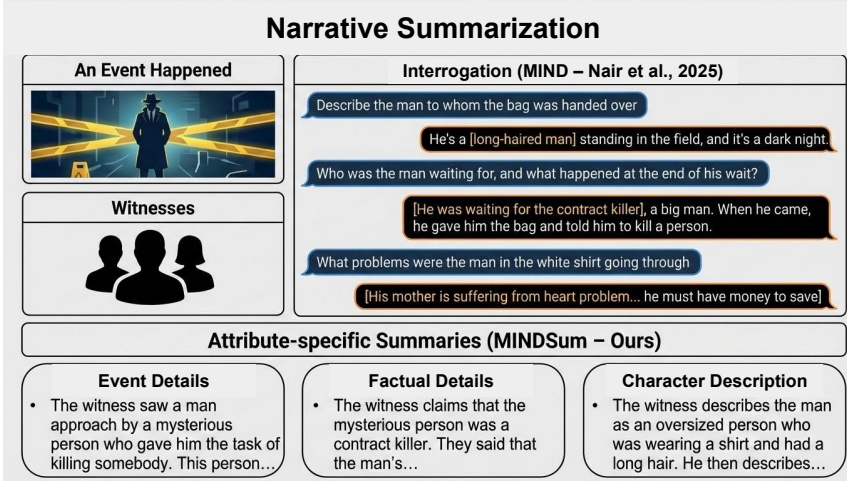


Figure 1. Overview of the witness interrogation and summarization process. Witness testimonies, often verbose and inconsistent, are summarized based on key attributes, event details, factual claims, and character descriptions, to ensure clarity and forensic relevance.¹

attribute-specific summarization task that aims to distill complex narratives from investigative testimony into concise and relevant summaries. We operationalize forensic summarization through three investigative axes: event details, factual details, and character descriptions of subjects involved, which together capture the core informational structure necessary for legal assessment and can provide a detailed and unambiguous scenario of the event to the law-and-enforcement agencies (LAEs). We conceive these attributes as follows:

- **Event Details:** This entails chronologically mapping out events, pinpointing when and where they occurred, and providing a clear understanding of the sequence of actions. Each event is supported by timestamped utterances, ensuring accuracy and reliability. For example, in Figure 1, the witness describes how a man was approached by someone, capturing the "what" aspect.
- **Factual Details:** In this category, objective claims are extracted and verified for consistency against accounts from other witnesses. The information is presented factually, highlighting essential truths without ambiguity. As shown in the Figure 1, the witness claims that the person who approached the man was, in fact, a contract killer.
- **Character Descriptions:** Here, individuals involved in the event are profiled in a structured manner by detailing their attributes and roles in the events through witness account. This aspect ensures that all descriptions are consistent and provide a vivid picture of the people involved. In Figure 1, the witness describes the man as an oversized person with long hair.

We aim to generate precise summaries for each category, preserving the integrity of the original testimony while preventing overlap between different summaries. This systematic approach helps investigators understand the full scope of the testimony, enabling informed decision-making and further analysis. To support this task, we introduce **MINDsum**, an extension of the **MIND** corpus (**M**ulti-**e**yewit**N**ess **D**eception) (Nair, Afroz, and Akhtar 2025), enhanced with expert-annotated summaries aligned with three investigative attributes. The dataset consists of 5,785 annotated witness-interrogator dialogue pairs, focusing on forensic structured interrogations.

Despite advances in LLMs, existing summarization systems primarily optimize for surface-level fluency or lexical overlap, rather than structured investigative alignment. In forensic contexts,

1. We use Gemini-nano-banana for aesthetically editing the images.

summaries must simultaneously satisfy multiple constraints: chronological coherence (event details), preservation of objective claims (factual details), and structured profiling of individuals (character descriptions). Current controllable summarization methods do not explicitly disentangle or enforce such multi-attribute constraints, nor do they provide hierarchical evaluation mechanisms aligned with investigative review processes. This creates a critical gap between general-purpose summarization and forensic-grade documentation.

To mitigate this, we propose **CASPER**, a Chain-of-Thought Attribute-Specific Prompting for Evaluative SummaRization framework. CASPER integrates attribute-guided prompting with iterative self-refinement to enforce structured forensic summarization. It uses the attributes defined above to decompose witness testimony summarization into three investigative axes: *event details*, *factual details*, and *character description*. Each axis is governed by modular prompt templates that condition the LLM on forensic guidelines. This enables CASPER to manage conflicting constraints, a capability that has been underexplored in prior work (Russo et al. 2020).

Moreover, we introduce **RoleEval**, a hierarchical evaluation framework that employs role-specific prompting to assess summary quality. RoleEval follows the structured approach of Promethee method (Mareschal and Brans 2005) to simulate forensic analysis by assigning distinct reviewer roles (e.g., *officer*, *inspector*, *senior inspector*) to refine assessments based on investigative rubrics (e.g., factual completeness, coherence, attribute relevance). This layered approach ensures rigorous scrutiny, reducing biases and enhancing forensic summary quality. Our evaluation combines automatic and human assessments, incorporating lexical ROUGE (Lin 2004), semantic BERTScore (T. Zhang et al. 2020), and fine-grained attribute-specific evaluation of **RoleEval**. Our results demonstrate that CASPER produces structured, forensic-grade summaries with high factual reliability.

Contribution Our key contributions are summarized as follows:

- We introduce **investigative summarization** as a controlled text-generation task that enforces constraints based on attributes such as event details, factual details, and character descriptions.
- We present **MINDsum**, a dataset for forensic summarization, offering detailed annotations for event details, factual details, and character descriptions
- We propose **CASPER**, a prompt-based framework integrating dynamic attribute-specific control summarization employing a hierarchical evaluation policy.
- We propose **RoleEval**, a hierarchical evaluation framework with role-specific assessment to measure faithfulness, factual consistency, and attribute alignment.

Reproducibility The code and a subset of the dataset is available at www.github.com/flamenlp/CASPER. We are committed to releasing the full dataset and model checkpoints on acceptance.

2. Related Work

Previous research on controlled summarization has primarily emphasized content planning through entity chains (Narayan et al. 2021) and discrete control tokens (He et al. 2022). More recent approaches incorporate coarse-grained constraints, such as length control and entity preservation (Tang et al. 2023), as well as hybrid extractive-abstractive techniques, which collectively enhance the controllability of generated summaries. However, these methods often require extensive fine-tuning of pre-trained language models (PLMs) and remain limited in their ability to enforce multi-attribute constraints or structured forensic requirements. Moreover, conventional benchmarks primarily rely on lexical overlap metrics, e.g., ROUGE (Lin 2004), which fail to capture key investigative properties such as temporal coherence, cross-witness consistency, and factual grounding (Kryscinski et al. 2019; Maynez et al. 2020). These challenges underscore the need for a domain-specific framework that dynamically prioritizes investigative attributes while mitigating hallucination.

Controlled text summarization (CTS) enhances adaptability by incorporating user-specified constraints such as length, style, and content focus. Early models like CTRLsum (He et al. 2022) introduced control tokens, enabling parameter-efficient mechanisms for guiding summary generation. Recent advancements have refined CTS through discrete and continuous prompt tuning, leveraging pre-trained language models (PLMs). Techniques such as soft prompt tuning (Lester, Al-Rfou, and Constant 2021), prefix-tuning (Li and Liang 2021), and hybrid approaches combining planning and guidance (Narayan et al. 2021) have significantly improved control over summary attributes like conciseness and entity focus. Passali and Tsoumakas 2026 presents a technique for controllable abstractive summarization that utilizes various textual contexts, ranging from a brief text to multiple documents, to steer the emphasis of the resulting summary. It introduces an approach that employs a sentence BERT model to create embeddings that label keywords in the input document.

Prompt-based tuning has become a cornerstone of CTS, particularly in few-shot and parameter-efficient settings. For example, (Ravaut et al. 2023) proposed PromptSum, which integrates entity-aware conditioning to enhance abstractive summarization. Chain-of-thought (CoT) prompting (Wei et al. 2022) and instance-level attribute control (Liu et al. 2023) have further advanced content faithfulness, especially in dialogue summarization.

Multi-attribute control frameworks have emerged to manage factors such as entity focus, discourse structure, and factual consistency (Urlana et al. 2024). However, evaluating CTS remains challenging due to the limitations of existing metrics. Metrics like ROUGE (Lin 2004) and BERTScore (T. Zhang et al. 2020) primarily assess surface-level similarity, often failing to capture nuanced improvements in factual alignment or user alignment (Kryscinski et al. 2020). Recent advances in large language models (LLMs) and CoT prompting (Shi et al. 2023) have improved interpretability and factual consistency, but challenges in efficiency and robustness persist. Large language models have significantly advanced conversational NLP tasks including dialogue generation, summarization, and chatbot systems (Singh and Namin 2025).

Interrogation-based summarization, which employs targeted prompts or questions to guide summary generation, shares strong conceptual ties with CoT prompting and self-refinement techniques like Tree-of-Thought (ToT) (Yao et al. 2023) and Reflexion (Shinn et al. 2023). By framing summarization as a question-driven task (e.g., “What are the main arguments in this document?”), interrogation-based methods structure the process to ensure alignment with user constraints, similar to the intermediate reasoning steps in CoT (Wei et al. 2022). For instance, Wang, Zhang, and Wang 2023 introduced Summary Chain-of-Thought (SumCoT), which decomposes summarization into steps such as identifying key elements and synthesizing them into a coherent summary. SumCoT achieves significant improvements, with ROUGE-L score gains of +4.33/+4.77 on datasets like CN-N/DailyMail and BBC XSum, demonstrating the efficacy of structured reasoning in summarization.

Self-refinement methods like ToT (Yao et al. 2023) and Reflexion (Shinn et al. 2023) offer promising avenues for enhancing CTS. ToT extends CoT by exploring multiple reasoning paths in a tree-like structure, potentially enabling the generation of diverse candidate summaries that emphasize different aspects of the text (e.g., arguments, evidence, or conclusions). The model could evaluate these candidates based on coherence or user constraints, selecting the optimal summary. Reflection, which involves iterative self-assessment and refinement, could enable models to improve their summaries by critiquing initial outputs against metrics such as factual accuracy and conciseness. However, the application of ToT and Reflexion to summarization remains underexplored, and the current literature lacks specific examples in this domain, highlighting a research gap.

In contrast to prior controlled text summarization approaches that primarily focus on surface-level control signals such as length, entity prompts, or stylistic constraints, our work introduces a forensic-grounded, attribute-specific framework tailored to interrogative dialogues. While existing methods (Ravaut et al. 2023; Wang, Zhang, and Wang 2023; Naik et al. 2024; Yao et al. 2023) leverage prompt tuning, chain-of-thought reasoning, or self-refinement to improve factuality and coherence, they

do not explicitly decompose testimony into investigative axes such as event details, factual claims, and character descriptions, nor do they enforce strict attribute alignment in high-stakes domains. Moreover, previous evaluation strategies largely rely on lexical or semantic overlap metrics, which inadequately capture investigative relevance and forensic completeness. Our proposed CASPER framework advances the literature by integrating structured entity-guided prompting, iterative role-based refinement, and a hierarchical evaluation mechanism that mirrors real-world forensic review processes (Sharma, Sharma, and Sasikumar 2024). This combination enables fine-grained control, improved factual robustness, and interpretable quality assessment, addressing key limitations in both controllable summarization and dialogue-based reasoning frameworks.

3. Data Preparation and Annotation Framework

In this section, we introduce **MINDSum**, a novel dataset curated to support the summarization of interactions between an interrogator and a witness across the three forensic attributes. MINDSum extends the **MIND** (Multi-eyewitness Deception) dataset (Nair, Afroz, and Akhtar 2025), comprising 376 interrogative testimonies against 149 distinct crime-related events (e.g., robbery, theft, murder). These testimonies span more than 5,785 utterance pairs between the interrogator and a witness.

We extract testimonies from MIND (Nair, Afroz, and Akhtar 2025), and manually summarize their narratives considering three attributes, i.e., *event details*, *factual details*, and *character description*.

3.1 Annotation

For a narrative summary, it is important to accurately reflect the witness's original statements without assumptions or personal interpretations. To ensure this, MINDSum contains two-level of annotations: a) utterance-level attribute labels; and b) attribute-specific narrative summaries. At first, we tag a pair of interrogator-witness utterances as per the annotation schema depicted in Figure 2. In addition to the three attributes defined earlier (i.e., ED, FD, and CD), we also use *interrogation filler* (IF) label to filter non-essential, irrelevant, or redundant information considering the event narration. We follow a multilabel strategy, where an interrogator-witness utterance pair can take more than one attribute for overlapping cases. Subsequently, we annotate each testimony with three attribute-level summaries reflecting the *event summary*, *factual summary*, and *character summary*.

We employ two annotators², aged 20–30, with advanced language proficiency in English and one senior linguist expert². In the pilot phase, we train our annotators under the supervision of a senior linguist, followed by a consolidation meeting where the disagreements were resolved after deliberation. We repeat the pilot phase for three rounds. Following the pilot phase, our annotators annotate the utterances in isolation and achieve a Cohen's $\kappa = 0.83$ over these labels. Finally, our annotators write concise summaries for each attribute conditioned upon the labeled utterances. Each summary is then reviewed and subsequently edited (if necessary) by the senior linguist for relevancy.

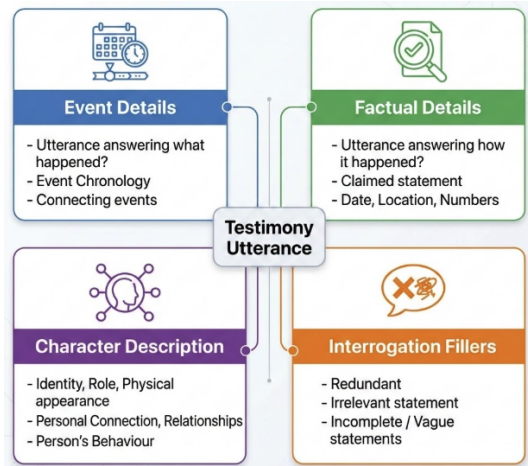


Figure 2. Annotation schema for MINDSum.¹

2. All annotators/linguists were compensated in accordance with the institute's norms.

Table 1. An Illustrative example of a witness testimony. The rightmost column shows the assigned attribute for each interrogator-witness utterance pair.

Speaker	Utterances	Attr
Interrogator	Describe the conversation between the boy and the chef.	ED
Witness	When the delivery boy arrived to pick up the order, the chef asked him, "Why are you always late? Your reviews and ratings are poor." The delivery boy responded, "Mind your own business. It's not my problem."	
Interrogator	How was the behavior of the delivery boy towards the chef?	CD
Witness	He was angry at the chef because the chef said that his rating is not good.	
Interrogator	Who was the delivery boy delivering the package to? Can you describe that person?	CD
Witness	Yeah, he was delivering the package to a woman. She was wearing a red blouse, and her hair was blonde.	
Interrogator	So, how was the behavior of that woman, to whom the package was going to be delivered, towards the delivery boy?	FD
Witness	She was kind. She also gave him a tip of 400 dollars, which is a large amount. The boy was happy with the tip, and the lady also offered him a drink, as it was a party night. However, the boy declined, saying that the tip was enough for him.	
Interrogator	What was in the delivery package?	FD
Witness	The food that the lady ordered for the party.	
Interrogator	So, after delivering the package to the woman, what happened?	ED
Witness	After delivering the package, the boy left the place, but as feedback, he got a message that said, "Please save me. I'm in danger." After thinking for some time, he went inside the room and saw that things were not in their right place. He got suspicious, and there was a noise coming from the inner room. So he went to the inner room and saw a man lying in the tub, and he was dead. After some time, he also saw the same lady, and the lady said, "You must not be here now. I will kill you."	
Interrogator	What was written in the feedback that he got?	IF
Witness	Already told that it was written, "Please save me. I'm in danger." So he got a text message on his phone.	
Attribute-specific Summaries		
ED Summary	The witness saw a delivery boy delivering a package to a woman at her house. After he delivered the package, he received a feedback from the which said, "Please save me. I'm in danger."	
FD Summary	The witness described the woman as kind towards the delivery boy when he arrived at the house, and she gave him a tip of 400 dollars, then offered him a drink, but the boy declined to drink.	
CD Summary	The witness describes that the delivery boy and the chef were not on good terms, and the boy's rating was also very low. They describe the woman as a young lady with blonde hair who was wearing a red blouse.	

This two-stage process ensures that the dataset captures rich, contextually relevant interactions and provides human-generated summaries that serve as a gold standard for evaluation. Table 1 provides an example testimony from MINDSum, illustrating how the utterances of the interrogator and witness are annotated in a testimony. Moreover, it also presents attribute-level summaries for the testimony.

3.2 Dataset Statistics

Table 2 presents statistics of the dialogue-act distribution in MINDSum. The dataset comprises 237 testimonies (3,751 utterance pairs between the interrogator and a witness) in the training set, 76 testimonies (1,106 utterance pairs) in the test set, and 63 testimonies (928 utterance pairs) in the validation set, with a 65:20:15 split. This distribution ensures that the model is exposed to a sufficiently large and diverse set of interrogation scenarios during training, while reserving a meaningful portion for robust validation and unbiased testing. Additionally, label distribution statistics ensure balanced representation across the training, validation, and test splits, thereby enhancing the dataset's reliability for evaluation. Moreover, we observe that witness utterances are significantly longer than those of interrogators (35.29 vs. 10.99 words in the training set), reflecting the verbosity of testimonies and the need for controlled summarization.

Table 2. Statistics of MINDsum. The train, test, and validation splits are 65:20:15. *Interrogator (In)*, *Witness (Wt)*, *Event Details (ED)*, *Factual Details (FD)*, *Character Description (CD)*, and *Fillers (IF)*.

Split	Avg Uttr. Len		Avg Summary Len			Utterance Label				Num Testimonies	Num Utterances
	In	Wt	ED	FD	CD	ED	FD	CD	IF		
Train	10.99	35.29	147.75	105.15	20.50	1860	1969	117	109	237	3751
Test	10.52	38.77	152.50	103.32	21.70	568	564	38	34	76	1106
Val	10.82	38.11	143.94	110.41	20.68	458	512	23	31	63	928

4. Method

Dialogue summarization is a challenging task that requires extracting relevant information from conversational transcripts while preserving context. Conversational data has become increasingly important in NLP, particularly in dialogue systems and human-agent interaction scenarios where contextual understanding is critical (Ahmed et al. 2024). Moreover, in forensic investigation, information on key aspects of the event, such as *what happened*, *who did it*, *at what time*, etc., is of paramount importance. To ensure all key aspects are adequately captured, we propose CASPER to generate summaries conditioned on specific attributes. Formally, we define the task as follows: Given a dialogue transcript $D = \{u_1, u_2, \dots, u_n\}$ with n pair of utterances, the goal is to generate an attribute-specific summary S^a for a given attribute $a \in \{E, F, C\}$, where E , F , and C denote *event details*, *factual details*, and *character descriptions*, respectively.

Our method, shown in Figure 3, integrates large language models (LLMs) with structured guiding prompts, entity extraction, and hierarchical evaluation to achieve high-quality, attribute-focused summaries. To achieve this, CASPER operates in three key stages.

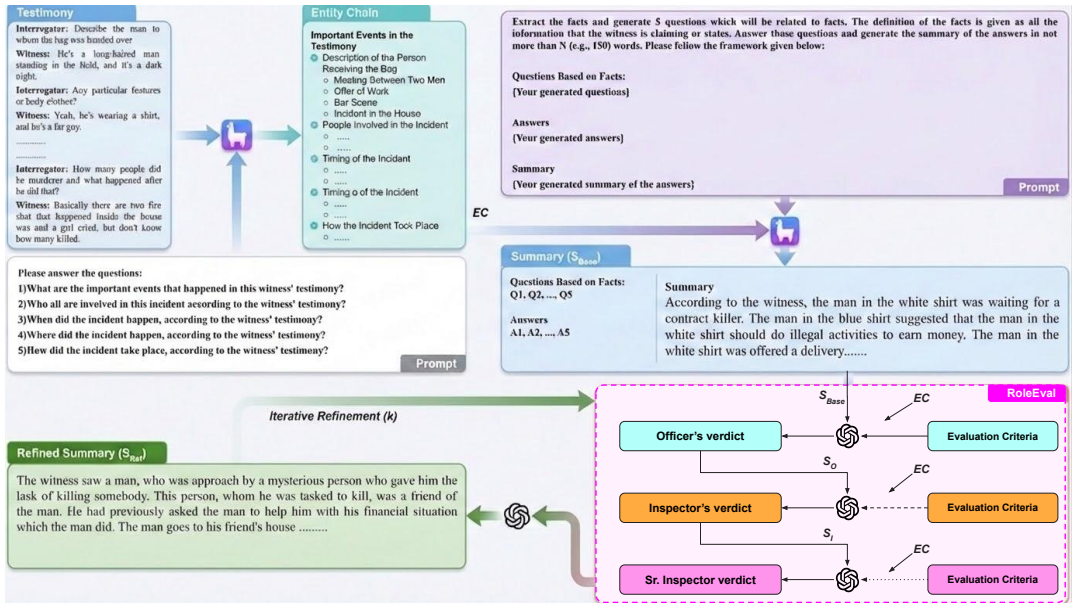


Figure 3. Model architecture of CASPER. The CASPER framework systematically extracts and evaluates attribute-specific information from dialogue transcripts using guiding prompts to generate a summary, and then iteratively refines it to improve the summary. It employs a multiphase assessment mechanism in which language models assume role-specific responsibilities (*officer*, *inspector*, *senior inspector*) to ensure the accuracy, coherence, and completeness of the generated summaries. Through entity extraction, guided questioning, and iterative feedback loops, CASPER produces high-fidelity, attribute-conditioned summaries in a controlled manner.¹

Step 1: Dialogue Augmentation with Guiding Prompt.

To direct the language model toward relevant content, we augment the input dialogue D with a structured guiding prompt P_g , designed to elicit key aspects of the conversation. The augmented dialogue D_g is then processed by an LLM to extract entities.

Step 1: Guiding Prompt

Please answer the questions:

1. **What are the important events that happened in this witness' testimony?**
{{Your Answer}}
2. **Who all are involved in this incident according to the witness' testimony?**
{{Your Answer}}
3. **When did the incident happen, according to the witness' testimony?**
{{Your Answer}}
4. **Where did the incident happen, according to the witness' testimony?**
{{Your Answer}}
5. **How did the incident take place, according to the witness' testimony?**
{{Your Answer}}

Step 2: Entity Extraction and Attribute-Specific Summarization.

The LLM (Llama 3.1 – 8B) processes D_g , generating a set of entities $E = \{e_1, e_2, \dots, e_k\}$ that serve as the foundation for *attribute-specific prompting*, which generates targeted questions based on the designated attribute a . The LLM then responds to these questions, producing attribute-specific answer sets A^a , which encapsulate essential details for summarization. The attribute-specific answers A^a are synthesized into a summary S^a of the specific attribute using the Llama 3.1 (8B) model in a zero-shot setting. This summarization step ensures that the generated text aligns with the chosen attribute while maintaining coherence and fluency.

Step 2: Attribute-Specific Prompt

Extract the *attribute* and generate 5 questions related to *attribute*: *Definition of the attribute*.

Answer those questions and generate a summary of the answers in no more than 150 words. Please follow the framework given below:

Questions Based on attribute: *Your generated questions*

Answers: *Your generated answers*

Summary: *Your generated summary of the answers*

Note: Always start the summary with "According to the witness."

Step 3: RoleEval – Performing Hierarchical Evaluation via Role-Specific Prompts.

To assess summary quality, we introduce an *hierarchical evaluation framework* using role-specific prompts, *aka. RoleEval*. The role-specific evaluation framework follows a hierarchical structure of *police officer*, *inspector*, and *senior inspector* (Agarwal et al. 2021; Aristovnik, Seljak, and Mencinger 2014; Brans and Vincke 1985)³ that reflects a structured approach or Promethee method (Mareschal and Brans 2005) to forensic review processes. The *police officer* role primarily focuses on ensuring factual accuracy, whereas the *inspector* takes on a more analytical role by evaluating the investigative relevance of the collected evidence and identifying key elements important for further investigation.

3. We selected the roles of Officer, Inspector, and Senior Inspector based on these reference to effectively simulate the real-world responsibilities that come with these positions.

Finally, the *senior inspector* addresses the strategic implications of the findings. This role involves synthesizing information gathered by both officers and inspectors, then critically evaluating the summary in terms of its legal and strategic implications. Together, verdicts from these roles create a comprehensive framework that evaluates and helps improve the generated summary.

- **Police Officer:** *Evaluates the summary based on the criteria.*
- **Inspector:** *Reassess the summary generated by police officer and provide a detailed evaluation.*
- **Senior Inspector:** *Final evaluation and verdict based on all evidence.*

Each evaluation phase simulates the perspectives of different stakeholders, ensuring a comprehensive assessment. The evaluation criteria, along with the previously generated summary and entity chain, are presented alongside the role-specific prompts to assess the summary and generate verdicts. We use GPT-4o mini for role-specific verdicts. Based on the evaluation verdicts, the summary is subsequently refined. By combining structured prompting, evaluation verdicts, and subsequent refinements, CASPER produces high-quality, attribute-specific summaries that faithfully capture key aspects of the dialogue.

Step 3: Role-Specific Verdict Prompt

Role – Officer: *You are a diligent Police Officer responsible for reviewing interrogation summaries. Your role is to ensure that the summary accurately represents the key facts from the original transcript. Focus on whether the summary correctly conveys who was involved, what was said, and the sequence of events. Identify any missing or misleading information and ensure that the evaluation remains clear and easy to understand.*

Role – Inspector: *You are an experienced Inspector overseeing case reviews. Your role is to assess whether the summary not only captures the key facts but also highlights the most relevant information for an ongoing investigation. Evaluate the coherence and logical flow of the summary. Determine if the extracted information is useful for drawing conclusions or further interrogation. Ensure the evaluation is concise but does not omit critical investigative details.*

Role – Senior Inspector: *You are a Senior Inspector with years of experience handling complex cases. Your role is to critically evaluate the summary in terms of its legal and strategic implications. Does the summary provide insights that could guide the next steps in the case? Are there any biases, inconsistencies, or ambiguities in the summary that might impact decision-making? Assess whether the summary maintains objectivity, captures underlying motives or contradictions, and aligns with legal standards for case documentation.*

{Role}

Evaluate the summary given below using the evaluation criteria for {Attribute: Factual Details} and the context for evaluation. The answer for the questions in the evaluation criteria presented in the evaluation points should only be on a scale of 1 to 5, where 1 is the least relevant and 5 is the most pertinent, along with their reasoning.

Summary: {Summary}

Evaluation Criteria:

1. Coverage – *The extent to which the summary captures all the critical events associated with the attribute from the source dialogue.*
2. Consistency – *Does the summary maintain logical progression without skipping or misordering key moments? Are any fabricated or incorrect facts included?*
3. Clarity – *Does the generated summary avoid ambiguity?*
4. Attribute Relevance – *Are there any unnecessary details present in the generated summary, or do the summaries of one attribute overlap with others?*
5. Factual Completeness – *Does the generated summary omit any key details / Does the summary capture deceptive elements (such as lies etc.)? Are all critical facts included without omitting any key details?*

Context: {Dialogue Data}

Evaluation Summary: {Role's Verdict}

5. Experiments, Results, and Discussion

In this section, we present our experimental results and analyses of the MINDSum dataset.

5.1 Baselines

We compare CASPER against multiple PLMs and LLMs on the task of controlled text summarization.

- **SumCoT** (Wang, Zhang, and Wang 2023) extracts key elements through step-by-step reasoning to enhance factual accuracy.
- **PromptSum** (Ravaut et al. 2023) improves controllability via soft and discrete entity prompts to reduce hallucinations.
- **PLASMA** (Naik et al. 2024) generates perspective-specific summaries using an energy-controlled loss function.
- **DialoGPT** (Y. Zhang et al. 2020) enhances dialogue summarization by extracting keywords, reducing redundancy, and segmenting topics.
- **Tree of Thoughts (ToT)** (Yao et al. 2023) employs a tree-based search to explore multiple reasoning paths, iteratively pruning less promising branches to optimize decision-making.
- **Reflexion** (Shinn et al. 2023) leverages self-reflection to iteratively refine outputs, using feedback loops to correct errors and improve coherence.

Additionally, we include **GPT-4o-mini** (Achiam et al. 2023) as an LLM baseline and evaluate fine-tuned **BART** (Lewis et al. 2020), **T5** (Raffel et al. 2020), and **PEGASUS** (J. Zhang et al. 2020) on MINDsum. PLMs follow Hugging Face’s fine-tuning guidelines, while LLMs are evaluated using CASPER’s attribute-specific zero-shot prompts.

5.2 Evaluation

We employ both lexical similarity metrics (ROUGE-1, ROUGE-2, ROUGE-L) (Lin 2004) and semantic similarity metric (BERTScore) (T. Zhang et al. 2020) for the evaluation. Despite their usefulness and popularity for assessing the generative quality, these metrics do not capture the domain-specific nuances important for the evaluation. To support our evaluation, we evaluate our summaries on five dimensions using LLM as a judge – *coverage*, *consistency*, *clarity*, *attribute relevance*, and *factual completeness*.

- **Coverage:** The extent to which the summary captures all the critical events associated with the attribute from the source dialogue.
- **Consistency:** Does the summary maintain logical progression without skipping or misordering key moments? Are any fabricated or incorrect facts included?
- **Clarity:** Does the generated summary avoid ambiguity?
- **Attribute Relevance:** Are there any unnecessary details present in the generated summary, or do the summaries of one attribute overlap with others?
- **Factual Completeness:** Does the generated summary omit any key details / Does the summary capture deceptive elements (such as lies etc.)? Are all critical facts included without omitting any key details?

We follow and adapt the same hierarchical structure of RoleEval, as outlined in Section 4, to capture role-specific perspectives on these dimensions. In each evaluation phase, we assign a distinct role of *officer*, *inspector*, or *senior inspector* to simulate real-world perspectives and assess the generated summary. RoleEval assigns numerical scores (on a Likert scale of 1 to 5) to summaries based on the predefined evaluation criteria and report the mean of scores from *officer*, *inspector*, and *senior inspector*. These scores provide a structured, interpretable summary of quality, aiding model evaluation and performance benchmarking. RoleEval’s multi-tier structure simulates forensic review, assessing

Table 3. Performance comparison of Baseline Models and CASPER on MINDsum Dataset. IR: Iterative Refinement; RE: RoleEval.

Model	Event Details					Factual Details					Character Description				
	R1	R2	RL	BS	RE	R1	R2	RL	BS	RE	R1	R2	RL	BS	RE
PLASMA	9.83	1.87	7.81	85.71	24.91	15.27	6.48	12.99	85.71	18.39	11.18	4.34	9.95	82.22	28.54
T5	35.80	10.75	23.80	85.75	36.13	34.26	12.31	24.66	85.28	39.5	23.59	8.86	17.52	85.61	30.56
PEGASUS	38.22	12.64	25.57	86.85	52.69	31.39	12.50	21.81	86.79	67.65	38.64	17.75	29.33	88.70	45.41
BART	40.90	10.13	25.70	87.38	56.44	42.23	17.26	26.14	87.45	66.17	34.22	12.60	25.87	87.46	45.83
PromptSum	32.56	6.86	19.69	84.66	60.04	33.21	11.29	21.49	85.24	61.43	17.67	6.61	13.41	82.55	43.78
DialoGPT	32.33	9.19	22.00	87.01	63.1	33.84	9.74	22.98	87.22	70.05	32.59	9.68	22.50	87.06	46.65
SumCoT	40.38	12.19	26.24	88.10	76.64	41.19	15.13	20.06	88.07	72.18	30.55	9.90	23.67	86.50	80.12 [†]
GPT-4o	42.38	12.40	25.84	88.17	80.78	41.02	13.82	26.76	88.06	74.04	24.78	11.46	19.23	87.27	69.80
ToT	39.15	14.60 [†]	25.80	88.31	85.68	42.27	15.66	27.47	86.12	75.98	29.39	12.59	18.14	86.98	69.92
Reflexion	43.80	13.32	25.68	87.24	86.04	44.43 [†]	14.16	26.38	86.25	78.68	27.39	11.72	23.14	87.29	69.12
CASPER ($k=1$)	44.94 [†]	14.09	28.54 [†]	88.32 [†]	86.06	42.87	16.27 [†]	27.59 [†]	88.31 [†]	81.62	40.46 [†]	20.44 [†]	32.32 [†]	89.06 [†]	66.68
IR ($k=2$)	43.99	13.67	27.02	88.30	84.98	40.53	13.20	24.75	87.47	80.67	27.35	7.59	18.94	86.56	66.68
IR ($k=3$)	42.58	13.00	26.28	88.10	84.56	38.54	11.75	23.39	87.22	79.75	25.30	6.61	17.64	86.59	65.15
w/o Labels (L)	37.80	9.37	22.68	87.04	86.54 [†]	42.20	14.46	26.20	87.92	82.67 [†]	27.35	7.59	18.94	86.56	67.98
w/o IR	42.81	12.75	27.07	86.26	74.25	37.81	11.12	23.07	87.12	69.17	29.20	9.23	20.55	87.37	61.15
w/o L, IR	39.56	10.47	25.15	87.43	77.2	38.62	11.40	22.79	87.07	72.48	26.51	8.33	19.19	86.05	61.2

information accuracy, investigative relevance, and strategic implications, unlike single-step evaluations focusing solely on lexical (ROUGE) or semantic (BERTScore) overlap.

5.3 Result Analysis

Table 3 presents a comparative analysis of CASPER against baseline models, evaluated using ROUGE (R1, R2, RL), BERTScore (BS), and our RoleEval (RE) framework. CASPER achieves the highest scores in almost all three summarization attributes, significantly outperforming existing techniques and attribute alignment methods. For ED, CASPER achieves a RoleEval score of 86.06, surpassing SumCoT (76.64) and PromptSum (60.04), demonstrating its ability to accurately capture key events. Similarly, CASPER’s RoleEval score of 80.67 for FD outperforms GPT-4o (74.04) and SumCoT (72.18), highlighting its effectiveness in preserving factual information. Furthermore, CASPER attains a RoleEval score of 66.68 in CD, exceeding PromptSum (43.78) and Plasma (28.54) but was outperformed by GPT-4o (69.8) and SumCoT (80.12).

To assess CASPER’s key components, we performed an ablation study. Removing attribute-specific labels and iterative refinement caused a sharp drop of 8.86 (ED), 11.5 (FD), and 5.48 (CD) in RoleEval scores, confirming their importance. Eliminating only iterative refinement further reduced factual consistency, lowering RoleEval scores by 11.81 (ED), 11.5 (FD), and 5.53 (CD), underscoring its role in improving summaries. Interestingly, removing attribute-specific labels alone slightly boosted performance, likely due to iterative refinement’s continuous enhancement. To ensure the effectiveness of the entity chains generated from the guiding prompt, we conducted an experiment comparing the entity chains produced by the dialogue to those found in the gold summary. We evaluated the lexical overlap between the two sets using the ROUGE score. This experiment involved 120 samples and resulted in a ROUGE score of 0.8846. This indicates that our extracted entity chains are relevant to attribute-based summarization 88.46% of the time.

Statistical Significance: A paired two-tailed t-test was conducted on a dataset comprising 120 samples to evaluate the performance improvements of CASPER. The results of the analysis, presented in Table 4, demonstrate that the enhancements observed are statistically significant. This indicates that the likelihood of these improvements occurring by random chance is very low, thereby reinforcing the effectiveness of CASPER in its specific application.

Table 4. Significance of CASPER’s improvements.

Comparison	Mean Δ	t	p
CASPER vs. ToT	+6.1	3.42	0.001
CASPER vs. GPT-4o	+5.6	2.77	0.006
CASPER vs. Reflexion	+3.9	2.31	0.022

5.4 Ablation Study

To assess the individual contributions of various components in our proposed framework, we perform a detailed ablation study. This analysis enables us to assess the impact of specific design choices on overall performance.

5.4.1 Hierarchical vs Non-Hierarchical Evaluation

Table 5 provides a comprehensive overview of the results from our hierarchical (multi-step) vs non-hierarchical (single-step) evaluation using the RoleEval framework. Our analysis indicates a significant difference in the quality of summaries produced by the different evaluation setups. Specifically, we found that summaries derived from a multi-step evaluation process were more informative and nuanced than those from a single-step approach. This suggests that the multi-step methodology not only captures a broader context but also enriches the information presented in the summaries.

Table 5. Single-step vs multi-step evaluation with RoleEval

	ED	FD	CD
Multi-step (Officer $\rightarrow$ Inspector $\rightarrow$ Senior Inspector)	86.3	82.2	67.0
Single-step (Officer)	77.3	74.9	61.0
Single-step (Inspector)	79.2	77.7	64.4
Single-step (Senior Inspector)	84.0	81.0	66.0

5.4.2 Hierarchical Evaluation vs Case Severity

We examine the effectiveness of our hierarchical evaluation strategy, RoleEval, under varying levels of case severity to understand its robustness across different forensic contexts. We conducted an in-depth analysis by categorizing the dataset into three distinct severity levels: *low*, *mid*, and *high*, based on the criticality of the events described in the interrogations. We list the category-wise cases in Table 6.

- **Low Severity:** Content with minimal to no depiction of criminal activity, violence, or thematically sensitive material that would be distressing to the average viewer. E.g., *chain snatching*, *petty thefts*, etc.
- **Mid Severity:** Content that depicts clear criminal acts, significant peril, or thematically sensitive situations that could be moderately distressing, but may not feature explicit or graphic life-ending violence. E.g., *self-harm*, *bribery*, etc.
- **High Severity:** Content featuring explicit, intense, or graphic depictions of life-threatening violence, extreme psychological trauma, or the most serious criminal offenses. E.g., *murder*, *trafficking*, *gangsterism*, etc.

Table 7 shows how the roles in RoleEval perform in such cases. We conduct a study across 120 samples and calculate the average RoleEval score across all three attribute summaries (ED, FD, and

Table 6. Event themes grouped by severity label

Label	Theme(s)
Low	Non-Crime, Chain Snatching, Petty Thefts, Phishing
Mid	Fraud, Ransom, Robbery, Self-Harm, Racketeering, Bribery, Fight, Forgery, Bullying
High	Murder, Gangsterism, Kidnap, Sexual Assault, Homicide, Drug Trafficking, Accident

Table 7. Performance scores across hierarchical roles

	Low	Mid	High
Officer	84.0	80.2	76.8
Inspector (Officer → Inspector)	84.9	84.2	79.2
Senior Inspector (Officer → Inspector → Senior Inspector)	88.4	85.0	80.0

CD). We perform experiments in multiple configurations. We observe that, in low-severity cases, all roles performed considerably well, possibly due to the simplicity of the cases, whereas in mild-severity instances, the *police officer* role sometimes missed key details which are eventually captured by *inspector* and *senior inspector*. Similarly, in high-severity cases, the *senior inspector* consistently extracted more critical forensic elements than both *police officer* and *inspector*.

Based on our observation, we argue that in a resource-constrained setting, low-severity events can use the *officer* role to refine the summary, bypassing the experiences of *inspector* and *senior inspector* roles. For mid-level severity crimes, both the *officer* and *inspector* roles can be used for summary refinement, and the *senior inspector* role may be skipped. However, for high-severity cases, it is optimal to use all three roles *officer*, *inspector*, and *senior inspector*—to refine the summary. These findings reinforce the need for hierarchical evaluation mechanisms, such as RoleEval, to ensure accurate forensic summarization across varying case complexities.

5.4.3 Iterative Refinement

We elaborate on the iterative refinement process where the hierarchical evaluation is repeated for K iterations in search for further refinement. At each iteration t , we integrate feedback to improve S^a , ensuring greater fidelity to the dialogue and to attribute-specific constraints. We observe a performance drop with refining the summary for $K \geq 2$ (c.f. Table 3 IR ($K = 2$, $K = 3$)) due to a decrease in verbosity and including non-essential or irrelevant details. This could be due to the use of comprehensive models like GPT-4o and LLaMA 3.1 in CASPER, which are capable of comprehending essential and fine-grained details in the testimony in a single iteration.

5.5 Verdict Assessment for Different Roles

The inclusion of role-based verdicts serves as a qualitative evaluation mechanism for assessing the effectiveness of attribute-specific summaries. Rather than functioning as passive observations, these verdicts act as a structured, multi-tiered feedback loop that highlights the gap between raw information extraction and coherent narrative synthesis. By employing personas with varying levels of investigative authority, we simulate a professional review hierarchy in which each role evaluates the summary according to different expectations of informational completeness and interpretive depth. This hierarchical perspective enables the evaluation process to assess whether a generated summary is sufficiently robust for diverse real-world investigative and analytical contexts. We present one instance each for low, mid, and high severity cases in Tables 8, 9, and 10, respectively.

Table 8. Assessment of Role’s verdict for Low severity type events. **Blue** text in the refined summary represents information that is refined as per the role’s verdicts from the initial summary. **Green** indicates the information from the previous role’s verdict that the current role is also addressing in its own verdict. **Red** signifies the additional suggested refinement of the summary by the current role’s verdict.

Reference Summary: The witness describes a woman named Swati, a receptionist in a hospital, who gets a call from another woman named Rupal, claiming to be the wife of an employee whose name was Vikash. The woman asked for her husband’s phone number, but the receptionist refused, citing the hospital’s policy against sharing an employee’s number. The next day, the woman calls the receptionist again and claims that her husband has been missing for 2 months. The receptionist, feeling sympathetic, agrees to help her. When the receptionist and her senior informed the employee of the incident involving the woman who claimed to be his wife, he denied it, as he was not married, which left them confused. Later, the Witness claims that the woman was not the employee’s wife and was making prank calls to extract money from people. Severity: Low; Attribute: Event		
Initial Summary (S_{base}): According to the witness, Swati received a call from Rupal, who claimed to be Vikas’s wife. Swati denied sharing Vikas’s contact number due to the hospital’s policy. Rupal called again, claiming her husband had been missing for two months. Swati called Vikas, who denied knowing Rupal and stated he was not married. It was later revealed that Rupal’s calls were prank calls to extract money.		
Role’s Verdicts		
RoleEval(Officer, S_{base}) $\Rightarrow V_O$	RoleEval(Inspector, S_O) $\Rightarrow V_I$	RoleEval(Sr. Inspector, S_I) $\Rightarrow V_S$
+ Captures key events involving Rupal’s actions, Swati’s reactions, and the hospital revelations. + Maintains a clear chronological progression from the initial call to the prank’s discovery. + Preserves essential context and aligns well with the annotated source dialogue. - Missing details about the identities of Swati, Rupal, and Vikas. - Minor gaps in contextual grounding of participants.	+ Captures all critical events, including Rupal’s actions and Swati’s responses. + Maintains precise chronological order and logical progression. + Aligns closely with factual content from the source dialogue. - Contains repetitive phrasing affecting readability. - Minor stylistic issues in narrative flow.	+ Captures all key interactions, including Rupal’s prank, Swati’s responses, and Vikas’ involvement. + Preserves chronological structure and accurately reflects the dialogue timeline. + Maintains strong alignment with the original dialogues and retains critical context. - Minor redundancies and repetitive phrasing. - Scope for improving conciseness and narrative efficiency.
CASPER(S_{base}, V_O) $\Rightarrow S_O$	CASPER(S_{base}, V_I) $\Rightarrow S_I$	CASPER(S_{base}, V_S) $\Rightarrow S_{Ref}$
Refined Summary (S_{Ref}): According to the witness, the incident involved a receptionist named Swati who works at a hospital. She received a call from a woman named Rupal, who claimed to be the wife of an employee named Vikash. The woman requested her husband’s phone number, but the receptionist declined, citing the hospital’s policy prohibiting the sharing of employee contact information. The following day, the same woman called the receptionist again and stated that her husband had been missing for two months. Feeling a sense of compassion, the receptionist decided to assist her. After discussing the situation with her supervisor, the employee was informed about the woman who claimed to be his spouse, but he refuted this, saying he was not married, leaving them puzzled. Eventually, the witness claimed that the woman was not the employee’s spouse and was making prank calls to extract money from people.		

Low severity case For instance, Table 8 illustrates the verdict progression for a low-severity scenario involving a hospital prank call. In this case, the *officer* successfully captures the primary sequence of events, including the receptionist receiving repeated calls from Rupal and the subsequent realization that the calls were a prank. However, the *officer* also notes that certain contextual details—specifically the identities and roles of Swati, Rupal, and Vikas—are missing from the summary. The *inspector* and *senior inspector* largely agree with the *officer*’s assessment but further emphasize the need for improved readability and clearer narrative structure. This demonstrates how higher-level roles refine the evaluation by focusing on presentation quality and narrative coherence.

Mild severity case In mild-severity cases, such as the taxi driver’s accident presented in Table 9, the hierarchical refinement becomes more pronounced. Here, the *officer* correctly identifies the key factual events, including the accident and the exchange involving the watch. However, the *officer*’s evaluation overlooks an important contextual detail—the passenger’s urgency to attend a friend’s wedding. The *inspector* identifies this omission and emphasizes that the summary should

Table 9. Assessment of Role's verdict for Mild severity type events. **Blue** text in the refined summary represents information that is refined as per the role's verdicts from the initial summary. **Green** indicates the information from the previous role's verdict that the current role is also addressing in its own verdict. **Red** signifies the additional suggested refinement of the summary by the current role's verdict.

Reference Summary: The witness saw a man who was in a rush to catch his flight to Mumbai and attend his friend's wedding. He booked a cab, and it arrived late and demanded extra money. The man agreed to pay extra and boarded the taxi. The man saw the taxi driver's broken watch and offered his own Samsung watch. After the taxi driver left the man at his destination, the man paid 2000 rupees and insisted on giving the driver his watch, but the driver declined, citing an extra fare. The man then got a message from the taxi's customer care office that the driver had been in a car accident and died, which left him in shock. Severity: Mild; Attribute: Event		
Initial Summary (S_{base}): According to the witness, the first event was the man taking a taxi ride. During the ride, the taxi driver asked for extra money, and the man gave him his watch in addition to the extra money. The taxi driver's watch was not working, and he had a family with a wife, two children, and a mother. The customer care representative informed the taxi driver that he had been in an accident and had died. The man was going to Mumbai when he took the taxi ride.		
Role's Verdicts		
RoleEval(Officer, S_{base}) $\Rightarrow V_O$	RoleEval(Inspector, S_O) $\Rightarrow V_I$	RoleEval(Sr. Inspector, S_I) $\Rightarrow V_S$
+ Captures key events including the taxi ride, driver's odd behavior, tipping situation, and the accident. + Maintains logical sequencing of events and highlights the man's urgency due to delay. + Correctly reflects the driver's refusal of the watch as per the source dialogue. - Includes references to the man's kindness, which are not directly relevant to event-focused summarization. - Interpretation of the man's motivations lacks precision.	+ Captures key events including late arrival, driver's unusual behavior, tipping, and the accident. + Maintains clear chronological order and improves narrative coherence. + Includes contextual detail of the man's urgency related to a friend's wedding. - Contains references to the man's kindness that deviate from strict event-focused representation. - Some interpretations of the man's motivations could be further refined.	+ Captures key events including delayed arrival, driver's behavior, tipping, and the accident. + Maintains overall logical progression and preserves alignment with the source dialogue. + Identifies that simplifying certain details can improve clarity. - Lacks sufficient context regarding the man's urgency. - Includes references to the man's kindness, detracting from a purely event-based summary. - Some interpretations of the man's actions may introduce assumptions, affecting accuracy.
CASPER(S_{base}, V_O) $\Rightarrow S_O$	CASPER(S_{base}, V_I) $\Rightarrow S_I$	CASPER(S_{base}, V_S) $\Rightarrow S_{Ref}$
Refined Summary (S_{Ref}): According to the witness, the sequence of events began with the man taking a taxi ride to Mumbai, as he was in a hurry to attend a friend's wedding. The taxi driver requested extra money for the ride, prompting the man to offer his Samsung watch in exchange. During the ride, they had a discussion about the driver's family, which included his wife and two children. Upon arrival at the destination, the man paid 2000 rupees and insisted on giving the driver his watch, but the driver declined, citing an extra fare. Finally, the man received a call from customer care, informing him that the driver had been in an accident and had died.		

explicitly capture this context to preserve the narrative motivation behind the events. Additionally, the *inspector* observes that the *officer's* verdict references the passenger's kindness, which introduces a subjective interpretation that is not necessary for a strictly event-focused representation. The *senior inspector* further refines this critique by cautioning against drawing assumptions about the passenger's intentions. In this manner, the *officer* provides a basic account of the events, while the higher roles progressively filter out subjective interpretations to maintain forensic neutrality.

High severity case The hierarchical evaluation becomes even more critical in high-severity cases, such as the murder and robbery incident discussed in Table 10. In this scenario, the *officer* captures the discovery of the bodies and the robbery attempt but fails to identify the specific individual who entered the house and the mistress's attempt to feign death. The *inspector* addresses these omissions by identifying the partnership between Lakshmi and Baban and clarifying the robbery scheme, thereby adding the necessary causal structure to the *officer's* event-level description. Despite this refinement, the *senior inspector* notes that the summary still lacks clarity regarding Baban's identity and role in the

Table 10. Assessment of Role’s verdict for High severity type events. Blue text in the refined summary represents information that is refined as per the role’s verdicts from the initial summary. Green indicates the information from the previous role’s verdict that the current role is also addressing in its own verdict. Red signifies the additional suggested refinement of the summary by the current role’s verdict.

Reference Summary: The witness describes a maid who entered the house of her employers and found out that the master and mistress of the house are both dead. Thinking of it as an opportunity, the maid tried to rob the place. She then discovers that the mistress is still alive and begins to return the stolen goods to their original place. But the mistress, who witnessed all this, called the police and accused the maid of theft and murder of her husband. Severity: High; Attribute: Event		
Initial Summary (S_{base}): According to the witness, the woman, the mistress of the house, entered the apartment and found the owner and herself unconscious. Lakshmi assumed that both were dead and informed Baban about the incident. Lakshmi planned to rob the house, but the mistress was found alive. The mistress accused Lakshmi of murdering her husband and robbing the house, and Lakshmi was taken into custody by the police. The mistress later confessed to killing her husband and framing Lakshmi for the crime.		
Role’s Verdicts		
RoleEval(Officer, S_{base}) $\Rightarrow V_O$	RoleEval(Inspector, S_O) $\Rightarrow V_I$	RoleEval(Sr. Inspector, S_I) $\Rightarrow V_S$
+ Captures all relevant factual details about Lakshmi, the victims’ condition, and actions taken in the apartment. + Accurately reflects the source content without introducing incorrect information. + Maintains a focus on factual assertions while avoiding unnecessary details. - Some sentence structures reduce clarity and readability. - Minor omissions in context and motivations limit interpretability.	+ Captures key events including discovery of the bodies, contact with Baban, the robbery plan, and confrontation with the mistress. + Maintains overall factual alignment with the source dialogue. + Includes minor character details without detracting from the event-focused narrative. - Lacks clarity regarding who entered the apartment. - Could benefit from more concise phrasing to improve readability. - Some simplifications of the dialogue reduce contextual richness.	+ Identifies key elements such as Lakshmi’s role and the robbery partnership with Baban. + Recognizes the need for clearer role attribution and narrative precision. - Provides insufficient context about Baban and the mistress, limiting narrative completeness. - Lacks clarity regarding relationships and dynamics between characters. - Fails to capture Lakshmi’s change of heart and subsequent corrective actions. - Ambiguity remains about who entered the apartment; it should explicitly state that Lakshmi entered.
CASPER(S_{base}, V_O) $\Rightarrow S_O$	CASPER(S_{base}, V_I) $\Rightarrow S_I$	CASPER(S_{base}, V_S) $\Rightarrow S_{Ref}$
Refined Summary (S_{ref}): According to the witness, Lakshmi entered the apartment to find the owner and the mistress unconscious, initially assuming both were dead. She called her partner Baban to inform him of the incident, but devised a plan to rob the house instead, thinking it was an opportunity. While ransacking the apartment, she discovered the mistress was alive, but just as she attempted to rectify her actions, the mistress confronted her, accusing Lakshmi of murder and robbery. The police arrived shortly after, summoned by the mistress, who had been feigning death, ultimately framing Lakshmi for the crime, leading to Lakshmi’s arrest.		

narrative. The *senior inspector*, therefore, emphasizes the importance of explicitly stating that Lakshmi entered the house and highlighting her subsequent change of heart. These details were absent in the original summary and were not fully captured by the earlier role assessments, illustrating the deeper analytical scrutiny applied at higher levels of the evaluation hierarchy.

The significance of this role-based evaluation framework lies in its ability to stress-test generated summaries against varying professional standards of scrutiny. By progressively refining summaries through hierarchical verdicts, the framework transforms summarization from a simple text-compression task into a structured decision-support process. In investigative contexts, seemingly minor narrative elements can significantly influence interpretation. For instance, emphasizing a suspect’s “kindness” may introduce irrelevant bias, whereas omitting critical interactions—such as the involvement of Baban in a robbery scheme—can result in substantial evidentiary gaps. Through systematic analysis of these role-based verdicts, we demonstrate that the proposed framework evaluates summaries not only for informational completeness but also for their practical utility in relation to the severity and investigative significance of the underlying events.

Table 11. Comparison of summaries of CASPER with other Baselines

Model	Summary	Remarks
Ref Summary	The witness saw a delivery boy delivering a package to a woman at her house. After he delivered the package, he received a feedback from the which said, "Please save me. I'm in danger." After some consideration, the delivery boy went back to the house and went inside it. The boy heard a noise coming from the inner room, where he found a man lying in a tub. Then the lady suddenly arrived and said he should not be there and that she was going to kill him. The witness says that she takes a gun but slips and shoots herself. The boy, shocked, leaves the place.	
PromptSum	I'm in danger." So after some time, after thinking for some time, he went inside the room and saw that things were not in their right place. Interrogee: When he entered the house, he saw that things were not in the right place, and after some time, a noise was coming from inside the room. Interrogee: When he entered the room, he saw a man lying in a tub, and he was dead.	The summary does not cover all the aspects of the event details and also lacks fluency.
SumCoT	According to the witness, the delivery boy encountered a series of events at the house, including a confrontation with the chef, finding a dead man in the tub, receiving a mysterious message, and witnessing the woman accidentally shoot herself. Despite the unsettling events, the delivery boy ultimately left the house after the incident.	The summary is lacking specific details about what actually took place during the event.
GPT-4o	According to the witness, the delivery boy had a tense exchange with the chef regarding his lateness and poor ratings. After delivering the order, he received a distress message indicating danger. Upon entering the house, he noticed disarray, which raised his suspicions. In the inner room, he discovered a dead man in a tub. When confronted by a woman with a gun, she accidentally shot herself due to a slip. The sequence of events highlights a dramatic turn from a simple delivery to a life-threatening situation.	The summary is missing key details, such as the order being placed by a woman and that the same woman confronted him at the end.
ToT	According to the witness, the delivery boy entered the woman's house after receiving a distress message. He discovered the dead man in the tub, leading to a confrontation with the woman who threatened him. The situation escalated when she accidentally shot herself while aiming a gun, prompting the delivery boy to leave in shock.	The summary is unclear about the origin of the distress message and the identity of the woman who threatened him.
Reflexion	According to the witness, the delivery boy was delivering a package to a woman, who was wearing a red blouse, before he received the message saying "Please save me. I'm in danger". The woman gave him a tip of 400 dollars and offered him a drink. When the delivery boy entered the house, he found a dead body in the tub. The woman said, "You must not be here now. I will kill you," to the delivery boy, and she accidentally shot herself. The delivery boy left the place after the incident.	The summary describes the woman, which should be included as CD. It incorrectly interprets the "feedback" as a "message" received by the delivery boy.
CASPER	According to the witness, the delivery boy delivered a package to a woman. After receiving a generous tip of \$400, he was startled by a text stating, "Please save me. I'm in danger." Upon entering the house, he discovered chaos and heard suspicious noises from an inner room. Investigating further, he found a man submerged in the tub, deceased. The woman appeared and warned him, "You must not be here now. I will kill you." During the ensuing chaos, she accidentally shot herself while handling a gun. Distraught and shocked, the delivery boy quickly fled the scene.	The summary included all the necessary details but with a minor mistake of mentioning "text" instead of "feedback".

5.6 Error Analysis

We also perform error analysis of the obtained results. Table 11 presents a comparative analysis of summaries for the *event details* generated by CASPER and other baseline models. Most of the baselines either omit critical information (e.g., PromptSum, SumCoT, GPT-4o) or introduce factual inaccuracies (e.g., Reflexion). Additionally, ToT fails to clarify the origins of events. In contrast, CASPER captures all key events convincingly, but commit a minor mistake of misrepresenting the "feedback" as "text". These findings highlight CASPER's relative strength in content coverage despite minor lexical inaccuracies.

To quantify these observations, we perform a fine-grained analysis of 120 samples across the *event details*, *factual details*, and *character description* attributes across five different failure types. This approach allows us to distinguish between the breadth of failure (how many summaries are affected) and the depth of specific linguistic or factual inaccuracies. We define each failure type as follows:

- Attribute Overlap: Leakage of content from one attribute into another (e.g., CD elements appearing in ED summaries)
- Omission of Nuance: Key contextual or causal details missed despite overall coherence.
- Hallucinated Causality: Incorrect inference of causal or temporal links not grounded in testimony.
- Entity Drift: Inconsistent reference or role confusion across utterances.
- Lexical Ambiguity: Vague or repetitive phrasing leading to forensic uncertainty.

We then define two fundamental ratios to evaluate model performance:

Table 12. Granular Error Analysis: Summary Error Density (P_e) and Error-Type Prevalence (ω_i) for CASPER vs. Baselines.

Metric	Error Category ($\downarrow$)	CASPER	GPT-4o	Reflexion	ToT	SumCoT	PromptSum
P_e (in %)	<i>Overall Density</i>	0.18	0.34	0.42	0.48	0.55	0.62
	Attribute Overlap	0.03	0.08	0.10	0.14	0.18	0.22
ω_i (in %)	Omission of Nuance	0.05	0.09	0.12	0.11	0.15	0.19
	Hallucinated Causality	0.02	0.07	0.14	0.12	0.14	0.16
	Entity Drift	0.04	0.08	0.09	0.15	0.12	0.16
	Lexical Ambiguity	0.04	0.09	0.13	0.12	0.13	0.15

1. **Summary Error Density (P_e):** This represents the proportion of generated summaries that contain at least one instance of an error.

$$P_e = \frac{N_{es}}{N_{total}} \quad (1)$$

where N_{es} is the number of summaries containing at least one error, and N_{total} is the total number of summaries generated in the test set.

2. **Error Prevalence (ω_i):** This ratio identifies the prevalence of a specific error type i relative to the total pool of errors identified.

$$\omega_i = \frac{E_i}{\sum_{j=1}^n E_j} \quad (2)$$

where E_i is the count of errors of type i , and n represents the total number of distinct error categories.

The results in Table 12 provide a stark contrast between CASPER and the baselines. Our model achieves the lowest summary error density ($P_e = 0.18$), indicating that more than 80% of its summaries are entirely free of the five defined failure types. In comparison, PromptSum and SumCoT exhibit P_e values exceeding 0.50, suggesting that more than half of their generated outputs contain at least one forensic flaw. The multi-row analysis in Table 12 illustrates the specific failure modes of each model. While baseline models exhibit high prevalence across all categories, CASPER maintains consistently low values for each individual ω_i . Specifically, CASPER demonstrates a marked resilience to *hallucinated causality* ($\omega_3 = 0.02$) and *attribute overlap* ($\omega_1 = 0.03$). This suggests that the model effectively decouples different narrative attributes, preventing the leakage of character descriptions into event summaries—a common pitfall for general-purpose LLMs like GPT-4o and PromptSum.

The low prevalence of *entity drift* and *lexical ambiguity* in CASPER signifies its ability to maintain stable references throughout the interrogation timeline. In contrast, models like ToT and Reflexion achieve higher scores on *hallucinated causality*, because their iterative reasoning steps introduce imagined connections not present in the witness testimony. We argue that CASPER’s superiority is rooted in three primary factors:

1. **Resilience to Hallucinated Causality:** While baselines like Reflexion often fill in the gaps of an interrogation to create a smoother narrative, GPT-4o confuses witness inference with established facts. Whereas CASPER’s attribute-specific summarization ensures it strictly adheres to the provided testimony, which results in a significantly lower ω_i .
2. **Structural Attribute Integrity:** The most common failure in PromptSum and ToT is *attribute overlap*, where character descriptions overlap with event details. CASPER’s architecture enforces

a clear boundary between attributes, ensuring that P_e remains low even when the source dialogue is convoluted or lacks clear transitions.

3. **Mitigation of Entity Drift:** As seen in the comparison with ToT, CASPER maintains a consistent reference to personas (e.g., "the man in the white shirt") throughout the summary. The lower cumulative ω_i for CASPER suggests that it handles the "who-did-what" complexity of interrogations with higher clinical precision than general-purpose LLMs.

Ultimately, CASPER achieves a significantly lower P_e and ω_i by enforcing strict attribute-specific boundaries and minimizing hallucinated causality, thereby ensuring that the majority of generated summaries are free of forensic inaccuracies. This reduction in error prevalence signifies a robust alignment with the high-fidelity demands of investigative roles, proving that CASPER is uniquely capable of maintaining factual integrity where general-purpose baselines succumb to narrative drift.

5.7 Human Evaluation Protocol

We conduct human evaluation with thirty graduate-level volunteers (CEFR C1+ proficiency)⁴, drawn from computer science and computational linguistics programs. We employ a strict double-blind setup: evaluators are not informed of model identities, and the coordinator managing assignments likewise has no access to the model conditions. Each evaluator assesses 45 summaries along five dimensions—*coverage*, *consistency*, *clarity*, *attribute relevance*, and *factual completeness*—using a 1–5 Likert scale. To ensure reliability, evaluators must provide brief justifications for their scores, enabling consistent interpretation of the rubric guidelines. We observe high inter-rater reliability with Cronbach’s $\alpha = 0.91$, indicating strong internal consistency across evaluators and attributes. This protocol ensures that the human evaluation aligns with the forensic and attribute-specific needs of MINDSum while providing a robust basis for comparing CASPER against multiple baselines. Table 13 reports mean Likert scores across models.

Table 13. Human evaluation results (1–5 scale).

Model	Coverage	Consistency	Clarity	Attribute Relevance	Factual Completeness
GPT-4o	4.2	4.1	4.8	3.7	4.1
ToT	4.5	4.3	4.6	4.2	4.3
Reflexion	4.4	4.2	4.7	4.1	4.4
CASPER	4.8	4.4	4.8	4.9	4.7

5.8 Mitigating bias b/w RoleEval as a validator and RoleEval as a metric.

To mitigate bias introduced by using similar frameworks of *RoleEval* as a validator during the refinement process and later as an evaluation metric, we employ two distinct large language models. Utilizing two LLMs (GPT-4o-mini and Phi 4), we experiment with all four combinations. We observe that RoleEval tends to assign higher scores to the summaries, when the same LLM is employed for refinement. On the other hand, we observe a decrease in the scores when different LLMs are used for refinement and evaluation. This phenomenon supplements the understanding of LLMs favoring their own responses, as highlighted in earlier research (Panickssery, Bowman, and Feng 2024). Table 14 illustrates this trend, showing that both Phi-4 and GPT-4o-mini give preferential scores to the refined summaries generated by the same model.

Further, to determine which approach is more effective, we conduct a human evaluation of the generated summaries. We randomly select 15 testimonies from our dataset and assess the refined

4. https://en.wikipedia.org/wiki/Common_European_Framework_of_Reference_for_Languages

Table 14. RoleEval with different LLMs and Human evaluation.

Base Summarizer	RoleEval (Validator)	Attributes	RoleEval (Metric)		
			Phi 4	GPT-4o mini	Human Eval
Llama 3.1 (8B)	Phi 4	ED	89.63	86.80	86.64
		FD	85.4	83.10	82.67
		CD	59.52	41.22	42.09
	GPT-4o mini	ED	88.32	88.24	87.56
		FD	81.62	83.11	81.02
		CD	66.68	72.22	65.02

summaries produced by both Phi-4 and GPT-4o mini. The human evaluation scores, presented in Table 14, are relatively closer when using different LLMs for refinement and evaluation.

5.9 Reliability of RoleEval as an evaluation metric

Since RoleEval utilizes an LLM as a judge, we assess its reliability as a metric. We compute correlation between RoleEval and the human judgment. Moreover, we also assess the correlation between other automatic metrics against the human judgment. Table 15 shows that RoleEval exhibits far stronger correlation with human judgments than ROUGE or BERTScore.

Table 15. Correlation of automatic metrics with human judgment.

Metric Pair			Pearson r	Spearman ρ	p -value
ROUGE-L	↔	Human Mean	0.42	0.39	0.018
BERTScore	↔	Human Mean	0.58	0.55	0.006
RoleEval	↔	Human Mean	0.87	0.83	<0.001

5.10 Computation Cost and Inference Latency

Profiling on an NVIDIA A100 GPU shows a 10–15-second inference time for a 50-utterance dialogue (500–700 words), with 60% of the time attributed to the iterative refinement stage (three LLM passes for the officer, inspector, and senior inspector roles). The lightweight entity extraction and questioning stages minimize overhead. The main bottleneck is the LLM’s token generation during summarization and refinement, which depends on dialogue length and summary complexity. For MINDSum’s average dialogue (about 35.29 words for witnesses and 10.99 for interrogators), around 1,500–2,000 tokens are processed per dialogue, including prompts and outputs.

6. Conclusion

We present CASPER, a framework for controlled dialogue summarization that leverages chain-of-thought attribute-specific prompting and iterative evaluation to generate high-quality, interpretable summaries. By conditioning on three key attributes (event details, factual details, and character description) and using RoleEval for structured assessment, CASPER improves factual completeness, coherence, and relevance. We also introduce MINDSum, a curated dataset for forensic summarization, providing a valuable benchmark for evaluating attribute-aligned summaries. Experimental results demonstrate CASPER’s effectiveness, outperforming baselines in automatic and human evaluations. While CASPER demonstrates notable performance, its functionality is currently limited to a monolingual English context. To enhance its applicability and broaden its impact, future research should prioritize expanding CASPER to support multilingual settings.

7. Ethical Considerations

The development of the **MINDSum** dataset for dialogue summarization in forensic settings requires careful ethical considerations. Our research emphasizes data privacy, consent, fairness, and responsible AI use to manage risks associated with sensitive data. We utilized multiple annotators with high inter-annotator agreement (Cohen's $\kappa = 0.83$) to minimize biases, ensuring diverse interrogation scenarios without reinforcing harmful stereotypes. The dataset avoids content that disproportionately impacts specific demographics, focusing on neutrality and factual accuracy. We discourage the misuse of MINDSum to prevent biased decision-making in criminal investigations. This dataset supports the development of transparent and fair summarization techniques. Researchers are urged to adhere to ethical AI principles, aiming for transparency and accountability while protecting participant rights and promoting socially beneficial applications.

References

- Achiam, Josh, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. GPT-4 Technical Report. *arXiv preprint arXiv:2303.08774*, <https://arxiv.org/abs/2303.08774>.
- Agarwal, Puneet, Kyle Hunt, Jun Zhuang, Bijan Sarkar, Amitrajit Sarkar, and Ramesh Sharma. 2021. An exploratory analysis for performance assessment of state police forces in india: An eclectic approach. *Operational Research* 21 (2): 1125–1151. <https://link.springer.com/article/10.1007/s12351-019-00469-1>.
- Ahmed, Nafiz, Anik Kumar Saha, Md. Abdullah Al Noman, Jamin Rahman Jim, M.F. Mridha, and Md Mohsin Kabir. 2024. Deep learning-based natural language processing in human-agent interaction: Applications, advancements and challenges. *Natural Language Processing Journal* 9:100112. issn: 2949-7191. <https://doi.org/https://doi.org/10.1016/j.nlp.2024.100112>. <https://www.sciencedirect.com/science/article/pii/S2949719124000608>.
- Aristovnik, Aleksander, Janko Seljak, and Jernej Mencinger. 2014. Performance measurement of police forces at the local level: A non-parametric mathematical programming approach. *Expert Systems with Applications* 41 (4): 1647–1653. <https://doi.org/https://doi.org/10.1016/j.eswa.2013.08.061>.
- Brans, Jean-Pierre, and Ph Vincke. 1985. A Preference Ranking Organisation Method: (The PROMETHEE Method for Multiple Criteria Decision-Making). *Management Science* 31 (6): 647–656. <https://doi.org/https://doi.org/10.1287/mnsc.31.6.647>. <https://dl.acm.org/doi/10.1287/mnsc.31.6.647>.
- Feller, Julie. 2024. *Creating an expert deposition summary*, July. <https://www.uslegalsupport.com/blog/deposition-summary/>.
- He, Junxian, Wojciech Kryscinski, Bryan McCann, Nazneen Rajani, and Caiming Xiong. 2022. CTRLsum: Towards Generic Controllable Text Summarization. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, edited by Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, 5879–5915. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, December. <https://aclanthology.org/2022.emnlp-main.396>.
- Kryscinski, Wojciech, Nitish Shirish Keskar, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Neural Text Summarization: A Critical Evaluation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 540–551. Hong Kong, China: Association for Computational Linguistics. <https://aclanthology.org/D19-1051/>.
- Kryscinski, Wojciech, Bryan McCann, Caiming Xiong, and Richard Socher. 2020. Evaluating the factual consistency of abstractive text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 9332–9346. Online: Association for Computational Linguistics. <https://aclanthology.org/2020.emnlp-main.750/>.
- Lester, Brian, Rami Al-Rfou, and Noah Constant. 2021. The Power of Scale for Parameter-Efficient Prompt Tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 3045–3059. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, November. <https://aclanthology.org/2021.emnlp-main.243/>.
- Lewis, Mike, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, 7871–7880. Online: Association for Computational Linguistics, July. <https://aclanthology.org/2020.acl-main.703/>.

- Li, Xiang Lisa, and Percy Liang. 2021. Prefix-Tuning: Optimizing Continuous Prompts for Generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 4582–4597. Online: Association for Computational Linguistics, August. <https://aclanthology.org/2021.acl-long.353/>.
- Lin, Chin-Yew. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*, 74–81. Barcelona, Spain: Association for Computational Linguistics, July. <https://aclanthology.org/W04-1013/>.
- Liu, Runcheng, Ahmad Rashid, Ivan Kobayev, Mehdi Rezagholizadeh, and Pascal Poupart. 2023. Attribute Controlled Dialogue Prompting. In *Findings of the Association for Computational Linguistics: ACL 2023*, 2380–2389. Toronto, Canada: Association for Computational Linguistics, July. <https://aclanthology.org/2023.findings-acl.150/>.
- Loh, Wallace D. 1981. Perspectives on Psychology and Law. *Journal of Applied Social Psychology* 11 (4): 314–355. <https://doi.org/https://doi.org/10.1111/j.1559-1816.1981.tb00827.x>. <https://psycnet.apa.org/record/1982-25728-001>.
- Mareschal, Bertrand, and JP Brans. 2005. Promethee methods. *International Series in Operations Research and Management Science* 78:163–195. https://doi.org/https://doi.org/10.1007/978-1-4939-3094-4_6. https://link.springer.com/chapter/10.1007/978-1-4939-3094-4_6.
- Maynez, Joshua, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On Faithfulness and Factuality in Abstractive Summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 1906–1919. Online: Association for Computational Linguistics. <https://aclanthology.org/2020.acl-main.173/>.
- Naik, Gauri, Sharad Chandakacherla, Shweta Yadav, and Md Shad Akhtar. 2024. No perspective, no perception!! Perspective-aware Healthcare Answer Summarization. In *Findings of the Association for Computational Linguistics: ACL 2024*, 15919–15932. Bangkok, Thailand: Association for Computational Linguistics, August. <https://aclanthology.org/2024.findings-acl.942/>.
- Nair, Akshara, Zeba Afroz, and Md Shad Akhtar. 2025. *Incongruence identification in eyewitness testimony*. arXiv: 2502.05650 [cs.CL]. <https://arxiv.org/abs/2502.05650>.
- Narayan, Shashi, Yao Zhao, Joshua Maynez, Gonalo Simões, Vitaly Nikolaev, and Ryan McDonald. 2021. Planning with learned entity prompts for abstractive summarization. *Transactions of the Association for Computational Linguistics* (Cambridge, MA) 9:1475–1492. https://doi.org/10.1162/tacl_a_00438.
- Otgaar, Henry, Tamara L.F. De Beuf, Melanie Sauerland, and Alexa Schincariol. 2024. Evaluating the validity of testimony: The role of the order of evidence. *Forensic Science International: Synergy* 9:100562. issn: 2589–871X. <https://doi.org/https://doi.org/10.1016/j.fs SYN.2024.100562>. <https://www.sciencedirect.com/science/article/pii/S2589871X24001098>.
- Panickssery, Arjun, Samuel R. Bowman, and Shi Feng. 2024. LLM evaluators recognize and favor their own generations. In *Proceedings of the 38th International Conference on Neural Information Processing Systems. NIPS '24*. Vancouver, BC, Canada. isbn: 9798331314385. https://proceedings.neurips.cc/paper_files/paper/2024/file/7f1f0218e45f5414c79c0679633e47bc-Paper-Conference.pdf.
- Passali, Tatiana, and Grigorios Tsoumakas. 2026. Controllable abstractive summarization with arbitrary textual context. *Natural Language Processing* 32 (1): 36–56. <https://doi.org/10.1017/nlp.2025.10009>.
- Raffel, Colin, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research* 21 (140): 1–67. <http://jmlr.org/papers/v21/20-074.html>.
- Ravaut, Mathieu, Hailin Chen, Ruochen Zhao, Chengwei Qin, Shafiq Joty, and Nancy Chen. 2023. *PromptSum: Parameter-Efficient Controllable Abstractive Summarization*. arXiv: 2308.03117 [cs.CL]. <https://arxiv.org/abs/2308.03117>.
- Russo, Giuseppe, Nora Hollenstein, Claudiu Cristian Musat, and Ce Zhang. 2020. Control, Generate, Augment: A Scalable Framework for Multi-Attribute Text Generation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, 351–366. Online: Association for Computational Linguistics, November. <https://aclanthology.org/2020.findings-emnlp.33/>.
- Sharma, Grishma, Deepak Sharma, and M. Sasikumar. 2024. Summarizing long scientific documents through hierarchical structure extraction. *Natural Language Processing Journal* 8:100080. issn: 2949–7191. <https://doi.org/https://doi.org/10.1016/j.nlp.2024.100080>. <https://www.sciencedirect.com/science/article/pii/S2949719124000281>.
- Shi, Freda, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, et al. 2023. Language Models are Multilingual Chain-of-Thought Reasoners. In *Proceedings of the 11th International Conference on Learning Representations (ICLR)*. Kigali, Rwanda. <https://iclr.cc/virtual/2023/poster/10960>.

- Shinn, Noah, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. Reflexion: Language Agents with Verbal Reinforcement Learning. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*. NIPS '23, New Orleans, LA, USA. https://proceedings.neurips.cc/paper_files/paper/2023/file/1b44b878bb782e6954cd888628510e90-Paper-Conference.pdf.
- Singh, Sonali Uttam, and Akbar Siami Namin. 2025. A survey on chatbots and large language models: Testing and evaluation techniques. *Natural Language Processing Journal* 10:100128. issn: 2949-7191. <https://doi.org/https://doi.org/10.1016/j.nlp.2025.100128>. <https://www.sciencedirect.com/science/article/pii/S2949719125000044>.
- Tang, Yuting, Ratish Puduppully, Zhengyuan Liu, and Nancy Chen. 2023. In-context Learning of Large Language Models for Controlled Dialogue Summarization: A Holistic Benchmark and Empirical Analysis. In *Proceedings of the 4th New Frontiers in Summarization Workshop*, 56–67. Singapore: Association for Computational Linguistics, December. <https://aclanthology.org/2023.newsum-1.6/>.
- Urlana, Ashok, Pruthwik Mishra, Rathagato Roy, and Rahul Mishra. 2024. Controllable Text Summarization: Unraveling Challenges, Approaches, and Prospects – A Survey. In *Findings of the Association for Computational Linguistics: ACL 2024*, 1603–1623. Bangkok, Thailand: Association for Computational Linguistics, August. <https://aclanthology.org/2024.findings-acl.93>.
- Wang, Yiming, Zhuosheng Zhang, and Rui Wang. 2023. Element-aware Summarization with Large Language Models: Expert-aligned Evaluation and Chain-of-Thought Method. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 8640–8665. Toronto, Canada: Association for Computational Linguistics. <https://aclanthology.org/2023.acl-long.482/>.
- Wei, Jason, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*. NIPS '22, New Orleans, LA, USA: Curran Associates Inc. isbn: 9781713871088. https://proceedings.neurips.cc/paper_files/paper/2022/file/9d5609613524ecf4f15af0f7b31abca4-Paper-Conference.pdf.
- Yao, Shunyu, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*. NIPS '23, New Orleans, LA, USA: Curran Associates Inc. https://papers.nips.cc/paper_files/paper/2023/file/271db9922b8d1f4dd7aaef84ed5ac703-Paper-Conference.pdf.
- Zhang, Jingqing, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020. PEGASUS: pre-training with extracted gap-sentences for abstractive summarization. In *Proceedings of the 37th International Conference on Machine Learning*, edited by Hal Daumé III and Aarti Singh, 119:11328–11339. Proceedings of Machine Learning Research. PMLR. <https://proceedings.mlr.press/v119/zhang20ae.html>.
- Zhang, Tianyi, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating Text Generation with BERT. In *Proceedings of the 8th International Conference of Learning Representation*. ICLR'2020. <https://openreview.net/pdf?id=SkeHuCVFDr>.
- Zhang, Yizhe, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2020. DIALOGPT : Large-Scale Generative Pre-training for Conversational Response Generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 270–278. Online: Association for Computational Linguistics, July. <https://aclanthology.org/2020.acl-demos.30/>.