JURIS: Bringing the Jury System to Multi-Agent Summarization

Ziling Li[1,2], **Junwei Zhang**[1,*], **Yixuan Yang**[1], **Yuqiang Han**[1,*] and **Xiaolin Li**[1,2,3,*]
[1]Hangzhou Institute of Medicine, Chinese Academy of Sciences
[2]Hangzhou Institute for Advanced Study, University of Chinese Academy of Sciences
[3]School of Statistics and Data Science, School of Information and Electronic Engineering, Zhejiang AIX College, Zhejiang Gongshang University
liziling23@mails.ucas.ac.cn, zhangjunwei@him.cas.cn, yangyixuan@him.cas.cn, hanyuqiang@him.cas.cn, xiaolinli@ieee.org

Abstract

Agents implemented using Large Language Models (LLMs) offer novel solutions for tasks such as document summarization. However, single-agent systems, limited by their own knowledge base, often exhibit information omissions or inconsistencies with the facts, while multi-agent systems lacking effective organization can fall into the trap of social cognitive defects such as flattery and premature consensus. Inspired by the U.S. jury system, we propose **JURIS**, a judicial-decision-inspired multi-agent collaborative framework. Specifically, the document summarization process simulates a lawsuit scenario: multiple generator agents condense the target text into multiple sentences from different perspectives, much like lawyers on both sides providing different evidence; multiple decision-making agents consider the overall picture based on defense and voting mechanisms to select candidate sentences, much like a jury adopting different evidence; finally, a chief editor agent compiles a summary of all selected sentences, much like a secretary summarizing a plan and a judge giving a final verdict. Extensive experiments on in-domain news and cross-domain benchmarks demonstrate that JURIS consistently and significantly outperforms strong single-agent and multi-agent baselines across automatic evaluation metrics, multi-dimensional quality assessments, and human judgments, validating the effectiveness of decision-driven structured multi-agent collaboration for document summarization.

1 Introduction

Document summarization aims to produce a concise and coherent representation of source documents while preserving essential information and factual consistency. Despite the rapid progress of LLMs, abstractive document summarization still faces persistent reliability challenges. Generated summaries frequently suffer from information omission, coverage imbalance, or factual inconsistency, limiting their applicability in knowledge-intensive and high-stakes scenarios [Feng *et al.*, 2024; Ramprasad *et al.*, 2024].

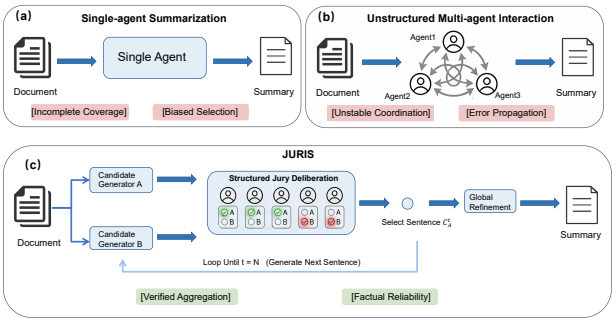

Figure 1: Comparison of summarization paradigms: (a) single-agent; (b) unstructured multi-agent; (c) judicial-decision-inspired JURIS with jury deliberation and chief editor agent refinement.

A key source of these limitations lies in the single-agent generation paradigm. When both content selection and surface realization are handled by a single model, the summarization process follows a single generative trajectory with limited external scrutiny. Once an incomplete or biased content selection is made at an early stage, subsequent generation often expands upon this partial interpretation, as the model lacks mechanisms to challenge or revise its own decisions from alternative perspectives [Belém *et al.*, 2025; Ji *et al.*, 2023]. As illustrated in Fig. 1(a), single-agent summaries tend to reflect a narrow viewpoint rather than a balanced synthesis of the source document.

To overcome these issues, recent studies have explored multi-agent collaboration, introducing multiple agents with diverse prompts or model backbones to improve robustness. Many of these approaches implicitly assume that increasing the number of participating agents leads to broader coverage and higher reliability [Wan *et al.*, 2025; Wang *et al.*, 2025]. However, empirical evidence shows that unstructured multi-agent interaction does not consistently yield robust improvements [Sharma *et al.*, 2023; Chan *et al.*, 2023]. As illustrated in Fig. 1(b), open-ended dialogue among agents is vulnerable to sociocognitive failure modes such as sycophancy, conformity bias, and premature consensus, where dominant opinions suppress minority yet valid information [Yao *et al.*, 2025].

*Corresponding author

In the absence of explicit decision structures, the potential benefits of agent diversity can be diminished or even negated. Consequently, simply increasing the number of interacting agents is insufficient to guarantee faithful and comprehensive summaries.

We argue that the core challenge is not the involvement of multiple agents, but how collective decisions are made. As shown in Fig. 1(c), we advocate a judicial-decision-inspired framework that introduces explicit deliberation and structured verdict formation. This perspective draws inspiration from the U.S. jury system, a well-established institution emphasizing independent judgment, adversarial evaluation, and verdict formation through structured critique and voting. Based on these principles, we propose **JURIS** (Joint Understanding and Reasoning via Iterative Selection), a multi-agent collaborative framework inspired by judicial decision-making, simulating a lawsuit scenario. In this framework, generator agents function like lawyers on opposing sides, offering candidate sentences from different perspectives as evidence. A group of decision-making agents, acting as the jury, evaluates the overall picture and selects sentences through defense and voting. Finally, a chief editor agent consolidates and polishes the selected sentences, akin to a judge issuing the final verdict and crafting the final summary.

To evaluate the effectiveness of the proposed framework, we conduct extensive experiments on standard in-domain news benchmarks and a challenging cross-domain dataset. Our contributions are threefold:

1. We propose JURIS, a decision-centric structured framework for organizing multi-agent collaboration, enabling systematic coordination of multiple agents to perform effective collective decision making;
2. We instantiate JURIS in the abstractive summarization setting by drawing inspiration from judicial jury deliberation and introducing sentence-level critique and selection to support balanced and reliable content aggregation;
3. Through comprehensive automatic, multi-dimensional, and human evaluations, we demonstrate that JURIS consistently outperforms strong single-agent and multi-agent baselines in terms of factual faithfulness, information coverage, and overall summary quality.

2 Related Work

LLM-based Summarization: Abstractive summarization has increasingly shifted toward in-context learning with large language models [Zhang *et al.*, 2024]. Recent prompt-engineering methods aim to improve information density and coverage, including recursive strategies such as Chain-of-Density [Adams *et al.*, 2023], as well as decomposition-based approaches that verify atomic claims through question answering [Sinha, 2025; Kamoi *et al.*, 2023]. However, a growing body of evidence indicates that single-agent summarization remains fragile, particularly under cross-domain conditions. Empirical analyses show that LLMs exhibit uneven attention allocation, systematically overlooking salient information located away from attention sinks [Liu *et al.*, 2024; Xiao *et al.*, 2023]. This limitation is further compounded by hallucination tendencies, where models rely on parametric knowledge rather than source-grounded evidence [Huang *et al.*, 2025]. In addition, alignment-focused fine-tuning often induces stylistic homogenization, resulting in summaries that are fluent but formulaic and lacking discourse variability [Wang *et al.*, 2024]. Collectively, these observations suggest that improving summarization reliability cannot be achieved solely through stronger prompting or isolated self-correction.

Complementing prior verification-oriented approaches, our work introduces a structured, decision-centric multi-agent formulation that emphasizes explicit selection and deliberation as part of the summarization process.

Multi-Agent Collaboration: Collaborative intelligence has emerged as an effective paradigm for addressing complex reasoning tasks. Frameworks such as AgentVerse [Chen *et al.*, 2024] and DyLAN [Liu *et al.*, 2023b] demonstrate that dynamic agent topologies can outperform monolithic models, while debate-based mechanisms encourage inter-agent critique to improve decision quality [Liang *et al.*, 2024; Chan *et al.*, 2023]. Despite these advances, recent studies increasingly report that unstructured collaboration is vulnerable to sociocognitive failure modes. Agents often exhibit sycophantic behavior, aligning with dominant or persuasive peers rather than factual evidence, which can lead to disagreement collapse and premature consensus [Sharma *et al.*, 2023; Yao *et al.*, 2025]. Such conformity effects are further amplified in homogeneous agent groups, giving rise to epistemic echo chambers in which correlated errors propagate systematically [Harman *et al.*, 2024; Cohen *et al.*, 2023]. In addition, loosely coupled dialogue-based interaction frequently incurs substantial inference latency without providing guarantees of convergence or stability [Tran *et al.*, 2025]. Together, these findings suggest that effective multi-agent collaboration requires explicit mechanisms to preserve independent judgment and constrain socially induced error reinforcement.

JURIS is also related to classical blackboard systems, in which multiple knowledge sources cooperatively contribute partial solutions to a shared problem-solving space [Penny, 1986]. While blackboard systems provide a general coordination paradigm for decomposed expertise and shared intermediate representations, JURIS adapts this idea to LLM-based summarization by replacing unconstrained shared-state updates with sentence-level candidate generation, jury-based critique and voting, and final editorial consolidation. This decision-centric design is tailored to improving factuality, coverage, and coherence in abstractive summarization.

In contrast to prior approaches centered on open-ended discussion, our work adopts a decision-centric view of multi-agent cooperation, emphasizing structured evaluation and aggregation over conversational consensus.

3 Methodology

3.1 Motivation

Judicial Jury Deliberation as Decision-Making Paradigm: Jury deliberation provides a well-established institutional framework for collective decision making in high-stakes settings. In criminal and civil trials, jurors are typically drawn at random from a broad population pool and are explicitly

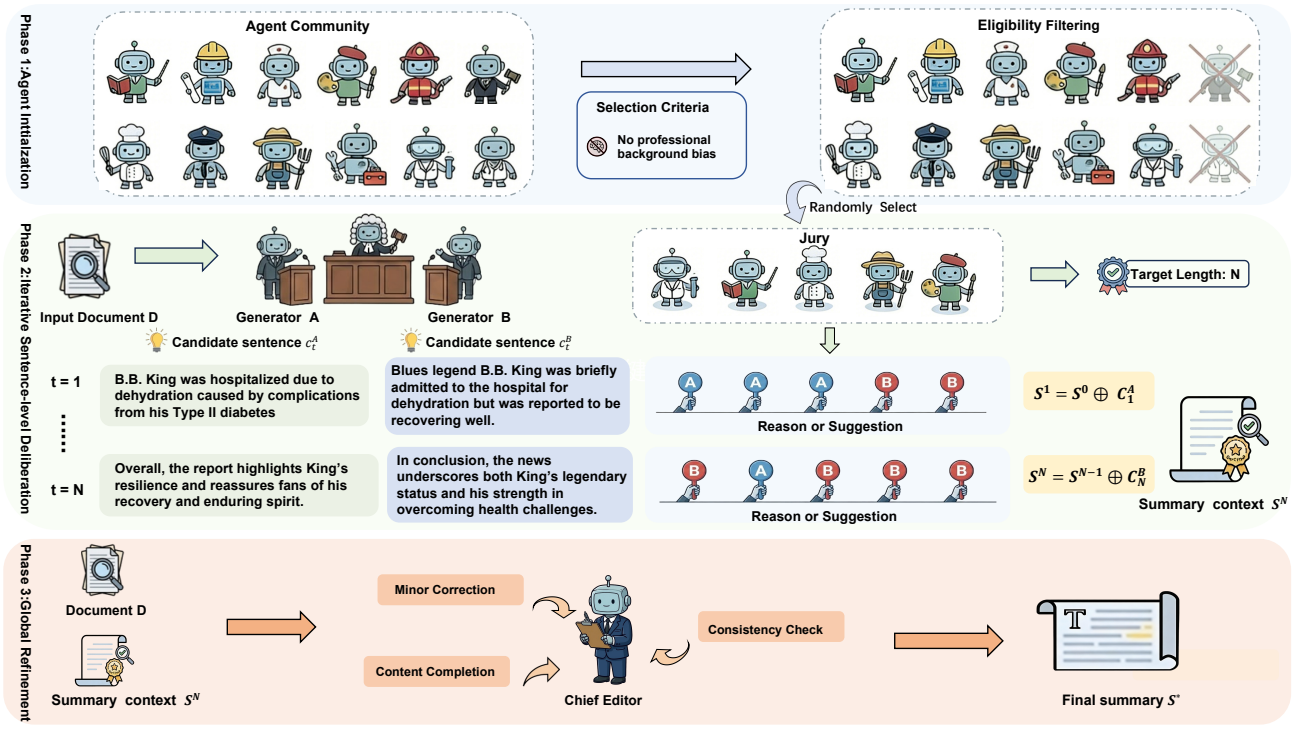

Figure 2: Overview of JURIS: heterogeneous jury planning, iterative sentence-level deliberation with generator agents providing candidates, and chief editor agent consolidation for coherent final summaries.

screened to exclude individuals with relevant professional expertise or conflicts of interest. This design is intended to prevent judgments from being dominated by specialized paradigms or authoritative viewpoints, and instead promotes independent assessment by multiple decision makers operating under the same evidentiary constraints.

Within this framework, jurors do not generate facts or legal arguments. Competing interpretations of the evidence are presented by opposing parties, and jurors are tasked with evaluating these alternatives based on credibility, consistency, and alignment with admissible evidence. Deliberation proceeds through reasoned discussion grounded in shared evidence, and the final verdict is produced through an explicit aggregation rule, most commonly majority voting. The resulting decision is therefore collective, auditable, and not reducible to informal conversational agreement.

An additional characteristic of jury deliberation is the separation between decision formation and decision articulation. While the verdict reflects the collective judgment of the jury, its final expression is often consolidated by a designated foreperson, whose role is to organize and communicate the outcome without altering the substance of the decision itself. This separation contributes to clarity and stability in the final outcome.

Implications for Multi-Agent LLM Collaboration: These characteristics motivate a decision-centric formulation of multi-agent collaboration with large language models. Different LLMs, due to variations in pre-training data, architecture, parameter scale, and alignment objectives, exhibit systematically different inductive biases. JURIS leverages this heterogeneity to construct a jury of agents whose diverse priors parallel the diversity of perspectives arising from jurors' distinct life experiences, while maintaining shared evaluative constraints.

JURIS uses two different language models as generator agents to independently produce candidate summary sentences for subsequent review and selection. Jury agents do not participate in generation; instead, they assess candidate sentences by grounding their judgments in the source document and applying a shared evaluation protocol. Candidate selection is performed through explicit voting, converting multiple assessments into a single decision at each step.

Because sentence-level decisions alone do not guarantee global discourse quality, JURIS further introduces a Chief Editor Agent to consolidate the selected summary sentences. In contrast to purely surface-level rewriting, this stage performs supplementary correction and revision over the chosen sentences, including necessary completion, clarification, and adjustment, while remaining consistent with the jury's decisions. Together, these components form a structured decision-making pipeline that operationalizes key principles of jury deliberation in the context of abstractive summarization.

Task Definition: Given a source document $\mathcal{D}$, the goal of abstractive summarization is to generate a summary $\mathcal{S} = \{s_1, s_2, \ldots, s_N\}$ consisting of N sentences. The objective is to maximize factual consistency with $\mathcal{D}$ and global discourse coherence, while reducing information omission and hallucination commonly observed in single-agent generation.

3.2 The JURIS Framework

JURIS casts document summarization as a structured collective decision-making process. Rather than relying on open-ended agent interaction, it explicitly decomposes summarization into three coordinated stages: (1) heterogeneous jury construction and planning, (2) iterative sentence-level deliberation with jury-based selection, and (3) global consolidation via a chief editor agent. Algorithm 1 summarizes the overall procedure.

Heterogeneous Jury Construction and Planning

To mitigate the inductive bias of any single language model, JURIS constructs a heterogeneous jury $\mathcal{J} = \{J_1, \ldots, J_K\}$ composed of agents instantiated from different model backbones, where K controls the jury size. The use of heterogeneous jurors introduces diverse priors arising from variations in pre-training data, architecture, and alignment objectives, enabling multi-perspective evaluation under shared evidentiary constraints. In the global jury assessment stage, the jury analyzes the source document $\mathcal{D}$ to estimate the target number of summary sentences N. Each juror independently proposes a candidate length based on document complexity and information density. The final target length is determined by aggregating these proposals using a majority-based rule, with optional truncation to remain within a predefined length range. This planning stage constrains the scope of subsequent generation and prevents uncontrolled summary expansion.

Iterative Sentence-Level Deliberation

Instead of generating the entire summary in a single pass, JURIS performs iterative sentence-level deliberation. At each step $t \in \{1, \ldots, N\}$, the framework executes a structured cycle of candidate generation, jury evaluation, and decision aggregation.

Candidate Generation: Two generator agents, G_A and G_B, instantiated from different language models, independently propose candidate sentences conditioned on the source document $\mathcal{D}$ and the previously accepted context $\mathcal{S}_{t-1}$:

$$c_t^X \leftarrow \text{Generate}(G_X, \mathcal{D}, \mathcal{S}_{t-1}), \quad X \in \{A, B\}. \tag{1}$$

Using distinct model backbones encourages diverse candidate proposals and reduces the likelihood that systematic biases shared by a single model dominate early content selection.

Jury Critique and Voting: Each juror $J_i \in \mathcal{J}$ evaluates the candidate sentences by producing a textual critique grounded in evidence from $\mathcal{D}$ and casting a discrete vote indicating preference. Jurors do not participate in generation and are restricted to assessment only, ensuring a clear separation between proposal and evaluation. This design preserves independent judgment and limits social influence among agents.

Majority-Based Selection: The accepted sentence s_t is selected via majority voting over juror preferences:

$$s_t = \arg \max_{c \in \{c_t^A, c_t^B\}} \sum_{J_i \in \mathcal{J}} \mathbb{I}(\text{Vote}(J_i) = c) \tag{2}$$

In the event of a tie, selection defaults to the candidate supported by stronger document grounding. The chosen sentence is appended to the summary context, forming $\mathcal{S}_t$, and propagated to the next deliberation step. This explicit decision mechanism filters unsupported or low-quality candidates before they can influence subsequent generation.

Algorithm 1 JURIS Iterative Deliberation

Input: Document $\mathcal{D}$, Jury $\mathcal{J}$, Generators G_A, G_B, Chief Editor E
Output: Final Summary $\mathcal{S}^*$

Initialize $\mathcal{S}_0 \leftarrow \emptyset$
Jury determines target length N based on $\mathcal{D}$
for $t = 1$ to N **do**
  $c_t^A \leftarrow G_A(\mathcal{D}, \mathcal{S}_{t-1})$
  $c_t^B \leftarrow G_B(\mathcal{D}, \mathcal{S}_{t-1})$
  for each juror $J_i \in \mathcal{J}$ **do**
    $R_i \leftarrow \text{Critique}(J_i, \mathcal{D}, c_t^A, c_t^B)$
    $v_i \leftarrow \text{Vote}(J_i, R_i)$
  end for
  $s_t \leftarrow \text{MajorityVote}(\{v_i\})$
  $\mathcal{S}_t \leftarrow \mathcal{S}_{t-1} \oplus s_t$
end for
$\mathcal{S}^* \leftarrow \text{GlobalRefinement}(E, \mathcal{S}_N, \mathcal{D})$
return $\mathcal{S}^*$

Global Coherence Optimization via Chief Editor Agent

While sentence-level deliberation ensures local factual reliability, the resulting sequence $\mathcal{S}_N$ may still exhibit fragmented discourse structure. To address this limitation, JURIS employs a dedicated Chief Editor Agent, denoted as E, to perform a final global refinement pass.

The chief editor operates exclusively on the jury-validated sentence sequence and is constrained to remain faithful to the source document $\mathcal{D}$. Its role is to perform supplementary correction and revision, including sentence completion, clarification, coreference resolution, and necessary local adjustments that improve coherence and readability. Crucially, this stage is not permitted to introduce unsupported facts or alter decisions previously endorsed through jury deliberation.

The output of this stage is the final summary $\mathcal{S}^*$, which integrates locally validated decisions into a coherent global narrative.

4 Experiments

We conduct extensive experiments to systematically evaluate the effectiveness, robustness, and generalization capability of the proposed JURIS framework. Our experimental design covers both in-domain and cross-domain summarization scenarios, and compares JURIS against strong single-agent and multi-agent baselines.

Specifically, we evaluate performance on news-domain benchmarks as well as a multi-domain summarization benchmark to assess generalization beyond news-style texts. We report results using standard automatic evaluation metrics, together with multi-dimensional quality assessments and human evaluation, in order to provide a comprehensive analysis of summarization quality. In addition, we perform ablation studies and parameter analyses to examine the contribution of individual components and design choices within JURIS.

4.1 Experimental Setup

Datasets:

CNN/DailyMail: We use the CNN/DailyMail dataset as the primary in-domain benchmark for news summarization [Hermann *et al.*, 2015]. To ensure controlled and reproducible evaluation, we randomly sample 2,000 document–summary pairs from the original corpus and partition them into two non-overlapping subsets of equal size, denoted as CNNDM1 and CNNDM2. This split enables robustness analysis across independent samples drawn from the same distribution and mitigates potential sampling bias.

MSumBench: To assess cross-domain generalization, we conduct experiments on the English subset of MSumBench [Min *et al.*, 2025]. This benchmark spans multiple domains, including news, medical texts, literature, reports, meetings, and interviews, and presents heterogeneous discourse structures beyond news-style summarization.

Baselines:

Single-agent LLMs: We evaluate two widely adopted open-source LLMs, **Mistral-7B** [Jiang *et al.*, 2023] and **Gemma3-12B-IT** [Team *et al.*, 2025], as single-agent baselines. These models represent complementary design choices in terms of parameter scale and architectural focus, and are commonly used in recent summarization studies. Both models are prompted in a zero-shot setting without external verification or refinement, reflecting the performance of standalone LLM-based summarization. Using the same backbone models in both single-agent and multi-agent settings enables controlled comparisons and isolates the impact of structured collaboration from model capacity.

Multi-agent Summarization Methods: We further compare JURIS with representative multi-agent summarization approaches that rely on iterative reasoning or question-driven decomposition. Specifically, we include **SummIt** [Zhang *et al.*, 2023], which performs iterative refinement through conversational interactions among agents, and **QA-Prompt** [Sinha, 2025], which decomposes summarization into a sequence of question-answering steps to improve coverage and factual consistency. These methods represent commonly adopted strategies for enhancing summarization quality via multi-agent or multi-step prompting, but lack an explicit jury-based verification mechanism.

Implementation Details:

We instantiate JURIS using a modular multi-agent architecture. All agents are instantiated from open-source language models and accessed through a unified inference interface.

Agent Configuration and Model Backbones. To promote model heterogeneity, we assign distinct model backbones to different agent roles:

- **Generators (G_A, G_B):** We employ a dual-stream generation strategy using **Mistral-7B** and **Gemma3-12B-IT**. This combination pairs a lightweight model with a higher-capacity model, encouraging complementary generation behaviors and diverse content proposals.
- **Jury Panel ($\mathcal{J}$):** The jury consists of a fixed set of five heterogeneous models: **Qwen2.5-7B** [Team, 2024], **Llama-3.1-8B** [Dubey *et al.*, 2024], **DeepSeek-R1-Distill-Llama-8B** [Guo *et al.*, 2025], **GLM-4-9B** [GLM *et al.*, 2024], and **Gemma3-12B-IT**. This selection spans a range of parameter scales (7B–12B) and pre-training corpora, reducing correlated errors and mitigating the risk of shared bias during deliberation. The jury size is fixed to $K = 5$, which is determined based on an empirical trade-off between performance and computational cost; as shown in the ablation study in Section 4.3.
- **Refiner:** The final global refinement stage is performed by **Mistral-7B**, acting as a Chief Editor Agent. We choose Mistral-7B as the Chief Editor Agent since it is already employed as one of the generator backbones and exhibits stable instruction-following behavior. This agent focuses on improving discourse coherence and resolving coreference issues in the jury-assembled summary sequence, while preserving all content previously validated through jury deliberation.

Evaluation Metrics:

To comprehensively assess summarization quality, we adopt a multi-level evaluation protocol combining automatic metrics, fine-grained quality assessment, and human evaluation.

Automatic Metrics: We report standard automatic metrics for summarization evaluation, including **ROUGE-1/2/L** to measure lexical overlap with reference summaries, **BLEU**, reported for completeness as a precision-oriented lexical metric, and **BERTScore** to capture semantic similarity based on contextual embeddings.

Multi-dimensional Evaluation: To go beyond surface-level overlap metrics, we employ UniSumEval [Lee *et al.*, 2024], a fine-grained evaluation framework for summarization. Following the G-Eval protocol [Liu *et al.*, 2023a], a GPT-4-based evaluator scores generated summaries along three core dimensions: **Faithfulness**, **Completeness**, and **Conciseness**.

Human Evaluation: We further conduct a ranking-based human evaluation protocol (denoted as **P-Eval**), where annotators with NLP backgrounds compare summaries in terms of overall quality, readability, and factual plausibility.

4.2 Main Results

Performance on Traditional Automatic Metrics

The left portion of Table 1 reports results on conventional automatic metrics, including ROUGE, BLEU, and BERTScore. Across all datasets, JURIS achieves the highest or near-highest ROUGE-1, ROUGE-2, and ROUGE-L scores among all compared methods. This indicates that JURIS effectively captures salient content and produces summaries that align well with reference texts at the lexical level.

JURIS also maintains competitive BLEU and BERTScore results, suggesting that its improvements extend beyond surface-level n-gram overlap to semantic similarity. Nevertheless, it is well recognized that overlap-based metrics alone are insufficient to fully assess factual reliability and information coverage in abstractive summarization. High ROUGE or BLEU scores do not necessarily imply faithful content selection, nor do they reveal whether critical information has been omitted or distorted. These limitations are particularly

Dataset	Model	Lexical Overlap (ROUGE)			Precision	Semantic	Quality-oriented Evaluation			
		R-1	R-2	R-L	BLEU	BERTScore	Faith.	Comp.	Conc.	P-Eval
CNNDM1	Mistral-7B	0.3295	0.1196	0.2013	0.0631	0.8655	95.95	46.87	94.00	87.3
	Gemma3-12B	0.3495	0.1162	0.2115	0.0663	0.8655	88.27	44.97	87.76	82.6
	SummIt	0.3211	0.1129	0.1970	0.0612	0.8625	93.51	16.43	60.91	85.5
	QA-Prompt	0.3397	0.1036	0.2108	0.0702	0.8677	97.79	15.15	61.67	87.9
	JURIS (Ours)	**0.3575**	**0.1202**	**0.2195**	**0.0756**	**0.8685**	**99.63**	**50.25**	**95.96**	**90.2**
CNNDM2	Mistral-7B	0.3222	**0.1197**	0.1988	0.0623	0.8654	98.34	46.85	92.39	84.6
	Gemma3-12B	0.3475	0.1134	0.2085	0.0647	0.8650	88.42	48.47	80.06	81.4
	SummIt	0.3190	0.1108	0.1948	0.0605	0.8622	94.01	13.22	38.65	82.3
	QA-Prompt	0.3367	0.1002	0.2088	0.0684	0.8672	97.69	15.06	60.77	87.7
	JURIS (Ours)	**0.3527**	0.1176	**0.2174**	**0.0745**	**0.8679**	**99.05**	**51.41**	**95.82**	**88.3**
MSUM	Mistral-7B	0.3069	0.1186	0.1832	0.0593	0.8506	97.30	14.63	47.45	83.4
	Gemma3-12B	0.2967	0.0966	0.1712	0.0480	0.8463	95.34	19.18	60.17	81.7
	SummIt	0.3090	0.1200	0.1843	0.0573	0.8515	92.20	19.09	78.82	85.3
	QA-Prompt	0.3055	0.1203	0.1841	0.0576	0.8528	97.12	12.91	51.91	88.4
	JURIS (Ours)	**0.3187**	**0.1257**	**0.1941**	**0.0636**	**0.8550**	**98.17**	**19.94**	**79.00**	**89.1**

Table 1: Main results on in-domain (CNNDM1, CNNDM2) and cross-domain (MSUM) benchmarks. We report ROUGE-1/2/L, BLEU (n-gram overlap), and BERTScore. Quality-oriented evaluation includes UniSumEval (Faithfulness, Completeness, Conciseness; 0–100 scale) and human preference (P-Eval).

pronounced in multi-agent generation settings, motivating the use of more fine-grained evaluation criteria.

Quality-oriented Evaluation

The right portion of Table 1 reports quality-oriented evaluation results, including Faithfulness, Completeness, Conciseness, and human preference scores. Across all datasets, JURIS consistently outperforms both single-agent baselines and existing multi-agent methods on all four dimensions.

Substantial gains are observed in Faithfulness and Completeness, indicating that JURIS more effectively suppresses hallucinations and reduces information omission during content selection. At the same time, strong Conciseness scores demonstrate that improved coverage is achieved without introducing excessive redundancy. The advantages of JURIS are further amplified on the cross-domain MSUM benchmark, highlighting its robustness under domain shifts.

Human evaluation results exhibit trends consistent with quality-oriented automatic metrics, suggesting that the improvements achieved by JURIS are also perceptible to human readers. Taken together, these results provide strong evidence that decision-centric, jury-guided deliberation enables more reliable and stable multi-agent summarization across domains.

4.3 Ablation Study

Impact of the Chief Editor

To evaluate the contribution of the **Chief Editor**, we remove this module from JURIS while keeping all other components unchanged. As illustrated in Figure 3, the variant equipped with the Chief Editor consistently outperforms its counterpart without this module across datasets and evaluation metrics. On the in-domain CNNDM benchmarks, the gains are stable and uniform, indicating that global refinement improves the integration of sentence-level decisions without altering their factual content. On the cross-domain MSUM dataset, the improvements are substantially larger, reflecting the increased difficulty of maintaining discourse coherence and content consistency under heterogeneous domain distributions.

Metric	In-Domain (News)		Cross-Domain (MSUM)	
	Fixed	Dyn.	Fixed	Dyn.
ROUGE-1	0.3575	**0.3684**	**0.3187**	0.3168
ROUGE-2	0.1202	**0.1233**	**0.1257**	0.1145
ROUGE-L	0.2195	**0.2333**	0.1941	**0.1972**
BLEU	0.0756	**0.0840**	0.0636	**0.0638**
BertScore	0.8685	**0.8720**	0.8550	**0.8571**

Table 2: Performance comparison between Fixed Jury and Dynamic Jury. While Dynamic Jury favors semantic diversity on in-domain news, Fixed Jury achieves superior factual stability (higher R-1/R-2) on the complex MSUM dataset.

These results indicate that the Chief Editor contributes beyond surface-level fluency enhancement. In addition to improving global coherence, it performs corrective and integrative refinement over previously validated content. Specifically, the Chief Editor resolves local inconsistencies, clarifies referential ambiguity, and supplements missing but already jury-approved information during global consolidation. By operating strictly on sentence-level outputs that have passed jury deliberation, this module enhances completeness and correctness without introducing unverified facts. This mechanism mitigates discourse fragmentation and residual inconsistencies that arise from directly concatenating independently validated content, highlighting the necessity of global consolidation in decision-centric summarization.

Effect of Jury Composition on Deliberation Stability

We further examine the effect of jury composition by comparing the proposed Fixed Jury strategy with a Dynamic Jury variant, in which jurors are randomly resampled at each deliberation step. As shown in Table 2, different jury composition strategies exhibit a clear trade-off between semantic diversity

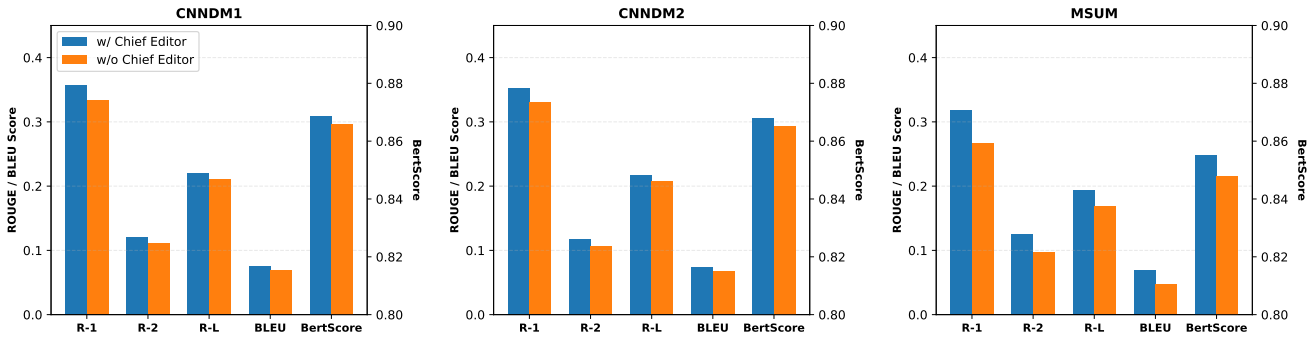

Figure 3: Ablation results across three datasets. The Left Axis denotes ROUGE/BLEU scores, while the Right Axis denotes BERTScore. The Chief Editor consistently improves performance, with more pronounced gains on the heterogeneous MSUM dataset.

and decision stability. On the in-domain news benchmark, the Dynamic Jury variant achieves modest gains on semantic similarity metrics, suggesting that frequent resampling of jurors can introduce additional evaluative perspectives when document structure and information distribution are relatively homogeneous. However, on the more challenging cross-domain MSUM dataset, Dynamic Jury leads to consistent degradation on precision-oriented metrics. This performance drop indicates that, in complex summarization scenarios, frequent changes in the evaluation panel disrupt fine-grained factual continuity across generated sentences. As jurors are replaced at each deliberation step, evaluation criteria may vary implicitly, weakening consistency in content selection and verification. In contrast, the Fixed Jury strategy provides a stable evaluation environment that preserves coherent decision standards across deliberation rounds. By maintaining a consistent set of evaluators, the framework better supports fact-level alignment and cross-sentence consistency, which are crucial in heterogeneous and information-dense documents.

These findings highlight an inherent trade-off between diversity and stability in multi-agent evaluation. While dynamic resampling may enhance semantic diversity under simpler conditions, a fixed jury configuration offers more robust and reliable decision making in complex summarization settings.

Parameter Analysis: Jury Size

We analyze the effect of jury size by progressively increasing the number of jurors to identify the optimal trade-off between performance and efficiency. This analysis is conducted on the CNNDM1 subset. Two aggregated indicators are considered: (1) The Combined Quality Score, defined as the sum of Faithfulness, Completeness, and Conciseness; and (2) Cost-Effectiveness, measured as the quality score divided by the total inference time. Since the computational overhead scales linearly with the number of jurors, this metric serves as a proxy for marginal utility, quantifying whether the incremental quality gains justify the additional latency.

Figure 4 illustrates the impact of jury size on both summary quality and computational efficiency. As the number of jurors increases, the Combined Quality Score consistently improves, indicating that larger juries strengthen collective decision-making and reduce individual model bias. However, the Cost-Effectiveness curve shows a clear diminishing-returns

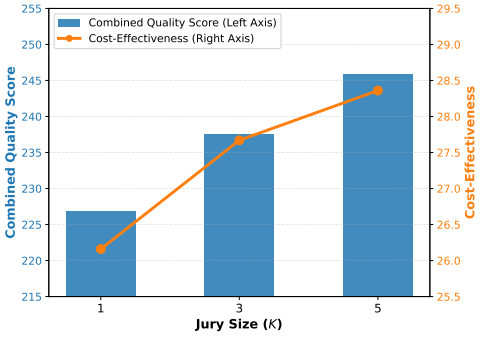

Figure 4: Trade-off analysis of jury size. The figure illustrates the relationship between summary quality and computational efficiency as the number of jurors increases.

pattern. While small increases in jury size bring noticeable quality gains, further expansion leads to much higher computational cost with only marginal additional improvement. Based on this trade-off, we set the jury size to $K = 5$, which provides a balanced compromise between performance and efficiency.

5 Conclusion

We proposed JURIS, a judicial-decision-inspired multi-agent framework that formulates abstractive document summarization as a structured collective decision-making process. By decomposing summarization into adversarial candidate generation, sentence-level deliberation, and global refinement, JURIS addresses key limitations of existing single-agent and unstructured multi-agent approaches, particularly in terms of factual consistency, content coverage, and discourse coherence. Extensive experiments on both in-domain and cross-domain benchmarks show that JURIS consistently outperforms strong baselines under automatic metrics, fine-grained quality evaluations, and human judgments, validating the effectiveness of explicit deliberation and role specialization in improving summarization reliability. More broadly, our work reframes multi-agent summarization from heuristic collaboration to a principled, decision-centric paradigm, where generation is guided by explicit evidence assessment and structured consensus building.

Ethical Statement

The authors declare that this study adheres to ethical standards. All data used for quantitative and qualitative analysis were obtained from publicly available online sources. The models used are based on authors' open-source code and are appropriately cited. Additionally, Google Gemini and Google Translate were employed solely for language polishing and translation to enhance the manuscript's clarity.

Acknowledgements

This work was supported in part by the National Key Research and Development Program of China (2022YFC3600902), the Key Research and Development Program of Zhejiang Province (2025C01129), and the National Science and Technology Major Project (2023ZD0509706).

References

[Adams *et al.*, 2023] Griffin Adams, Alex Fabbri, Faisal Ladhak, Eric Lehman, and Noémie Elhadad. From sparse to dense: Gpt-4 summarization with chain of density prompting. In *Proceedings of the 4th New Frontiers in Summarization Workshop*, pages 68–74, 2023.

[Belém *et al.*, 2025] Catarina G Belém, Pouya Pezeshkpour, Hayate Iso, Seiji Maekawa, Nikita Bhutani, and Estevam Hruschka. From single to multi: How llms hallucinate in multi-document summarization. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 5276–5309, 2025.

[Chan *et al.*, 2023] Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. Chateval: Towards better llm-based evaluators through multi-agent debate. *arXiv preprint arXiv:2308.07201*, 2023.

[Chen *et al.*, 2024] Weize Chen, Yusheng Su, Jingwei Zuo, Cheng Yang, Chenfei Yuan, Chi-Min Chan, Heyang Yu, Yaxi Lu, Yi-Hsin Hung, Chen Qian, et al. Agentverse: Facilitating multi-agent collaboration and exploring emergent behaviors. In *ICLR*, 2024.

[Cohen *et al.*, 2023] Roi Cohen, May Hamri, Mor Geva, and Amir Globerson. Lm vs lm: Detecting factual errors via cross examination. *arXiv preprint arXiv:2305.13281*, 2023.

[Dubey *et al.*, 2024] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv e-prints*, pages arXiv–2407, 2024.

[Feng *et al.*, 2024] Huawen Feng, Yan Fan, Xiong Liu, Ting-En Lin, Zekun Yao, Yuchuan Wu, Fei Huang, Yongbin Li, and Qianli Ma. Improving factual consistency of news summarization by contrastive preference optimization. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 11084–11100, 2024.

[GLM *et al.*, 2024] Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Dan Zhang, Diego Rojas, Guanyu Feng, Hanlin Zhao, et al. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *arXiv preprint arXiv:2406.12793*, 2024.

[Guo *et al.*, 2025] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

[Harman *et al.*, 2024] Joel Harman, Alessandro Soro, and Selen Türkay. Unmasking metadata bias: Evaluating the impact on large language model text analysis. In *Proceedings of the 36th Australasian Conference on Human-Computer Interaction*, pages 525–536, 2024.

[Hermann *et al.*, 2015] Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. Teaching machines to read and comprehend. *Advances in neural information processing systems*, 28, 2015.

[Huang *et al.*, 2025] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55, 2025.

[Ji *et al.*, 2023] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM computing surveys*, 55(12):1–38, 2023.

[Jiang *et al.*, 2023] Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b, 2023.

[Kamoi *et al.*, 2023] Ryo Kamoi, Tanya Goyal, Juan Diego Rodriguez, and Greg Durrett. Wice: Real-world entailment for claims in wikipedia. *arXiv preprint arXiv:2303.01432*, 2023.

[Lee *et al.*, 2024] Yuho Lee, Taewon Yun, Jason Cai, Hang Su, and Hwanjun Song. Unisumeval: Towards unified, fine-grained, multi-dimensional summarization evaluation for llms. *arXiv preprint arXiv:2409.19898*, 2024.

[Liang *et al.*, 2024] Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Shuming Shi, and Zhaopeng Tu. Encouraging divergent thinking in large language models through multi-agent debate. In *Proceedings of the 2024 conference on empirical methods in natural language processing*, pages 17889–17904, 2024.

[Liu *et al.*, 2023a] Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-eval: Nlg evaluation using gpt-4 with better human alignment. *arXiv preprint arXiv:2303.16634*, 2023.

[Liu *et al.*, 2023b] Zijun Liu, Yanzhe Zhang, Peng Li, Yang Liu, and Diyi Yang. Dynamic llm-agent network: An

llm-agent collaboration framework with agent team optimization. *arXiv preprint arXiv:2310.02170*, 2023.

[Liu *et al.*, 2024] Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2024.

[Min *et al.*, 2025] Hyangsuk Min, Yuho Lee, Minjeong Ban, Jiaqi Deng, Nicole Hee-Yeon Kim, Taewon Yun, Hang Su, Jason Cai, and Hwanjun Song. Towards multi-dimensional evaluation of llm summarization across domains and languages. *arXiv preprint arXiv:2506.00549*, 2025.

[Penny, 1986] Nii H Penny. Blackboard systems: The blackboard model of problem solving and the evolution of blackboard architectures. *The AI Magazine*, 1986.

[Ramprasad *et al.*, 2024] Sanjana Ramprasad, Elisa Ferracane, and Zachary C Lipton. Analyzing llm behavior in dialogue summarization: Unveiling circumstantial hallucination trends. *arXiv preprint arXiv:2406.03487*, 2024.

[Sharma *et al.*, 2023] Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R Johnston, et al. Towards understanding sycophancy in language models. *arXiv preprint arXiv:2310.13548*, 2023.

[Sinha, 2025] Neelabh Sinha. Qa-prompting: Improving summarization with large language models using question-answering. *arXiv preprint arXiv:2505.14347*, 2025.

[Team *et al.*, 2025] Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.

[Team, 2024] Qwen Team. Qwen2.5: A party of foundation models, September 2024.

[Tran *et al.*, 2025] Khanh-Tung Tran, Dung Dao, Minh-Duong Nguyen, Quoc-Viet Pham, Barry O'Sullivan, and Hoang D Nguyen. Multi-agent collaboration mechanisms: A survey of llms. *arXiv preprint arXiv:2501.06322*, 2025.

[Wan *et al.*, 2025] David Wan, Justin Chen, Elias Stengel-Eskin, and Mohit Bansal. Mamm-refine: A recipe for improving faithfulness in generation with multi-agent collaboration. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 9882–9901, 2025.

[Wang *et al.*, 2024] Xinpeng Wang, Bolei Ma, Chengzhi Hu, Leon Weber-Genzel, Paul Röttger, Frauke Kreuter, Dirk Hovy, and Barbara Plank. My Answer is C: First-token probabilities do not match text answers in instruction-tuned language models. *arXiv preprint arXiv:2402.14499*, 2024.

[Wang *et al.*, 2025] Weixuan Wang, Minghao Wu, Barry Haddow, and Alexandra Birch. Learning to summarize by learning to quiz: Adversarial agentic collaboration for long document summarization. *arXiv preprint arXiv:2509.20900*, 2025.

[Xiao *et al.*, 2023] Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. *arXiv preprint arXiv:2309.17453*, 2023.

[Yao *et al.*, 2025] Binwei Yao, Chao Shang, Wanyu Du, Jianfeng He, Ruixue Lian, Yi Zhang, Hang Su, Sandesh Swamy, and Yanjun Qi. Peacemaker or troublemaker: How sycophancy shapes multi-agent debate. *arXiv preprint arXiv:2509.23055*, 2025.

[Zhang *et al.*, 2023] Haopeng Zhang, Xiao Liu, and Jiawei Zhang. Summit: Iterative text summarization via chatgpt. *arXiv preprint arXiv:2305.14835*, 2023.

[Zhang *et al.*, 2024] Tianyi Zhang, Faisal Ladhak, Esin Durmus, Percy Liang, Kathleen McKeown, and Tatsunori B Hashimoto. Benchmarking large language models for news summarization. *Transactions of the Association for Computational Linguistics*, 12:39–57, 2024.