MAJI: A Multi-Agent Workflow for Augmenting Journalistic Interviews[*]

Kaiwen Guo[†]
Cornell University
Ithaca, New York
kg597@cornell.edu

Yimeng Wu
Sohu.com
Beijing, China
yimengwu@sohu-inc.com

Abstract

Journalistic interviews are creative, dynamic processes where success hinges on insightful, real-time questioning. While Large Language Models (LLMs) can assist, their tendency to generate coherent but uninspired questions optimizes for probable, not insightful, continuations. This paper investigates whether a structured, multi-agent approach can overcome this limitation to act as a more effective creative partner for journalists. We introduce MAJI, a system designed for this purpose, which employs a divergent-convergent architecture: a committee of specialized agents generates a diverse set of questions, and a convergent agent selects the optimal one. We evaluated MAJI against a suite of strong LLM baselines. Our results demonstrate that our multi-agent framework produces questions that are more coherent, elaborate, and original (+36.9% for our best model vs. a standard LLM baseline), exceeded strong LLM baselines on key measures of creative question quality. Most critically, in a blind survey, professional journalists preferred MAJI's selected questions over those from the baseline by a margin of more than two to one. We present the system's evolution, highlighting the architectural trade-offs that enable MAJI to augment, rather than simply automate, journalistic inquiry.

1 Introduction

The practice of journalism is a cornerstone of an informed society, with the interview serving as a primary tool for information gathering and narrative construction. While interviews are often prepared with a structured outline, the most compelling insights emerge from unscripted moments. A journalist's ability to react to new information and identify novel angles in real-time separates a standard interview from a revelatory one. This dynamic process, however, presents a significant cognitive load.

Recent advancements in Large Language Models (LLMs) have opened new avenues for assisting in complex, language-based tasks (Touvron et al., 2023; OpenAI et al., 2024). A straightforward approach might involve prompting an LLM with the conversation history and asking for the next question. However, this often yields generic or predictable questions (Gordin et al., 2023), as LLMs tend to optimize for the most probable continuation rather than the most insightful or creative one. Recognizing this limitation, recent research has proposed a range of methods to enhance the originality and depth of LLM-generated interview questions. For example, works such as Spangher et al. (2025), Lin et al. (2025b), and Tian et al. (2024) introduce agentic workflows and creative reasoning strategies to move beyond surface-level responses. Our work builds on this line of research, further exploring how a multi-agent, divergent-convergent architecture can systematically augment the creative process in journalistic interviews.

Formally, this task can be seen as a conditional language generation problem. A standard LLM approach maximizes the likelihood of the next question Q_{t+1} given the transcript history T_t:

$$Q_{t+1} = \arg\max_{Q} P(Q|T_t)$$

This formulation often leads to probable but uninspired responses. We propose reframing the problem as maximizing a utility function $U(Q)$ that captures the goals of journalistic inquiry:

$$Q_{t+1} = \arg\max_{Q} U(Q|T_t, O, P)$$

where utility depends on the question's insight and relevance to the interview's strategic `Outline` O and the interviewee's `Persona` P. MAJI is pro-

[*] The code for this project is available at: `https://github.com/sohu-mint-lab/MAJI-IJCNLP-AACL-2025/`

[†] Work done during an internship at Sohu.com.

Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 1061–1083
December 20-24, 2025 ©2025 Association for Computational Linguistics

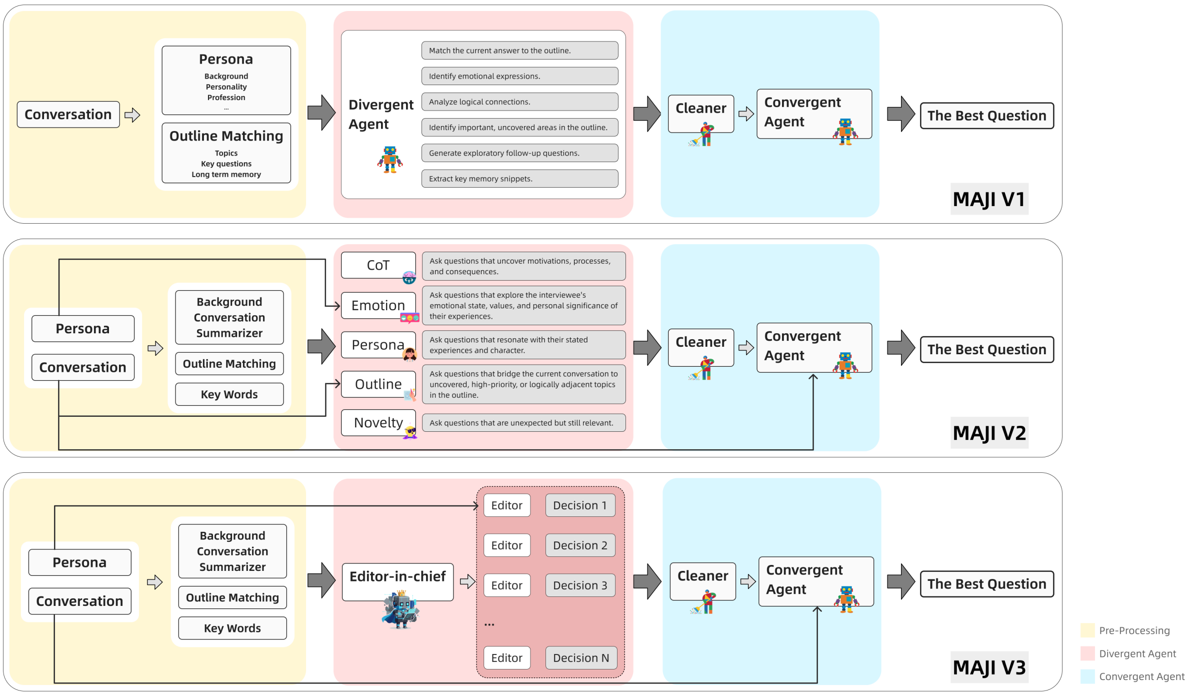
Conversation
Persona
Background
Personality
Profession
...
Outline Matching
Topics
Key questions
Long term memory
Divergent Agent
Match the current answer to the outline.
Identify emotional expressions.
Analyze logical connections.
Identify important, uncovered areas in the outline.
Generate exploratory follow-up questions.
Extract key memory snippets.
Cleaner
Convergent Agent
The Best Question
MAJI V1
Persona
Conversation
Background Conversation Summarizer
Outline Matching
Key Words
CoT
Ask questions that uncover motivations, processes, and consequences.
Emotion
Ask questions that explore the interviewee's emotional state, values, and personal significance of their experiences.
Persona
Ask questions that resonate with their stated experiences and character.
Outline
Ask questions that bridge the current conversation to uncovered, high-priority, or logically adjacent topics in the outline.
Novelty
Ask questions that are unexpected but still relevant.
Cleaner
Convergent Agent
The Best Question
MAJI V2
Persona
Conversation
Background Conversation Summarizer
Outline Matching
Key Words
Editor-in-chief
Editor
Decision 1
Editor
Decision 2
Editor
Decision 3
...
Editor
Decision N
Cleaner
Convergent Agent
The Best Question
MAJI V3
Pre-Processing
Divergent Agent
Convergent Agent

Figure 1: The architectural evolution of MAJI across its three versions. V1 (left) established a simple two-agent divergent-convergent model. V2 (center) introduced a specialized committee of agents for greater diversity. V3 (right) explored dynamic agent generation for adaptive strategies.

posed to address this more complex optimization problem.

To address this gap, we introduce MAJI (Multi-Agent Workflow for Journalism Interview), a system designed not to replace the journalist, but to augment their creative process. The human-in-the-loop workflow, depicted in Figure 2, positions MAJI as an assistant that provides suggestions while the journalist retains full control. MAJI is built on the psychological principle of divergent-convergent thinking, a cornerstone of creative problem-solving (Guilford, 1950). Our hypothesis is that by decomposing question generation into specialized sub-tasks, a multi-agent system (MAS) can produce more diverse and insightful questions than a single model. The use of MAS has been shown for complex logical tasks (Wooldridge, 2009; Wang et al., 2023; Wu et al., 2023; Qian et al., 2023; Li et al., 2023a), and we apply this paradigm to a creative domain (Lin et al., 2025a; Xi et al., 2025; Zhou et al., 2023; Li et al., 2023b).

This paper details the design and evolution of the MAJI framework across three major versions, from a simple proof-of-concept to a more complex and dynamic system capable of dynamically generating its own specialized agents. We conducted a rigorous evaluation using a simulated interview environment inspired by prior work in computational journalism (Diakopoulos, 2019; Lu et al., 2024). Our contributions are threefold:

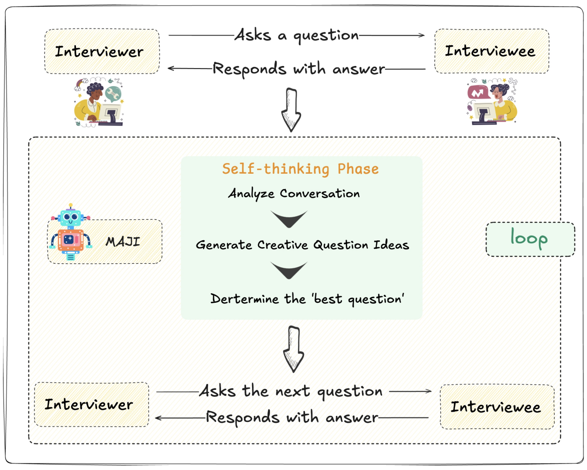

Figure 2: High-level overview of the MAJI-assisted interview workflow. MAJI operates in a continuous loop, observing the conversation and providing question suggestions to the journalist, who retains final control over the direction of the interview.

1. **A Novel Multi-Agent Framework:** We proposed and implemented a divergent-convergent multi-agent architecture for the

creative task of interview question generation.

2. **Empirical Evaluation:** We conducted a multi-metric comparison of MAJI against strong LLM baselines, using quantitative, qualitative, and comparative metrics assessed by both LLM-as-judge and professional journalists.

3. **Architectural Insights:** Through a controlled architectural ablation, we empirically demonstrate the superiority of a specialized multi-agent committee for the divergent phase of question generation. We further document MAJI's evolution to provide insights into the design trade-offs between fixed-specialist and dynamically-generated agent committees.

Our findings show that MAJI consistently generates questions that are more insightful, original, and contextually relevant than those from a well-prompted, powerful LLM. This work demonstrates the potential of structured multi-agent systems to move beyond simple automation and act as powerful creative partners in complex professional domains like journalism.

2 Related Work

Prior systems concentrate on data mining, fact checking and bias detection (Cohen et al., 2011; Diakopoulos, 2019; Hamborg et al., 2018), leaving the *real-time* interviewing stage under-explored. Datasets such as *NewsInterview* (Lu et al., 2024) highlight this gap; MAJI directly addresses it by operating live. Techniques like Chain-of-Thought (Wei et al., 2022), Tree-of-Thought (Yao et al., 2023) and retrieval-augmented generation improve single-model reasoning but still rely on one monolithic agent. Multi-agent frameworks such as AutoGen (Wu et al., 2023) and CAMEL (Li et al., 2023a) show that dividing labour can yield stronger reasoning; MAJI adapts this insight to creative follow-up generation. Psychology links creativity to alternating idea generation and selection (Guilford, 1950). Agent committees have applied this pattern to storytelling and design (Yao et al., 2019; Li et al., 2023b). MAJI is the first to embed the paradigm in journalistic interviewing and to validate its impact with professional users.

3 The MAJI Framework

MAJI is designed to mirror and support the cognitive workflow of a journalist. The framework is built on several core concepts: foundational inputs that set the stage, a multi-agent system that drives the question generation process, and an iterative loop that allows the system to learn and adapt as the interview progresses.

3.1 Foundational Concepts

The interview process begins with the journalist framing the initial context. This human-in-the-loop step is crucial for grounding the AI's subsequent contributions. Three key pieces of information establish this foundation:

- **Persona:** A detailed profile of the interviewee, including their background, personality, profession, and the primary purpose of the interview. This is crafted by the journalist to guide the system's interaction style. For example, a persona for a professional mermaid performer might highlight their artistic motivations and physical challenges, shaping questions toward these themes.

- **Outline:** A structured list of topics and key questions that the journalist intends to cover. This serves as the strategic backbone of the interview, ensuring coverage of critical areas while allowing flexibility for emergent insights.

- **Dynamic Background Summary:** To track evolving context, MAJI uses a dedicated agent to maintain a `BackgroundSummary`. It has a `long_term_summary` for foundational information (e.g., interviewee's career trajectory) and a `short_term_summary` for the last five conversational turns. This provides all agents with a continuously updated, concise view of the conversation history.

These inputs ensure MAJI remains grounded in the journalist's objectives and the interview's evolving context, enabling questions that are both strategically aligned and responsive to in-session dynamics.

3.2 The Divergent-Convergent Workflow

The core of MAJI is an implementation of the divergent-convergent thinking model, executed by specialized agents in a multi-stage workflow. This process, detailed for our primary MAJI V2 system in Algorithm 1, ensures a balance between creative exploration and strategic focus. MAJI's architecture operationalizes this principle. A committee of specialized *divergent agents* brainstorms potential questions, each focusing on a distinct creative vector, such as emotional depth, causal reasoning, adherence to the interviewee's persona, or pure novelty. Their suggestions are then processed by an *Editor Agent* to refine and deduplicate the pool of ideas. Finally, a *convergent agent*, acting as an Editor-in-Chief, selects the single best question that aligns with the conversation's flow and the journalist's strategic goals for the interview. The key stages are:

Algorithm 1 MAJI V2 Question Generation Workflow

```
Input: Outline $O$, Persona $P$, Transcript $T$
Output: Next question $Q$
$S \leftarrow$ BackgroundAgent$(T, P)$    ▷ Update summaries
$K \leftarrow$ KeywordsAgent$(T, S)$    ▷ Extract keywords
$M \leftarrow$ OutlineMatcherAgent$(O, T)$    ▷ Map to outline
$C \leftarrow \emptyset$    ▷ Initialize candidate pool
for each DivergentAgent$_i$ in $\{\mathrm{D}_1, \ldots, \mathrm{D}_n\}$ do
    $C_i \leftarrow$ DivergentAgent$_i(K, S, M, P)$    ▷ Propose questions
    $C \leftarrow C \cup C_i$
end for
$C' \leftarrow$ EditorAgent$(C)$    ▷ Refine candidates
$Q \leftarrow$ ConvergentAgent$(C', T, P, O)$    ▷ Select optimal question
return $Q$
```

1. **Context Analysis (Pre-Divergence):** Before any new questions are brainstormed, a set of pre-processing agents analyzes the latest turn in the conversation to establish a shared understanding of the current state. This includes the `BackgroundAgent` updating the long- and short-term summaries, the `KeywordsAgent` extracting the most salient terms from the latest response, and the `OutlineMatcherAgent` assessing which parts of the interview outline have been covered. In MAJI V2, the `KeywordsAgent` and `OutlineMatcherAgent` play crucial bridging roles between context understanding and creative generation. The `KeywordsAgent` extracts topical signals from each new conversational turn, ensuring that subsequent questions remain grounded in the immediate dialogue while maintaining thematic continuity across the session. Meanwhile, the `OutlineMatcherAgent` enforces structural discipline by aligning the conversation with the predefined outline—identifying which planned topics have been addressed and highlighting those still pending. Together, these agents provide the divergent specialists with a contextual and structural map, allowing their creative exploration to remain both relevant and strategically focused.

2. **Divergent Thinking:** With the updated context, the committee of specialized divergent agents generates a wide range of potential follow-up questions in parallel. This parallel, specialized approach is designed to produce a candidate pool with high "Flexibility," ensuring a rich set of creative options.

3. **Editing & Curation:** The raw list of questions from the divergent phase is often redundant. The `EditorAgent` uses sentence-transformer embeddings to identify and merge semantically similar questions (similarity threshold: 0.85). This step curates a clean, concise list of unique candidate questions.

4. **Convergent Selection:** Finally, the `ConvergentAgent` takes the curated list of questions and, guided by the journalist's stated strategic preference, selects the single best question to ask next. This selection process is not based on arbitrary heuristics, but on an implicit model of journalistic utility, as detailed below.

3.2.1 Convergent Utility Maximization

The core of the convergent step is the maximization of a utility function U. Rather than being a simple, hard-coded formula, this function is a conceptual model of question quality that the `ConvergentAgent` is prompted to approximate. We can formally define this utility for a candidate question $Q \in C'$ as a weighted sum of scores from various quality dimensions:

$$U(Q|\cdot, S_p) = \sum_{i=1}^{N} w_i(S_p) \cdot \phi_i(Q|T_t, O, P)$$

where each component represents a desirable attribute of a question:

- $\phi_i(Q|\cdot)$ are scoring functions that evaluate different facets of a question's quality, such as its *coherence*, *emotional depth*, *outline progression*, *persona alignment*, and *novelty*. These facets directly correspond to the specializations of the divergent agents.

- $w_i(S_p)$ are weights that are dynamically modulated by the journalist's Strategic Preference S_p. For example, if the journalist sets the preference to "focus on emotion," the weight w_{emo} for the emotional depth score ϕ_{emo} is implicitly increased. If the preference is "balanced," the weights are distributed more evenly.

In our implementation, the ConvergentAgent (a powerful LLM) does not compute explicit scores. Instead, it performs a holistic evaluation, directly approximating the $\arg\max$ operation by using the strategic preference S_p to guide its selection from the candidate set C'. The prompt instructs it to "select the single best question" that "aligns with the preference," effectively performing this weighted optimization. This leverages the LLM's nuanced reasoning to model the complex utility of a journalistic question.

This structured workflow ensures that the final question is not merely a probable continuation, but a strategically selected option from a creatively diverse and well-curated set of possibilities.

4 System Architecture and Evolution

The MAJI framework was developed and refined over three major versions (Figure 1). Each version represents a significant step in the architectural design, moving from a foundational implementation to a complex and dynamic system. In essence, these versions can be viewed as a theoretical ablation study, with each successive version adding architectural complexity to examine its impact on performance.

4.1 MAJI V1: A Foundational Divergent-Convergent Model

MAJI V1 serves as the foundational implementation of our proposed divergent-convergent architecture. In the divergent step, it uses a single DivergentAgent to generate a diverse pool of candidate questions using a broad LLM prompt. This raw pool of suggestions is then processed through a two-stage convergent step. First, an EditorAgent refines the question pool by removing duplicates and merging semantically similar ideas. Then, a powerful LLM-based ConvergentAgent, guided by the journalist's stated strategic preference, selects the single best question. This version establishes a robust baseline for the effectiveness of the core Divergent-Convergent workflow, against which more specialized architectures can be compared.

4.2 MAJI V2: The Specialized Agent Committee

MAJI V2 builds upon the V1 architecture to test the hypothesis that specializing the divergent step can yield higher-quality questions. The key architectural change is the replacement of V1's single, monolithic DivergentAgent. Instead, V2 implements a robust "agent committee" , where a fixed committee of five specialized divergent agents (ChainOfThought, Emotion, Outline, Persona, and Novelty) brainstorms questions in parallel, each focusing on a distinct creative vector. The convergent selection steps are identical to those in MAJI V1. This design further decouples the creative task, allowing for a more comprehensive and targeted exploration of potential follow-up questions than the generalist approach in V1. An example is in A.1

4.3 MAJI V3: Dynamic Agent Generation

MAJI V3 is an experimental evolution that dynamically devises its own interview strategy, where an agent plans the divergent phase. The key innovation is the introduction of an agent that plans the divergent phase itself. The V3 architecture modifies the divergent step:

1. The fixed committee of divergent agents is removed.
2. A new agent, the EditorInChiefAgent, is introduced. Based on a set of pre-defined heuristics, its role is to analyze the full conversation context and devise a *plan* for the divergent phase, as defined in our V3 data models.
3. This plan takes the form of a DivergentAgentPlan, which contains a list of DivergentAgentSpec objects. Each spec defines the name and instructions for a temporary, single-use divergent agent tailored to the immediate needs of the conversation. For example, if the interviewee seems evasive, it might create a "Probing_Clarification_Agent." If the

conversation is stalling, it might create a "Hypothetical_Scenario_Agent."

4. The system then dynamically instantiates these agents from the specs and runs them in parallel to generate questions.

The rest of the pipeline is unchanged. V3 is more autonomous, with the LLM defining generation strategy. However, this complexity introduced challenges. The `EditorInChiefAgent`'s heuristic-based approach is a limitation, and its performance did not surpass V2. A specific example is in Appendix A.2.

5 Evaluation

We designed an evaluation framework to assess the performance of the MAJI system. The evaluation aims to answer three key questions:

1. How does MAJI's question generation quality compare to a strong, conventionally-prompted LLM baseline across a diverse dataset?
2. How does the architectural choice (integrated vs. decoupled architecture, fixed vs. dynamic agent committee) impact the quality of both the candidate pool and the final selected question?
3. What are the specific strengths and weaknesses of each system, particularly regarding the trade-off between creativity and conversational coherence?

5.1 Experimental Setup

We evaluated MAJI against strong baselines on real-world interviews from the *NewsInterview* dataset (Lu et al., 2024) and proprietary sources. All systems used `gpt-4.1-mini`. Baselines included `LLM-Base`, `LLM-CoT`, `LLM-ToT`, and `LLM-RAG`. We used Prometheus 2 for validation and adjusted originality scores with a threshold-based method for realism (Kim et al., 2024). Further details are in Appendix A.4.

5.2 Metrics

Our evaluation uses a combination of metrics calculated per-question and per-interview, averaged across the dataset.

5.2.1 Per-Question Metrics

These metrics are evaluated for each generated question. Our primary judge (GPT-4o) scored questions on six criteria: **Coherence** (logical connection), **Elaboration** (encouraging detailed responses), **Originality** (novelty), **Context Relevance** (relation to the last turn), **Outline Relevance** (alignment with the interview plan), and **Persona Alignment** (matching interviewee character). Our benchmark judge (Prometheus 2) provided scores for similar criteria, adding measures for **Insight** (probing deeper), **Conversational Synthesis** (integrating prior conversation), and **Strategic Progression** (advancing interview goals).

To better assess originality, we used a threshold-based adjustment grounded in semantic similarity. Cosine similarity scores were derived from SentenceTransformer embeddings (`paraphrase-multilingual-MiniLM-L12-v2`), computed between the candidate question and prior suggestions. Since semantically distinct questions often yield non-zero similarity, raw scores can be misleading. Our method applies a 0.6 threshold: similarities below this are scored as 1.0 (completely original), and the rest are scaled proportionally. This adjustment yields more realistic originality estimates while preserving relative rankings across candidates.

5.2.2 Per-Interview Metrics

To assess overall strategic performance, we calculated the **Insight Trajectory**, measuring if a system asks more insightful questions in the second half of an interview compared to the first.

6 Results

Our evaluation demonstrates the effectiveness of the MAJI framework. The following sections analyze the quality of the final selected questions, the raw brainstormed candidates, and the value added by the convergent selection process to provide a clear, multi-faceted view of our architecture's performance.

6.1 Primary Analysis: Final Question Quality

Our primary evaluation focuses on the quality of the single question chosen by each system to be asked next (Table 1 and Table 2).

Both our primary judge (GPT-4o) and the benchmark judge (Prometheus 2) consistently rank MAJI as the top-performing systems, excelling on metrics like **Coherence**, **Elaboration**, **Originality**,

Table 1: Evaluation of Final Selected Questions on NewsInterview Dataset (Judged by GPT-4o). This table shows the quality of the single question chosen by each system to be asked next. All models are compared against the LLM-Base baseline. Significance from a two-tailed t-test is denoted by asterisks: * p < 0.05, ** p < 0.01, *** p < 0.001. Originality scores have been adjusted using a threshold-based method to account for baseline similarity between questions. Best score is in **bold**.

Metric	MAJI V1	MAJI V2	MAJI V3	LLM-Base	LLM-CoT	LLM-ToT	LLM-RAG
Coherence	0.770***	**0.795*****	0.791***	0.704	0.701	0.770***	0.735***
Elaboration	0.864	0.928***	**0.932*****	0.871	0.873	0.901***	0.871
Originality	0.639***	**0.764*****	0.736***	0.558	0.586***	0.666***	0.611***
Context Relevance	0.375***	**0.434*****	0.414***	0.319	0.319	0.373***	0.328
Outline Relevance	**0.778*****	0.656	0.661	0.670	0.673	0.668	0.658*
Persona Alignment	0.839***	0.868	0.865	0.874	**0.880**	**0.880**	0.872

Table 2: Evaluation of Final Selected Questions on NewsInterview Dataset (Judged by Prometheus 2). This table shows scores from a standardized, third-party model, validating the primary results. All models are compared against the LLM-Base baseline. Significance from a two-tailed t-test is denoted by asterisks: * p < 0.05, ** p < 0.01, *** p < 0.001. Best score is in **bold**.

Metric	MAJI V1	MAJI V2	MAJI V3	LLM-Base	LLM-CoT	LLM-ToT	LLM-RAG
Coherence	0.600	**0.780*****	0.771***	0.629	0.648	0.762***	0.631
Elaboration	0.780	0.898***	**0.908*****	0.774	0.822***	0.839***	0.793
Originality	0.601	**0.733*****	0.727***	0.586	0.584	0.654***	0.608
Context Relevance	0.693*	0.818***	**0.870*****	0.635	0.676	0.807***	0.659
Outline Relevance	0.458	0.586***	**0.624*****	0.417	0.485***	0.589***	0.425
Insight	0.641	0.815***	**0.835*****	0.688	0.741*	0.783***	0.703
Conversational Synthesis	0.577***	**0.743*****	0.739***	0.664	0.697*	0.730***	0.691
Persona Alignment	0.681***	0.837	**0.861*****	0.821	0.835	0.813	0.802

Context Relevance and **Outline Relevance** (p < 0.001). This finding confirms that the structured, multi-agent decomposition offers benefits beyond complex prompting alone. The result is particularly notable when compared against LLM-ToT, which represents an integrated, single-prompt baseline for divergent-convergent reasoning.

6.2 Analysis of the Divergent Step: Candidate Pool Quality

To understand why MAJI produces superior final questions, we first analyze the quality of the raw brainstormed candidate pools generated before any selection occurs (Table 4 and Table 7). This analysis reveals the strength of a decoupled divergent step in fostering creative exploration.

On both datasets, the divergent agents in MAJI consistently produced candidate pools with the highest scores on **Coherence**, **Elaboration**, **Originality**, **Context Relevance**. For instance, on both datasets, MAJI V1's pools scored 0.742 and 0.547 on Originality, outperforming LLM-ToT (0.635 and 0.428) by 16.9% and 27.8%. This indicates that by separating the task of idea generation from final selection, the divergent agents are free to explore a wider and more novel conceptual space, providing a richer set of raw materials for the subsequent selection step. This conclusion is further supported by an analysis of semantic diversity within the candidate pools generated by these systems (Section 6.5).

6.3 Analysis of the Convergent Step: From Candidate Pool to Final Selection

The final step is evaluating the effectiveness of the convergent agent. We measured the "value added" by this agent by calculating the change in score from the average quality of the brainstormed pool to the quality of the final selected question (Table 8).

The Convergent Agent shared by MAJI V2 and V3 demonstrates a consistent ability to improve the candidate pool across multiple dimensions simultaneously. On the Proprietary dataset, for instance, MAJI V2's selection process increased **Coherence** by +21.3% and **Context Relevance** by +20.8%, while still boosting Originality by +11.2%. MAJI V3 shows a even greater increment, with a +26.6% increase in **Coherence** and a +16.0% boost in **Originality**. This shows that the convergent agent is performing a sophisticated selection that synthesizes strategic and creative goals, rather than merely filtering.

In contrast, both MAJI V1 and LLM-ToT reveal limitations in their respective architectures. MAJI V1's selection process consistently involved a severe trade-off, sacrificing a significant amount of the pool's **Originality** (e.g., -21.0% on the proprietary dataset) to gain necessary **Outline Relevance**. This is because its single-prompt divergent step produces a pool where creativity and strategic alignment are poorly correlated. LLM-ToT, which acts as an integrated, single-prompt baseline for divergent-convergent reasoning, also demonstrates this selection capability with large relative improvements. However, this compressed design appears to limit its divergent phase, causing it to start from a much lower-quality brainstormed pool as stated in the previous analysis.

6.4 Inter-Judge Reliability and Analysis

To assess the consistency between our two judges' summary evaluations, we performed a correlation analysis on the aggregated final scores presented in Table 1 and 2. We calculated Spearman's rank correlation (ρ) to measure the association between the average scores assigned by each judge across all models and metrics. The overall agreement on these aggregated scores was statistically significant ($\rho = 0.565$, $p < 0.001$), indicating a moderate positive correlation and confirming that the models which performed well on average with one judge also tended to perform well with the other. A detailed breakdown of the correlation for each individual metric is provided in Appendix A.6.

6.5 Candidate Pool Analysis: Strategic Variety

Beyond the quality of suggested questions, a key advantage of the divergent-convergent architecture is its ability to generate a diverse portfolio of candidate questions.

We analyzed the diversity of the raw candidate question pools generated across our evaluation dataset by comparing the MAJI versions against the LLM-ToT baseline (Table 3). Diversity was measured using **Semantic Spread**, which calculates the average distance of each candidate's embedding—generated using a sentence-transformer model—from the centroid of all candidates in a given round. A higher spread indicates a greater variety in the conceptual space covered by the suggestions.

The results show a clear architectural progression. The divergent-convergent structure of MAJI V1 (0.302) produced a 35.4% higher spread than the multi-path reasoning of LLM-ToT (0.226), demonstrating the effect of separating idea generation from selection. Furthermore, the introduction of the specialized agent committee in MAJI V2 (0.393) increased the strategic variety by an additional 30.1% over V1. This result quantitatively demonstrates that a fixed committee of specialized agents, each exploring distinct creative vectors like emotional depth or logical causality, provides the journalist with a more varied portfolio of options at each turn.

MAJI V3's dynamic agent generation also produced a high spread (0.356) but did not surpass the structured approach of V2's fixed committee. This suggests that a predefined set of specialized roles provides more consistent conceptual exploration in this context. A detailed breakdown of the analysis is available in Appendix A.7.

Table 3: Mean Semantic Spread of candidate pools across the evaluation dataset. A higher value indicates greater diversity and conceptual coverage. The best-performing model is in **bold**.

Model	Mean Spread	Std. Dev.
MAJI V2	**0.393**	0.087
MAJI V3	0.356	0.094
MAJI V1	0.302	0.138
LLM-ToT	0.226	0.096

6.6 Human and Strategic Evaluation

We also conducted a human evaluation with 30 professional journalists and analyzed each system's strategic performance. Journalists preferred MAJI V2's questions nearly half the time (48.9%), more than double the rate of the baseline. This expert preference quantitatively aligns with the rankings from our LLM judges on key creative metrics, providing further validation of our automated evaluation. Furthermore, analysis of the "Insight Trajectory" showed that MAJI V1 had the highest rate of improvement over an interview (Table 11). While MAJI V2 and V3 started from and maintained a much higher insight baseline, V1's simpler architecture appeared effective at building conversational momentum. Full details are in Appendix A.8.

7 Discussion

The results of our experiments provide several key insights into the design of AI systems for creative professional domains.

7.1 The Power of Specialization

The performance difference between MAJI and advanced LLM baselines validates our core hypothesis: decomposing a complex creative task into specialized sub-tasks yields superior results. While advanced prompting like Tree-of-Thought improves LLM performance, MAJI V2's architecture with dedicated agents for emotion, logic, and novelty, consistently produced questions judged as more insightful and original by both AI and human experts. The adjusted originality scores show MAJI V2 achieves a 36.9% improvement over LLM-base on the NewsInterview dataset (0.764 vs 0.558) and a 44.1% improvement on the proprietary dataset (0.608 vs 0.422). This cross-dataset consistency across languages validates that the multi-agent architecture's benefits are not domain-specific but a fundamental advantage. This suggests for creative augmentation, a specialized multi-agent architecture is more promising than prompting strategies for single models.

7.2 The "Creative Partner" vs. "Coherent Assistant"

A crucial finding is the trade-off between creative value and conversational coherence. The baseline LLMs, optimized for next-token prediction, naturally excel as Coherent Assistants.

MAJI, conversely, acts as a creative partner. It is less constrained by the most probable conversational path and more focused on generating high-quality, novel, and insightful questions. The strong preference from professional journalists underscores the value of this approach; they want a tool that expands their creative options, not just one that affirms their instincts. This focus on quality comes at a computational cost, representing a trade-off between speed and insight (see Appendix A.10).

7.3 Error Analysis

Although MAJI's architecture explicitly avoids redundancy, we observe a trade-off between creativity and relevance. Divergent agents occasionally produce off-topic or abstract suggestions. This is by design, as it broadens potential questioning paths, even if many are discarded.

7.4 V3 and the Challenge of Meta-Cognition

MAJI V3's experiment in dynamic agent generation provides insight into the challenges of AI meta-cognition. While this approach produced highly novel and relevant questions, it was less consistent than MAJI V2. The preference for V2 in our human evaluation suggests that for professional use, V2's reliable creativity is currently more valuable than V3's experimental novelty. This highlights a key challenge: building an agent that can effectively strategize about creative strategy is a difficult, higher-order task. For now, a carefully designed, fixed architecture is more robust. While V3's dynamic agent generation is heuristic-based, this is a necessary stepping stone in creative domains where learning-based planning is an open challenge.

7.5 Broader Implications for Human-AI Creative Partnerships

MAJI's principles are not limited to journalism. The divergent-convergent framework, with specialized agent committees, is a generalizable template for augmenting human creativity in any domain involving iterative ideation and strategic selection. Applications could include helping scientists brainstorm hypotheses, assisting marketing teams with slogans, or helping attorneys explore case strategies. This work contributes to a broader vision of AI not as a replacement for human intellect, but as a structured tool for amplifying it.

8 Conclusion

We introduced MAJI, a multi-agent system that assists journalists in generating creative and insightful interview questions. By decomposing the task into a divergent–convergent workflow with specialized agents, MAJI advances beyond standard LLM prompting. In evaluations with 30 professional journalists, MAJI V2 produced questions preferred more than twice as often as a strong baseline, confirming its alignment with professional judgment.

Our results show that a structured multi-agent architecture can effectively augment complex creative tasks, trading some predictability for greater insight and originality. MAJI V2's "agent committee" offers a promising model for AI partners that help professionals overcome cognitive fixation and explore wider idea spaces. Insights from the experimental V3 suggest future directions toward more autonomous, strategy-driven systems.

Limitations

While this study provides strong evidence for MAJI's effectiveness, we acknowledged several limitations that provide avenues for future work. First, our evaluation is conducted on a dataset of professionally curated interviews. The system's robustness on noisier, out-of-domain data—such as unedited live transcripts or interviews on highly specialized topics—remains to be tested. Second, as our qualitative analysis of V3's failure case demonstrates, the system can occasionally produce awkward or contextually inappropriate suggestions. A more detailed qualitative error analysis would be beneficial for identifying and mitigating these failure modes. Our originality metric has been improved through a threshold-based adjustment method to account for baseline similarity between questions, but could be further enhanced to better distinguish semantic novelty from lexical paraphrasing. Third, the latency of the MAJI system, particularly V2 and V3, is a significant consideration (see Appendix A.10). While we argue this is a justifiable trade-off for question quality, further optimization is required to make the system more responsive.

Although a learning-based planner may offer better adaptability in theory, the lack of structured supervision and reward signals in open-ended domains like interviewing makes such training infeasible at present. We therefore use heuristic planning as an exploratory first step in operationalizing strategic meta-reasoning in creative workflows.

Building on these points, future work will focus on expanding our evaluation to broader datasets, refining V3's heuristic planner into a learning-based agent (e.g., using techniques like verbal reinforcement learning (Shinn et al., 2023) or RLHF (Christiano et al., 2017)), conducting component-wise ablation studies to quantify each agent's contribution, developing a real-time user interface for live interviews, and performing a systematic qualitative error analysis to guide future improvements. Future work should also explore model distillation to create smaller, faster versions of the agents, caching strategies for recurring sub-problems, and asynchronous processing to reduce the perceived latency for the user.

Ethical Considerations

The deployment of AI tools like MAJI in journalism necessitates careful consideration of ethical implications. First, there is a risk that the underlying LLM could introduce subtle biases into the question generation process, potentially reflecting political or confirmation biases from its training data and steering conversations in unintended directions. We suggested ongoing monitoring and fine-tuning with diverse datasets to mitigate this. Second, while MAJI is designed to augment, not replace, the journalist, there is a risk of over-reliance, which could diminish the journalist's critical thinking and rapport-building skills. The system should be framed as a supportive tool, with the final decision always resting with the human journalist, who must also approve any intermediate strategic pivots suggested by the AI. Finally, the privacy and consent of the interviewee are paramount. All data used for training and running the system must be handled with explicit consent and robust anonymization procedures, as was done in this study (Alzoubi et al., 2024).

Acknowledgments

During the preparation of this work, the authors utilized an AI-powered tool to assist with programming, text editing, and formatting. All AI-generated content, including code and text, was carefully reviewed, revised, and validated by the authors to ensure its accuracy and alignment with the research goals.

References

Omar Alzoubi, Normahfuzah Ahmad, and Norsiah Abdul Hamid. 2024. Artificial intelligence in newsrooms: Ethical challenges facing journalists. *Studies in Media and Communication*, 12:401.

Paul F. Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems 30*, pages 4299–4307. Curran Associates, Inc.

Sarah Cohen, James T. Hamilton, and Fred Turner. 2011. Computational journalism. *Commun. ACM*, 54(10):66–71.

Nicholas Diakopoulos. 2019. Automating the news: How algorithms are rewriting the media. In *Computational Journalism*, Cambridge, MA, USA. Harvard University Press.

Matan Gordin, Eric Horvitz, and Jaime Teevan. 2023. Evaluating creativity in large language models: A review. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, pages 313–324.

J. P. Guilford. 1950. Creativity. *Am. Psychol.*, 5(9):444–454.

Felix Hamborg and 1 others. 2018. Automated identification of media bias in news articles. *Int. J. Data Sci. Anal.*, 6(4):327–339.

Seungone Kim, Juyoung Suk, Shayne Longpre, Bill Yuchen Lin, Jamin Shin, Sean Welleck, Graham Neubig, Moontae Lee, Kyungjae Lee, and Minjoon Seo. 2024. Prometheus 2: An open source language model specialized in evaluating other language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4334–4353, Miami, Florida, USA. Association for Computational Linguistics.

Guohao Li, Hasan Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. 2023a. Camel: Communicative agents for "mind" exploration of large language model society. *Advances in Neural Information Processing Systems*, 36:51991–52008.

Yifu Li, Jialu Li, Yufang Lai, and Nanyun Peng. 2023b. Accord: A multi-document approach to generating diverse and coherent summaries. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8341–8356, Toronto, Canada. Association for Computational Linguistics.

Yi-Cheng Lin, Kang-Chieh Chen, Zhe-Yan Li, Tzu-Heng Wu, Tzu-Hsuan Wu, Kuan-Yu Chen, Hung-yi Lee, and Yun-Nung Chen. 2025a. Creativity in LLM-based multi-agent systems: A survey. *CoRR*, abs/2505.21116.

Yi-Cheng Lin, Kang-Chieh Chen, Zhe-Yan Li, Tzu-Heng Wu, Tzu-Hsuan Wu, Kuan-Yu Chen, Hung yi Lee, and Yun-Nung Chen. 2025b. Creativity in llm-based multi-agent systems: A survey. *Preprint*, arXiv:2505.21116.

Michael Lu, Hyundong Justin Cho, Weiyan Shi, Jonathan May, and Alexander Spangher. 2024. Newsinterview: a dataset and a playground to evaluate llms' ground gap via informational interviews. *Preprint*, arXiv:2411.13779.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Chen Qian, Xin Cong, Cheng Yang, Weize Chen, Yusheng Su, Juyuan Liu, Yufan Liu, and Zipeng Yu. 2023. Chatdev: Communicative agents for software development. *CoRR*, abs/2307.07924.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. Reflexion: Language agents with verbal reinforcement learning. In *Advances in Neural Information Processing Systems 36*. Curran Associates, Inc.

Alexander Spangher, Tenghao Huang, Philippe Laban, and Nanyun Peng. 2025. Creative planning with language models: Practice, evaluation and applications. In *Proceedings of the 2025 Annual Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 5: Tutorial Abstracts)*, pages 1–9, Albuquerque, New Mexico. Association for Computational Linguistics.

Yufei Tian, Abhilasha Ravichander, Lianhui Qin, Ronan Le Bras, Raja Marjieh, Nanyun Peng, Yejin Choi, Thomas L. Griffiths, and Faeze Brahman. 2024. MacGyver: Are large language models creative problem solvers? In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Mexico City, Mexico. Association for Computational Linguistics.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *Preprint*, arXiv:2302.13971.

Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, Wayne Xin Zhao, Zhewei Wei, and Ji-Rong Wen. 2023. A survey on large language model based autonomous agents. *CoRR*, abs/2308.11432.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems 35*. Curran Associates, Inc.

Michael Wooldridge. 2009. *An introduction to multiagent systems*. John Wiley & Sons, Hoboken, NJ, USA.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, and Li Li. 2023. Autogen: Enabling next-gen LLM applications via multi-agent conversation framework. *CoRR*, abs/2308.08155.

Zekun Xi, Wenbiao Yin, Jizhan Fang, Jialong Wu, Runnan Fang, Ningyu Zhang, Jiang Yong, Pengjun Xie, Fei Huang, and Huajun Chen. 2025. OmniThink: Expanding knowledge boundaries in machine writing through thinking. *CoRR*, abs/2501.09751.

Lili Yao, Nanyun Peng, Dietrich Klakow, Mark Riedl, and Weischedel Ralph Wang. 2019. Plan-and-write: Towards better automatic storytelling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7454–7461.

Shunyu Yao, 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. *Preprint*, arXiv:2305.10601.

Wangchunshu Zhou, Peng Wang, Yifei Li, Wenyong Zhu, and Bill Yuchen Lin. 2023. Agents: An open-source framework for autonomous language agents.

A Appendix

A.1 MAJI V2 Specialization Example

To illustrate the power of specialization in the MAJI V2 committee, consider a response from a professional mermaid performer: *"When I'm down there, everything goes silent. It's just me and the water, and sometimes I forget there's an audience. It's a very physically demanding job, but the peace I feel is worth it."*

The specialized divergent agents might respond as follows:

- `EmotionDivergentAgent`: "You mentioned a feeling of 'peace.' Can you describe the contrast between that inner peace and the intense physical demands of the job?"

- `ChainOfThoughtDivergentAgent`: "What specific physical training was required to allow you to reach that point where you can find peace despite the physical exertion?"

- `PersonaDivergentAgent`: "As an artist, how does that feeling of solitary peace underwater influence the performance that the audience eventually sees?"

- `NoveltyDivergentAgent`: "If you could perform in any body of water in the world, real or mythical, where would you choose to best capture that feeling of peace?"

This example shows how the agent committee generates a rich, multi-faceted set of candidate questions, giving the journalist a far more powerful set of options than a single, generic follow-up.

A.2 MAJI V3 Failure Case Example

The increased autonomy of MAJI V3's dynamic agent generation, while powerful, could sometimes lead to strategic missteps. For example, in an interview with a climate scientist who briefly mentioned enjoying hiking early in the conversation, a dynamically generated 'Personal_Connection_Agent' later interrupted a dense discussion on carbon sequestration models to ask, "You mentioned hiking—what's the most beautiful trail you've ever been on?" While a valid question in isolation, its poor timing demonstrated a failure in strategic conversational awareness—a key failure case for V3's heuristic planner, which occasionally struggled to weigh the global strategic context against a locally-optimized creative idea.

Beyond isolated timing errors, V3 also demonstrated a tendency toward *loss of concentration*. Whereas V1, V2, and the baseline models tended to remain within a shared conceptual trajectory—probing deeper along an established thematic path—V3's planner could divert attention toward tangential but momentarily interesting details. Once a divergent agent was spawned around such a detail, its contribution could dominate the candidate pool, resulting in a final question that shifted the conversation's focus away from its intended direction. This fragmentation of topical focus underscores a key challenge in dynamic agent systems: while autonomy encourages creative exploration, without stronger strategic gating it can erode the coherence and sustained depth of an interview.

A.3 Side-by-Side Generations and Choice Set Example

During the previously discussed interview with a professional mermaid performer, both MAJI V1, V2, V3 and LLM Baselines provided their next question recommendation. The interviewee is a female mermaid performer and social media influencer. As noted in the main text, her work environment involves irregular tourist-driven schedules, limited rest days, and salary deductions during menstrual leave. Wages in the industry have also declined in recent years. This example illustrates the different results between MAJI and LLM Baselines.

V1 Selected Question *Do you think there are any inconveniences or risks associated with using tampons during performances and training?*

V2 Divergent Pool (Specialized Agent Committee)

- `ChainOfThought Agent`: Why has the practice of pay deductions for leave and a 4–5 day work-break arrangement become common in the mermaid industry? What is the reasoning behind this system?

- Emotional Agent: What specific impacts has this tiered working schedule had on your mental and physical health? How do you adjust and cope with this intense work pace?
- Outline Agent: How do you think this limited rest and leave system affects new mermaid performers? Does it influence talent flow or staff stability within the industry?
- Persona Agent: As an industry opinion leader, what are your thoughts on the role of social media in promoting fair norms and protecting practitioners' rights? Have you ever encountered pressure or backlash for speaking out?
- Novelty Agent: If you could redesign the profession of "mermaid performer," which three aspects of the job would you most like to change or redefine?

The ConvergentAgent selected the Logic_and_Causality question as optimal for its analytical framing and structural relevance to economic conditions:

> *Why has the practice of pay deductions for leave and a 4–5 day work-break arrangement become common practice in the mermaid industry? What is the logic and reasoning behind this arrangement?*

V3 Dynamic Plan (Adaptive Divergent Agents)

- Work_Stress_Analyst: Could you describe your typical work schedule? For example, what time do you start preparing and what time do performances typically end?
- Industry_Advocacy_Champion: Given that wages continue to be suppressed, have you organized or participated in any activities to promote wage standards or labor-rights protection?
- Health_Risk_Investigator: Have you observed any impact of current rest policies on the adaptation and career longevity of new performers? Do these conditions cause some to leave the industry early?
- Economic_Reality_Interviewer: When did this trend of salary reductions begin? Is there any consistent industry data to support this phenomenon?

The ConvergentAgent selected the Work_Stress_Analyst question as the most contextually grounded opener:

> *Could you describe your typical work schedule? For example, what time do you start preparing and what time do performances typically end?*

Baseline Systems

- LLM-Base: How will the other performers arrange their work during menstruation? Is there a special substitute or shift system?
- LLM-CoT: You mentioned that taking leave results in salary deductions. Does this system affect your and your colleagues' enthusiasm for work?
- LLM-ToT: Have you ever experienced discomfort or emergencies while performing underwater with a tampon? How did you handle them?
- LLM-RAG: Besides using tampons, do you have any special hygiene or safety measures when going into the water?

A.4 Experimental Setup Details

- **Dataset:** The evaluation was conducted on a combined corpus from the public *NewsInterview* dataset (entirely in English) and proprietary transcripts from a media tech company (70% Chinese, 30% English). This multilingual setup was designed to test the robustness of the MAJI framework across different languages. The public dataset was filtered to include only those conversations with exactly two speakers and more than 50 conversational exchanges to focus on substantial, dyadic conversations suitable for our framework. The proprietary dataset includes interview transcripts with detailed personas and outlines. This combined dataset contains professionally conducted interviews covering a diverse range of topics, from profiling a professional mermaid performer to discussing political opinions on U.S. elections. This provides a diverse and realistic set of conversational contexts for the systems to respond to, moving beyond a single-interview analysis. Since the public NewsInterview dataset does not include structured outlines, we used GPT-4o to generate a plausible outline for each of these interviews based on the full transcript content. These generated outlines were then provided to all systems.

- **Inputs:** For each turn in every interview, the systems were provided with the same set of inputs: the interviewee's `Persona`, the interview `Outline`, and the full conversation `transcript` up to that point.

- **Models:** All agent systems (MAJI V2, V3) and the baseline systems use `gpt-4.1-mini` as the underlying LLM to ensure a fair comparison of architectural benefits versus model capabilities.

- **Baselines:** We compared MAJI against a suite of strong baselines representing common and advanced prompting techniques that do not employ agent-based architectures.
  - **LLM-Base:** A single, well-prompted call to the base LLM, including the full context.
  - **LLM-CoT:** A baseline using Chain-of-Thought prompting (Wei et al., 2022) to encourage step-by-step reasoning before generating a question.
  - **LLM-ToT:** A baseline using a Tree-of-Thought approach (Yao et al., 2023), where the model explores multiple reasoning paths.
  - **LLM-RAG:** A baseline augmented with a Retrieval-Augmented Generation mechanism. This system uses a sentence-transformer model to find and retrieve the most semantically similar turns from earlier in the same conversation, providing the LLM with relevant long-term context that might have been lost.

- **Benchmark Judge (Prometheus 2):** To validate our findings against a standardized, third-party metric, we also used Prometheus 2, a state-of-the-art open-source evaluation model. This "black-box" judge acts as an impartial adjudicator, scoring the final selected question from each system. Using a benchmark judge mitigates the risk of "own-model-bias" and strengthens the credibility of our results. Its distinct metrics, such as *conversational synthesis* and *strategic progression*, also provide a valuable alternative perspective on performance.

- **Baseline Prompts**: The core instruction for all baselines was: "You are an expert journalist. Based on the provided Persona, Outline, and Transcript, generate the best possible next question to ask. Your goal is to be insightful, creative, and strategic." For CoT and ToT, additional instructions for step-by-step reasoning and exploring alternatives were included based on their respective papers.

A.5 Supplemental Evaluation Tables

This appendix contains supplemental tables supporting the main paper's analysis. Table 4 and Table 7 provide the raw quality scores for the brainstormed candidate pools on both the NewsInterview and proprietary datasets, respectively. Table 5 and Table 6 show the final selected question scores on the proprietary dataset, demonstrating the cross-dataset robustness of our findings. Finally, Table 8 presents the full data for the convergent selection uplift analysis discussed in Section 6.3.

A.6 Detailed Inter-Judge Reliability and Analysis

We further analyze the inter-judge reliability by metrics to validate our automated evaluation results.

A per-metric analysis reveals a clear and insightful pattern. The judges demonstrated exceptionally high and significant agreement on metrics evaluating creative quality and immediate contextual relevance: **Elaboration** ($\rho = 0.99$), **Originality** ($\rho = 0.96$), and **Context Relevance** ($\rho = 0.90$). This high level of consensus indicates that LLMs can reliably and consistently assess these core attributes of a question.

In contrast, the judges' agreement diminished significantly on metrics requiring more holistic or long-term strategic judgment. We found no meaningful correlation for **Persona Alignment** ($\rho = 0.04$) and **Outline Relevance** ($\rho = -0.29$). This divergence suggests that while LLMs are consistent in evaluating a question's creativity and local fit, their assessment of its alignment with a broader persona or interview plan remains a subjective and model-dependent task.

A.7 Detailed Candidate Pool Analysis

The quantitative analysis of semantic diversity within the candidate question pools, introduced

Table 4: Evaluation of All Brainstormed Suggestions on `NewsInterview` Dataset (Judged by GPT-4o). This table shows the average quality of the entire pool of questions generated by the divergent phase, before any selection occurs. All models are compared against the LLM-Base baseline. Significance from a two-tailed t-test is denoted by asterisks: * p < 0.05, ** p < 0.01, *** p < 0.001. This measures the raw creative output. Originality scores have been adjusted using a threshold-based method. Best score is in **bold**.

Metric	MAJI V1	MAJI V2	MAJI V3	LLM-Base	LLM-CoT	LLM-ToT	LLM-RAG
Coherence	**0.809*****	0.728***	0.702*	0.682	0.679	0.704***	0.702***
Elaboration	0.874	**0.899*****	**0.899*****	0.863	0.871**	0.881***	0.863
Originality	0.742***	**0.745*****	0.705***	0.592	0.599	0.635***	0.621***
Context Relevance	**0.389*****	0.369***	0.343***	0.298	0.296	0.313***	0.304*
Outline Relevance	0.661	0.635***	0.665	0.659	0.670***	**0.672*****	0.655
Persona Alignment	0.846***	0.870*	0.873*	0.881	**0.884**	0.882	0.880

Table 5: Evaluation of Final Selected Questions on `Proprietary` Dataset (Judged by GPT-4o). All models are compared against the LLM-Base baseline. Significance from a two-tailed t-test is denoted by asterisks: * p < 0.05, ** p < 0.01, *** p < 0.001. Originality scores have been adjusted using a threshold-based method. Best score is in **bold**.

Metric	MAJI V1	MAJI V2	MAJI V3	LLM-Base	LLM-CoT	LLM-ToT	LLM-RAG
Coherence	0.740	**0.820****	0.723*	0.688	0.725***	0.780***	0.767***
Elaboration	0.874	**0.953****	0.925***	0.875	0.890*	0.890*	0.871
Originality	0.432	**0.608*****	0.561***	0.422	0.455*	0.534***	0.512***
Context Relevance	0.348	**0.430****	0.380	0.351	0.368	0.411***	0.391*
Outline Relevance	**0.761****	0.618*	0.635**	0.668	0.666	0.645*	0.650
Persona Alignment	0.746	**0.834**	0.786**	0.823	0.823	0.811	0.803

Table 6: Evaluation of Final Selected Questions on `Proprietary` Dataset (Judged by Prometheus 2). All models are compared against the LLM-Base baseline. Significance from a two-tailed t-test is denoted by asterisks: * p < 0.05, ** p < 0.01, *** p < 0.001. Best score is in **bold**.

Metric	MAJI V1	MAJI V2	MAJI V3	LLM-Base	LLM-CoT	LLM-ToT	LLM-RAG
Coherence	0.536	**0.734*****	0.674	0.591	0.597	0.690*	0.646
Elaboration	0.780	**0.872***	0.860*	0.759	0.768	0.775	0.754
Originality	0.525	0.570	**0.588***	0.536	0.538	0.526	0.469
Context Relevance	0.605	**0.864*****	0.703	0.639	0.655	0.740	0.652
Outline Relevance	0.499	**0.658**	0.639	0.599	0.591	0.654	0.566
Insight	0.619	**0.802***	0.748*	0.672	0.701	0.675	0.618
Conversational Synthesis	0.365	**0.530***	0.402	0.390	0.388	0.420	0.335
Persona Alignment	0.372	**0.590***	0.525	0.496	0.545	0.512	0.462

Table 7: Evaluation of All Brainstormed Suggestions on `Proprietary` Dataset (Judged by GPT-4o). All models are compared against the LLM-Base baseline. Significance from a two-tailed t-test is denoted by asterisks: * p < 0.05, ** p < 0.01, *** p < 0.001. Originality scores have been adjusted using a threshold-based method. Best score is in **bold**.

Metric	MAJI V1	MAJI V2	MAJI V3	LLM-Base	LLM-CoT	LLM-ToT	LLM-RAG
Coherence	**0.768*****	0.676*	0.571	0.618	0.610	0.632	0.662***
Elaboration	0.873	**0.905*****	0.891***	0.860	0.869	0.865	0.850*
Originality	**0.547*****	**0.547*****	0.488***	0.395	0.383	0.428*	0.462***
Context Relevance	**0.367*****	0.356***	0.288	0.287	0.280	0.300	0.320***
Outline Relevance	0.655	0.637***	0.647**	0.675	**0.683*****	0.680	0.663
Persona Alignment	0.759***	0.808	0.804	**0.814**	0.812	0.795***	0.795

Table 8: Effectiveness of the Convergent Selection Process Across Datasets. This table shows the score change (Final Score - Brainstormed Avg. Score) after the convergent selection step. Each cell displays results for the `NewsInterview` dataset (top) and the Proprietary dataset (bottom). Percentage change is in parentheses. Positive values demonstrate the Convergent Agent's value-add. Key improvements are in **bold**.

Metric	MAJI V1	MAJI V2	MAJI V3	LLM-ToT
Coherence	-0.039 (-4.8%) -0.028 (-3.6%)	+0.067 (+9.2%) +0.144 (+21.3%)	**+0.089 (+12.7%)** **+0.152 (+26.6%)**	+0.066 (+9.4%) +0.148 (+23.4%)
Elaboration	-0.010 (-1.1%) +0.001 (+0.1%)	+0.029 (+3.2%) **+0.048 (+5.3%)**	**+0.033 (+3.7%)** +0.034 (+3.8%)	+0.020 (+2.3%) +0.025 (+2.9%)
Originality	-0.103 (-13.9%) -0.115 (-21.0%)	+0.019 (+2.6%) +0.061 (+11.2%)	+0.031 (+4.4%) +0.073 (+15.0%)	**+0.031 (+4.9%)** **+0.106 (+24.8%)**
Context Relevance	-0.014 (-3.6%) -0.019 (-5.2%)	+0.065 (+17.6%) +0.074 (+20.8%)	**+0.071 (+20.7%)** +0.092 (+31.9%)	+0.060 (+19.2%) **+0.111 (+37.0%)**
Outline Relevance	**+0.117 (+17.7%)** **+0.106 (+16.2%)**	+0.021 (+3.3%) -0.019 (-3.0%)	-0.004 (-0.6%) -0.012 (-1.9%)	-0.004 (-0.6%) -0.035 (-5.1%)
Persona Alignment	-0.007 (-0.8%) -0.013 (-1.7%)	**-0.002 (-0.2%)** **+0.026 (+3.2%)**	-0.008 (-0.9%) -0.018 (-2.2%)	**-0.002 (-0.2%)** +0.016 (+2.0%)

Table 9: Inter-candidate similarity statistics. The table shows the percentage of question pairs that are highly diverse (sim < 0.5), diverse (sim < 0.6), similar (sim > 0.7), and highly similar (sim > 0.8).

Model	Mean Sim.	Std. Dev.	% Highly Div. (<0.5)	% Div. (<0.6)	% Sim. (>0.7)	% Highly Sim. (>0.8)
LLM-ToT	0.462	0.144	61.1%	82.5%	4.8%	0.8%
MAJI V3	0.478	0.171	56.4%	75.5%	10.5%	3.4%
MAJI V2	0.532	0.170	42.7%	64.0%	17.3%	5.5%
MAJI V1	0.546	0.177	38.8%	59.7%	19.0%	7.6%

in Section 6.5, was conducted on the raw, brainstormed suggestions generated by each model prior to any filtering or selection. The analysis used our proprietary dataset, which contains a total of 309 interview rounds. To compute embeddings for all candidate questions, we used the sentence-transformer model `paraphrase-multilingual-MiniLM-L12-v2`.

A.7.1 Semantic Spread

Semantic Spread measures the conceptual volume of the candidate pool by calculating the mean Euclidean distance of each question's embedding from the geometric centroid of all candidates in a given round. As shown in Table 3, MAJI V2 achieved the highest average spread, indicating it consistently generated the most varied portfolio of choices.

A.7.2 Inter-Candidate Similarity

We also calculated the mean pairwise cosine similarity between all candidates generated in the same round for each system (Table 9). A lower mean similarity score suggests that individual suggestions are, on average, more distinct from one another.

A.7.3 Analysis and Interpretation

LLM-ToT achieved the lowest average inter-candidate similarity (0.4615), indicating highly distinct question directions within each round. Its low semantic spread (0.2261) reflects a balanced distribution pattern: the model's use of three fixed reasoning paths generates questions that, while dissimilar to each other, maintain consistent distances from each round's topical centroid. This contrasts with MAJI V2's highest semantic spread (0.3926), which results from its specialist agents creating denser exploration of the question space, including intermediate conceptual positions, though with moderately higher mean similarity (0.5321).

MAJI V3 achieved the optimal balance, combining low similarity (0.4782, second only to LLM-ToT) with substantial semantic spread (0.3558) and the highest proportion of highly diverse pairs (56.4% with similarity < 0.5). Importantly, V3 minimized redundancy with only 3.4% very similar pairs (> 0.8), compared to 7.6% for V1 and 5.5% for V2.

To ensure fair comparison, similarities were computed only between questions from different specialist sources, excluding intentional within-source variants in MAJI V2/V3's multi-agent architecture (where each specialist generates broad/depth/balanced variations). Two-sample t-tests confirmed all pairwise differences were statistically significant ($p < 0.001$), with effect sizes from negligible (V1 vs V2: Cohen's $d = -0.084$) to medium (V1 vs LLM-ToT: $d = -0.533$). The progression V1 → V2 → V3 demonstrates systematic improvement, with V3 reducing mean similarity by 12.5% while achieving 18% higher semantic spread compared to V1.

A.8 Human and Strategic Evaluation Results

A.8.1 Human Evaluation: The Professional's Choice

To complement our automated metrics, we conducted a qualitative survey with 30 professional journalists from a media tech company, all with over five years of experience and a focus on online media. Participants were presented with 25 conversational snippets. These snippets were randomly drawn from 5 interviews, themselves randomly selected from the NewsInterview portion of our dataset. For each snippet, the journalists were provided with the full conversational context up to that point, as well as the interviewee's Persona and the interview Outline to ensure they had the same information as the AI systems. For this study, we report preference shares as the primary outcome. While inter-annotator agreement metrics like Fleiss' Kappa are valuable for tasks with objective ground truths, their interpretation is less straightforward for subjective, creative-preference tasks where there is no single 'best' answer. Therefore, we did not compute IAA for this study, but note that future work using a rated scale (e.g., 1-5) instead of a forced choice could more meaningfully incorporate such agreement analyses.

Participants were asked a forced-choice question: "If you were the interviewer, which question would you have chosen to ask next?" This study provides strong evidence of MAJI's value and usability for professional journalists. The results (Table 10) show a decisive preference for MAJI V2, which was chosen nearly half the time.

Table 10: Human Journalist Preference. Results from a blind survey of 30 professional journalists asked to choose the best question from MAJI V2, MAJI V3, and a representative LLM ToT Baseline across 25 conversational snippets.

System	Preference Share (%)
MAJI V2	**48.9%**
MAJI V3	29.5%
LLM ToT	21.6%

A.8.2 Insight Trajectory: Improving Over Time

Beyond the quality of individual questions, we analyzed each system's ability to improve its performance over the course of an interview (Table 11). The results highlight an interesting trade-off. MAJI V1 demonstrated a remarkable ability to improve its insight score, achieving a statistically significant 39.0% improvement rate—far surpassing all other models. Conversely, while MAJI V2 and V3 show smaller relative improvement, this is because they start from a much higher performance baseline. MAJI V3 achieves the highest absolute insight scores in the second half of the interview, a statistically significant improvement over the LLM ToT baseline (p < 0.01). Their sustained high quality underscores their superiority, even if their rate of improvement is less dramatic.

Table 11: Insight trajectory. Average insight scores in the first and second halves of interviews. Significance vs. LLM-Base: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

System	Initial	Final	Improvement (%)
MAJI V1	2.11^{***}	2.60	**39.02**
MAJI V2	2.57	3.02^{*}	29.57
MAJI V3	**2.62**	$\mathbf{3.05}^{**}$	25.47
LLM-RAG	2.48	2.62	17.82
LLM-Base	2.56	2.76	18.07
LLM-ToT	2.48	2.83	31.01
LLM-CoT	2.52	2.83	22.96

A.9 Dataset Topics

The two datasets used in our evaluation cover a wide range of subjects, providing a robust testbed for our system.

NewsInterview Dataset Topics

The topics in the public *NewsInterview* dataset span a broad range of journalistic beats. The main categories are summarized in Table 12.

Table 12: Topics in the NewsInterview Dataset.

Category	Description
Politics & Government	U.S./international politics, elections, impeachment, and national security.
Economy, Trade, & Employment	International trade, economic crises, financial regulation, and employment.
Law, Justice, & Human Rights	Criminal justice, human rights issues, press freedom, and whistleblowing.
Health & Social Issues	Healthcare policy, gun control, gender/racial issues, and social dynamics.
Climate, Env., & Disasters	Climate change, disaster management, conservation, and environmental policy.
Arts, Culture, & Literature	Literature, poetry, music history, philosophy, and cultural identity.
Science, Tech., & Education	Biology, astrophysics, engineering, internet technology, and education.
Sports & Ethics	Professional sports, commentary, and ethical debates in sports.

Proprietary Dataset Topics

The proprietary dataset from the media tech company contains interviews with a more personal and narrative focus. The topics are summarized in Table 13.

Table 13: Topics in the Proprietary Interview Dataset.

Category	Description
Social & Personal Issues	Personal narratives, family dynamics, and cultural identity.
Health & Wellness	Experiences with the healthcare system and personal well-being.
Professional Life	Career paths, workplace experiences, and industry insights.
Disaster & Adversity	Personal accounts of overcoming natural disasters or adversity.

A.10 Latency Analysis

MAJI V2 requires an average of 16.59 seconds per question, compared to 1.42 seconds for baseline models. This higher latency reflects MAJI's sequential, multi-agent architecture, which explicitly trades speed for strategic depth. Unlike single-shot LLMs, MAJI decomposes the task into multiple specialized agents followed by editing and convergence steps, each run in sequence.

While not instantaneous, this latency remains acceptable in a human-in-the-loop interview setting, where brief pauses between questions are natural. We position MAJI as a near real-time assistant—optimized for insight and originality over immediacy—designed to support, not replace, the journalist's creative process. Future work will explore techniques such as model distillation, agent parallelization, and incremental reasoning to reduce latency.

Table 14: Average latency per question. Measured from the start of MAJI's pipeline to final output.

System	Avg. Latency (s)
LLM Baselines (Avg.)	1.42
MAJI V1	4.22
MAJI V2	16.59
MAJI V3	20.48

A.11 In-the-Field User Study

To assess MAJI's real-world applicability, we conducted a live deployment with a professional journalist during a 15-minute interview. The MAJI V2 system was accessed via a web-based dashboard on a MacBook Pro M2, positioned adjacent to the interview screen. It continuously updated suggestions based on transcribed utterances (via Whisper X), using GPT-4.1-mini with a context window of the last two speaker turns.

Over the course of the interview, the journalist asked 24 questions: 7 (29%) were adopted directly from MAJI's output, and 5 (21%) were minor variations. The remaining 12 were entirely original. Adopted questions were described as 'creative'—e.g., "Have you ever changed your reporting strategy based on your interviewee's mood?"—as opposed to clarifying prompts.

Despite MAJI's average 16 second generation latency, the journalist found the system non-disruptive. Two coping strategies helped: (1) high-quality suggestions remained relevant across multiple turns, and (2) the journalist could "buy time" by summarizing prior content while waiting for the next batch.

This study demonstrates that MAJI can be successfully integrated into real-time journalistic workflows. The journalist reported that MAJI enhanced creativity under deadline pressure. While this pilot involved a single professional, MAJI's architecture generalizes to other interactive formats. Ethical consent was obtained prior to deployment, and no interviewee data was retained.

A.12 Data Statement

Our study utilizes a combination of publicly available and proprietary datasets. The public data is drawn from the *NewsInterview* dataset (Lu et al., 2024), which consists of previously published interview transcripts. As such, it does not contain personally identifiable information beyond what was already made public by the original news organizations.

Our proprietary data consists of interview transcripts from a media tech company. As detailed in our Ethical Considerations, this data was used with explicit consent, and robust anonymization procedures were applied to protect the privacy of all individuals involved.

We acknowledge the ethical complexities of using real-world interview data. Some interviews may touch on sensitive topics, and we have handled this data with care. Furthermore, while the datasets we used are intended for research, we recognize that the copyright of the original material resides with the news organizations. Our use of this data is strictly for non-commercial research purposes to advance the understanding of computational tools in journalism.

A.13 Annotator and Participant Statement

The human evaluation and user study involved professional journalists who participated on a voluntary basis. The 30 journalists who participated in the qualitative survey (Appendix A.8) and the journalist who participated in the in-the-field user study (Appendix A.11) were colleagues from a media tech company. We are grateful for their time and expert feedback, which was essential for validating the practical applicability of our work. No monetary compensation was provided. All participants were informed about the research goals and how their feedback would be used.

A.14 Computational Resources

All experiments were run with OpenAI resources where a total of 150 dollars was spent.

B System Prompts

This section contains the core prompts used for the LLM baselines and the various MAJI agents. Each prompt is enclosed in a code block for clarity, with placeholders like {placeholder} representing dynamically populated data. Prompts are organized by system version for ease of reference.

B.1 LLM Baselines

Figure 1: LLM-Base Prompt

```
You are a professional interviewer. Based
    on the following information, please
    generate {num_questions} suitable next
    questions for the interview.

[Interviewee Information]
{persona_str}

[Interview Outline]
{outline_str}

[Conversation History]
{history}

Please output the list of questions in JSON
    array format, for example:
["Question 1", "Question 2", "Question 3"]
```

Figure 2: LLM-CoT Appended Instruction

```
First, think step-by-step about the
    interview's goal, the interviewee's
    personality, and the recent
    conversation flow. Consider what topics
    are yet to be covered and what previous
    points could be explored deeper. Based
    on this reasoning, then generate the
    questions.
```

Figure 3: LLM-ToT Appended Instruction

```
Explore multiple reasoning paths to decide
    on the best questions.
1. Path 1: Focus on deepening the last
    topic.
2. Path 2: Focus on transitioning to a new
    topic.
3. Path 3: Focus on the interviewee's
    emotional state.
Evaluate these paths and generate a final
    list of questions that synthesizes the
    best options.
```

Figure 4: LLM-RAG Appended Instruction

```
[Retrieved from long-term memory]
Here are some potentially relevant snippets
    from earlier in the conversation:
{retrieved_context}

Based on the conversation history AND the
    retrieved memories, generate the next
    questions.
```

B.2 MAJI V1 Agents

Figure 5: MAJI V1: DivergentAgent Prompt Summary

```
You are the interview's thinking engine,
    responsible for divergent analysis.
    Your core tasks are:
 1. Match the current answer to the outline.
 2. Identify emotional expressions.
 3. Analyze logical connections.
 4. Identify important, uncovered areas in
     the outline.
 5. Generate exploratory follow-up
     questions.
 6. Extract key memory snippets.
You must follow strict matching criteria
    and output a valid JSON containing a
    `DivergentAnalysis` object. Follow-up
    questions should be natural, fluent,
    and avoid simply repeating the outline.
```

Figure 6: MAJI V1: EditorAgent Prompt

```
You are an expert editor. Your task is to
    review a list of proposed interview
    questions from different AI agents.
    Your goal is to clean up this list by
    removing duplicates and combining very
    similar questions.
```

Figure 7: MAJI V1: ConvergentAgent Prompt

```
You are the Editor-in-Chief of this
    interview, responsible for selecting
    the single best question to ask next.
    You will be given a list of candidate
    questions from various specialist
    agents. Your decision should be guided
    by the user's stated preference for the
    interview's direction.
```

B.3 MAJI V2 Agents

Figure 8: MAJI V2: BackgroundAgent Prompt

```
You are an AI assistant that maintains a
    dynamic background summary for an
    ongoing interview. Your task is to
    integrate the latest conversation turn
    into the existing background summary.
```

Figure 9: MAJI V2: KeywordsAgent Prompt

```
You are an AI assistant that extracts
    critical keywords from the latest
    conversation turn. Use the provided
    background summary to identify keywords
    that are not only salient to the
    current turn but also connect to the
    broader conversation context.
```

Figure 10: MAJI V2: OutlineMatcherAgent Prompt

```
You are a precise AI analyst. Your sole job
    is to match the user's latest response
    to a specific question in the provided
    interview outline. You must determine
    the best match and assess how well the
    response covers the question.
```

B.3.1 Divergent Agent Committee

All divergent agents share a common preamble, followed by their specific specialization instructions.

Figure 11: MAJI V2: Divergent Agent Common Preamble

```
You are a creative and insightful interview
    question generator. Your goal is to
    propose at least one, and up to three,
    thoughtful follow-up questions based on
    the provided context. Your output MUST
    be a valid JSON object. Do not simply
    repeat questions from the outline.
```

Figure 12: ChainOfThoughtDivergentAgent Specialization

```
Your Specialization: Logic and Causality
Focus on the 'why' and 'how'. Analyze the
    logical flow of the conversation. Ask
    questions that uncover motivations,
    processes, and consequences. Connect
    ideas that were mentioned but not
    explicitly linked. Do chain-of-thought
    reasoning for each question.
```

Figure 13: EmotionDivergentAgent Specialization

```
Your Specialization: Emotional Depth
Focus on the feelings and emotions behind
    the words. Ask questions that explore
    the interviewee's emotional state,
    values, and personal significance of
    their experiences. Listen for subtext
    and unspoken feelings.
```

Figure 14: OutlineDivergentAgent Specialization

```
Your Specialization: Structured Progression
Your goal is to ensure the interview covers
    all essential topics from the outline.
    Ask questions that bridge the current
    conversation to uncovered,
    high-priority, or logically adjacent
    topics in the outline. Your questions
    should be inspired by the outline but
    phrased naturally in the context of the
    conversation.
```

Figure 15: PersonaDivergentAgent Specialization

```
Your Specialization: Role-playing
Think from the interviewee's perspective.
    Based on their persona (background,
    personality, goals), what question
    would they find most engaging or
    relevant? Ask questions that resonate
    with their stated experiences and
    character.
```

Figure 16: NoveltyDivergentAgent Specialization

```
Your Specialization: Creative Surprise
Your goal is to introduce novel angles and
    break patterns. Ask questions that are
    unexpected but still relevant. Think
    about metaphors, hypothetical
    scenarios, or connections to broader
    themes that haven't been touched upon.
    Challenge assumptions.
```

B.4 MAJI V3 Agents

Figure 17: MAJI V3: EditorInChiefAgent Prompt

```
You are the Editor-in-Chief of a dynamic
    interview system. Your role is to
    analyze the state of the conversation
    and devise a strategy for which *types*
    of questions to ask next. Based on the
    persona, summary, keywords, and outline
    coverage, generate a diverse and
    creative set of 2-4 divergent agent
    specifications. Each specification
    should include a unique, descriptive
    name and a clear set of instructions
    for that agent to follow.
```

B.5 Evaluation Prompts

This section contains the prompts used for evaluating the quality of generated questions, including both LLM-as-judge prompts and Prometheus evaluation rubrics.

B.5.1 LLM-as-Judge Prompts

Figure 18: QualitativeJudgeAgent Prompt

```
You are an expert conversational analyst.
    Your task is to evaluate a single
    proposed interview question based on
    the conversation's context. Provide
    scores from 0.0 to 1.0 for the
    following two subjective qualities:
1. Conversational Flow: Does the question
    feel like a natural, smooth
    continuation of the dialogue, or is it
    abrupt and jarring?
2. Elaboration: Does the question encourage
    the interviewee to provide a detailed,
    in-depth, and comprehensive answer,
    rather than a short or simple one?
```

```
Your output MUST be a single JSON object
     with the keys: `flow`, `elaboration`.
```

Figure 19: PersonaJudgeAgent Prompt

```
You are an expert profiler and interviewer.
     You will be given an interviewee's
     persona and a proposed question. Your
     task is to evaluate how well the
     question aligns with the interviewee's
     stated background, personality, and
     goals. A high score means the question
     would be engaging, relevant, and
     interesting *to this specific person*.
     A low score means it is too generic,
     irrelevant, or misaligned with their
     character.

Your output MUST be a single JSON object
     with the key: `alignment_score` (a
     float from 0.0 to 1.0).
```

Figure 20: CoherenceJudgeAgent Prompt

```
You are an expert in discourse analysis.
     You will be given the last question
     asked, the answer given, and a new
     proposed question. Your task is to
     evaluate the logical and thematic
     coherence of the new question as a
     follow-up. A high score means the
     question is a sensible, well-connected
     continuation of the dialogue. A low
     score means it feels abrupt, random,
     disconnected, or ignores the context of
     the previous answer.

Your output MUST be a single JSON object
     with the key: `coherence_score` (a
     float from 0.0 to 1.0).
```

Figure 21: InsightJudgeAgent Prompt

```
You are an expert conversation analyst.
     Your task is to categorize a proposed
     interview question based on the full
     context of the interview history.
     Analyze how the question relates to the
     entire dialogue, not just the last turn.

Categories:
- `Connecting`: The question links the
     current topic to a significantly
     earlier part of the conversation (more
     than 2-3 turns ago).
- `Challenging`: The question identifies
     and probes a potential contradiction,
     inconsistency, or assumption in the
     interviewee's statements.
- `Motivational`: The question explores the
     deep-seated 'why' behind an answer,
     focusing on core values, goals, or
     driving forces.
- `Hypothetical`: The question poses a
     creative 'what if' scenario to explore
     the interviewee's principles or
     thinking process.
- `SurfaceLevel`: A standard, logical
     follow-up that explores the immediate
     topic but lacks a deeper connection or
     creative angle.

Your output MUST be a single JSON object
     with the keys: `insight_category` and
     `reasoning`.
```

Figure 22: PlanEvaluatorAgent Prompt

```
You are an expert evaluator of AI agent
     systems. Your task is to assess the
     quality of a *plan* for generating
     interview questions, not the questions
     themselves. The plan consists of a list
     of specialist agents that will be
     created to handle the current
     situation. Evaluate the plan based on
     the conversational context.

1. Plan Relevance: How well does the chosen
     set of agents address the immediate
     needs of the conversation? (e.g., if
     the user is being emotional, is there
     an 'Emotion' agent planned?).
2. Plan Creativity: How creative is the
     plan? Does it propose novel specialists
     to find unique angles, or is it a
     generic, boilerplate plan?

Your output MUST be a single JSON object
     with the keys: `plan_relevance` and
     `plan_creativity`.
```

Figure 23: CategorizerAgent Prompt

```
You are an expert in conceptual analysis.
     You will be given a list of interview
     questions. Your task is to assign a
     single, concise conceptual category
     label to each question. For example,
     'Career Motivation', 'Work-Life
     Balance', 'Technical Skills'. You MUST
     return a list of strings, where each
     string is the category for the
     corresponding input question. The list
     must have the same number of items as
     the input list.
```

B.5.2 Prometheus Evaluation Rubrics

The following rubrics were used with the Prometheus 2 evaluation model to provide standardized, third-party assessment of question quality.

Figure 24: Context Relevance Rubric

```
Criteria: How well does the question
     logically follow from the interviewee's
     previous answer?

Score Descriptions:
1: Not at all relevant.
2: Slightly relevant.
```

```
3: Moderately relevant.
4: Relevant.
5: Highly relevant.
```

Figure 25: Insight Rubric

```
Criteria: Does the question probe deeper,
    encouraging novel reflection?

Score Descriptions:
1: Surface-level.
2: Asks for basic elaboration.
3: Encourages some reflection.
4: Prompts connection of ideas.
5: Deeply insightful.
```

Figure 26: Strategic Progression Rubric

```
Criteria: Does the question creatively
    bridge the current dialogue with the
    intended interview structure (outline),
    or does it just bluntly repeat an
    outline point?

Score Descriptions:
1: Completely ignores or contradicts the
    outline's direction.
2: Bluntly asks a question from the outline
    without connecting it to the
    conversation.
3: Loosely connects to an outline topic but
    the transition is awkward.
4: Smoothly transitions to an outline
    topic, clearly building on the last
    answer.
5: Artfully weaves an outline topic into
    the conversation, making the transition
    feel both natural and strategic.
```

Figure 27: Persona Alignment Rubric

```
Criteria: How well-suited is this question
    to the interviewee's specific
    background, expertise, and known
    interests as described in their
    persona? A good question is tailored to
    elicit a unique and insightful answer
    based on the interviewee's specific
    experiences.

Score Descriptions:
1: Generic question, irrelevant to the
    interviewee's specific persona.
2: Vaguely related to the interviewee's
    field, but not tailored to their
    specific role or accomplishments.
3: Asks about a topic relevant to the
    interviewee, but it's a standard
    question that doesn't probe their
    unique expertise.
4: The question is well-tailored, touching
    on specific aspects of the
    interviewee's known experience or
    expertise.
5: Excellent question that targets the core
    of the interviewee's unique expertise
    or perspective, making it highly likely
    to elicit a novel and insightful
    response.
```

Figure 28: Conversational Synthesis Rubric

```
Criteria: Does the question connect the
    interviewee's most recent answer with
    earlier parts of the conversation,
    weaving together themes, or does it
    treat each turn as an isolated event?

Score Descriptions:
1: Feels completely disconnected from the
    rest of the conversation history.
2: Vaguely references something said
    earlier, but the connection is weak.
3: Makes a simple, direct link to an
    immediately preceding turn.
4: Connects the current answer to a broader
    theme discussed earlier in the
    conversation.
5: Masterfully synthesizes multiple points
    from the conversation history to create
    a deeply contextualized and insightful
    question.
```

Figure 29: Perspective Diversity Rubric

```
Criteria: How diverse are these questions
    in their angle of approach and topic?
    Do they explore different facets of the
    previous answer, or are they all very
    similar to each other?

Score Descriptions:
1: All questions are essentially
    rephrasings of the same core idea.
2: Most questions are similar, with only
    minor variations in phrasing.
3: Some questions show different angles,
    but most are still on the same theme.
4: The questions explore a good variety of
    different topics and perspectives.
5: The questions are highly diverse, each
    approaching the conversation from a
    unique and creative angle.
```

Figure 30: Relative Comparison Rubric

```
Criteria: Which of the two proposed
    questions is a better, more insightful,
    and more natural follow-up to the
    conversation?

This rubric is used for pairwise
    comparisons between questions from
    different systems, with the judge
    selecting either response A, response
    B, or declaring a tie.
```