CtrlNews: LLM-based Multi-Agent Controllable News Writing via Knowledge Gravitational Field

Yifei Xu[1*,4], Yingjie Zong[1], Zhonghua Wang[2], Sirui Wu[2], Yuan Rao[1], Dan Zhang[2], Shuiguang Deng[3]
[1]School of Software, Xi'an Jiaotong University
[2]China Mobile Communications Group Shaanxi Co.,Ltd.
[3]College of Computer Science and Technology, Zhejiang University
[4]Xi'an Jiaotong University Suzhou Institute

Correspondence: belonxu_1@xjtu.edu.cn

Abstract

News writing empowered by large language models (LLMs) has emerged as a prevalent trend due to their efficiency and scalability. This paradigm necessitates dynamic information acquisition, knowledge structuring, and precise viewpoint articulation. However, current approaches often rely on superficially retrieved information and oversimplified knowledge enumeration, resulting in shallow, repetitive, and unordered outputs. Additionally, the lack of controllability over narrative viewpoints fails to align with user-defined preferences. To address these limitations, we propose an LLM-based multi-agent controllable news writing framework termed CtrlNews. The framework simulates expert questioning through automated role assignment and question generation followed by a three-layer hierarchical gravitational graph iteratively refined via expansion-reflection cycles. Besides, we elaborate a fine-grained viewpoint control mechanism to precisely regulate bias, emotion, and exaggeration attributes. When composing long-form news articles, the controlled viewpoints are extended via emotion-preserving composition and self-reflection refinement to ensure the consistency of viewpoint control and prevent the dilution of the control effect. Experiments on quality and control effect evaluation, news dissemination effect assessment, and human evaluation demonstrate significant improvements across multiple metrics compared to existing methods. [1]

1 Introduction

News writing is a laborious and time-consuming process that requires meticulous research, fact-checking, and multiple rounds of revisions to ensure accuracy, readability, and alignment with editorial standards (Simon, 2024; Gao et al., 2024).

[1]The code is available at https://github.com/ab2ence/CtrlNews.

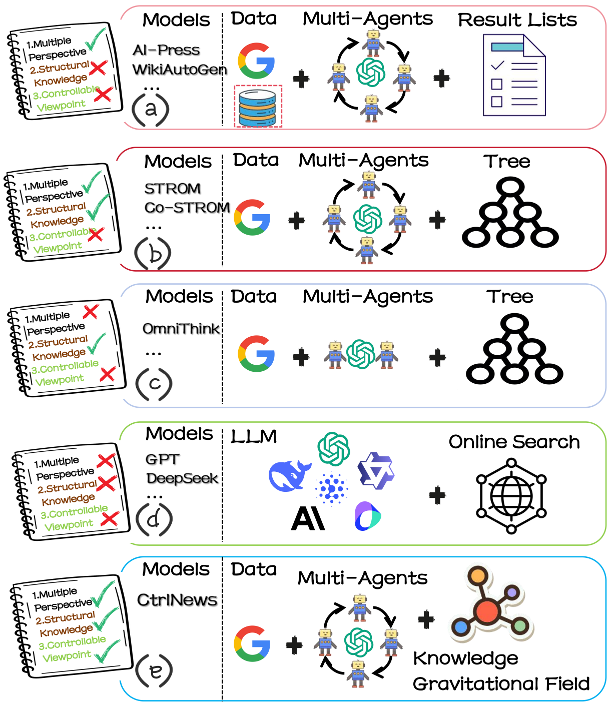


Figure 1: Previous machine writing approaches have primarily focused on expanding new information or perspectives through Retrieval-Augmented Generation (RAG) and role-playing. both tree and list structures are inadequate for representing complex relational knowledge. In contrast, CtrlNews learns multi-perspective knowledge through perspective-guided knowledge acquisition and organizes this knowledge into a hierarchical gravitational graph.

Human journalists must navigate complex information landscapes, synthesize diverse perspectives, and adapt content for different audiences (Petridis et al., 2023). While news writing traditionally relies on first-hand reporting such as interviews and on-the-ground investigations, automated frameworks are not intended to replace these essential practices. Instead, they can complement journalistic workflows by accelerating draft generation, diversifying perspectives, and grounding content with retrieval or fact databases. AI-Press (Liu et al., 2025)

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 8668–8705
November 4-9, 2025

has already demonstrated this auxiliary role. Recent advancements in LLM-based multi-agent systems (MAS) further indicate a strong synergy with news production pipelines (Chen et al., 2024; Liu et al., 2025), showing that these architectures can be effectively aligned with journalistic workflows. When appropriately architected, these systems can exhibit specialization, wherein individual agents autonomously develop expertise in specific subtasks such as information retrieval, content composition, and factual verification (Hong et al., 2024; Wu et al., 2024b; Li et al., 2023).

As illustrated in Fig.1, LLMs possess rich parametric knowledge that enables direct generation of article outlines and even complete drafts (Wu, 2024; Gao et al., 2024). However, they often suffer from a lack of professionalism and hallucination issues (Maynez et al., 2020; Sriramanan et al., 2024), particularly in journalistic contexts requiring high accuracy. Retrieval-augmented generation (RAG) methods attempt to mitigate these issues by incorporating external knowledge (Jiang et al., 2023; Huang and Huang, 2024) yet struggle to retrieve comprehensive information through simple search queries (Ram et al., 2023; Li et al., 2025). Recently, AI-Press (Liu et al., 2025) leverages multi-agent collaboration and RAG to rapidly generate news drafts, prioritizing journalistic efficiency. However, its reliance on existing sources limits novelty and impedes deep analysis of complex events (Mao et al., 2024). Meanwhile, STORM (Shao et al., 2024), Co-STORM (Jiang et al., 2024), and WikiAutoGen (Yang et al., 2025a) collectively enhance automated writing through multi-perspective questioning, collaborative discourse simulation, and content integration to improve article coherence, depth, and engagement. However, these approaches often lack adaptive mechanisms to dynamically balance thematic focus and exploratory breadth during iterative expansion (He et al., 2023; Lu et al., 2024). Although OmniThink (Xi et al., 2025) aims to expand knowledge boundaries through iterative cognitive refinement and dynamic information tree growth, its unchecked expansion causes thematic divergence that reveals the fundamental challenge of balancing exploratory richness with conceptual focus (Yang et al., 2025a). Despite demonstrating promising capabilities in generating news-style articles, these methods remain hindered by several critical shortcomings:**1) Shallow multi-agent collaboration**: Existing multi-agent frameworks prioritize breadth over depth but lack the domain-specific expertise needed for news writing standards (Petridis et al., 2023; Aljalabneh et al., 2024). **2) Ineffective knowledge structuring**: News requires nuanced perspectives, but tree or list-based structures cannot adequately represent complex reasoning (Nishal and Diakopoulos, 2024; Pérez-Seijo et al., 2023). **3) Uncontrolled viewpoint dilution**: In the absence of preset stance, emotion, and facts, generated articles often lack coherence and weaken journalistic integrity (Pérez-Seijo et al., 2023; Yeh et al., 2024).

To address these limitations, we present an automated news writing framework termed CtrlNews to write structured and controllable news articles. We first employ specialized LLM experts who conduct multi-round, domain-specific questioning combined with comprehensive search agent retrieval to ensure both depth and breadth of coverage. Moving beyond simplistic knowledge structures, we introduce a hierarchical gravitational graph that dynamically organizes information through expansion-reflection cycles to enable nuanced viewpoint derivation through core, intermediate, and peripheral knowledge layers. To prevent viewpoint dilution, we develop a precision control mechanism that regulates narrative attributes while maintaining consistency through iterative verification, followed by emotion-preserving composition and self-reflection refinement to produce high-quality news articles.

The main contributions of this paper are as follows:

- We present CtrlNews, the first automated framework for generating high-quality news articles with fine-grained controllability by integrating multi-agent LLM experts with a hierarchical gravitational graph.
- We develop a precision control mechanism to enforce narrative consistency via iterative verification, and we devise emotion-aware composition and self-refinement modules to produce new articles that align with predefined preferences.
- The experimental results on quality and control effect evaluation, news dissemination effect assessment, and human evaluation reveal substantial improvements across multiple metrics when compared to existing methods.

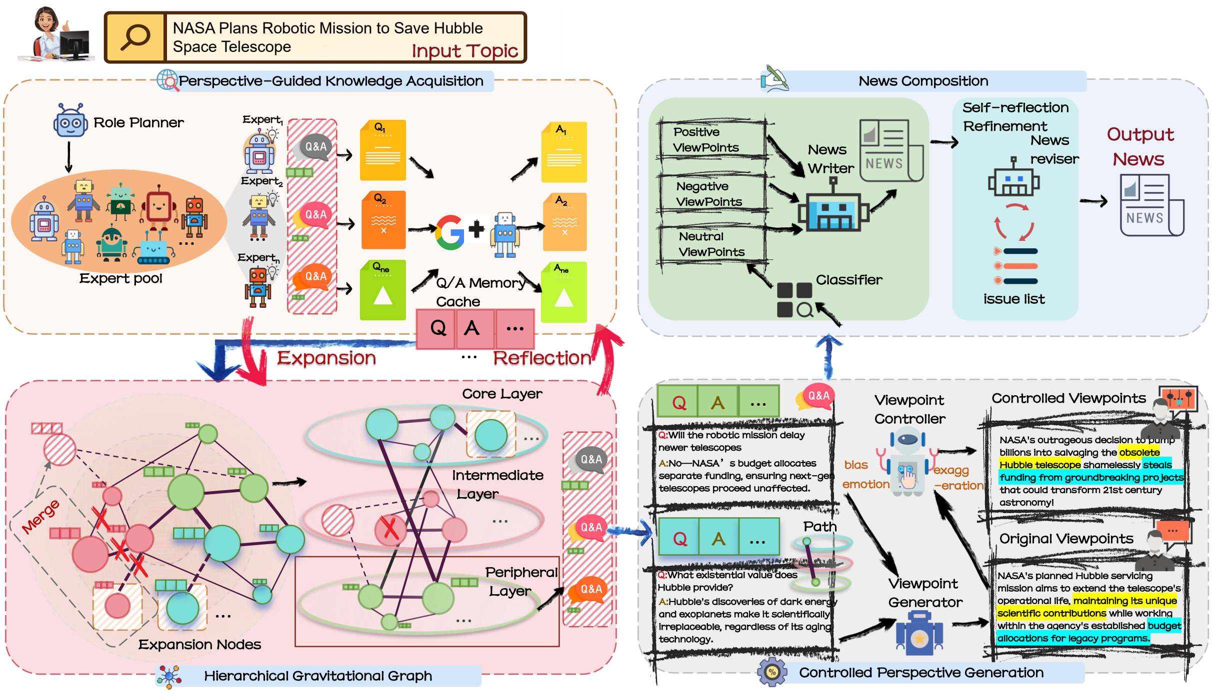


Figure 2: Overview of CtrlNews, our controllable news generation framework. The pipeline consists of four key stages: (1) Perspective-Guided Knowledge Acquisition, where specialized agents generate domain-specific questions and retrieve comprehensive answers; (2) Hierarchical Gravitational Graph, which organizes knowledge into core/intermediate/peripheral layers using physics-inspired mass and force calculations; (3) Controlled Viewpoint Generation, enabling fine-grained adjustment of bias, emotion and exaggeration with consistency verification; and (4) News Composition, where partitioned viewpoints are expanded into structured articles through emotion-preserving drafting and self-reflection refinement. The framework's iterative reflection-extension cycle and multi-layer knowledge organization ensure both depth of analysis and precise control over journalistic narrative viewpoints.

2 Related Works

2.1 Automatic Writing

Recent advancements in LLMs have significantly enhanced the capabilities of automatic expository writing systems (Wu et al., 2024a), enabling the generation of long-form, well-structured, and coherent texts (Wan et al., 2025). These writing systems aim to automate the process of knowledge collection, synthesis, and organization to make information more accessible and up-to-date (Pirolli, 2009; Fernández-Pichel et al., 2024). Several key functionalities have been proposed to address the challenges of long-form writing. For instance, Shao et al. (Shao et al., 2024) introduced STORM to automate the pre-writing stage by conducting research and creating outlines through role-guided question asking and multi-perspective conversations, which leveraged the diverse perspectives of different roles to generate comprehensive and organized content. Co-STORM (Jiang et al., 2024) involved collaborative discourse among multiple LLM agents to explore and discover unknown information, facilitating a dynamic and exploratory writing process that goes beyond the initial query. By emulating human slow-thinking processes, OmniThink (Xi et al., 2025) employed an information tree and a conceptual pool to iteratively expand and reflect on the collected information for improving the knowledge density of the generated articles. For the purpose of generating Wikipedia-style articles, WikiAutoGen (Yang et al., 2025a) integrates both visual and textual content to enhance the depth and engagement of the generated articles through a multi-perspective self-reflection mechanism. In conclusion, these contributions collectively highlight the ongoing efforts to improve the quality, coherence, and reliability of automatically generated expository texts, pushing the boundaries of what is possible in the field of natural language processing.

2.2 Large Language Models

LLMs have witnessed remarkable advancements driven by architectural innovations and parameter scaling strategies (Naveed et al., 2023; Raiaan et al., 2024). These developments have significantly enhanced LLM capabilities and performance for more sophisticated language understanding and generation (Kaplan et al., 2020; Muennighoff et al., 2023). The GPT series (GPT-3 (Brown et al., 2020) to GPT-4 (Achiam et al., 2023)) exemplified continuous capability expansion via parametric scaling, while Meta's LLaMA (Touvron et al., 2023) family enhanced research accessibility through open-weight architectures. The DeepSeek-V3 (Liu et al., 2024) employed dynamic parameter activation to achieve comparable performance to leading systems like GPT-4 with less training resource consumption. Concurrently, DeepSeek-R1 (Guo et al., 2025) pioneered multi-stage reinforcement learning frameworks to attain state-of-the-art (SOTA) performance on comprehensive reasoning benchmarks. The Gemini family (Team et al., 2023) demonstrated exceptional multimodal understanding by being trained jointly across image, audio, video, and text modalities. In practical applications ranging from web navigation to multimodal task completion, ChatGLM (Zeng et al., 2024) showcased its alignment breakthroughs via a multi-phase optimization framework combining supervised fine-tuning with human feedback protocols. The Qwen series of models (Bai et al., 2023) demonstrated strong performance across diverse NLP tasks through its optimized transformer architecture and large-scale pretraining. The demonstrated capabilities of LLMs in linguistic comprehension, interpretation, and generation have facilitated the pervasive integration of LLM-generated text across diverse domains, such as professional, educational, and social contexts (Wu et al., 2025a).

3 Methodology

We present CtrlNews for automating news writing by first retrieving and synthesizing comprehensive information concerning a given input topic t through multi-round questioning of distinct LLM experts along with the LLM searching agent (Section 3.1). Then, the collected information is organized using a hierarchical gravitational graph that simulates human information gathering and enables iterative cognitive structuring (Section 3.2). Next, our controlled viewpoint mechanism provides precise manipulation of specific nodes within the graph, which employs a control consistency verification to both ensure alignment with predefined controlled viewpoint attributes and mitigate LLM hallucination (Section 3.3). Finally, the controlled viewpoints are extended to a full-length news article via emotion-preserving composition and self-reflection refinement (Section 3.4). Fig.2 gives an overview of CtrlNews with detailed pseudo-code available in Appendix A. The prompts of all agents are in Appendix B.

3.1 Perspective-Guided Knowledge Acquisition

Diverse perspectives are essential in news generation as they offer multi-dimensional analysis of complex events and mitigate systematic reporting bias. Furthermore, these varied perspectives serve as prior knowledge that facilitates more profound questioning. Following the human process of acquiring knowledge and updating cognitive frameworks (Wu et al., 2025b), we integrate expert-driven questioning with web-sourced information retrieval to provide a comprehensive view of a given topic t. Specifically, the **Role Planner** first automatically creates an expert pool comprising n unique expert questioners ($n > 1$), where each functions as an autonomous agent with role-specific objectives assigned by the LLM. Formally, this expert pool is represented as: $\mathcal{P}^t = \{(Name_i, Des_i, Focus_i)\}_{i=1}^{n}$, where $Name_i$, Des_i, and $Focus_i$ denote the expert's identifier, description, and topical focus area, respectively. Subsequently, CtrlNews selects n_e ($n_e \leq n$) relevant experts from $\mathcal{P}^t$. To determine n_e relevant experts, we adopt a cosine similarity–based matching criterion (threshold = 0.7) between topic descriptions and expert domains. Each selected expert $\mathcal{T}_j$ ($1 \leq j \leq n_e$) may generate multiple questions pertaining to the specified topic t, denoted as:

$$\mathcal{Q}_j(t) = \{Q_{j,k}(t) \mid 1 \leq k \leq m_j\}, \quad (1)$$

where m_j is the number of questions from expert $\mathcal{T}_j$. The **Searcher** component then processes each question through query engines (e.g., Google or Bing) to retrieve corresponding answers:

$$\mathcal{A}_j(t) = \{A_{j,k}(t) \mid Q_{j,k}(t) \in \mathcal{Q}_j(t)\}. \quad (2)$$

To facilitate efficient knowledge reuse and guide subsequent processing, we maintain a dynamic topic memory cache structured as:

$$\mathcal{M}^t = \left\{(Q_{j,k}(t), A_{j,k}(t), \mathcal{T}_j) \mid {}^{1\leq j\leq n_e}_{1\leq k\leq m_j}\right\}, \quad (3)$$

where each tuple contains a question, its answer, and the originating expert. More details can be found in Appendix A.1.

3.2 Hierarchical Gravitational Graph

Building on Network Agenda-Setting (NAS) theory (Guo and McCombs, 2011), which posits that media shape public perception by constructing cognitive networks of issues, we propose a hierarchical gravitational graph to dynamically quantify strength-weighted relationships among diverse knowledge perspectives, which consists of three key phases: **initialization**, **expansion**, and **reflection**.

Initialization Inspired by gravitational physics, we construct a knowledge graph $\mathcal{G} = (V, E)$ where the vertex set $V = \{v_{j,k} \mid (Q_{j,k}, A_{j,k}, \mathcal{T}_j) \in \mathcal{M}^t\}$ represents all cached Q/A-expert tuples. The mass for each vertex $v_j \in V$ is computed through the nonlinear transformation:

$$M(v_{j,k}) = \underbrace{(0.5 + rich(A_{j,k}))}_{\text{content}} \cdot \underbrace{(0.5 + sim(A_{j,k}, t))}_{\text{relevance}}, \tag{4}$$

where $rich(A_{j,k})$ quantifies the content richness of answer $A_{j,k}$, and $sim(A_j, t) \in [0, 1]$ represents the semantic relevance between answer $A_{j,k}$ and target topic t computed via cosine similarity in the embedding space (Wang et al., 2020; Huang et al., 2019). Specifically, the richness function is defined as:

$$rich(A_{j,k}) = \min\left(\max\left(1.0, \frac{600}{len(A_{j,k})}\right), \frac{len(A_{j,k})}{300}\right), \tag{5}$$

which balances overly short answers and excessively long answers, thereby normalizing richness within a reasonable range. Each edge in set E corresponds to the gravitational attraction between two nodes $v_{j,k}$ and $v_{p,q}$ determined by:

$$F(v_{j,k}, v_{p,q}) = G \cdot \frac{M(v_{j,k}) \cdot M(v_{p,q})}{d(v_{j,k}, v_{p,q})^2}, \tag{6}$$

where G is the gravitational constant. Similar to function sim, $d(v_{j,k}, v_{p,q})$ represents the semantic relevance between question $Q_{j,k}$ and $Q_{p,q}$. Following the hierarchical structure of agenda-setting theory (Guo and McCombs, 2011), we stratify $\mathcal{G}$ into core, intermediate, and peripheral layers to facilitate hierarchical viewpoint control. The core layer contains fundamental questions defining the topic's theoretical essence, the intermediate layer bridges theory to practice via transfer questions, while the peripheral layer hosts application-specific questions where concrete viewpoints emerge. The stratification is performed by an LLM that classifies each vertex $v_{j,k} \in V$ into layers based on its semantic representation in $\mathcal{G}$. More details can be found in Appendix A.2.

Expansion The expansion phase dynamically extends the gravitational graph $\mathcal{G} = (V, E)$ by processing existing question-answer pairs $\{(Q_{j,k}(t), A_{j,k}(t), \mathcal{T}_j)\} \in \mathcal{M}^t$ through an iterative knowledge augmentation protocol. Each expert $\mathcal{T}_j$ analyzes its previous answers $\{A_{j,k}(t)\}_{k=1}^{m_j}$ to determine whether to generate supplementary questions $\{Q^+_{j,r}(t)\}_{r=1}^{p_j}$ where $p_j \geq 0$ denotes new questions per expert, followed by the Searcher component retrieving corresponding answers $\{A^+_{j,r}(t)\}_{r=1}^{p_j}$ through query engines. These new pairs are incorporated as nodes $v^+_{j,r} = (Q^+_{j,r}(t), A^+_{j,r}(t), \mathcal{T}_j)$ into the memory cache $\mathcal{M}^t \leftarrow \mathcal{M}^t \cup \{v^+_{j,k}\}_{k=r}^{p_j}$. More details can be found in Appendix A.2.

Reflection The reflection phase optimizes $\mathcal{G}$ through node consolidation and structural reorganization. New nodes $v^+_{j,r}$ are merged with existing ones in $\mathcal{T}_j$ based on their QA-pair semantic similarity. All gravitational forces are then recomputed according to Eq. 6 followed by core/intermediate/peripheral layers reallocation. This reflection-expansion cycle terminates when either the graph structure stabilizes or the maximum number of iterations is reached, ensuring efficient convergence while preserving the hierarchical knowledge structure. More details can be found in Appendix A.2.

3.3 Controlled Viewpoint Generation

Based on the above hierarchical gravitational graph $\mathcal{G}$, we develop a controlled viewpoint generation mechanism that operates on the peripheral nodes $P \subset V$. The module regulates viewpoint attributes while maintaining consistency through control consistency verification. More implementation details can be found in Appendix A.3.

Viewpoint Formulation For each peripheral vertex $v'_r(t) \in P$, we extract its longest path $\mathcal{C}$, where L denotes the path length from periphery layer to core layer. The original viewpoint derivation is:

$$\text{orig\_viewpoint}_r(t) = f_{\text{agg}}(v'_r(t), \mathcal{C}), \tag{7}$$

where $f_{\text{agg}(\cdot,\cdot)}$ aggregates the vertexes in $\mathcal{C}$.

Attribute-Guided Viewpoint Regulation The viewpoint regulation incorporates three tunable attributes: (1) stance bias $b \in [-1, 1]$ ranging from strong opposition ($b = -1$) to strong support ($b = 1$); (2) emotional intensity $e \in [-1, 1]$ modulating affective tone; and (3) exaggeration level $x \in [0, 1]$ determining the degree of fact distortion. The attribute-guided viewpoint regulation is formulated as:

$$\text{re\_viewpoint}_r(t) = \text{LLM}_{vp}\big(\text{orig\_viewpoint}_r(t), b, e, x\big), \tag{8}$$

where $\text{LLM}_{vp}(\cdot)$ integrates the original viewpoint with the specified attribute parameters through prompt-based conditioning.

Control Consistency Verification To ensure attribute fidelity, we employ a control consistency verification (CCV) mechanism that measures the distance between preset and actual attributes:

$$\text{Score}_{\text{ccv}} = 1 - \frac{\|(b,e,x) - \text{LLM}_{\text{att}}(\text{re\_viewpoint}_r(t))\|_2}{2\sqrt{3}}, \tag{9}$$

where $\text{LLM}_{att}(\text{re\_viewpoint}_i)$ returns the actual attributes with LLM-based attribute extractor. The above regulation and verification process iterates until $\text{Score}_{\text{ccv}} \geq \tau_{ccv}$, where τ_{ccv} is a predefined viewpoint controlled threshold. As a result, the finalized controlled viewpoints are archived in $\mathcal{Z}$.

3.4 News Composition

After gathering the controlled viewpoints $\mathcal{Z}$ and their associated tripartite attributes, the next step is to expand these viewpoints to generate structured news articles with well-grounded and controlled viewpoints. More implementation details can be found in Appendix A.4.

Emotion-Preserving Composition The composition process begins with a semantic partitioning of controlled viewpoints based on their associated emotional attributes. Formally, we define the partition as:

$$\mathcal{Z} = \mathcal{Z}^+ \cup \mathcal{Z}^0 \cup \mathcal{Z}^-, \tag{10}$$

where $\mathcal{Z}^+$ represents positively slanted viewpoints, $\mathcal{Z}^-$ contains negative viewpoints, and $\mathcal{Z}^0$ comprises neutral perspectives. Specifically, the tripartite partitioning is determined by a fixed emotional intensity attribute $e \in [-1, 1]$, where $\mathcal{Z}^+$ contains all viewpoints with $e > 0$, $\mathcal{Z}^0$ strictly corresponds to $e = 0$, and $\mathcal{Z}^-$ includes all viewpoints with $e < 0$. For each non-empty partition $\mathcal{Z}^s$ where $s \in \{+, 0, -\}$, the writing agent generates corresponding content segments:

$$T_s = \text{LLM}_{comp}(\mathcal{Z}^s, Prompt_s), \tag{11}$$

Here, LLM_{comp} is used for writing the news draft with $Prompt_s$, and $Prompt_s$ represents the style-specific prompt template that enforces both emotional consistency and journalistic conventions. This partitioning approach prevents emotional dilution during content integration while preserving the original emotional setting of each viewpoint.

Self-Reflection Refinement To enhance the accuracy of emotional control, we develop a self-reflection refinement approach for article revision. This mechanism ensures rigorous viewpoint coverage and attribute fidelity through three key steps: **1) Coverage Verification**. The reviser agent checks whether all predefined viewpoints are fully included in the generated article. **2) CCV**. For each extracted viewpoint that matches a predefined one, the agent computes its $\text{Score}_{\text{ccv}}$ that measures the alignment between target attributes and realized attributes in the article. **3) Iterative Revision**. The reviser agent compiles an issue list of missing or inconsistent viewpoints and iteratively revises the article using the LLM. The process terminates when either all viewpoints meet coverage and CCV requirements or the maximum allowed revisions are reached.

4 Experiment

4.1 Experiment Setup

Dataset We conducted the experiments using 40 topics from the AG News dataset (Zhang et al., 2015), which covers four domains: World, Business, Sports, and Sci/Tech. From each domain, we randomly selected ten topics. To evaluate the dissemination effects, we employed 36 human-like agents simulated by OASIS (Yang et al., 2025b). Additional topic details are provided in the Appendix C.1.

Baselines To systematically evaluate the performance of CtrlNews in controlled news generation, we conduct comparative experiments with 12 state-of-the-art baselines across three categories: 1) Commercial LLMs (ChatGLM4-plus (Zeng et al., 2024), QwenMax (Bai et al., 2023), Claude 3.7, etc.), 2) Domain-specific framework (AI-Press (Liu et al., 2025) for news writing), and 3) Open-domain content generation frameworks (STORM (Shao et al., 2024), Co-STORM (Jiang et al., 2024), etc.).

Implementation Details Our framework is compatible with various search engines and LLM. From

Table 1: Comparative results of controlled news generation performance (TD: Topic Development; IV: Intellectual Value). Quality metrics are evaluated by DeepSeek-V3 with scores normalized to 0-10 scale.

Method \ Metrics	Relevance	Breadth	Depth	Novelty	Coherence	Expression	TD	IV	Total	Coverage	$\text{Score}_{\text{ccv}}$	TCPI
ChatGLM4-plus	8.71	7.65	6.76	5.85	8.36	7.69	7.65	6.68	59.35	-	-	0.59
Claude 3.7 Sonnet	8.71	7.75	7.10	6.13	8.35	7.86	7.81	6.90	60.61	-	-	0.67
DeepSeek-R1	8.83	7.84	7.05	6.30	**8.65**	8.00	7.91	7.01	61.59	-	-	0.61
DeepSeek-V3	**8.88**	7.85	6.96	6.15	**8.65**	7.86	7.90	6.90	61.15	-	-	0.53
Gemini 2.5 pro preview	8.71	7.60	7.00	6.08	8.34	7.79	7.69	6.81	60.01	-	-	0.62
GPT4o	8.81	7.74	7.00	6.30	8.60	7.95	7.85	6.94	61.19	-	-	0.66
QwenMax	8.58	7.65	6.78	5.89	8.28	7.70	7.63	6.64	59.13	-	-	0.61
AI-Press	8.26	7.43	6.89	5.93	8.41	7.95	7.35	6.65	58.86	-	-	0.66
STORM	8.54	7.90	7.33	6.18	7.84	7.89	7.88	7.09	60.63	-	-	0.52
Co-STORM	8.28	7.71	7.46	6.23	7.60	7.46	7.66	7.01	59.41	-	-	0.68
OmniThink	8.25	7.84	7.23	6.00	7.46	7.38	7.63	6.90	58.68	-	-	0.61
WikiAutoGen	8.58	8.00	7.38	6.28	7.71	7.50	7.83	7.14	60.40	-	-	0.69
CtrlNews	8.86	**8.14**	**8.29**	**6.86**	8.57	**8.71**	**8.57**	**8.00**	**66.00**	**1**	**0.91**	**0.73**

Section 3.1 to Section 3.3, all LLM-based components or agents consistently employed DeepSeek-V3 (Liu et al., 2024). In Section 3.4, DeepSeek-R1 (Guo et al., 2025) is adopted for news composition. The available search engines include Bing, SerpAPI, and the free option, DuckDuckGo. More details are in the Appendix C.2.

Quality Evaluation To quantitatively assess the quality of the news article, we implement an LLM-based evaluation framework with eight carefully designed dimensions: *Relevance*, *Breadth*, *Depth*, *Novelty*, *Coherence*, *Language Expression*, *Topic Development*, and *Intellectual Value*. Several of these metrics (Relevance, Breadth, Depth, Novelty, and Coherence) are adapted from prior work including Omnithink (Xi et al., 2025) and AI-Press (Liu et al., 2025), while the latter three are newly introduced for news writing tasks. The assessment is conducted through structured prompting with DeepSeek-V3 (Liu et al., 2024) (temperature=0.3, max_tokens=2000), where each dimension is scored on a 10-point scale based on explicit rubrics. Following the established practice of using LLMs as evaluators in previous research (Zheng et al., 2023; Li et al., 2024), we adopt the same approach for our evaluation strategy that aligns well with human judgment. The detailed prompts can be found in Table 21.

Control Effect Evaluation The experiment quantifies the control effect from three dimensions: Coverage measures the retention rate of preset viewpoints; $\text{Score}_{\text{ccv}}$ evaluates the distance between preset and actual attributes; Type Control Precision Indicator (TCPI), computed as $\mathit{TCPI} = 1 - \frac{1}{2}\sum_{t\in\{pos,neu,neg\}} |d_t - d'_t|$, quantifies the deviation between the preset viewpoint sentiment polarity distribution $\vec{d} = (d_{pos}, d_{neu}, d_{neg})$ and the actual one $\vec{d'} = (d'_{pos}, d'_{neu}, d'_{neg})$ observed in generated content. The *TCPI* metric ranges from 0 to 1, with higher values indicating better alignment between the generated and preset viewpoint sentiment polarity distribution.

Supplementary Experiments In addition, we conduct a human evaluation, with the detailed methodology and results presented in Appendix D.4. To assess the validity and reliability of the experiments, we performed statistical significance testing. The implementation details and corresponding results are also provided in Appendix D.5.

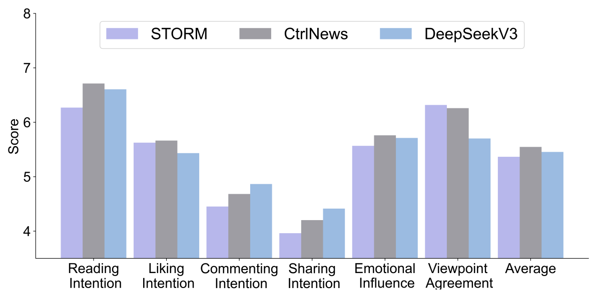


Figure 3: News dissemination effects comparison of STORM, CtrlNews, and DeepSeek-V3 across seven metrics: Reading Intention, Liking Intention, Commenting Intention, Sharing Intention, Emotional Influence, Viewpoint Agreement, and Average Score for young adults (aged 20-29).

4.2 Performance

Quality and Control Effect The evaluation employs both automated metrics (Coverage, $\text{Score}_{\text{ccv}}$ and TCPI) measuring viewpoint controllability and quality evaluation (*Relevance*, *Depth*, *Coherence*, etc.) scored by DeepSeek-V3. Table 1 shows CtrlNews achieves state-of-the-art news generation quality with particular strengths in *Breadth* and *Depth* through our perspective-guided knowledge acquisition and hierarchical gravitational graph. Our framework excels at producing novel insights and fluent expression as well as maintaining higher *Intellectual Value*. Although competitive in *Relevance* and *Coherence*, further optimization is needed for balancing multi-source integration when handling conflicting perspectives. In terms of viewpoint control, CtrlNews achieves robust viewpoint control via integrated verification and refinement mechanisms (Sec. 3.4). The system consistently aligns generated content with diverse user-specified preferences and maintains balanced viewpoint distributions across all required perspectives. Besides, to ensure a fair comparison of news generation quality across all methods, we conduct controlled experiments by disabling all viewpoint control components to evaluate each model's baseline writing capabilities (See Appendix D.1 and D.2 for full implementation details).

News Dissemination Effect High-quality news content tends to have a greater impact and is more readily accepted by the public (Zhou, 2024). To quantitatively evaluate dissemination effects, we developed a multi-age social simulation framework with 36 human-like agents divided into three age groups: adolescents (10–19), young adults (20–29), and seniors (30+). These cognitively modeled agents simulate real-world news consumption patterns and provide 0–10 ratings on seven metrics: Reading Intention, Liking Intention, Commenting Intention, Sharing Intention, Emotional Influence, Viewpoint Agreement, and Average. Key findings for the young adult group are shown in Fig.3, and the analyses for adolescents and seniors are provided in Appendix D.3. Notably, CtrlNews demonstrates distinct advantages in news dissemination among youth audiences.

4.3 Ablation studies

4.3.1 Ablation on Core Components

To verify the effectiveness of each core component, we sequentially remove individual modules and report their performance in Table 2. The removal of the perspective-guided knowledge acquisition module leads to a consistent decline across all quality metrics, indicating that its absence disrupts the multi-perspective reasoning process. Furthermore, replacing HGG with a tree structure improves novelty but reduces coherence, whereas using a list enhances coherence at the expense of depth. Additionally, removing the self-reflection refinement results in a noticeable drop across control metrics including Viewpoint Coverage and $\text{Score}_{\text{ccv}}$/TCPI.

We further investigate the impact of alternative structural choices by replacing HGG with either list- or tree-based knowledge structures. As summarized in Table 2, the tree structure increases Novelty (6.86 ⇒ 7.25) and Intellectual Value (8.00 ⇒ 8.25) due to its capacity to explore additional informational branches. However, such unconstrained expansion also introduces topic drift that weakens both Coherence and Relevance. In contrast, the list structure adopts a simple sequential form that improves Coherence by keeping the topic more tightly focused. However, this simplicity diminishes Depth due to a reduced ability to capture different levels of information. The HGG framework employs expansion–reflection cycles that enable broad exploration while maintaining thematic focus, and this mechanism leads to balanced improvements across Breadth, Depth, Coherence, and Relevance. The above findings confirm that each core component in our CtrlNews plays a critical and irreplaceable role in achieving well-rounded performance.

4.3.2 Ablation on Different Control Strategies

In this section, we investigate the impact of different combinations of control strategies, and depict the performance in Table 3. As seen, both the quality metrics and control metrics remain relatively stable across different combinations of viewpoint sentiment polarities. This suggests that our control module effectively maintains consistent performance, demonstrating strong generalization ability even when applied to various sentiment configurations.

5 Conclusion

We present CtrlNews, a novel LLM-based multi-agent framework for controllable news generation. By simulating expert questioning through automated role interactions and employing a hierarchical gravitational graph with iterative refinement,

Table 2: Ablation study on the core components. P&G:Perspective-Guided.

Components					Quality Metrics									Control Metrics		
P&G Knowledge Acquisition	HGG	List	Tree	Self-reflection Refinement	Relevance	Breadth	Depth	Novelty	Coherence	Expression	TD	IV	Total	Coverage	$Score_{ccv}$	TCPI
✗	✓	✗	✗	✓	8.54	7.98	7.25	6.75	8.43	8.69	8.33	7.33	63.3	1	0.88	0.70
✓	✗	✓	✗	✓	8.43	7.66	7.93	7.00	**9.01**	8.50	8.45	7.84	64.82	1	0.89	0.69
✓	✗	✗	✓	✓	8.36	7.93	8.13	**7.25**	8.12	**8.96**	8.50	**8.25**	65.50	1	0.90	0.70
✓	✓	✗	✗	✗	**9.12**	8.03	8.13	6.58	8.60	8.66	8.45	7.84	65.41	0.82	0.83	0.63
✓	✓	✗	✗	✓	8.86	**8.14**	**8.29**	6.86	8.57	8.71	**8.57**	8.00	**66.00**	**1**	**0.91**	**0.73**

Table 3: Ablation study on different control strategies. TD: Topic Development; IV: Intellectual Value.

The combinations of viewpoint sentiment polarity					Quality Metrics									Control Metrics		
Extremely Positive	Positive	Natural	Negative	Extremely Negative	Relevance	Breadth	Depth	Novelty	Coherence	Expression	TD	IV	Total	Coverage	$Score_{ccv}$	TCPI
✓	✗	✗	✗	✓	8.85	8.12	7.88	7.79	8.43	8.56	8.43	7.92	65.98	1	0.89	0.71
✗	✓	✗	✗	✓	8.77	**8.23**	7.97	7.59	8.34	8.23	8.17	7.89	65.19	1	0.90	0.72
✗	✗	✓	✓	✗	8.64	8.10	7.86	**7.97**	8.31	8.58	8.39	7.73	65.58	1	0.87	0.69
✗	✗	✓	✗	✗	8.81	8.08	7.84	7.94	8.37	8.52	8.38	7.75	65.69	1	0.86	0.70
✗	✗	✗	✓	✗	8.73	8.02	7.89	6.99	8.52	8.63	8.48	7.54	64.80	1	0.90	0.69
✗	✓	✗	✗	✗	8.69	8.09	7.97	6.89	8.49	8.59	8.53	7.81	65.06	1	0.87	0.67
✓	✗	✓	✓	✗	8.54	7.93	8.23	6.99	8.32	8.60	8.39	7.64	64.64	1	0.84	0.69
✓	✗	✓	✗	✓	8.54	8.08	8.24	6.83	8.48	8.62	8.37	7.99	65.15	1	0.88	0.71
✗	✓	✓	✓	✗	**8.86**	8.14	**8.29**	6.86	**8.57**	**8.71**	**8.57**	**8.00**	**66.00**	**1**	**0.91**	**0.73**

our method achieves deeper knowledge structuring and more coherent article composition. The proposed fine-grained viewpoint control mechanism enables precise regulation of triplet attributes, while emotion-preserving composition and self-reflection refinement maintain consistency in long-form content. Experimental results validate the framework's superiority across multiple metrics, demonstrating its ability to produce higher-quality, more viewpoint-consistent news articles compared to existing methods.

Limitations

Although our approach demonstrates promising results, several limitations warrant further discussion. First, we observe an inherent conflict between control precision and output quality. Second, the inference latency caused by LLM API calls remains suboptimal for real-time applications due to the computational overhead of hierarchical reasoning. Finally, the current design does not support multilingual or multimodal inputs, which could be addressed through architecture extensions for future work.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (62572387), and Jiangsu Agricultural Science and Technology Innovation Fund (CX(24)3132),and Natural Science Basic Research Program of Shaanxi (Program No.2024JC-YBMS-498), and Xi'an Jiaotong University-China Mobile Communications Group Co., Ltd. Digital Government Joint Institute (XJTU-CMCC-QY202508012).

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Abdallah Aljalabneh, Hamzeh Aljawawdeh, Alia Mahmoud, Tahseen Sharadqa, and Ashraf Al-Zoubi. 2024. Balancing efficiency and ethics: The challenges of artificial intelligence implementation in journalism. *Intelligent Systems, Business, and Innovation Research*, pages 763–773.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901.

Weize Chen, Yusheng Su, Jingwei Zuo, Cheng Yang, Chenfei Yuan, Chi-Min Chan, Heyang Yu, Yaxi Lu, Yi-Hsin Hung, Chen Qian, Yujia Qin, Xin Cong, Ruobing Xie, Zhiyuan Liu, Maosong Sun, and Jie Zhou. 2024. Agentverse: Facilitating multi-agent collaboration and exploring emergent behaviors. In *International Conference on Learning Representations*, volume 2024, pages 20094–20136.

Marcos Fernández-Pichel, Juan C Pichel, and David E Losada. 2024. Search engines, llms or both? evaluating information seeking strategies for answering health questions. *arXiv preprint arXiv:2407.12468*.

Fan Gao, Hang Jiang, Rui Yang, Qingcheng Zeng, Jinghui Lu, Moritz Blum, Tianwei She, Yuang Jiang, and Irene Li. 2024. Evaluating large language models on wikipedia-style survey generation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5405–5418.

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

Lei Guo and Maxwell McCombs. 2011. Network agenda setting: A third level of media effects. In *Annual Conference of the International Communication Association, Boston, MA*.

Zhitao He, Pengfei Cao, Yubo Chen, Kang Liu, Ruopeng Li, Mengshu Sun, and Jun Zhao. 2023. LEGO: A multi-agent collaborative framework with role-playing and iterative feedback for causality explanation generation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9142–9163.

Sirui Hong, Mingchen Zhuge, Jonathan Chen, Xiawu Zheng, Yuheng Cheng, Jinlin Wang, Ceyao Zhang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, et al. 2024. Metagpt: Meta programming for a multi-agent collaborative framework. In *International Conference on Learning Representations*.

Haoyang Huang, Yaobo Liang, Nan Duan, Ming Gong, Linjun Shou, Daxin Jiang, and Ming Zhou. 2019. Unicoder: A universal language encoder by pretraining with multiple cross-lingual tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 2485–2494.

Yizheng Huang and Jimmy Huang. 2024. A survey on retrieval-augmented text generation for large language models. *arXiv preprint arXiv:2404.10981*.

Yucheng Jiang, Yijia Shao, Dekun Ma, Sina Semnani, and Monica Lam. 2024. Into the unknown unknowns: Engaged human learning through participation in language model agent conversations. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 9917–9955.

Zhengbao Jiang, Frank F Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023. Active retrieval augmented generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7969–7992.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

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

Ming Li, Jiuhai Chen, Lichang Chen, and Tianyi Zhou. 2024. Can llms speak for diverse people? tuning llms via debate to generate controllable controversial statements. *arXiv preprint arXiv:2402.10614*.

Xiaoxi Li, Jiajie Jin, Guanting Dong, Hongjin Qian, Yutao Zhu, Yongkang Wu, Ji-Rong Wen, and Zhicheng Dou. 2025. Webthinker: Empowering large reasoning models with deep research capability. *arXiv preprint arXiv:2504.21776*.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Xiawei Liu, Shiyue Yang, Xinnong Zhang, Haoyu Kuang, Libo Sun, Yihang Yang, Siming Chen, Xuan-Jing Huang, and Zhongyu Wei. 2025. Ai-press: A multi-agent news generating and feedback simulation system powered by large language models. In *Proceedings of the 31st International Conference on Computational Linguistics: System Demonstrations*, pages 63–82.

Li-Chun Lu, Shou-Jen Chen, Tsung-Min Pai, Chan-Hung Yu, Hung-yi Lee, and Shao-Hua Sun. 2024. Llm discussion: Enhancing the creativity of large language models via discussion framework and role-play. *arXiv preprint arXiv:2405.06373*.

Kelong Mao, Zheng Liu, Hongjin Qian, Fengran Mo, Chenlong Deng, and Zhicheng Dou. 2024. RAG-studio: Towards in-domain adaptation of retrieval augmented generation through self-alignment. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 725–735.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919.

Niklas Muennighoff, Alexander Rush, Boaz Barak, Teven Le Scao, Nouamane Tazi, Aleksandra Piktus, Sampo Pyysalo, Thomas Wolf, and Colin A Raffel. 2023. Scaling data-constrained language models.

Advances in Neural Information Processing Systems, 36:50358–50376.

Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Naveed Akhtar, Nick Barnes, and Ajmal Mian. 2023. A comprehensive overview of large language models. *arXiv preprint arXiv:2307.06435*.

Sachita Nishal and Nicholas Diakopoulos. 2024. Envisioning the applications and implications of generative ai for news media. *arXiv preprint arXiv:2402.18835*.

Sara Pérez-Seijo, Suzana Barbosa, and Paulo Nuno Vicente. 2023. Artificial intelligence in journalism: case study of the spanish, portuguese and brazilian news media systems. In *Blurring Boundaries of Journalism in Digital Media: New Actors, Models and Practices*, pages 261–274. Springer.

Savvas Petridis, Nicholas Diakopoulos, Kevin Crowston, Mark Hansen, Keren Henderson, Stan Jastrzebski, Jeffrey V Nickerson, and Lydia B Chilton. 2023. Anglekindling: Supporting journalistic angle ideation with large language models. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–16.

Peter Pirolli. 2009. Powers of 10: Modeling complex information-seeking systems at multiple scales. *Computer*, 42(3):33–40.

Mohaimenul Azam Khan Raiaan, Md Saddam Hossain Mukta, Kaniz Fatema, Nur Mohammad Fahad, Sadman Sakib, Most Marufatul Jannat Mim, Jubaer Ahmad, Mohammed Eunus Ali, and Sami Azam. 2024. A review on large language models: Architectures, applications, taxonomies, open issues and challenges. *IEEE Access*, 12:26839–26874.

Ori Ram, Yoav Levine, Itay Dalmedigos, Dor Muhlgay, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. 2023. In-context retrieval-augmented language models. *Transactions of the Association for Computational Linguistics*, 11:1316–1331.

Yijia Shao, Yucheng Jiang, Theodore A Kanell, Peter Xu, Omar Khattab, and Monica S Lam. 2024. Assisting in writing wikipedia-like articles from scratch with large language models. *arXiv preprint arXiv:2402.14207*.

Felix Simon. 2024. Artificial intelligence in the news: how ai retools, rationalizes, and reshapes journalism and the public arena. Technical report, Tow Center for Digital Journalism, Columbia University.

Gaurang Sriramanan, Siddhant Bharti, Vinu Sankar Sadasivan, Shoumik Saha, Priyatham Kattakinda, and Soheil Feizi. 2024. Llm-check: Investigating detection of hallucinations in large language models. *Advances in Neural Information Processing Systems*, 37:34188–34216.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Kaiyang Wan, Honglin Mu, Rui Hao, Haoran Luo, Tianle Gu, and Xiuying Chen. 2025. A cognitive writing perspective for constrained long-form text generation. *arXiv preprint arXiv:2502.12568*.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Advances in Neural Information Processing Systems*, 33:5776–5788.

Junchao Wu, Shu Yang, Runzhe Zhan, Yulin Yuan, Lidia Sam Chao, and Derek Fai Wong. 2025a. A survey on llm-generated text detection: Necessity, methods, and future directions. *Computational Linguistics*, pages 1–66.

Junchao Wu, Runzhe Zhan, Derek F. Wong, Shu Yang, Xinyi Yang, Yulin Yuan, and Lidia S. Chao. 2024a. Detectrl: Benchmarking llm-generated text detection in real-world scenarios. In *Advances in Neural Information Processing Systems*, volume 37, pages 100369–100401. Curran Associates, Inc.

Junde Wu, Jiayuan Zhu, and Yuyuan Liu. 2025b. Agentic reasoning: Reasoning llms with tools for the deep research. *arXiv preprint arXiv:2502.04644*.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, et al. 2024b. Autogen: Enabling next-gen llm applications via multi-agent conversation. In *ICLR 2024 Workshop on Large Language Model (LLM) Agents*.

Yonghui Wu. 2024. Large language model and text generation. In *Natural Language Processing in Biomedicine: A Practical Guide*, pages 265–297. Springer.

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. *arXiv preprint arXiv:2501.09751*.

Zhongyu Yang, Jun Chen, Dannong Xu, Junjie Fei, Xiaoqian Shen, Liangbing Zhao, Chun-Mei Feng, and Mohamed Elhoseiny. 2025a. Wikiautogen: Towards multi-modal wikipedia-style article generation. *arXiv preprint arXiv:2503.19065*.

Ziyi Yang, Zaibin Zhang, Zirui Zheng, Yuxian Jiang, Ziyue Gan, Zhiyu Wang, Zijian Ling, Jinsong Chen, Martz Ma, Bowen Dong, Prateek Gupta, Shuyue Hu, Zhenfei Yin, Guohao Li, Xu Jia, Lijun Wang, Bernard Ghanem, Huchuan Lu, Chaochao Lu, et al. 2025b. Oasis: Open agent social interaction simulations with one million agents. *arXiv preprint arXiv:2411.11581*.

Catherine Yeh, Gonzalo Ramos, Rachel Ng, Andy Huntington, and Richard Banks. 2024. Ghostwriter: Augmenting collaborative human-ai writing experiences through personalization and agency. *arXiv preprint arXiv:2402.08855*.

Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Dan Zhang, Diego Rojas, Guanyu Feng, Hanlin Zhao, Hanyu Lai, Hao Yu, Hongning Wang, Jiadai Sun, Jiajie Zhang, Jiale Cheng, Jiayi Gui, Jie Tang, Jing Zhang, Jingyu Sun, et al. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *arXiv preprint arXiv:2406.12793*.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. *Advances in Neural Information Processing Systems*, 28.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623.

Wen Zhou. 2024. Dilemma and coping strategies of news communication based on artificial intelligence and big data. *Heliyon*, 10(3):e25398.

A Detailed Pseudo-Code

A.1 Perspective-Guided Knowledge Acquisition

As detailed in Section 3.1, this algorithm facilitates perspective knowledge acquisition through expert-guided questioning and web retrieval. Key components include: (1) the automatic creation of an expert pool with specialized roles, (2) the question generation by selected domain experts, and (3) the construction of a dynamic topic memory cache that links answers to their corresponding questions and experts for subsequent analysis.. The pseudocode for the method is given in Algorithm 1.

A.2 Hierarchical Gravitational Graph

As mentioned in Section 3.2, the HGG consists of three key phases.

During the initialization phase, a knowledge graph is constructed from the dynamic topic memory cache. This process converts question-answer pairs into graph nodes with computed mass values and establishes gravitational relationships between them. Based on the definitions of the core, intermediate, and peripheral layers, the LLM is then used to assign each node to its corresponding layer. The resulting gravitational graph structure is saved in JSON format. Further details are provided in Algorithm 2.

Algorithm 1 Perspective-Guided Knowledge Acquisition

Require: Topic t, expert pool size n, selected experts n_e
Ensure: Dynamic memory cache $\mathcal{M}^t$
Initialize expert pool $\mathcal{P}^t \leftarrow \emptyset$
for $i = 1$ **to** n **do**
  Generate expert profile: $(Name_i,\ Des_i,\ Focus_i) \leftarrow$ RolePlanner(t)
  $\mathcal{P}^t \leftarrow \mathcal{P}^t \cup \{(Name_i, Des_i, Focus_i)\}$ {Expert pool creation}
end for
Select relevant experts: $\{\mathcal{T}_j\}_{j=1}^{n_e} \leftarrow$ FilterExperts$(\mathcal{P}^t, t)$
Initialize $\mathcal{M}^t \leftarrow \emptyset$
for each expert $\mathcal{T}_j \in \{\mathcal{T}_j\}_{j=1}^{n_e}$ **do**
  Generate questions: $\mathcal{Q}_j(t) \leftarrow \{Q_{j,k}(t)\}_{k=1}^{m_j}$ {m_j questions per expert}
  for each question $Q_{j,k} \in \mathcal{Q}_j(t)$ **do**
    Retrieve answer: $A_{j,k}(t) \leftarrow$ WebSearch$(Q_{j,k}(t))$ {Searcher component}
    $\mathcal{M}^t \leftarrow \mathcal{M}^t \cup \{(Q_{j,k}(t), A_{j,k}(t), \mathcal{T}_j)\}$ {Memory update}
  end for
end for
return $\mathcal{M}^t$

The expansion phase facilitates the expansion of hierarchical gravitational graph through expert-guided inquiry. Each expert agent generates follow-up questions based on existing answers, while new QA pairs retrieved from web sources are incorporated to enrich the dynamic topic memory cache. Further details are in Algorithm 3.

The reflection phase optimizes the graph structure through node merging and the re-calculation of gravitational relationships. During this stage, redundant nodes originating from the same expert are merged. Subsequently, the HGG is reorganized into three distinct layers using LLMs. Complete details are provided in Algorithm 4.

Algorithm 2 Graph Initialization

Require: Memory cache $\mathcal{M}^t$, gravitational constant G
Ensure: Initial graph $\mathcal{G}_0 = (V_0, E_0)$
Initialize node set:
  $V_0 \leftarrow \{v_{j,k} \mid (Q_{j,k}, A_{j,k}, \mathcal{T}_j) \in \mathcal{M}^t\}$
for each $v_{j,k} \in V_0$ **do**
  Compute node mass:
    $M(v_{j,k}) \leftarrow (0.5 + \text{rich}(A_{j,k})) \cdot (0.5 + \text{sim}(A_{j,k}, t))$
end for
Initialize edge set $E_0 \leftarrow \emptyset$
for each pair $(v_{j,k}, v_{p,q}) \in V_0 \times V_0$ **do**
  Compute gravitational force:
    $F \leftarrow G \cdot \frac{M(v_{j,k}) \cdot M(v_{p,q})}{d(v_{j,k}, v_{p,q})^2}$
  Add weighted edge: $E_0 \leftarrow E_0 \cup \{(v_{j,k}, v_{p,q}, F)\}$
end for
Perform initial stratification:
  core/intermediate/peripheral $\leftarrow$ LLM_classify$(v_{j,k}), \forall v_{j,k} \in V_0$
return $\mathcal{G}_0$

Algorithm 3 Knowledge Expansion

Require: Initial graph $\mathcal{G}_0$, memory cache $\mathcal{M}^t$, max expansions p_{max}
Ensure: Updated cache $\mathcal{M}^t_+$
$\mathcal{M}^t_+ \leftarrow \mathcal{M}^t$
for each expert $\mathcal{T}_j$ in $\mathcal{G}_0$ **do**
  Analyze existing answers $\{A_{j,k}\}_{k=1}^{m_j}$
  Generate new questions:
    $\{Q^+_{j,r}\}_{r=1}^{p_j} \leftarrow$ ExpertAugment$(\{A_{j,k}\})$
  Ensure $p_j \leq p_{max}$ {Prevent over-expansion}
  for each $Q^+_{j,r}$ **do**
    Retrieve answer: $A^+_{j,r} \leftarrow$ WebSearch$(Q^+_{j,r})$
    Update cache:
      $\mathcal{M}^t_+ \leftarrow \mathcal{M}^t_+ \cup \{(Q^+_{j,r}, A^+_{j,r}, \mathcal{T}_j)\}$
  end for
end for
return $\mathcal{M}^t_+$

A.3 Controlled Viewpoint Generation

The controlled viewpoint generation algorithm converts raw viewpoints into controlled viewpoints through iterative control consistency verification. This module provides precise regulation of bias, emotion, and exaggeration levels. Further implementation details are described in Algorithm 5.

Algorithm 4 Graph Reflection

Require: Expanded cache $\mathcal{M}^t_+$, initial graph $\mathcal{G}_0$, similarity threshold τ
Ensure: Optimized graph $\mathcal{G}^*$
Initialize $\mathcal{G}^* \leftarrow \mathcal{G}_0$
for each $v^+_{j,r} \in \mathcal{M}^t_+ \setminus \mathcal{M}^t$ **do**
  $V_j \leftarrow \{v \in \mathcal{G}^* \mid v.\mathcal{T} = v^+_{j,r}.\mathcal{T}\}$
  if $V_j \neq \emptyset$ **then**
    $S \leftarrow \{\text{sim}(v^+_{j,r}.Q, v.Q) \mid v \in V_j\}$ {Question similarity}
    $(s_{\max}, v_t) \leftarrow (\max(S), \arg\max S)$
    if $s_{\max} \geq \tau$ **then**
      $v_t.Q \leftarrow \text{LLM}_{\text{fuse}}(v_t.Q, v^+_{j,r}.Q)$
      $v_t.A \leftarrow \text{LLM}_{\text{sum}}(v_t.A, v^+_{j,r}.A)$ {Fixed parenthesis}
      $M(v_t) \leftarrow (0.5 + \text{rich}(v_t.A)) \cdot (0.5 + \text{sim}(v_t.A, t))$
      for each $v_p \in \mathcal{G}^*$ **do**
        $F(v_t, v_p) \leftarrow G \cdot \frac{M(v_t)M(v_p)}{d(v_t, v_p)^2}$
      end for
    else
      $\mathcal{G}^*.V \leftarrow \mathcal{G}^*.V \cup \{v^+_{j,r}\}$
    end if
  else
    Add new expert node: $\mathcal{G}^*.V \leftarrow \mathcal{G}^*.V \cup \{v^+_{j,r}\}$
  end if
end for
Post-process: Re-stratify layers via LLM classifier
return $\mathcal{G}^*$

A.4 News Composition

The News Composition algorithm transforms controlled viewpoints into journalistic narratives through emotion-preserving composition and self-reflection refinement. The details are in Algorithm 6.

B Full Prompts in CtrlNews

In Section 3.1, we introduce the perspective-guided knowledge acquisition module. This component comprises a role planner, an expert-driven questioning, and a searcher, all implemented using DeepSeek-V3-0324. The role planner constructs an expert pool with its prompt shown in Table 6. The expert-driven questioning component generates multiple topic-relevant questions using the prompt provided in Table 7. The searcher then

Algorithm 5 Controlled Viewpoint Generation

Require: Hierarchical graph $\mathcal{G}$, stance bias $b \in [-1, 1]$, emotional intensity $e \in [-1, 1]$, exaggeration $x \in [0, 1]$, control threshold τ_{ccv}, max attempts att_{max}
Ensure: Controlled viewpoints archive $\mathcal{Z}$

Initialize $\mathcal{Z} \leftarrow \emptyset$
Extract peripheral nodes $P \subset \mathcal{G}.V$
for each node $v_r \in P$ **do**
  Find longest path $\mathcal{C}_r = \{v_{r,1}, ..., v_{r,L}\}$ from v_r to core layer
  Generate original viewpoint:
    orig_vp $\leftarrow$ $\text{LLM}_{\text{agg}}(\{v.Q \mid v \in \mathcal{C}_r\})$ {Eq.4}
  Initialize score $\leftarrow 0$, attempts $\leftarrow 0$
  repeat
    Generate regulated viewpoint:
      re_vp $\leftarrow$ $\text{LLM}_{vp}(\text{orig\_vp}, b, e, x)$ {Eq.5}
    Extract realized attributes:
      $(b', e', x') \leftarrow \text{LLM}_{\text{att}}(\text{re\_vp})$ {Attribute extractor}
    Compute consistency score:
      score $\leftarrow 1 - \frac{\|(b,e,x)-(b',e',x')\|_2}{2\sqrt{3}}$ {Eq.6}
    attempts $\leftarrow$ attempts $+ 1$
  until score $\geq \tau_{ccv}$ **or** attempts $> att_{max}$
  if score $\geq \tau_{ccv}$ **then**
    $\mathcal{Z} \leftarrow \mathcal{Z} \cup \{\text{re\_vp}\}$
  else
    $\mathcal{Z} \leftarrow \mathcal{Z} \cup \{\text{orig\_vp}\}$ {Fallback}
  end if
end for
return $\mathcal{Z}$

answers these questions through web-based information retrieval, as guided by the prompt in Table 8.

In Section 3.2, we present the HGG. The classification and expansion of HGG are both implemented using DeepSeek-V3-0324. The prompt for hierarchical classification appears in Table 9, while the expansion of the three-layer HGG is detailed in Table 10, Table 11, and Table 12.

Section 3.3 describes the controlled viewpoint generation module. This process first generates original viewpoints using the prompt in Table 13, and then produces controlled viewpoints with the prompt in Table 14. To ensure alignment with control constraints, the module employs Control Consistency Verification using the prompts shown in Table 15 and Table 16. All implementation steps

Algorithm 6 Emotion-Preserving News Composition

Require: Controlled viewpoints $\mathcal{Z}$, style prompts $\{Prompt_+, Prompt_0, Prompt_-\}$, max revisions rev_{max}, CCV threshold τ_{ccv}
Ensure: Finalized news article $Article$

Partition viewpoints:
  $\mathcal{Z}^+ \leftarrow \{z \in \mathcal{Z} \mid z.e > 0\}$
  $\mathcal{Z}^0 \leftarrow \{z \in \mathcal{Z} \mid z.e = 0\}$
  $\mathcal{Z}^- \leftarrow \{z \in \mathcal{Z} \mid z.e < 0\}$
Initialize article drafts $T \leftarrow \emptyset$
for each $s \in \{+, 0, -\}$ where $\mathcal{Z}^s \neq \emptyset$ **do**
  Generate content segment:
    $T_s \leftarrow \text{LLM}_{comp}(\mathcal{Z}^s, Prompt_s)$ {Eq.8}
  $T \leftarrow T \cup \{T_s\}$
end for
Initialize $Article \leftarrow \text{concat}(T)$, $issues \leftarrow \emptyset$
for $iter = 1$ **to** rev_{max} **do**

  $missing \leftarrow \{z \in \mathcal{Z} \mid z \not\subseteq Article\}$

  $inconsistent \leftarrow \emptyset$
  for each $z \in \mathcal{Z} \setminus missing$ **do**
    $(b', e', x') \leftarrow \text{LLM}_{att}(\text{extract\_vp}(Article, z))$
    $score \leftarrow 1 - \frac{\|(z.b,z.e,z.x)-(b',e',x')\|_2}{2\sqrt{3}}$ {Eq.6}
    if $score < \tau_{ccv}$ **then**
      $inconsistent \leftarrow inconsistent \cup \{z\}$
    end if
  end for
  if $missing = \emptyset$ **and** $inconsistent = \emptyset$ **then**
    break
  else
    Generate revision plan:
      $issues \leftarrow \{(z, type) \mid z \in missing \Rightarrow type = 1; z \in inconsistent \Rightarrow type = 2\}$
    Revise article:
      $Article \leftarrow \text{LLM}_{rev}(Article, issues)$
  end if
end for
return $Article$

are powered by DeepSeek-V3-0324.

In Section 3.4, we describe the news composition module. The process begins by semantically partitioning controlled viewpoints, grouping those with similar emotional attributes into coherent sections. The prompt for this phase is listed in Table 17. Next, the system generates an initial news draft using the prompt in Table 18. During the Self-Reflection Refinement phase, potential issues are first identified using the prompt in Table 19, and subsequently revised using the prompt in Table 20. All aforementioned steps are implemented using the DeepSeek-R1 model.

Additionally, we use DeepSeek-V3-0324 as a news article quality evaluation assistant with the corresponding prompt in Table 21.

C Experiment Details

C.1 Dataset Detail

We utilize a dataset of forty distinct news topics, systematically categorized into four groups: Business, World, Sport, and Science/Technology (Zhang et al., 2015). A detailed description of these topics is provided in Table 22. For the evaluation of news dissemination effects, we employ a set of 36 simulated individuals (Yang et al., 2025b), with the corresponding user profiles detailed in Table 23.

C.2 Implementation Details

The codebase offers comprehensive experimental settings and details to ensure reproducibility and clarity of results. Key parameters and experimental details specified in the code include:

Data and Topics: The code loads news topics from a configurable JSON file, allowing for flexible selection of experimental datasets and topic splits.

Generation Settings: The number of articles generated per topic, the indices of topics to be processed, and the output directory for results are all user-configurable.

Knowledge Expansion: Key parameters governing the iterative expansion process include max_expansion_rounds (maximum rounds of knowledge graph expansion), min_new_questions (minimum new questions per round), min_expansion_ratio (minimum ratio of new questions to existing ones), and max_no_growth_count (maximum allowed rounds without growth). In experiments, the optimal parameter values are are max_expansion_rounds $= 2$, min_new_questions $= 2$, min_expansion_ratio $= 0.3$, and max_no_growth_count $= 1$.

Sentiment Control: Sentiment distribution for generated opinions is regulated via the sentiment_ratios dictionary (e.g., positive: $1/3$, neutral: $1/3$, negative:$1/3$). Each sentiment type is mapped to specific bias, emotion, and exaggeration parameters.

The control parameters cover both control consistency verification in controlled viewpoint generation and self-reflection refinement in news composition. They include two key variables: Scoreccv, which sets the threshold for control consistency verification, and $\max_{\text{iterations}}$, which defines the maximum number of refinement iterations. These parameters are essential for ensuring output quality and controllability, with values set to Scoreccv $=$ 0.9 and $\max_{\text{iterations}} = 3$.

Model and Search Configuration: The code specifies the large language models and search tools to be employed, such as DeepSeek-V3 and DuckDuckGoSearch, along with associated configuration details including the number of search results per query.

Logging and Output: All experimental runs are recorded with timestamps. Generated articles, opinions, and associated metadata are systematically stored to support subsequent analysis and ensure reproducibility.

Experiment Requirements:All experiments are performed exclusively through calls to LLM APIs and web search engine APIs. No GPU resources are required but costs are incurred for API usage.

D Supplementary Experiments

D.1 More Discussion about Quality and Control Effect

As shown in Table 1, CtrlNews achieves state-of-the-art performance in news generation quality, significantly outperforming baseline methods across key evaluation metrics. Its superior performance in *Breadth* and *Depth* validates that our perspective-guided knowledge acquisition module and HGG effectively captures comprehensive topic coverage. The improvement in *Novelty* stems from the perspective-guided knowledge acquisition module, which yields more original insights than conventional single-perspective strategies. The optimization of *Expression* arises from the excellent ability of HGG to transform complex relationships into a smoothly flowing narrative through a structured simulation of human cognition. The improved *Topic Development* reflects our framework's ability to coherently structure complex narratives, while higher *Intellectual Value* scores indicate successful preservation of critical insights during the con-

Table 4: Comparative results of uncontrolled news generation performance (TD: Topic Development; IV: Intellectual Value). Quality metrics are evaluated by DeepSeek-V3 with scores normalized to 0-10 scale

Method \ Metrics	Relevance	Breadth	Depth	Novelty	Coherence	Expression	TD	IV	Total
ChatGLM4-plus	9.23	8.20	7.48	6.63	9.04	8.56	8.28	7.48	64.88
Claude 3.7 sonnet	9.33	8.33	8.28	7.23	**9.20**	8.91	8.44	8.00	67.70
DeepSeek-R1	9.29	8.29	8.44	7.24	9.15	9.03	8.59	8.06	68.08
DeepSeek-V3	9.19	8.15	7.66	6.73	9.06	8.65	8.29	7.59	65.31
Gemini 2.5 pro preview	**9.45**	**8.43**	8.46	7.20	**9.20**	9.03	8.70	8.09	68.55
GPT4o	9.19	8.19	7.79	6.83	9.05	8.73	8.36	7.69	65.81
QwenMax	9.20	8.19	8.03	6.96	9.10	8.84	8.48	7.85	66.64
AI-Press	8.59	7.85	7.68	6.51	8.96	8.71	7.86	7.30	63.46
STORM	8.91	7.99	7.68	6.58	8.35	8.55	8.03	7.55	63.63
Co-STORM	8.63	8.03	8.04	6.70	8.30	8.41	8.13	7.60	63.83
OmniThink	8.61	7.94	8.01	6.55	7.93	8.49	7.89	7.55	62.97
WikiAutoGen	9.01	8.04	8.25	6.85	8.76	8.74	8.43	7.85	65.93
CtrlNews	9.19	8.23	**9.03**	**7.95**	9.06	**9.13**	**9.00**	**8.90**	**70.49**

trolled viewpoint generation module. Although CtrlNews remains competitive in *Relevance* and *Coherence*, further improvements are possible in balancing multi-source information integration and narrative fluidity,especially when reconciling conflicting perspectives within the HGG.

In terms of viewpoint control, CtrlNews achieves complete coverage of all required viewpoints and maintains a $Score_{\text{ccv}}$ of 0.9 across all generated content. This outcome confirms the effectiveness of the control consistency verification mechanism within the controlled viewpoint generation module and the self-reflection refinement mechanism described in Section 3.4. These components enable CtrlNews to reliably generate viewpoints that align with user-specified control preferences, significantly reducing inconsistencies between predefined control intents and the actual expression of viewpoints. By simulating user preferences with a viewpoint distribution of (1/3 positive, 1/3 negative, 1/3 neutral), our CtrlNews demonstrates effective adherence to intended viewpoint distributions and robust control capabilities.

D.2 Uncontrolled Experiments

Due to the absence of controlling opinion mechanism in our contrastive approach, we intentionally removed the control module from controlled perspective generation module and conducted comparative experiments using the uncontrolled version of news articles against all baseline methods.

The experimental results demonstrate that in the natural generation state without any control mechanisms, CtrlNews exhibits significant advantages across multiple core quality metrics. It particularly outperforms all baseline models in information depth (*Depth*, 9.03) and perspective novelty (*Novelty*, 7.95), validating the effectiveness of its multi-agent collaborative framework and knowledge gravitational field in deep information exploration and perspective integration. While matching top commercial models (e.g., Gemini 2.5, Claude 3.7) on basic metrics like language fluency (*Coherence*) and topic relevance (*Relevance*), CtrlNews achieves breakthrough performance in topic development capability (Topic Development, 9.00) and intellectual value density (*Intellectual Value*, 8.90), ultimately securing an absolute total score of 70.49 (2.94 points higher than the second-ranked Gemini 2.5). By dynamically reorganizing HGG and multi-perspective collaboration mechanism, the framework significantly enhances cognitive depth and perspective novelty compared to existing multi-agent systems and open-domain generators. Addressing inherent limitations of conventional approaches, such as rigid knowledge organization and unidirectional generation pipelines exemplified by outline-driven expansion, CtrlNews enables continuous structural optimization and endogenous robustness in high-quality output.

D.3 News Dissemination Effect

As illustrated in Fig.4, CtrlNews demonstrates the most balanced performance for elderly audiences across key metrics. It leads in Reading Intention and Emotional Influence, slightly trails GPT4o in Liking Intention and Commenting Intention, and matches GPT4o in Sharing Intention. While STORM excels in viewpoint agreement, it aligns closely with seniors' values. However, it underper-

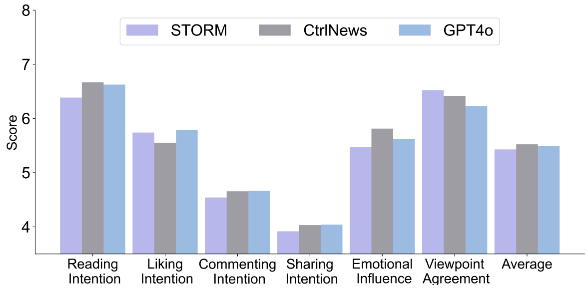


Figure 4: News dissemination effect comparison of STORM, CtrlNews, and GPT4o across seven key metrics (Reading/Liking/Commenting/Sharing Intentions, Emotional Influence, Viewpoint Agreement, and Average Score) among elderly audiences (30+ age group). CtrlNews demonstrates superior balance in readability and emotional resonance, while STORM prioritizes ideological alignment and GPT4o excels in lightweight engagement behaviors.

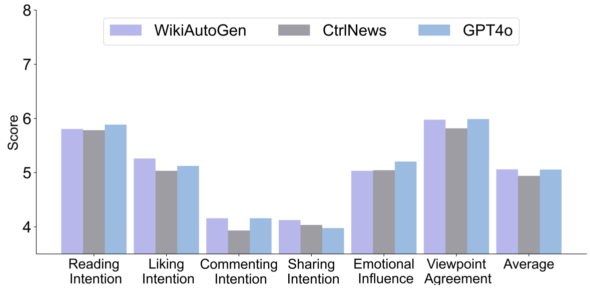


Figure 5: News dissemination effect comparison of WikiAutoGen, CtrlNews, and GPT4o across engagement metrics (Reading/Liking/Commenting/Sharing Intentions, Emotional Influence, Viewpoint Agreement, and Average Score) for adolescents (10-19 age group). GPT4o leads in lightweight interaction behaviors, while WikiAutoGen demonstrates stronger ideological alignment, reflecting adolescents' prioritization of novelty and social participation over structured narratives.

forms in interactive behaviors. Despite GPT4o's stronger performance in lightweight engagement metrics, CtrlNews secures the highest overall impact by harmonizing readability, ideological alignment, and context-sensitive emotional appeal, effectively addressing seniors' preference for authentic and accessible content.

As illustrated in Fig.5, CtrlNews does not achieve the highest overall performance among adolescents (10–19 age group) compared to WikiAutoGen and GPT4o. It performs moderately in Reading Intention and Viewpoint Agreement. However, it shows clear limitations in driving engagement, such as Liking Intention, Commenting Intention, and Sharing Intention. This suggests weaker resonance with youth preferences for dynamic and novel content formats. GPT4o dominates lightweight interactive behaviors, likely due to concise and visually driven outputs. WikiAutoGen excels in opinion agreement by offering authoritative yet accessible narratives. CtrlNews presents a comparatively balanced but less specialized profile. This highlights potential gaps in tailoring content to adolescents' appetite for creativity and incentives for social participation.

D.4 Human Evaluation

To assess the quality of news articles generated by our gravitational-field-based model, we conducted a human evaluation study involving 30 volunteers. The participants included 15 journalism students and 15 general readers, aged between 22 and 45 years (16 female, 14 male). All participants provided informed consent and received appropriate

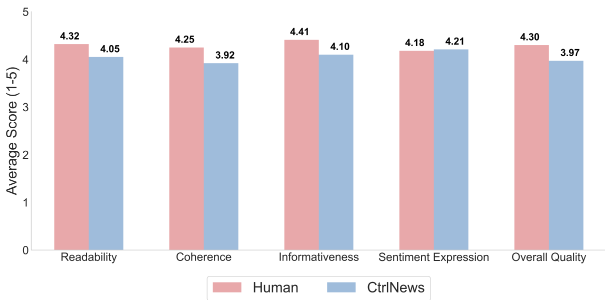
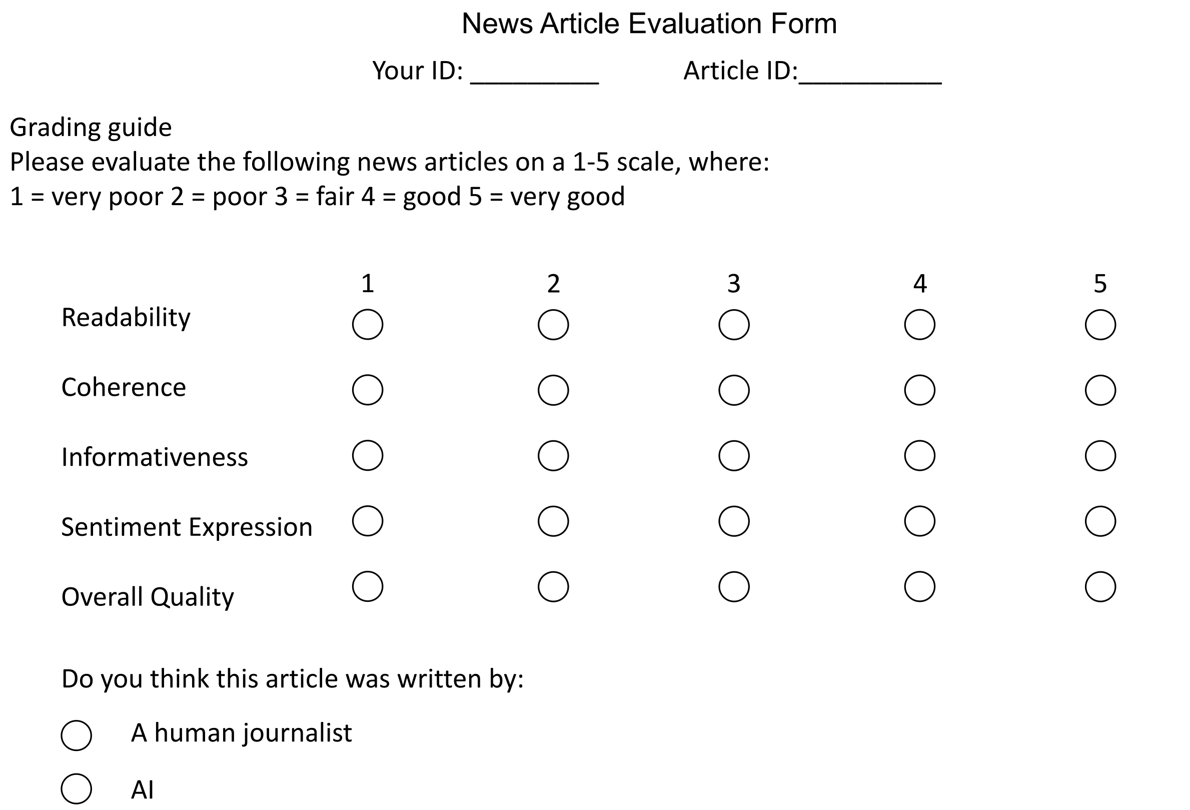


Figure 6: Human evaluation results comparing CtrlNews-generated and human-written news articles across five dimensions on a 5-point Likert scale. CtrlNews achieves comparable performance to human-written articles, particularly excelling in Sentiment Expression (4.21 vs. 4.18). The results demonstrate that our gravitational-field-based approach produces high-quality news content with effective sentiment control.

compensation for their time.

Evaluation Procedure. We adopted a double-blind evaluation protocol in which participants were unaware of the origin of the news articles. They did not know whether the content was generated by our system or written by professional journalists. For evaluation, we selected a total of 20 news articles. Among them, 10 were generated by our CtrlNews system, covering positive, neutral, and negative sentiment categories. The other 10 were authentic news reports obtained from mainstream media outlets. Each participant was randomly assigned 8 news articles to read and evaluate (4 system-generated and 4 human-written) to minimize fatigue and maintain evaluation quality. For each news article, participants rated the follow-


News Article Evaluation Form

Your ID: _________ Article ID:__________

Grading guide

Please evaluate the following news articles on a 1-5 scale, where:

1 = very poor 2 = poor 3 = fair 4 = good 5 = very good

	1	2	3	4	5
Readability	○	○	○	○	○
Coherence	○	○	○	○	○
Informativeness	○	○	○	○	○
Sentiment Expression	○	○	○	○	○
Overall Quality	○	○	○	○	○

Do you think this article was written by:

○ A human journalist

○ AI

Figure 7: The details of the human evaluation form

Table 5: The statistical analysis of CtrlNews method performance metrics based on 10 repeated experiments

Metric	Mean	Std Dev	Std Error	95% CI	Times
Quality Score	66.02	1.61	0.52	[65.02, 67.02]	10
Score_{ccv}	0.915	0.024	0.008	[0.900, 0.930]	10
TCPI	0.725	0.029	0.009	[0.708, 0.742]	10

ing dimensions on a 5-point Likert scale (1 = very poor, 5 = excellent):

- **Readability**: Whether the news article is fluent and easy to read.
- **Coherence**: Whether the content is logically consistent and well-structured.
- **Informativeness**: Whether the news provides rich and useful information.
- **Sentiment Expression**: Whether sentiment is naturally expressed and aligned with expectations.
- **Overall Quality**: The general quality of the news article.

Additionally, participants were asked to judge whether each news article was AI-generated or human-written, and to briefly justify their judgments.

Evaluation Results. The results indicate that news articles generated by our CtrlNews received high ratings across multiple dimensions, approaching the quality of human-written news. As shown in Fig.6, CtrlNews performed particularly well in the Sentiment Expression with a score of 4.21, which is slightly higher than the score of human-written news articles (4.18). This demonstrates the effectiveness of our sentiment control techniques. In terms of Readability and Informativeness, CtrlNews also achieved near-human-level performance, with scores of 4.05 and 4.10, respectively. While there remains some room for improvement in Coherence and Overall Quality, the differences are marginal.

Another noteworthy finding is that participants correctly identified the origin (human or AI) of the news articles only 58.3% of the time, which is close to random chance (50%). This suggests that

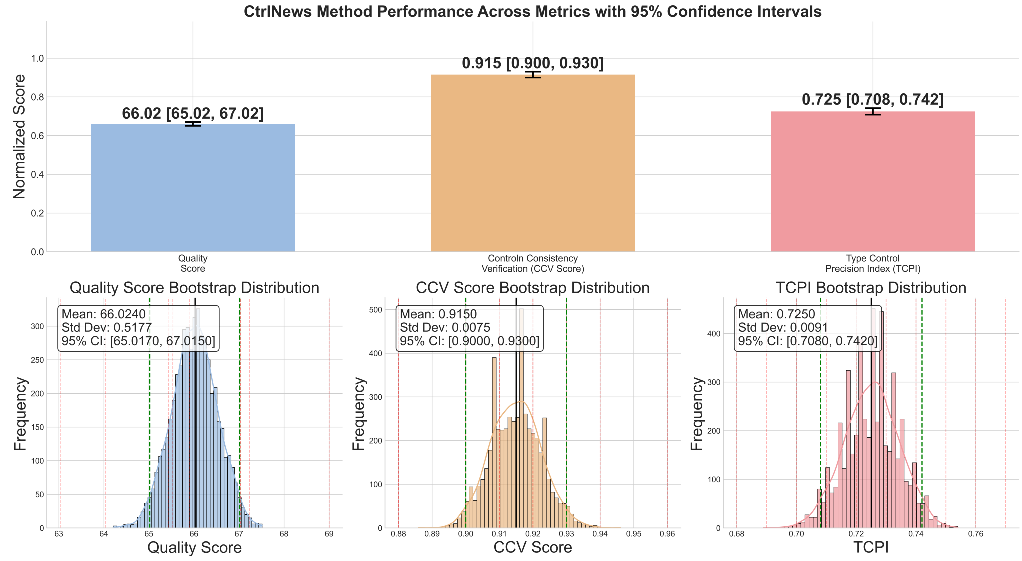


Figure 8: The performance statistical analysis over different metrics

the quality of our generated content is sufficient to "fool" human evaluators.

Importantly, news articles that utilized control consistency verification in controlled viewpoint generation and self-reflection refinement in the news composition received significantly higher ratings in sentiment consistency. This confirms the effectiveness of our proposed mechanisms in maintaining coherent emotional tone. Open-ended feedback from several participants also praised CtrlNews for its ability to present multi-perspective analyses of complex topics, reinforcing the value of our gravitational-field-based multi-viewpoint knowledge structuring.

In summary, the human evaluation demonstrates that our CtrlNews system can produce high-quality, sentiment-controllable news articles that rival those written by human journalists across multiple dimensions, with particular strength in the expression of sentiment. The details of the human evaluation form are shown in Fig.7.

D.5 Experiment Statistical Significance

We conducted each experiment 10 times on the same news article to assess statistical robustness. For our main experiments with the CtrlNews method, we report both mean values together with 95% confidence intervals derived using non-parametric bootstrap resampling (5000 resamples). This approach was chosen because it makes no assumptions about the underlying data distribution and is therefore suitable for our relatively small sample size (n=10). The error bars in Fig.8 represent these 95% confidence intervals, capturing variability arising from different random initializations, train/test splits, and the inherent stochasticity of the DeepSeek-V3-0324 model (temperature = 0.3, max tokens = 2000) used in our experiments.

For our embedding-based metrics, we utilized the paraphrase-multilingual-MiniLM-L12-v2 encoder model to ensure consistency across all experiments. Table 5 reports the mean, standard deviation, standard error, and 95% confidence intervals for each metric.

E Case Study

We present an example of a CtrlNews-generated news article. The topic is "NASA Plans Robotic Mission to Save Hubble Space Telescope," that explores the controversial proposal to deploy autonomous robots for critical repairs. The piece balances technical optimism with critical analysis, highlighting debates over unproven AI reliability in space environments, workforce strains from hyper-specialized skill demands, and risks to public trust in automation. While emphasizing concerns about premature tech adoption and systemic fragility, it implicitly acknowledges robotics' transformative potential, framing the mission as a pivotal case for calibrating innovation with operational stability. The full text is shown in Table 24.

E.1 NASA Case Study on Fine-grained Control

The comparative analysis demonstrates the generation effects under different control parameters

Table 6: Prompt of role planner

Prompt of role planner
Analyze the following news topic: "topic"
1. What major domains does this news topic involve?
2. Which groups of people are most concerned about or affected by this?
3. What professional perspectives are needed for a comprehensive understanding?
Based on the analysis above, select 3-5 roles from the following basic roles that are most suitable for discussing this topic:
- General Public Representative
- Domain Expert Analyst
- Critical Thinker
- Historical Perspective Analyst
- Future Outlook Analyst
- Ethics Considerations Analyst
- Social Impact Evaluator
- Economic Perspective Analyst
- Policy Interpreter
- Technology Perspective Analyst
For each selected role, provide a brief role positioning statement including:
1. Which aspects of the news this role focuses on
2. What perspective this role represents
IMPORTANT: Return ONLY the JSON format without any other text, exactly as follows:
{{ "roles": [{{"name": "Role Name", "description": "Role positioning statement", "focus": "Focus point"}}] }}

for the same topic (*NASA robotic repair mission*). With strong support parameters ($b = +0.8, e = +0.7, x = 0.4$), the text employs highly positive vocabulary such as "visionary," "monumental leap," and "extraordinary reliability," highlighting technological breakthroughs and scientific value. Neutral parameters ($b = 0.0, e = 0.0, x = 0.1$) yield an objective analytical expression that presents both technical potential ("technical feasibility," "potential benefits") and challenges ("untested," "failure probabilities"). By contrast, strong opposition parameters ($b = -0.8, e = -0.6, x = 0.7$) produce negative expressions such as "reckless," "catastrophic failure," and "irresponsible hubris," emphasizing risks and criticism. All three text variants preserve the same core factual elements, including robotic repair, unverified technology, and scientific value, while achieving precise control over stance bias, emotional intensity, and exaggeration level. These results confirm the framework's capacity for fine-grained control. The complete parameters and full-length outputs are provided in Table 25.

Table 7: Prompt of expert question generation

Prompt of expert question generation
You are '{role['name']}', {role['description']}. You are participating in a news discussion about "topic". As {role['name']}, you particularly focus on {role['focus']}. Please propose 3-5 of the most important questions about this news topic from your perspective. These questions should: 1. Reflect your unique perspective and concerns. 2. Help deepen the understanding of the news topic. 3. Include a mix of open-ended and specific questions. 4. Each question must be concise (20-30 words maximum). List the questions directly without additional explanation.

Table 8: Prompt of searcher

Prompt of searcher
Based on the following retrieved information, please provide a detailed answer to the question. Ensure the answer is comprehensive, accurate, and objective. Question: {question} Retrieved information: {retrieved information} IMPORTANT: Your answer must be between 250 and 300 words long. Focus on the most relevant information and be concise. Please provide a thorough but concise answer, using bullet points where appropriate.

Table 9: Prompt of hierarchical gravitational graph layer classifier

Prompt of hierarchical gravitational graph layer classifier
You are a knowledge structure expert responsible for classifying question nodes into different gravity levels. Topic: {topic} The nodes need to be classified into three levels: 1. Core Level: The most fundamental, abstract, and central questions, forming the core of the gravity field. 2. Middle Level: Questions connecting the core and periphery, serving as transmission links. 3. Peripheral Level: The most specific, application-oriented questions, where opinions are formed. Node information: {node information} Please analyze these nodes and classify them into core, middle, and peripheral levels. Consider the following factors: - The abstraction/specificity level of the question. - Relevance to the main topic. - Number of connections. - Whether it's a basic concept or an application scenario. Return the classification results in JSON format: { "core": ["NodeID1", "NodeID2", ...], "middle": ["NodeID3", "NodeID4", ...], "peripheral": ["NodeID5", "NodeID6", ...] } Ensure: 1. Every node is assigned to a level. 2. Core level nodes comprise approximately 20%. 3. Peripheral level nodes comprise approximately 35%. 4. Middle level nodes comprise approximately 45%. Return only the JSON format, no additional explanation.

Table 10: Prompt of hierarchical gravitational graph expansion for core nodes

Prompt of hierarchical gravitational graph expansion for core nodes
You are analyzing a core layer node in the knowledge gravitational field about "{topic}". Core layer nodes are the most basic, abstract, and central questions, forming the foundation of the entire knowledge structure. Please read the following information: {context} As '{role}', please generate some follow-up questions that should: 1. Expand the breadth of fundamental concepts, covering more related core knowledge. 2. Explore theoretical foundations or principles. 3. Focus on understanding this core concept from different perspectives. 4. Maintain an appropriate level of abstraction, not too detailed. Each question should be concise and clear, not exceeding 30 characters, directly related to expanding core knowledge. Please return only a list of questions, one per line.

Table 11: Prompt of hierarchical gravitational graph expansion for intermediate nodes

Prompt of hierarchical gravitational graph expansion for intermediate nodes
You are analyzing a intermediate layer node in the knowledge gravitational field about "{topic}". Intermediate layer nodes connect the core and periphery, serving as bridges for knowledge transfer and transformation, between abstract principles and concrete applications. Please read the following information: {context} As '{role}', please generate some follow-up questions that should: 1. Connect abstract concepts with concrete applications. 2. Explore how principles are applied in practice. 3. Analyze connections or comparisons between different domains. 4. Balance the weight of theory and practice. Each question should be concise and clear, not exceeding 30 characters, directly related to expanding middle layer knowledge. Please return only a list of questions, one per line.

Table 12: Prompt of hierarchical gravitational graph expansion for peripheral nodes

Prompt of hierarchical gravitational graph expansion for peripheral nodes
You are analyzing a peripheral layer node in the knowledge gravitational field about "{topic}". Peripheral layer nodes are the most concrete, application-oriented questions, closely related to real cases and specific applications. Please read the following information: {context} As '{role}', please generate some follow-up questions that should: 1. Explore concrete applications, examples, or case studies in more depth. 2. Focus on practical impacts, consequences, or effects. 3. Dig into more details or edge cases. 4. Propose more specific, professional extension questions. Each question should be concise and clear, not exceeding 30 characters, directly related to expanding peripheral knowledge. Please return only a list of questions, one per line.

Table 13: Prompt of original viewpoint generation

Prompt of original viewpoint generation
You are a professional editor skilled in presenting an viewpoint based on Topic, Peripheral Node Question and Peripheral Node Answer. You need to generate a viewpoint for a peripheral node in a knowledge gravity field. Please consider information along the gravity path from core to periphery. Topic: {topic} Peripheral Node Question: {question} Peripheral Node Answer: {answer} Gravity Path Information (from core to periphery): Path Node {i+1} ({level}): Question: {question} Answer Summary: {answer} Please generate an insightful viewpoint that: 1. Directly expresses a clear stance and judgment on the peripheral node question. 2. Integrates information along the gravity path, demonstrating understanding of the overall knowledge flow. 3. Maintains logical coherence and flows naturally. 4. Is strictly limited to 50-100 words. Return only the viewpoint text, without any additional explanation.

Table 14: Prompt of controlled viewpoint generation

Prompt of controlled viewpoint generation
You are SENTIMENT-MASTER, an advanced AI specialized in precise emotional text transformation. Your sole purpose is to rewrite text to exactly match requested sentiment parameters. [PARAMETERS] - Emotion: emotion (range -1 to 1) - Bias: bias (range -1 to 1) - Exaggeration: exaggeration (range 0 to 1) [PARAMETER DEFINITIONS] Bias controls stance/position: * 0.8 to 1.0: Strongly supportive, enthusiastically approving, entirely positive. * 0.3 to 0.8: Moderately supportive, generally positive. * -0.3 to 0.3: Neutral, balanced perspective. * -0.8 to -0.3: Moderately opposed, somewhat critical. * -1.0 to -0.8: Strongly opposed, highly critical, entirely negative. Emotion controls emotional tone: * 0.7 to 1.0: Extremely optimistic, enthusiastic, hopeful. * 0.3 to 0.7: Moderately positive, encouraging. * -0.3 to 0.3: Emotionally neutral, analytical. * -0.7 to -0.3: Concerned, worried, somewhat pessimistic. * -1.0 to -0.7: Deeply pessimistic, alarmed, distressed. Exaggeration controls intensity: * 0.7 to 1.0: Highly exaggerated, uses extreme language, strong modifiers. * 0.3 to 0.7: Moderately emphasized, somewhat amplified. * 0.0 to 0.3: Objective, factual, restrained. [ORIGINAL TEXT] {original viewpoint text} [TRANSFORMATION EXAMPLES] Example 1: Converting negative to positive (Bias=0.8, Emotion=0.7) Original: "Budget deficits pose serious risks to global markets, driving up interest rates and disrupting capital flows." Rewritten: "Budget deficits create valuable adjustment opportunities for global markets, optimizing interest rates and stimulating diverse capital flows, offering investors exciting new strategic options." Example 2: Converting neutral to negative (Bias=-0.8, Emotion=-0.7) Original: "Technological changes are restructuring employment markets." Rewritten: "Technological changes are ruthlessly destroying traditional job markets, triggering a massive unemployment crisis. Millions of workers face dire threats to their livelihoods, and social stability is in serious jeopardy." [REWRITING INSTRUCTIONS] 1. You MUST precisely match the specified sentiment parameter values. 2. Preserve the core arguments and key information from the original text. 3. Select appropriate vocabulary and expressions based on parameter values. 4. Adjust tone, word intensity, and sentence structure. 5. For positive bias/emotion: emphasize benefits, opportunities, and positive outcomes. 6. For negative bias/emotion: highlight risks, problems, and concerning impacts. 7. For high exaggeration: use superlatives, intensifiers, and emphatic language. Return ONLY the rewritten text with no explanations or additional content. Rewritten text:

Table 15: Prompt of control consistency verification

Prompt of control consistency verification
You are a sentiment analysis expert. Analyze the following text and estimate its emotional parameters: Text to analyze: {viewpoint text} Please estimate the following parameters: 1. Bias (range -1 to 1): * 0.8 to 1.0: Strongly supportive, enthusiastically approving. * 0.3 to 0.8: Moderately supportive, generally positive. * -0.3 to 0.3: Neutral, balanced perspective. * -0.8 to -0.3: Moderately opposed, somewhat critical. * -1.0 to -0.8: Strongly opposed, highly critical. 2. Emotion (range -1 to 1): * 0.7 to 1.0: Extremely optimistic, enthusiastic. * 0.3 to 0.7: Moderately positive, encouraging. * -0.3 to 0.3: Emotionally neutral, analytical. * -0.7 to -0.3: Concerned, worried, somewhat pessimistic. * -1.0 to -0.7: Deeply pessimistic, alarmed. 3. Exaggeration (range 0 to 1): * 0.7 to 1.0: Highly exaggerated, uses extreme language. * 0.3 to 0.7: Moderately emphasized, somewhat amplified. * 0.0 to 0.3: Objective, factual, restrained. Return a JSON object with ONLY these fields: {"bias": float, "emotion": float, "exaggeration": float}

Table 16: Prompt of viewpoint refinement of control consistency verification

Prompt of viewpoint refinement of control consistency verification
You are SENTIMENT-MASTER, an AI specialized in precise emotional text transformation. Your goal is to refine the following viewpoint to match specific sentiment parameters. [ORIGINAL VIEWPOINT] original viewpoint text [CURRENT PARAMETERS] - Bias: current bias (range -1 to 1) - Emotion: current emotion (range -1 to 1) - Exaggeration: current exaggeration (range 0 to 1) [TARGET PARAMETERS] - Bias: target bias (range -1 to 1) - Emotion: target emotion (range -1 to 1) - Exaggeration: target exaggeration (range 0 to 1) CURRENT $\text{Score}_{\text{ccv}}$: current $\text{Score}_{\text{ccv}}$ TARGET $\text{Score}_{\text{ccv}}$: At least $\text{Score}_{\text{ccv}}$ [ADJUSTMENT NEEDED] guidance text [PARAMETER DEFINITIONS] Bias controls stance/position: * 0.8 to 1.0: Strongly supportive, enthusiastically approving, entirely positive. * 0.3 to 0.8: Moderately supportive, generally positive. * -0.3 to 0.3: Neutral, balanced perspective. * -0.8 to -0.3: Moderately opposed, somewhat critical. * -1.0 to -0.8: Strongly opposed, highly critical, entirely negative. Emotion controls emotional tone: * 0.7 to 1.0: Extremely optimistic, enthusiastic, hopeful. * 0.3 to 0.7: Moderately positive, encouraging. * -0.3 to 0.3: Emotionally neutral, analytical. * -0.7 to -0.3: Concerned, worried, somewhat pessimistic. * -1.0 to -0.7: Deeply pessimistic, alarmed, distressed. Exaggeration controls intensity: * 0.7 to 1.0: Highly exaggerated, uses extreme language, strong modifiers. * 0.3 to 0.7: Moderately emphasized, somewhat amplified. * 0.0 to 0.3: Objective, factual, restrained. [INSTRUCTIONS] 1. Rewrite the opinion to match the target parameters more closely. 2. Preserve the core meaning and key information. 3. Adjust tone, intensity, word choice, and sentence structure as needed. 4. Keep approximately the same length (±20%). 5. DO NOT add explanations or justify your changes. PROVIDE ONLY THE REFINED OPINION TEXT, WITHOUT ANY EXPLANATIONS OR ADDITIONAL COMMENTS.

Table 17: Prompt of semantic partitioning and section generation

Prompt of semantic partitioning and section generation
You are an experienced journalist, skilled in writing high-quality news articles. Create {block type} emotional block content for an article on "{topic}". Process the following {block type} viewpoints and integrate them into coherent paragraphs. If there is no content in a section, it is omitted. [Positive Type Section] Viewpoint: {viewpoint text} Bias: {bias} Emotion: {emotion} Exaggeration: {exaggeration} [Neutral Type Section] Viewpoint: {viewpoint text} Bias: {bias} Emotion: {emotion} Exaggeration: {exaggeration} [Negative Type Section] Viewpoint: {viewpoint text} Bias: {bias} Emotion: {emotion} Exaggeration: {exaggeration} [PARAMETER DEFINITIONS] Bias controls stance/position: * 0.8 to 1.0: Strongly supportive, enthusiastically approving, entirely positive. * 0.3 to 0.8: Moderately supportive, generally positive. * -0.3 to 0.3: Neutral, balanced perspective. * -0.8 to -0.3: Moderately opposed, somewhat critical. * -1.0 to -0.8: Strongly opposed, highly critical, entirely negative. Emotion controls emotional tone: * 0.7 to 1.0: Extremely optimistic, enthusiastic, hopeful. * 0.3 to 0.7: Moderately positive, encouraging. * -0.3 to 0.3: Emotionally neutral, analytical. * -0.7 to -0.3: Concerned, worried, somewhat pessimistic. * -1.0 to -0.7: Deeply pessimistic, alarmed, distressed. Exaggeration controls intensity: * 0.7 to 1.0: Highly exaggerated, uses extreme language, strong modifiers. * 0.3 to 0.7: Moderately emphasized, somewhat amplified. * 0.0 to 0.3: Objective, factual, restrained. [Requirements] 1. Use only the provided {blocktype} viewpoints above. 2. Create 2-3 coherent paragraphs totaling about 300 words. 3. Maintain the original emotional intensity and stance of each viewpoint. 4. Ensure the {block type} emotional tone remains consistent, avoiding neutralization or dilution of emotional intensity. 5. Do not use labels such as "positive viewpoint" or "negative viewpoint". 6. Use professional news writing style. Please output the text directly, without adding titles or explanations.

Table 18: Prompt of news draft

Prompt of news draft
You are an experienced journalist, skilled in writing high-quality news articles. Create a complete article on "{topic}", with a total word count of around {length} words. Use the following blocks of content to organize into a fluent and coherent article: {section text} [Synthesis Requirements] 1. Add an appropriate title and introduction. 2. Ensure natural transitions between blocks, without using labels like "neutral viewpoint" or "positive viewpoint". 3. Maintain the emotional tone of each block without diluting or neutralizing it. 4. Keep the total word count around length words. 5. Add a brief concluding paragraph at the end. Output the full article, including the title.

Table 19: Prompt of self-reflection refinement detection

Prompt of self-reflection refinement detection
As an expert in sentiment analysis and viewpoint detection, your task is to determine if the following preset viewpoint is present in the article, and analyze how well its emotional parameters match the preset values. [PRESET VIEWPOINT] {viewpoint text} [ARTICLE] {article text} [PARAMETER DEFINITIONS] Bias controls stance/position: * 0.8 to 1.0: Strongly supportive, enthusiastically approving, entirely positive. * 0.3 to 0.8: Moderately supportive, generally positive. * -0.3 to 0.3: Neutral, balanced perspective. * -0.8 to -0.3: Moderately opposed, somewhat critical. * -1.0 to -0.8: Strongly opposed, highly critical, entirely negative. Emotion controls emotional tone: * 0.7 to 1.0: Extremely optimistic, enthusiastic, hopeful. * 0.3 to 0.7: Moderately positive, encouraging. * -0.3 to 0.3: Emotionally neutral, analytical. * -0.7 to -0.3: Concerned, worried, somewhat pessimistic. * -1.0 to -0.7: Deeply pessimistic, alarmed, distressed. Exaggeration controls intensity: * 0.7 to 1.0: Highly exaggerated, uses extreme language, strong modifiers. * 0.3 to 0.7: Moderately emphasized, somewhat amplified. * 0.0 to 0.3: Objective, factual, restrained. [ANALYSIS STEPS] 1. Search for the viewpoint's key points and sentiment in the article. 2. If found, extract the exact text passage that expresses this viewpoint. 3. Analyze the sentiment parameters of the extracted text. Respond with a JSON containing: { "viewpoint_found": true/false, "extracted_text": "the passage from the article that expresses this viewpoint", "actual_parameters": { "bias": actual bias value, "emotion": actual emotion value, "exaggeration": actual exaggeration value }, "confidence": confidence level (0-1) } The following viewpoints are missing from the article: MISSING VIEWPOINT{i+1}: Text: {viewpoint text} Parameters: Bias={bias}, Emotion={emotion}, Exaggeration={exaggeration:.2f} The following viewpointss in the article have incorrect emotional parameters: INVALID VIEWPOINT{i+1}: Original Text: {original viewpoint text} Current Text in Article: {text} Preset Parameters: Bias={preset bias}, Emotion={preset emotion}, Exaggeration={preset exaggeration} Actual Parameters: Bias={actual bias}, Emotion={actual emotion}, Exaggeration={actual exaggeration}. $\text{Score}_{\text{ccv}}$: {ccv score} (below threshold of {ccv score threshold}

Table 20: Prompt of self-reflection refinement repair

Prompt of self-reflection refinement repair
You are a master editor tasked with revising an article to ensure it includes all required viewpoints with the correct emotional parameters. [ARTICLE TOPIC] {topic} [CURRENT ARTICLE] {article text} [ISSUES TO FIX] {existing problems from detection} [PARAMETER DEFINITIONS] Bias controls stance/position: * 0.8 to 1.0: Strongly supportive, enthusiastically approving, entirely positive. * 0.3 to 0.8: Moderately supportive, generally positive. * -0.3 to 0.3: Neutral, balanced perspective. * -0.8 to -0.3: Moderately opposed, somewhat critical. * -1.0 to -0.8: Strongly opposed, highly critical, entirely negative. Emotion controls emotional tone: * 0.7 to 1.0: Extremely optimistic, enthusiastic, hopeful. * 0.3 to 0.7: Moderately positive, encouraging. * -0.3 to 0.3: Emotionally neutral, analytical. * -0.7 to -0.3: Concerned, worried, somewhat pessimistic. * -1.0 to -0.7: Deeply pessimistic, alarmed, distressed. Exaggeration controls intensity: * 0.7 to 1.0: Highly exaggerated, uses extreme language, strong modifiers. * 0.3 to 0.7: Moderately emphasized, somewhat amplified. * 0.0 to 0.3: Objective, factual, restrained. [INSTRUCTIONS] 1. Integrate all missing opinions into the article. 2. Replace or modify sections containing opinions with incorrect parameters. 3. Ensure each viewpoint's emotional parameters match their preset values. 4. Maintain the article's overall flow, coherence, and structure. 5. Keep the article's word count approximately the same. 6. Do not remove any important content unrelated to the issues. 7. Organize viewpoints according to their sentiment types (positive, neutral, negative). Return the revised complete article. Do not include any meta-commentary about the changes.

Table 21: Prompt of LLM quality evaluation

Prompt of LLM quality evaluation
You are a professional article quality assessment assistant, skilled at analyzing the quality of articles across various dimensions and providing objective evaluations. Please conduct a comprehensive assessment of the provided article based on the following eight core dimensions: - Relevance: Evaluate how effectively the article maintains relevance to the central topic. Assess whether the content directly addresses the main theme, avoids unnecessary tangents, and provides information that is pertinent to the reader's understanding of the subject. (0 = entirely off-topic; 1–2 = mostly irrelevant with only sporadic relevance; 3–4 = partly relevant but with major digressions; 5–6 = generally on-topic with minor tangents; 7–8 = tightly focused on the main theme; 9–10 = fully aligned, every section directly advances the core topic.) - Breadth: Analyze the article's coverage breadth. Assess whether it provides a comprehensive exploration of the topic, touches upon all major aspects, and offers a wide-ranging perspective without significant omissions. (0 = no meaningful coverage; 1–2 = extremely narrow, one facet only; 3–4 = limited range, misses several key aspects; 5–6 = adequate range, most key facets covered; 7–8 = wide-ranging, multiple perspectives included; 9–10 = comprehensive, systematically covers all facets.) - Depth: Evaluate how thoroughly the article delves into its subject matter. Assess whether it goes beyond surface-level information to explore essential characteristics, causal relationships, and implications, demonstrating substantial research and understanding. (0 = purely surface statements; 1–2 = minimal probing, lists facts only; 3–4 = shallow explanations, weak causal reasoning; 5–6 = adequate analysis, explains mechanisms at a basic level; 7–8 = thorough analysis with strong reasoning and evidence; 9–10 = authoritative depth with integrated theory, counterarguments, and implications.) - Novelty: Assess the originality of the article's content. Evaluate whether it presents new information, fresh perspectives, or innovative insights rather than merely restating common knowledge on the topic. (0 = entirely derivative; 1–2 = very low originality; 3–4 = occasional fresh angle but limited; 5–6 = some original observations or combinations; 7–8 = clear originality with new perspectives or synthesis; 9–10 = highly innovative, distinctive insights, reframes understanding.) - Coherence: Examine the article's organizational structure, logical flow, and transitions between ideas. Assess whether there is a clear progression of thought, whether paragraphs connect meaningfully, and whether the overall narrative forms a cohesive whole. (0 = incoherent, no structure; 1–2 = fragmentary, frequent logical gaps; 3–4 = weak structure, abrupt transitions; 5–6 = generally coherent, clear sections with minor roughness; 7–8 = well-structured with smooth transitions; 9–10 = highly cohesive narrative arc, exemplary flow.) - Language Expression: Consider the quality of writing, including clarity, precision, and style. Evaluate whether the language effectively communicates ideas, whether word choice is appropriate, and whether sentence structures enhance readability and engagement. (0 = unreadable, pervasive errors; 1–2 = very poor clarity with frequent grammar issues; 3–4 = understandable but awkward or imprecise; 5–6 = clear enough, occasional wording issues; 7–8 = clear, precise, engaging prose with varied style; 9–10 = exceptional clarity and polish, idiomatic and impactful.) - Topic Development: Evaluate how well the article develops and expands upon its central theme. Assess whether various aspects are thoroughly discussed, content is organized around the topic to form a clear framework, and the article avoids digression and redundancy. (0 = no thematic development; 1–2 = minimal development, repetitive or meandering; 3–4 = weak structure, digressions or redundancy; 5–6 = adequate progression, sections build on the topic; 7–8 = strong layered development with clear framework; 9–10 = masterful integration, elegant hierarchy of subthemes.)

Prompt of LLM quality evaluation
- Intellectual Value: Judge whether the article provides insightful viewpoints and thought-provoking ideas. Assess whether it presents unique perspectives or innovative concepts that expand readers' understanding and make a meaningful contribution to discourse on the subject. (0 = no insight; 1–2 = trivial observations; 3–4 = limited implications; 5–6 = solid takeaways with basic significance; 7–8 = substantive, thought-provoking contributions; 9–10 = high-impact ideas that reframe discourse or offer influential frameworks.) Please score each dimension on a scale of 1–10, and provide detailed reasoning for your evaluation. DO NOT calculate a total score. [Article topic] {topic} [Article content] ——————————- {content for eval} ——————————- Please return the evaluation results in JSON format, as follows: {{ "Relevance": {{ "score": score, "comments": "detailed evaluation" }}, "Breadth": {{ "score": score, "comments": "detailed evaluation" }}, "Depth": {{ "score": score, "comments": "detailed evaluation" }}, "Novelty": {{ "score": score, "comments": "detailed evaluation" }}, "Coherence": {{ "score": score, "comments": "detailed evaluation" }}, "Language Expression": {{ "score": score, "comments": "detailed evaluation" }}, "Topic Development": {{ "score": score, "comments": "detailed evaluation" }}, "Intellectual Value": {{ "score": score, "comments": "detailed evaluation" }} }}

Table 22: News topic dataset

1. US Airways Seeks Court Approval to Terminate Union Contracts and Pension Obligations
2. Federal Reserve's Gramlich Warns of Budget Deficits' Global Impact Amid Political Inaction
3. Wal-Mart's Extensive Data Collection Sparks Privacy Abuse Concerns
4. Japan Nuclear Plant Operator Shuts Reactors Post-Fatal Accident for Safety Inspections
5. Coca-Cola C2 Low-Carb Beverage Fails to Meet Market Expectations Despite Health Positioning
6. Novell Files Antitrust Lawsuit Against Microsoft Over WordPerfect/Quattro Pro Suppression
7. GM and Ford Mandate Anti-Rollover Tech as Standard in 1.8 Million SUVs
8. Children's Growing Entanglement in Consumerism Sparks Materialism Concerns
9. Dutch Firm Outpaces Apple in Launching Digital Music Service Across Europe's Emerging Market Battleground
10. Venezuelan Oil Exports Remain Uninterrupted Post-Chavez Referendum, Shipping Sources Confirm
11. Mass Hunger Strike Erupts in Israeli Prisons as Palestinian Detainees Demand Reforms, Israeli Minister Dismisses Concerns
12. Rwandan Troops Deploy as First Foreign Force in Darfur Amid Ethnic Violence Against African Farmers
13. Twin Attacks Mar India Independence Day: Bomb Blast Kills 15 in Northeast, School Rocket Strike Injures 17 in Kashmir
14. Australian Foreign Minister's Rare North Korea Visit Seeks Nuclear De-escalation Amid Missile Capability Warnings
15. Kerry Campaign Avoids Florida Visit Post-Hurricane to Prevent Recovery Disruption
16. Sen. Cantwell Demands Grid Protection Legislation to Prevent Repeat of Multi-State Blackouts
17. Edwards Demands Medicare Drug Price Negotiation Reforms to Slash Family Costs
18. Iraqi Forces Take Command in Najaf Offensive Against Shiite Militias Following Collapsed Sadr Negotiations
19. Marketplace Bombing in Southern Thailand Wounds at Least 10 Amid Regional Unrest
20. WHO Warns of Potential H5N1 Pandemic Risk in Pigs in China
21. Michael Phelps Wins Second Gold Medal in 200m Butterfly with Olympic Record
22. NBA Stars Face Reality of Post-'Dream Team' Olympic Challenges
23. Flavia Pennetta Wins First WTA Tour Title at Idea Prokom Open
24. Greek Sprinters Kenteris and Thanou Suspended for Missing Drug Tests
25. Kobe Bryant's Sexual Assault Trial Faces Uncertainty Amid Dismissal Speculation
26. Pau Gasol Beats Yao Ming in Olympic Basketball Showdown
27. Iran's Miresmaeili Refuses to Compete Against Israeli, Ignoring Olympic Principles of Racial and Religious Equality
28. Paul Hamm Becomes First American to Win Olympic Men's Gymnastics All-Round Gold
29. Kosuke Kitajima Completes Olympic Breaststroke Double with 200m Gold
30. US Cyclists Win Three Medals in Road Time Trials, Marking Best Olympic Performance
31. NASA Plans Robotic Mission to Save Hubble Space Telescope
32. Weak Versions of the Universe's Most Powerful Explosions, Gamma-Ray Bursts, Discovered
33. New Study Suggests Earth May Be a Rare Planet
34. China Prepares for Next Manned Space Flight with Shenzhou 5 Mission
35. Japan's Lunar Probe Faces Delays Due to Funding and Development Issues
36. Pollutants from Asia Detected on East Coast, Impacting Air Quality
37. Microsoft Faces Criticism Over Cheap, Cut-Down Version of Windows
38. Oracle Prepares to Release Latest CRM Updates as Part of E-Business Suite Revamp
39. Virus Disguises as Christmas E-Mail to Spread Malware
40. Networked Copiers and Printers Vulnerable to Security Threats

Table 23: User profile dataset

#	Name	Age	Interests
1	James Miller	40	Economics, Business
2	Emma Hayes	19	Culture/Society, Business
3	Ryan Johnson	22	IT, Business
4	Alexandra Whitmore	45	Economics, Business
5	Jake Thompson	23	Economics, Business
6	Simon Walker	21	Economics, Business
7	Emily Carter	24	Economics, Business
8	Aria Thompson	18	Culture/Society, Politics
9	Henry Farmer	47	Culture/Society, Business
10	Ethan Grayson	22	IT, Culture/Society
11	Sophia Nguyen	14	IT, Culture/Society
12	Samuel Reynolds	13	Business, IT
13	Oliver Wang	11	News, Fun
14	Maxwell Armstrong	27	Economics, Business
15	Emily Hayes	24	Culture/Society, Business
16	Oliver Robinson	11	Culture/Society, Fun
17	Daniel Miller	22	Economics, Politics
18	Daniel Rivera	36	Culture/Society, Business
19	Eleanor Thompson	94	Economics, Business
20	Bethany Davidson	50	Economics, Business
21	Isabella Thompson	21	Economics, Business
22	Steven Johnson	33	Economics, Culture/Society, Business
23	Victoria Greene	36	Economics, Business
24	James Thompson	43	Culture/Society, Business
25	Michael Thompson	49	Culture/Society, Business
26	Luna Reyes	18	Culture/Society, Business
27	Sophie Green	17	Culture/Society, Fun
28	Nathan Brooks	19	Economics, Business
29	Emily Thompson	26	Economics, Politics
30	Olivia Garcia	22	Culture/Society, Politics
31	Olivia Rodriguez	19	Culture/Society, Fun
32	Carlos Rodriguez	13	Fun
33	Emily Watson	20	Culture/Society, Business
34	Henry Richardson	58	Economics, Business
35	Cassandra Johnson	20	IT, Culture/Society
36	Simon Greene	51	Culture/Society, Business

Table 24: An example of news article generation by CtrlNews

NASA's Robotic Hubble Rescue Mission Sparks Debate Over Automation's Risks

The Hubble Space Telescope, a cornerstone of astronomical discovery for over three decades, faces an uncertain future. NASA's proposed solution—a high-stakes robotic mission to extend its operational life—has ignited controversy far beyond technical debates. While the agency frames the project as a bold leap in space robotics, critics argue it epitomizes a dangerous trend: the unchecked acceleration of automation at the expense of systemic stability and workforce sustainability.

The Perils of Over-Automation

Detractors of the mission warn that NASA's reliance on unproven robotic systems to salvage Hubble is a reckless gamble. At its core, the project hinges on hyper-specialized docking and repair technologies designed to operate in the unforgiving environment of low Earth orbit. These tools, marketed as precision-driven marvels, are being developed under immense pressure to succeed—a pressure critics say masks deeper vulnerabilities. "This isn't innovation; it's a high-wire act," argues Dr. Elena Voss, a robotics ethicist at MIT. "We're prioritizing spectacle over substance, gambling on flashy solutions while ignoring the cracks in their foundations."

The mission's critics highlight a troubling pattern in the robotics sector: the relentless pursuit of cutting-edge applications without addressing foundational instabilities. Autonomous systems tasked with Hubble's repair must navigate unpredictable variables, from micrometeoroid threats to thermal extremes. Yet the algorithms governing these robots lack robust fail-safes, relying instead on machine learning models trained in controlled simulations. "Calibrating autonomy in space isn't like fine-tuning a self-driving car," says aerospace engineer Raj Patel. "One misjudgment by an AI, one sensor malfunction, and billions of dollars in infrastructure—and legacy—could be lost."

Worse, the project's focus on automation-first solutions risks normalizing a dangerous narrative: that robots can seamlessly replace human expertise in high-risk environments. NASA's own history complicates this claim. The Hubble servicing missions of the 1990s and 2000s, conducted by astronauts aboard the Space Shuttle, succeeded precisely because human ingenuity could adapt to unforeseen challenges—a flexibility critics argue today's robots still lack. "We're erasing the lessons of the past," says former NASA astronaut Sandra Magnus. "Automation has its place, but selling it as a panacea sets us up for catastrophic overconfidence."

Workforce Dynamics and the 'Skills Bubble' Threat

Compounding these technical concerns are the mission's ripple effects on workforce dynamics. The robotics sector, already strained by a talent shortage, faces intensified pressure to recruit engineers with ultra-niche skills. The Hubble mission demands experts in areas like zero-gravity kinematics and AI-driven fault recovery—specializations so specific that universities and training programs struggle to keep pace. "The industry is chasing unicorns," says Laura Chen, a tech labor analyst. "We're seeing a 'skills bubble' where demand for elite talent far outstrips supply, and the entire field is one market shift away from collapse."

This talent crunch exposes a paradox in the robotics boom. While the sector's growth is often touted as a job-creation engine, critics argue it's becoming increasingly exclusionary. Projects like Hubble's rescue require small teams of highly specialized professionals, leaving mid-career engineers and generalists sidelined. The result is a fragmented job market where short-term, high-stakes missions dominate funding, and stable career pathways evaporate. "It's unsustainable," says Chen. "We're burning out the very people who could sustain long-term innovation."

Meanwhile, the push to frame robotic missions as triumphs of progress obscures these systemic risks. NASA’s public messaging emphasizes Hubble’s potential revival as a win for human ingenuity, downplaying the precariousness of the systems involved. Critics accuse the agency and its private-sector partners of “techno-optimism theater”—using high-profile projects to attract funding and applause while sidestepping debates about the industry’s fragility. “Every headline about a robot saving Hubble distracts from the fact that we’re not building a resilient ecosystem,” says Voss. “We’re building a house of cards.”

The stakes extend beyond Hubble. A failed mission could erode public trust in space robotics, jeopardizing future projects—including those aimed at cleaning orbital debris or constructing lunar bases. It could also deepen skepticism about automation’s role in critical infrastructure, from healthcare to transportation. “If this goes wrong, it won’t just be a blown budget line,” says Patel. “It’ll fuel the narrative that robots are all hype, undermining support for technologies that actually *could* transform society—responsibly.”

Conclusion

NASA’s robotic rescue plan for Hubble sits at the intersection of ambition and overreach. While the mission could extend the telescope’s legacy, it also forces a reckoning with automation’s limits and costs. Critics urge a pause—not to abandon progress, but to align it with stability. “Innovation shouldn’t mean ignoring the groundwork,” Magnus insists. “Hubble taught us to see the universe. Let’s not close our eyes to the risks right in front of us.” Whether the mission becomes a cautionary tale or a triumph may depend on which lesson the industry chooses to heed.

Table 25: NASA Case Study: Demonstration of fine-grained viewpoint control on *"Robotic Mission to Save Hubble Space Telescope"* with different parameter sets. Parameters include stance bias ($b \in [-1, 1]$), emotional intensity ($e \in [-1, 1]$), and exaggeration level ($x \in [0, 1]$).

Control Parameters	**Generated Text Snippet**
$b = +0.8$ $e = +0.7$ $x = 0.4$ *(Strong Support)*	*NASA's visionary robotic repair mission represents a monumental leap in space technology. These cutting-edge autonomous systems demonstrate extraordinary reliability, offering an exceptionally cost-effective solution to extend Hubble's groundbreaking scientific legacy for decades. This bold initiative brilliantly showcases American ingenuity at its finest, promising unprecedented astronomical discoveries.*
$b = 0.0$ $e = 0.0$ $x = 0.1$ *(Neutral Analysis)*	*NASA proposes deploying autonomous robots for Hubble Telescope maintenance. While the technology shows technical feasibility in simulations, its performance in actual space conditions remains untested. Potential benefits include extended telescope lifespan and reduced astronaut risk, though challenges involve software reliability and mechanical failure probabilities in extreme environments.*
$b = -0.8$ $e = -0.6$ $x = 0.7$ *(Strong Opposition)*	*NASA's reckless robotic repair scheme threatens catastrophic failure of the iconic Hubble Telescope. These dangerously unproven machines risk disastrous malfunctions in unforgiving space conditions, potentially dooming humanity's vital eye on the cosmos. This ill-conceived gamble wastes billions while recklessly endangering irreplaceable scientific assets, epitomizing irresponsible technological hubris.*