Agent Newsroom: Efficient Chronological Report Generation via Dynamic Multi-Agent Collaboration

Zhenhua Wang[1,2] **Chunlei Wang**[2] **Yue Geng**[2] **Bang Wang**[1,*]
[1]School of Electronic Information and Communications,
Huazhong University of Science and Technology
[2]North China Institute of Computer Systems Engineering
{wangzhenhua, wangbang}@hust.edu.cn

Abstract

Many real-world applications require generating a chronological report from an evolving document stream; Timeline Summarization (TLS) provides a standard testbed for this setting. While large language models (LLMs) improve event synthesis, most LLM-based TLS systems remain monolithic: they repeatedly process overlapping evidence and often mirror the corpus' bursty reporting patterns, producing redundant timelines with temporal/topical imbalance and high cost. We propose **MAS-TLS**, a multi-agent framework that casts TLS as a *newsroom-like* collaboration. A master editor steers balanced coverage by allocating system-visible evidence with a coverage–diversity objective; specialist reporter agents independently draft time-anchored, evidence-grounded events while cross-reviewing to limit redundancy; an adjudication round reconciles competing drafts and consolidates duplicates into a global timeline; and a non-stationary Bayesian controller adaptively staffs agents under token/time budgets. Experiments on three benchmarks show that MAS-TLS improves semantic coverage and temporal grounding while substantially reducing token usage and latency.

1 Introduction

Understanding complex knowledge domains requires unraveling the chronological evolution of events. Timeline Summarization (TLS) (Yan et al., 2011b; Chen et al., 2019; Gholipour Ghalandari and Ifrim, 2020) addresses this by identifying and chronologically organizing significant events from massive corpora. Such chronological summary reports provide compact situational awareness and support downstream decision-making. Recently, Large Language Models (LLMs) have delivered impressive performance on TLS benchmarks (Sojitra et al., 2024; Song et al., 2025), demonstrating

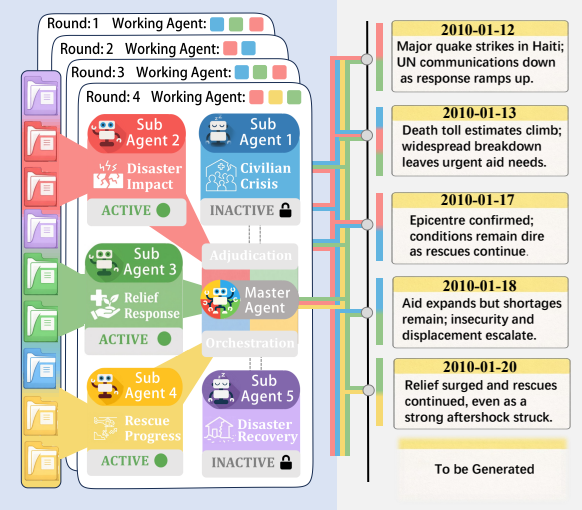

Figure 1: **Conceptual illustration of MAS-TLS.** Left: The Master Agent dynamically schedules role-specialized sub-agents over massive corpora. Right: The final output of TLS—the global timeline.

that their semantic reasoning capabilities can substantially push the field forward.

Despite these advances, most LLM-based TLS systems are still designed as single-model, non-interactive pipelines over the full corpus or long retrieved contexts. This design yields two recurring limitations in practice: **First, they incur high token and time costs.** Recent frameworks either insert multiple LLM calls across the pipeline or repeatedly prompt a single model with long, overlapping contexts, amplifying redundant processing and increasing wall-clock latency (Sojitra et al., 2024; Hu et al., 2024; Bao et al., 2025; Wu et al., 2025). Moreover, a single global reasoning scope offers no reliable way to separate sub-stories or deduplicate overlapping candidates, causing the same evidence to be re-processed with low marginal gain. **Second, they do not reliably maintain temporal and topical balance.** In real-world text collections, evidence is unevenly distributed over time and across themes (Yu et al., 2021).

*Corresponding author.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 25063–25084

While prior work attempts to mitigate such imbalance by adjusting temporal granularity (Zhang et al., 2025a) or imposing topical constraints (Qorib et al., 2025), most systems still lack explicit mechanisms for sub-story disentanglement and cross-candidate arbitration, making it difficult to balance competing narratives and maintain stable event pacing.

We argue that these limitations stem from a structural mismatch: TLS requires multiple complementary decisions (coverage, time anchoring, deduplication, and pacing), yet it is often realized as a single-model pipeline with limited explicit decomposition and coordination. A promising solution lies in **Multi-Agent Systems (MAS)**. While MAS has shown strong effectiveness in other domains (Hong et al., 2023; Zhang et al., 2024; Qian et al., 2024; Zhang et al., 2025b), its potential for summarization remains underexplored. Existing works (Wang et al., 2025; Kim and Kim, 2025) take early steps toward agentic summarization, but they typically rely on manually designed prompts and fixed interaction patterns, rather than coordinated, budget-aware agent behaviors. To the best of our knowledge, TLS has not yet been explicitly formulated as a coordinated and dynamic multi-agent problem.

To bridge this gap, we propose **MAS-TLS**, a hierarchical multi-agent framework inspired by a modern newsroom: a chief editor allocates coverage across reporters, reporters investigate in parallel and cross-check each other, and the newsroom iteratively curates a coherent timeline under a shared budget. Rather than a fixed pipeline, MAS-TLS runs an iterative *editorial cycle*: a Master Agent assigns system-visible evidence to role-specialized sub-agents (reporters) using a submodular coverage–diversity objective to promote balanced perspectives; sub-agents independently investigate their assigned evidence and draft time-anchored, evidence-grounded event candidates while maintaining cross-agent distinctiveness; the Master performs collaborative scrutiny by adjudicating competing proposals with cross-review, consolidating duplicates and retaining the most supported events to form the global timeline; finally, a non-stationary Bayesian bandit controller adaptively staffs the next-round agent committee under a token/time budget.

We conduct extensive experiments on three mainstream TLS benchmarks. Empirical results demonstrate that MAS-TLS is: **(1) Effective:** it achieves competitive-to-superior performance on challenging datasets with consistent gains in semantic coverage (AR) and temporal precision (Date-F1); **(2) Efficient:** it reduces token consumption by 31%–86% and inference time by 5%–52% compared to monolithic baselines; **(3) Balanced:** it mitigates corpus-induced bias, producing timelines with more balanced temporal pacing and topical coverage closer to professional summaries; **(4) Transferable:** it remains robust across different LLM backbones and generalizes across languages.

In short, the main contributions of our work can be outlined as follows:

1. **Framework Innovation:** To our knowledge, we are among the first to formulate TLS as a coordinated, dynamic multi-agent collaboration problem, moving beyond static summarization pipelines.

2. **Algorithmic Mechanism:** We propose **MAS-TLS**, a hierarchical newsroom-style multi-agent framework that (i) partitions evidence for balanced coverage, (ii) promotes cross-agent distinctiveness to reduce redundancy, (iii) adjudicates competing candidates to form a consistent timeline, and (iv) dynamically schedules agents under a token/time budget.

3. **Empirical Validation:** Extensive evaluations demonstrate that MAS-TLS achieves competitive-to-superior performance and reduces token usage by 31%–86%, delivering temporally balanced narratives with robust transferability.

2 Related Works

2.1 Timeline Summarization Task

Timeline Summarization (TLS) organizes evolving events into chronological narratives (Allan et al., 2001; Wang et al., 2014; Chen et al., 2019; Gholipour Ghalandari and Ifrim, 2020), typically formulated as a specialized multi-document summarization task (Martschat and Markert, 2018). Standard approaches prioritize extracting pivotal dates (Tran et al., 2015b; Steen and Markert, 2019) or pinpointing milestone events (Li et al., 2021; Chen et al., 2023). Recently, the field has been dominated by LLM-driven frameworks, which leverage semantic reasoning (Hu et al., 2024) and retrieval augmentation to scale across open domains (Wu

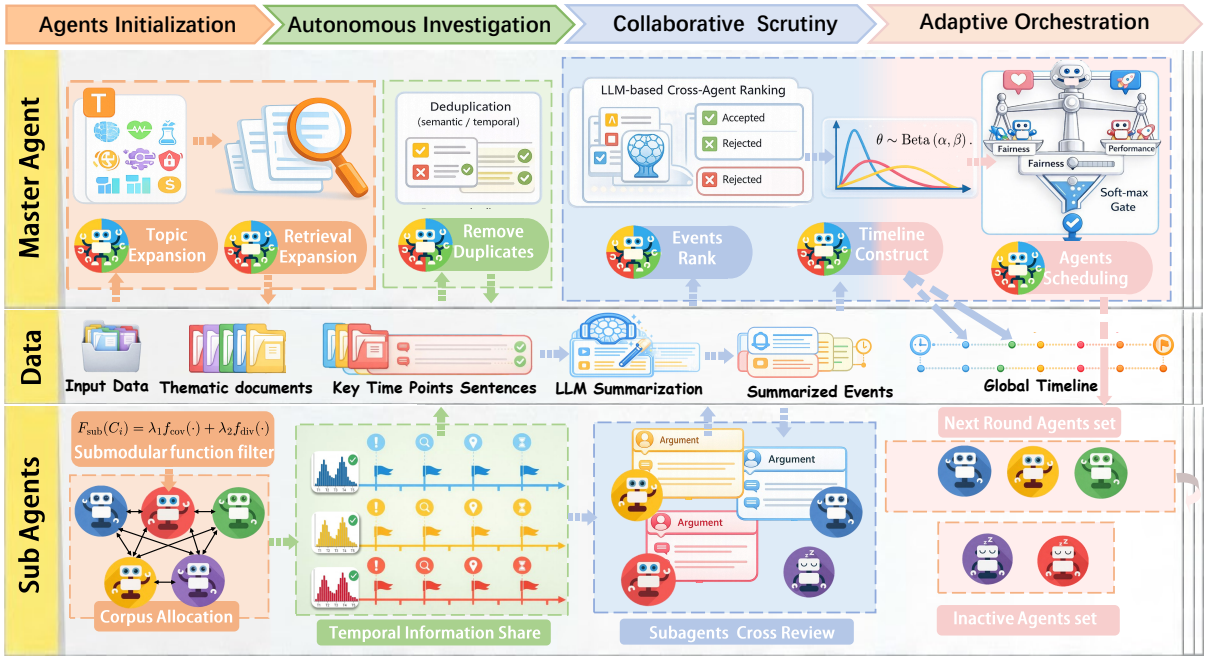

Figure 2: **MAS-TLS overview.** The figure is organized horizontally by an iterative workflow and vertically by roles: a **master agent** (top) coordinates multiple **sub-agents** (bottom) via a **data layer** of system-visible intermediate artifacts (middle). Sub-agents mine evidence-grounded candidates in parallel, and the master adjudicates cross-agent proposals to build the global timeline and schedule the next-round committee.

et al., 2025; Bao et al., 2025). Subsequent investigations adjusting granularity (Zhang et al., 2025a) or introducing constraints (Qorib et al., 2025) have provided further fresh insights. However, current high-performing methods remain confined to single-model optimization, whereas our MAS-TLS framework decomposes timeline generation into specialized, interacting agents under unified coordination.

2.2 Multi-Agent LLMs for Summarization

Multi-agent systems (MAS) have shown promise in handling long-context summarization by distributing reasoning loads across specialized roles (Hong et al., 2023). Hierarchical frameworks like Chain-of-Agents (Zhang et al., 2024) and NEXUS-SUM (Kim and Kim, 2025) employ sequential collaboration to extend context windows for book-length narratives. To improve evidence faithfulness and coverage, recent approaches introduce adversarial quizzing (Wang et al., 2025), multi-perspective personas (Luo et al., 2025), and agentic evaluation loops (Jeong et al., 2025; Chang et al., 2024; Fang et al., 2024, 2025). While pioneering, many prior multi-agent summarization systems primarily rely on prompt-engineered role specialization and pre-defined interaction patterns, with limited adaptive scheduling or explicit coordination mechanisms tailored to TLS. In this work, we study a coordinated, dynamic multi-agent formulation of timeline summarization and instantiate it as a newsroom-inspired framework.

3 Methodology

Task Setup. Given a query Q and a timestamped corpus $\mathcal{C} = \{d_1, \ldots, d_N\}$, Timeline Summarization (TLS) generates a timeline $\mathcal{Y} = \{(t_1, s_1), \ldots, (t_L, s_L)\}$, where each event contains a date t_l and a summary sentence s_l. Our goal is to produce a timeline that aligns with the reference $\mathcal{Y}^*$ in both semantic content and temporal grounding.

System Setup. MAS-TLS consists of a *Master Agent* A_M and specialized *Sub-Agents* $\mathcal{A} = \{a_1, \ldots, a_M\}$. The interaction protocol Φ follows a star topology: sub-agents operate on assigned sub-corpora $\mathcal{C}_i$ to propose time-anchored candidates, while the master performs centralized adjudication and maintains per-agent utility estimates $\mathbf{\Theta}^{(r)}$ to gate and schedule agents across rounds.

Guided by Φ, MAS-TLS constructs $\mathcal{Y}$ through a collaborative editorial cycle. As illustrated in Figure 2, the system proceeds in four steps: Editorial Assignment (Sec. 3.1), Autonomous Investigation (Sec. 3.2), Collaborative Scrutiny (Sec. 3.3), and Adaptive Orchestration (Sec. 3.4).

3.1 Master–Sub Agents Initialization

The initialization stage constructs the subtopic–subcorpus local state $\mathcal{S}_i^{(0)} = (\tau_i, \mathcal{C}_i)$ for each sub-agent a_i. Unlike prior approaches that perform static retrieval stage (Wang et al., 2023; Sojitra et al., 2024)—often missing global coordination for temporal spread—our Master Agent A_M performs a retrieval-augmented, global allocation to promote semantic coverage and temporal balance.

Specifically, we first expand an over-complete set of auxiliary sub-queries $\{\tau_j\}$ with an LLM to drive retrieval, and use them to retrieve a shared candidate pool $\mathcal{C}_{\text{cand}} \subseteq \mathcal{C}$ for subsequent global allocation. To make the decomposition MAS-compatible, the Master Agent jointly initializes $(\tau_i, \mathcal{C}_i)$ via a budgeted soft partition that approximately maximizes a submodular objective (Lin and Bilmes, 2011):

$$F_{\text{sub}}(\mathcal{C}_i) = \lambda_1 f_{\text{cov}}(\mathcal{C}_i) + \lambda_2 f_{\text{div}}(\mathcal{C}_i). \quad (1)$$

We jointly allocate documents from $\mathcal{C}_{\text{cand}}$ to all sub-agents by approximately maximizing $\sum_{i=1}^{M} F_{\text{sub}}(\mathcal{C}_i)$ under a per-agent budget, allowing each document to be assigned to multiple sub-agents.

The coverage term f_{cov} ensures broad relevance via a saturated facility-location form:

$$f_{\text{cov}}(\mathcal{C}_i) = \sum_{d_q \in \mathcal{C}_{\text{cand}}} \min\big(g(d_q, \mathcal{C}_i), \eta\big), \quad (2)$$

$$g(d_q, \mathcal{C}_i) = \sum_{d_p \in \mathcal{C}_i} \text{sim}(d_p, d_q) \cdot e^{-\alpha |t(d_p) - t(d_q)|}.$$

Here, $e^{-\alpha|\Delta t|}$ acts as a soft temporal consistency term, down-weighting semantically similar but chronologically distant evidence.

To encourage temporal spread and stabilize event pacing, f_{div} applies a group-wise concave aggregation over temporal bins $\{G_b\}_{b=1}^{B}$:

$$f_{\text{div}}(\mathcal{C}_i) = \sum_{b=1}^{B} \sqrt{\sum_{d_p \in G_b \cap \mathcal{C}_i} \sum_{d_q \in \mathcal{C}_{\text{cand}}} \text{sim}(d_p, d_q)}. \quad (3)$$

The resulting allocation determines $(\tau_i, \mathcal{C}_i)$, defining the initial evidence scope for each sub-agent a_i in subsequent coordination.

3.2 Autonomous Temporal Reasoning within Sub-Agents

Armed with the sub-corpus $\mathcal{C}_i$, sub-agent a_i enters the proposal phase. Unlike standard frequency-based heuristics inherit corpus burstiness (Gholipour Ghalandari and Ifrim, 2020), a_i selects complementary time anchors by balancing local salience with cross-agent distinctiveness, reducing redundant candidates and stabilizing event pacing.

First, a_i defines a local probability distribution $P_i(t)$ based on normalized sentence frequencies within $\mathcal{C}_i$:

$$P_i(t) = \frac{\text{Freq}_i(t) + \epsilon}{\sum_{t'} \big(\text{Freq}_i(t') + \epsilon\big)}. \quad (4)$$

To reduce redundancy, a_i also takes into account the temporal distributions $\{P_j\}_{j \neq i}$ of other sub-agents (updated each round), discouraging multiple agents from converging on the same bursty periods. Adopting a mutual-information–inspired formulation (Konan et al., 2022), the agent computes a divergence score $D_i(t)$ that quantifies how strongly its focus departs from the distributions of its peers:

$$D_i(t) = \frac{1}{M-1} \sum_{j \neq i} \log \frac{P_i(t)}{\max(P_j(t), \epsilon)}. \quad (5)$$

A high $D_i(t)$ indicates a locally salient timestamp that is under-covered by other agents, encouraging complementary temporal coverage instead of duplicated peaks.

These signals are combined into a final temporal score:

$$\text{Score}_i(t) = P_i(t)\, e^{\beta D_i(t)}, \quad (6)$$

where $\beta > 0$ controls the influence of cross-agent distinctiveness.

Finally, sub-agent a_i selects the peak timestamp $t_i^{(r)} = \arg\max_t \text{Score}_i(t)$ and prompts its underlying LLM to generate a candidate event $o_i^{(r)} = (t_i^{(r)}, s_i^{(r)})$, submitting this proposal to the master agent for multi-agent adjudication.

3.3 Collaborative Cross-Review and Master Adjudication

With candidate events $o_i^{(r)}$ prepared, we perform a lightweight cross-candidate arbitration to (i) remove redundant/overlapping proposals and (ii) enable efficient scrutiny without quadratic, multi-round interactions. Instead of pairwise multi-round discussions (Koupaee et al., 2025), we adopt a single-pass cross-review protocol: each sub-agent submits its proposal and then, after observing all peer proposals, produces two concise signals— a

supportive argument S_i^{sup} (salience & time consistency) and a critical argument S_i^{cri} (overlap/redundancy or weaknesses of competing candidates).

The master agent aggregates $\mathcal{I}^{(r)} = \{(o_i^{(r)}, S_i^{\text{sup}}, S_i^{\text{cri}})\}_{i \in \mathcal{A}^{(r)}}$ and ranks candidates with an LLM-based judge:

$$\pi^{(r)} = \text{LLM}_{\text{rank}}(\mathcal{I}^{(r)}), \tag{7}$$

It then selects top-ranked events that are non-redundant with the current timeline $\mathcal{Y}^{(r-1)}$ (e.g., filtering near-duplicates or overlapping time anchors) to update $\mathcal{Y}$. Selected agents are marked **Accepted** and others **Rejected**; this binary outcome is used as the reward signal for budget-aware Bayesian scheduling in the next phase.

3.4 Adaptive Sub-Agent Orchestration

Since agent utility can vary across rounds, we cast sub-agent scheduling as a non-stationary selection problem. Inspired by dynamic expert selection frameworks (Zhang et al., 2025c), we model this process as a Non-Stationary Bayesian Multi-Armed Bandit (MAB) task (Besbes et al., 2014). In each round, A_M dynamically gates the committee $\mathcal{A}^{(r)}$ to control token/time overhead by balancing exploitation (activating proven agents) with fairness-driven exploration. To track evolving utility, A_M maintains a Beta belief distribution $\text{Beta}(m_i, f_i)$ for each agent. Upon receiving the binary reward (Accepted/Rejected) from Sec. 3.3, A_M updates the effective counts with a forgetting factor $\gamma \in (0, 1]$ to decay historical evidence:

$$\begin{aligned} m_i^{(r)} &= \gamma m_i^{(r-1)} + \mathbb{1}[a_i \in \text{Accepted}], \\ f_i^{(r)} &= \gamma f_i^{(r-1)} + \mathbb{1}[a_i \in \text{Rejected}]. \end{aligned} \tag{8}$$

A_M then samples a base score $\hat{\theta}_i^{(r)}$ via Thompson Sampling. Simultaneously, to avoid repeatedly selecting the same agents and improve topical balance, A_M manages a fairness credit $q_i^{(r)}$ that accumulates during inactivity ($a_i \notin \mathcal{A}^{(r-1)}$) and depletes upon selection:

$$q_i^{(r)} = \begin{cases} q_i^{(r-1)} + \rho & \text{inactive} \\ \max(0, q_i^{(r-1)} - (1-\rho)) & \text{active} \end{cases}. \tag{9}$$

The final activation probability is derived from a composite score

$$s_i^{(r)} = \hat{\theta}_i^{(r)} + \lambda_{\text{fair}} q_i^{(r)} \tag{10}$$

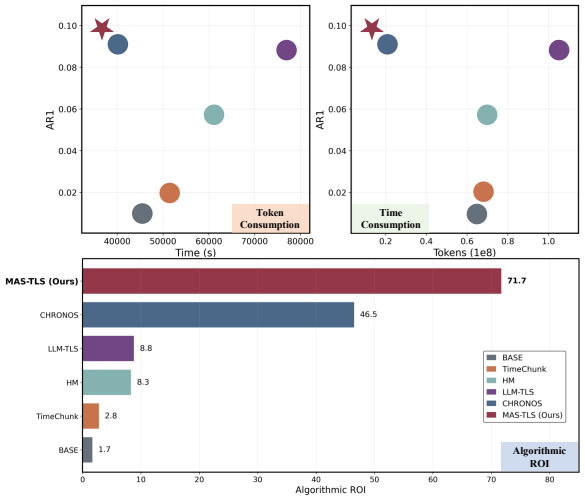

Figure 3: **Efficiency Analysis (Entities). Top:** AR-1 versus wall-clock time and token usage. **Bottom:** algorithmic ROI measured as AR-1 per 10^{10} tokens. MAS-TLS achieves a favorable quality–efficiency trade-off among compared methods.

via Softmax: $p_i^{(r)} \propto \exp(s_i^{(r)}/u)$. To allow flexible committee sizing, A_M employs a Bernoulli gating mechanism $z_i^{(r)} \sim \text{Bern}(\min(1, K_c p_i^{(r)}))$, thereby instantiating $\mathcal{A}^{(r+1)} = \{a_i \mid z_i^{(r)} = 1\}$ for the next round.

4 Experiments

4.1 Experimental Setup

Datasets. We evaluate MAS-TLS on three widely used datasets: **T17** (Tran et al., 2013) for general news, **Crisis** (Tran et al., 2015a) for dense conflict narratives and **Entities** (Gholipour Ghalandari and Ifrim, 2020) for long-span biographical timelines. Detailed statistics are provided in App. C.1.

Baselines. We compare MAS-TLS against two categories of baselines (details in APP. A): (1) **Traditional**: Graph/clustering-based models including TRAN (Tran et al., 2013) MAR (Martschat and Markert, 2018), DATE (Gholipour Ghalandari and Ifrim, 2020), SDF (La Quatra et al., 2021), CLUST (Gholipour Ghalandari and Ifrim, 2020), EGC (Li et al., 2021), (2) **LLM-based**: frameworks including TimeChunk (Sojitra et al., 2024) HM (Zhang et al., 2025a), LLM-TLS (Hu et al., 2024), and CHRONOS (Wu et al., 2025). A direct prompting baseline (**BASE**) is included to benchmark the backbone's raw capability.

Metrics. We adopt a multi-dimensional evaluation protocol: (1) **Quality**: Following the standard *Tilse* framework (Martschat and Markert, 2017), we

Table 1: **Main Results.** Performance comparison across three benchmarks. We report **AR-1/AR-2** for semantic coverage and **Date-F1** for temporal precision. MAS-TLS achieves competitive performance across both semantic and temporal metrics. Best results are **bolded**.

Model	Crisis			Entities			T17		
	AR-1	AR-2	Date-F1	AR-1	AR-2	Date-F1	AR-1	AR-2	Date-F1
BASE	0.016	0.005	0.192	0.011	0.003	0.045	0.047	0.011	0.163
TRAN (Tran et al., 2013)	0.052	0.012	0.289	0.036	0.011	0.185	0.094	0.022	0.517
MAR (Martschat and Markert, 2018)	0.075	0.016	0.281	0.042	0.009	0.167	0.105	0.030	0.544
DATE (Gholipour Ghalandari and Ifrim, 2020)	0.089	0.026	0.295	0.057	0.017	0.205	0.120	0.035	0.544
SDF (La Quatra et al., 2021)	0.086	0.018	0.302	0.051	0.014	0.197	0.120	0.035	0.553
CLUST (Gholipour Ghalandari and Ifrim, 2020)	0.061	0.013	0.226	0.051	0.015	0.174	0.082	0.020	0.407
EGC (Li et al., 2021)	0.079	0.015	0.291	-	-	-	0.103	0.024	0.550
TimeChunk (Sojitra et al., 2024)	0.025	0.008	0.258	0.019	0.005	0.050	0.051	0.014	0.176
HM (Zhang et al., 2025a)	0.096	0.025	0.313	0.058	0.022	0.209	0.108	0.028	0.541
LLM-TLS (Hu et al., 2024)	0.111	0.030	0.330	0.092	0.040	0.240	0.119	0.035	0.547
CHRONOS (Wu et al., 2025)	0.108	0.031	0.323	0.094	0.036	0.241	0.116	**0.036**	0.549
MAS-TLS (Ours)	**0.113**	**0.033**	**0.337**	**0.099**	**0.041**	**0.251**	**0.123**	0.035	**0.556**
Δ vs. Best Baseline	+1.8%	+6.5%	+2.1%	+5.3%	+2.5%	+4.1%	+2.5%	-2.8%	+0.5%

Table 2: **Efficiency and ROI analysis (Entities).** We report token usage and wall-clock time over the Entities dataset, AR-1, and efficiency metrics including AR-1 per 10^{10} tokens and AR-1 per hour.

Method	Resource Consumption		Metric	Efficiency ROI (Higher is Better)	
	Tokens (M) ↓	Time (s) ↓	AR-1 ↑	AR / 10^{10} Toks ↑	AR / Hour ↑
LLM-TLS	103.9	77,204	0.092	8.8	0.0043
BASE	65.5	45,908	0.011	1.7	0.0009
HM	68.6	61,432	0.057	8.3	0.0033
TimeChunk	67.4	51,267	0.019	2.8	0.0013
CHRONOS	20.2	39,009	0.094	46.5	0.0087
MAS-TLS (Ours)	**13.8**	**37,078**	**0.099**	**71.7**	**0.0096**
Δ vs. Best Baseline	31.7% ↓	5.0% ↓	5.3% ↑	54.2% ↑	10.3% ↑

report **AR-1/2** for semantic coverage and **Date-F1** for temporal precision (detailed in APP. C.2) (2) **Efficiency**: We track total **token usage** and **inference time** to quantify computational costs. (3) **Balance**: We introduce **CV** and **JS Divergence** (Yan et al., 2011a) to quantify both temporal burstiness inheritance and topical skew, comparing system distributions against the corpus and gold references; we also report **Gini** as a topical skewness diagnostic (detailed in APP. C.3).

Implementation Details. We employ Qwen-3-32b as the default backbone. We set sub-agents $M = 5$, coordination $\beta = 0.1$, forgetting factor $\gamma = 0.95$, and fairness $\lambda_{\text{fair}} = 0.05$. All experiments are conducted on dual NVIDIA A800 GPUs.

4.2 Results: Effectiveness

Table 1 compares MAS-TLS with 11 baselines across three TLS benchmarks. Overall, MAS-TLS achieves strong semantic coverage while improving temporal precision, showing consistent advantages across datasets with different timeline characteristics.

Finding ①: MAS-TLS improves both semantic coverage (AR) and temporal accuracy (Date-F1) across benchmarks. As shown in Table 1, MAS-TLS obtains the best AR-1 on Crisis, Entities, and T17, and achieves the highest Date-F1 on all three datasets. This pattern suggests that the gains are not confined to a single benchmark, but remain stable across both dense crisis narratives and long-span biographical timelines. On the most challenging Entities benchmark, MAS-TLS improves AR-1 by **5.3%** over the strongest baseline and increases Date-F1 by **4.1%**. Compared with the single-pass BASE model, the consistent gains across AR and Date-F1 support the effectiveness of the proposed collaborative multi-agent workflow with explicit

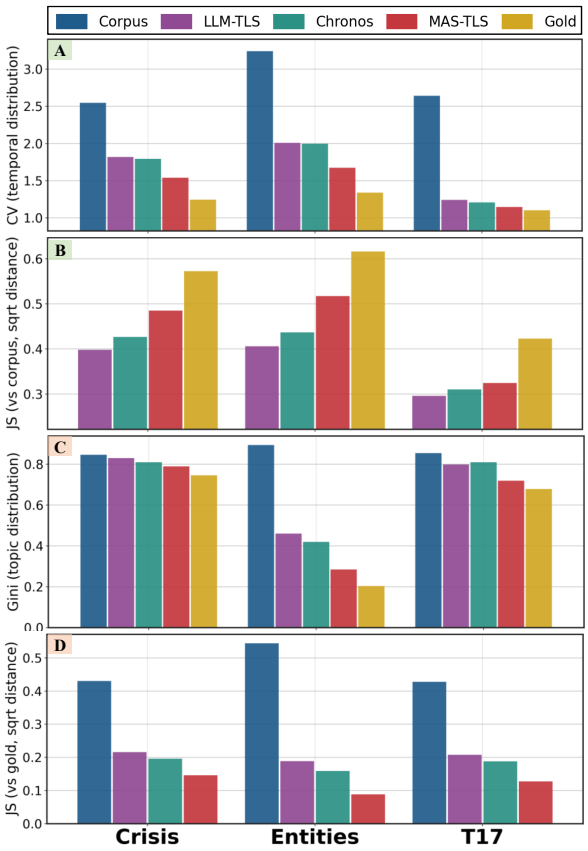

Figure 4: **Temporal and topical balance.** (A) interval CV, (B) temporal JS to corpus, (C) topical Gini, (D) topical JS to gold. MAS-TLS is closest to gold overall.

time-aware coordination, rather than attributing the improvements solely to a different backbone.

4.3 Results: Efficiency

We evaluate computational efficiency on the Entities benchmark for LLM-based methods, reporting (1) token usage, (2) wall-clock inference time, and (3) return on investment (ROI).

Finding ②: MAS-TLS is resource-friendly, achieving favorable ROI with substantially fewer tokens and lower latency. As shown in Table 2 and Figure 3, MAS-TLS achieves the best AR-1 (0.099) with the fewest tokens (13.8M), yielding the highest ROI (71.7). In contrast, LLM-TLS reaches a comparable AR-1 (0.092) but requires 103.9M tokens (7.5× more), resulting in a much lower ROI (8.8). Moreover, MAS-TLS reduces wall-clock time to 37,078s, which is 52% faster than LLM-TLS (77,204s), while remaining competitive with the efficiency-oriented CHRONOS (39,009s). Overall, these results are consistent with our design that combines early evidence filtering with budget-aware scheduling to reduce redundant context and generation during inference.

Table 3: Entities ablation with quality, balance and cost

Variant	AR-1 ↑	Temp. CV ↓	Tokens (M) ↓
Single Agent	0.084	2.410	**9.0**
w/o F_{sub}	0.093	1.868	15.8
w/o $D_i(t)$	0.088	1.862	14.1
w/o LLM_{rank}	0.086	1.714	9.4
w/o Scheduling	0.091	1.902	18.5
MAS-TLS	**0.099**	**1.671**	13.8

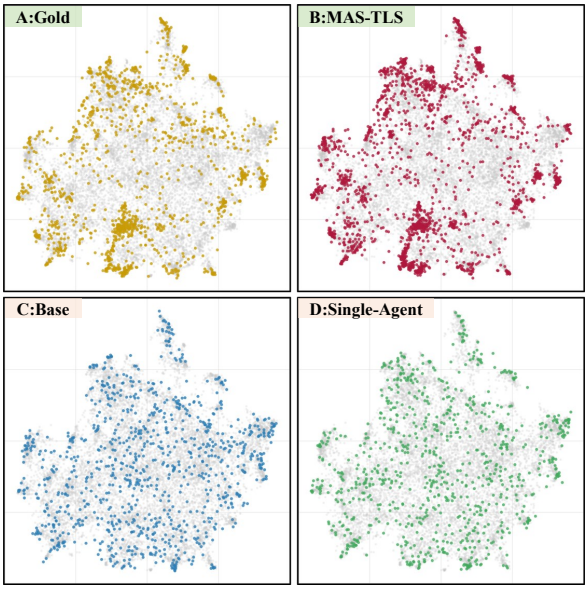

Figure 5: **Semantic-space reweighting.** Gray: corpus; color: generated timeline (or gold). MAS-TLS shows a redistribution pattern closer to gold than other methods.

4.4 Results: Temporal and Topical Balance

To diagnose whether a system inherits skew from the input stream, we analyze temporal and topical balance. In Figure 4 we measure temporal pacing stability with **CV** and corpus coupling with the **corpus–timeline JS distance**. For topics, we report the **topical JS distance to gold** and **Gini** in a corpus-built topic space (details in App. C.3). We further include a lightweight semantic-space visualization to qualitatively illustrate how each method reweights probability mass relative to the corpus in Figure 5 (details in App. C.4).

Finding ③: MAS-TLS produces a more balanced narrative in both time and topics. As shown in Figure 4 (A,B), MAS-TLS achieves the lowest inter-event CV and remains closest to the gold reference, while exhibiting weaker coupling to corpus reporting peaks (larger corpus–timeline JS distance), suggesting a more stable and gold-consistent temporal allocation. In Figure 4 (C,D), it attains the smallest topical JS distance to gold

and reduces topical concentration relative to the corpus (lower Gini). The semantic-space reweighting maps in Figure 5 provide qualitative evidence in the same direction, showing a redistribution trend closer to the gold reference than other baselines. The observed improvements align with the intended roles of our modules: allocation and distinctiveness promote balanced evidence, adjudication reconciles competing candidates, and scheduling controls budget while sustaining coverage.

5 Analysis

5.1 Ablation Study

We ablate MAS-TLS on ENTITIES (Table 3), reporting quality (AR), temporal balance (Temp. CV), and cost (Tokens).

Architecture Analysis (Single vs. Multi-Agent). We compare MAS-TLS with a *Single Agent* baseline (App. B.1) that runs a monolithic, non-interactive pipeline over the full corpus without role specialization. MAS-TLS achieves higher quality and substantially better temporal balance (CV: 1.671 vs. 2.410), with higher token cost due to multi-agent interactions. This pattern is also reflected in the semantic-space reweighting visualization (Figure 5), where the Single Agent variant shows a larger deviation from the gold redistribution pattern than MAS-TLS.

Component Ablations. We perform an ablation study on four key mechanisms by replacing each with a simpler ablated variant: (1) **w/o** $F_{\textbf{sub}}$, substituting the submodular allocation in Eq. 1 with random partitioning; (2) **w/o** $D_i(t)$, setting the MI-inspired distinctiveness term in Eq. 5 to zero so that timestamp selection relies only on local frequency; (3) **w/o** $\textbf{LLM}_{\textbf{rank}}$ in Eq. 7, bypassing the LLM-based adjudication step; and (4) **w/o Scheduling**, freezing the Bayesian update in Eq. 8 to enforce a static full committee.

As shown in Table 3, removing $D_i(t)$ or $\textbf{LLM}_{\textbf{rank}}$ leads to the largest AR-1 drops, indicating that cross-agent distinctiveness and adjudication are important for selecting complementary, non-redundant events. Removing **Scheduling** primarily affects efficiency: disabling the Bandit scheduler increases token consumption by **34%**, suggesting that dynamic gating is important for resource-aware scaling. Finally, **w/o** $F_{\textbf{sub}}$ reduces semantic coverage and worsens temporal balance while increasing token usage, supporting the role of structured evidence allocation in organizing useful evidence for downstream agents.

5.2 Case Study

In this section, we qualitatively analyze the coordination mechanisms of the agentic newsroom using the 2011 Yemen crisis as a testbed. Figure 6 visualizes the scheduling dynamics alongside the generated timeline.

As shown by the activation matrix and reliability trajectories (Right), MAS-TLS counteracts topic inertia across scheduling rounds and adjusts its focus as the crisis evolves. In the early rounds, protest-related agents dominate, reflecting the initial concentration of evidence around demonstrations and violent repression. In **Round 4**, however, as the narrative shifts from street protests toward political transition, the system down-weights *Agent 1 (Protest)* while boosting *Agent 3 (Political)*. This transition suggests that the scheduler is able to move beyond high-frequency early signals and reallocate attention to newly emerging but consequential developments.

Later, in **Rounds 7–8**, the fairness mechanism triggers concurrent activation of *Agent 4 (Tribal)* and *Agent 5 (Mediation)*, so multiple specialized agents propose complementary evidence in the same round for master adjudication. This helps the system capture intertwined sub-stories in the late-stage crisis and avoid collapsing the timeline into a single dominant thread, resulting in a more balanced and coherent chronological narrative.

5.3 Transferability

We evaluate transferability along three dimensions: (1) **backbone generality**, (2) **cross-lingual transfer**, and (3) **hyperparameter sensitivity**. Detailed results are reported in Tables 5 and 6, and Figure 7.

Finding ④: MAS-TLS transfers reliably across backbones and languages, with interpretable hyperparameter trade-offs. Across the evaluated backbones, MAS-TLS maintains consistent performance trends relative to LLM-based baselines, with less than 8% relative variation in AR-1 across backbones. Notably, the framework remains competitive on both closed-source and open-weight models, suggesting that its gains are not tied to a particular model family. We observe similar trends in a cross-lingual setting, where semantic coverage improves by about 9% relative to the strongest baseline while remaining competitive on temporal

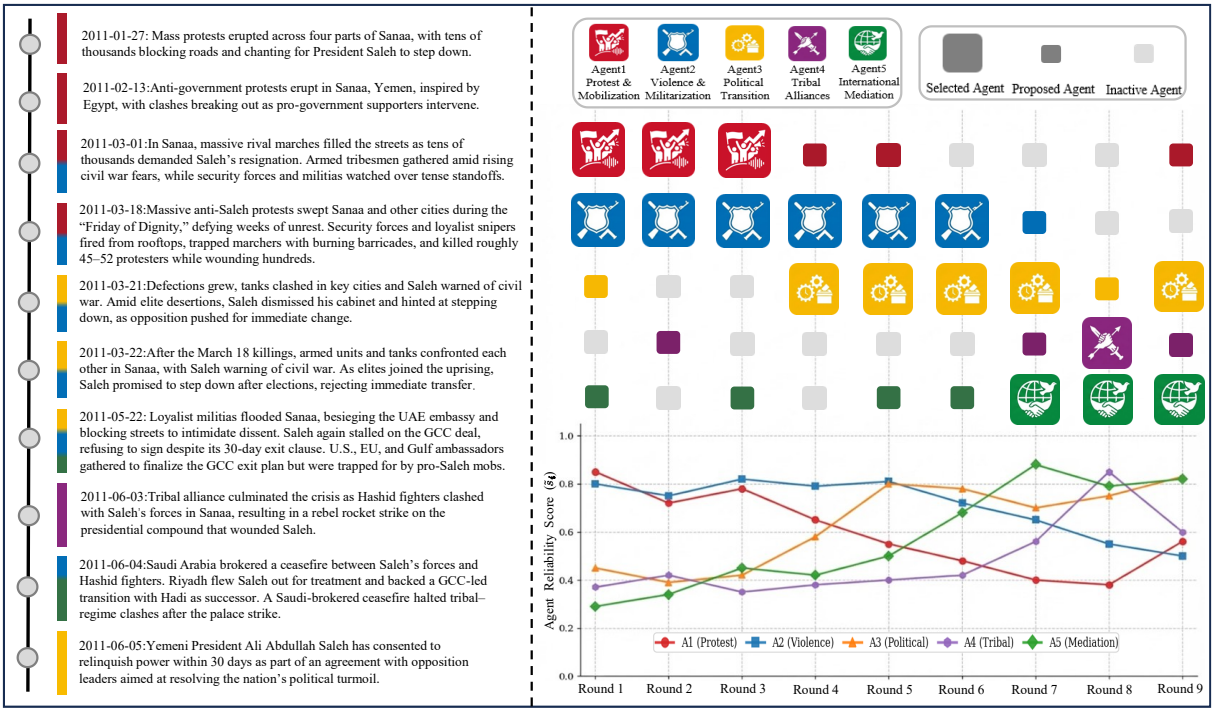

Figure 6: Visualizing the scheduling dynamics on the Yemen 2011 timeline. **Left:** The generated timeline, where the colored bars represent the specific topic distribution at each timestamp, visually confirming the topical balance. **Right:** The agent activation matrix (top) and reliability scores (bottom) organized by scheduling rounds, illustrating how MAS-TLS rebalances narrative focus via adaptive handoffs and fairness-driven collaboration.

precision. In addition, the sensitivity analysis indicates interpretable trade-offs with respect to key hyperparameters: agent number affects decomposition granularity, the temporal coefficient controls complementary time coverage, and the scheduler parameters balance responsiveness with fair exploration. Overall, these results support the transferability of our workflow across model and language choices, with predictable quality–cost trade-offs.

5.4 Evidence Support

We audit evidence support and date consistency using a system-visible evidence protocol. Each sampled event is assessed in terms of both content support and time support, producing the four-way breakdown in Figure 8. The results show that MAS-TLS yields a low unsupported rate (about 5.3–7.9% across benchmarks), and that most generated events are fully supported by the available evidence.

Removing LLM_{rank} leads to similar distributions, with unsupported rates changing by less than 3% overall, suggesting that adjudication does not trade off grounding for AR gains. It improves selection without increasing unsupported events. Details are in APP. D.5.

6 Conclusion

In this paper, we formulate TLS as a dynamic, collaborative multi-agent process, shifting the modeling focus from monolithic sequence generation to coordinated decision making. We propose MAS-TLS, a hierarchical master–sub agent framework inspired by a newsroom workflow, where a master agent coordinates specialized sub-agents to reduce redundancy and rectify narrative imbalance via submodular allocation, cross-agent adjudication, and Bayesian scheduling. Empirical results confirm that MAS-TLS achieves a competitive trade-off, delivering strong performance with significantly reduced costs, while also improving temporal and topical balance. We believe that MAS-TLS contributes to the development of transparent, adaptive, and resource-efficient collective intelligence for complex information streams, and may inspire future research on structured multi-agent systems for timeline summarization and broader information-stream processing tasks.

Acknowledgement

This work was partially supported by the National Natural Science Foundation of China under Grant No.62172167.

Limitations

Our study focuses on retrospective, news-style timeline summarization benchmarks, and we have not evaluated settings such as real-time streams, substantially different domains (e.g., scientific or social media corpora), or broader multilingual coverage. While the proposed framework is modular, deploying it in new settings may require revisiting upstream retrieval choices and budget configurations.

In the current implementation, coordination is instantiated through fixed agent roles/prompts, a predetermined interaction protocol, and explicit objectives for evidence allocation, redundancy control, and scheduling. These design choices improve interpretability and controllability, but we leave learning adaptive coordination policies and automatically selecting interaction depth as future work.

Finally, MAS-TLS inherits dependence on the quality of retrieved evidence and temporal cues. We mitigate this by evidence-grounded generation and an adjudication step, but we do not claim to eliminate all unsupported inferences. More comprehensive human-centered evaluation of faithfulness and usability would further strengthen conclusions.

Ethics Statement

This paper has no particular ethical consideration.

References

James Allan, Rahul Gupta, and Vikas Khandelwal. 2001. Temporal summaries of new topics. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 10–18.

Chenlong Bao, Shijie Li, Minghao Hu, Ming Qiao, Bin Zhang, Jin-Tao Tang, Shasha Li, and Ting Wang. 2025. R2A-TLS: Reflective retrieval-augmented timeline summarization with causal-semantic integration. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 752–766, Suzhou, China. Association for Computational Linguistics.

Omar Besbes, Yonatan Gur, and Assaf Zeevi. 2014. Stochastic multi-armed-bandit problem with non-stationary rewards. *Advances in neural information processing systems*, 27.

Yapei Chang, Kyle Lo, Tanya Goyal, and Mohit Iyyer. 2024. Booookscore: A systematic exploration of book-length summarization in the era of LLMs. In *The Twelfth International Conference on Learning Representations*.

Xiuying Chen, Zhangming Chan, Shen Gao, Meng-Hsuan Yu, Dongyan Zhao, and Rui Yan. 2019. Learning towards abstractive timeline summarization. In *IJCAI*, pages 4939–4945.

Xiuying Chen, Mingzhe Li, Shen Gao, Zhangming Chan, Dongyan Zhao, Xin Gao, Xiangliang Zhang, and Rui Yan. 2023. Follow the timeline! generating an abstractive and extractive timeline summary in chronological order. *ACM Transactions on Information Systems*, 41(1):1–30.

Dengzhao Fang, Jipeng Qiang, Xiaoye Ouyang, Yi Zhu, Yunhao Yuan, and Yun Li. 2025. Collaborative document simplification using multi-agent systems. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 897–912.

Jiangnan Fang, Cheng-Tse Liu, Jieun Kim, Yash Bhedaru, Ethan Liu, Nikhil Singh, Nedim Lipka, Puneet Mathur, Nesreen K Ahmed, and Franck Dernoncourt. 2024. Multi-LLm text summarization. *arXiv preprint arXiv:2412.15487*.

Demian Gholipour Ghalandari and Georgiana Ifrim. 2020. Examining the state-of-the-art in news timeline summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1322–1334, Online. Association for Computational Linguistics.

Sirui Hong, Mingchen Zhuge, Jonathan Chen, Xiawu Zheng, Yuheng Cheng, Jinlin Wang, Ceyao Zhang, Zili Wang, Steven Ka Shing Yau, and Zijuan Lin. 2023. Metagpt: Meta programming for a multi-agent collaborative framework. In *The Twelfth International Conference on Learning Representations*.

Qisheng Hu, Geonsik Moon, and Hwee Tou Ng. 2024. From moments to milestones: Incremental timeline summarization leveraging large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7232–7246.

Yeonseok Jeong, Minsoo Kim, Seung-won Hwang, and Byung-Hak Kim. 2025. Agent-as-judge for factual summarization of long narratives. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 23602–23619, Suzhou, China. Association for Computational Linguistics.

Hyuntak Kim and Byung-Hak Kim. 2025. NexusSum: Hierarchical LLM agents for long-form narrative summarization. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10120–10157, Vienna, Austria. Association for Computational Linguistics.

Sachin G Konan, Esmaeil Seraj, and Matthew Gombolay. 2022. Iterated reasoning with mutual information in cooperative and byzantine decentralized teaming. In *International Conference on Learning Representations*.

Mahnaz Koupaee, Jake W. Vincent, Saab Mansour, Igor Shalyminov, Han He, Hwanjun Song, Raphael Shu, Jianfeng He, Yi Nian, Amy Wing-mei Wong, Kyu J. Han, and Hang Su. 2025. Faithful, unfaithful or ambiguous? multi-agent debate with initial stance for summary evaluation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 12209–12246, Albuquerque, New Mexico. Association for Computational Linguistics.

Moreno La Quatra, Luca Cagliero, Elena Baralis, Alberto Messina, and Maurizio Montagnuolo. 2021. Summarize dates first: A paradigm shift in timeline summarization. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 418–427.

Manling Li, Tengfei Ma, Mo Yu, Lingfei Wu, Tian Gao, Heng Ji, and Kathleen McKeown. 2021. Timeline summarization based on event graph compression via time-aware optimal transport. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6443–6456.

Hui Lin and Jeff Bilmes. 2011. A class of submodular functions for document summarization. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 510–520, Portland, Oregon, USA. Association for Computational Linguistics.

Yun Luo, Yingjie Li, Xiangkun Hu, Qinglin Qi, Fang Guo, Qipeng Guo, Zheng Zhang, and Yue Zhang. 2025. PerSphere: A comprehensive framework for multi-faceted perspective retrieval and summarization. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 21790–21805, Vienna, Austria. Association for Computational Linguistics.

Sebastian Martschat and Katja Markert. 2017. Improving rouge for timeline summarization. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 285–290.

Sebastian Martschat and Katja Markert. 2018. A temporally sensitive submodularity framework for timeline summarization. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 230–240, Brussels, Belgium. Association for Computational Linguistics.

Chen Qian, Wei Liu, Hongzhang Liu, Nuo Chen, Yufan Dang, Jiahao Li, Cheng Yang, Weize Chen, Yusheng Su, Xin Cong, Juyuan Xu, Dahai Li, Zhiyuan Liu, and Maosong Sun. 2024. ChatDev: Communicative agents for software development. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15174–15186, Bangkok, Thailand. Association for Computational Linguistics.

Muhammad Reza Qorib, Qisheng Hu, and Hwee Tou Ng. 2025. Just what you desire: Constrained timeline summarization with self-reflection for enhanced relevance. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 25065–25073.

Daivik Sojitra, Raghav Jain, Sriparna Saha, Adam Jatowt, and Manish Gupta. 2024. Timeline summarization in the era of llms. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2657–2661.

Jiayu Song, Mahmud Elahi Akhter, Dana Atzil-Slonim, and Maria Liakata. 2025. Temporal reasoning for timeline summarisation in social media. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 28085–28101, Vienna, Austria. Association for Computational Linguistics.

Julius Steen and Katja Markert. 2019. Abstractive timeline summarization. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 21–31.

Giang Tran, Mohammad Alrifai, and Eelco Herder. 2015a. Timeline summarization from relevant headlines. In *Advances in Information Retrieval: 37th European Conference on IR Research, ECIR 2015, Vienna, Austria, March 29-April 2, 2015. Proceedings 37*, pages 245–256. Springer.

Giang Tran, Eelco Herder, and Katja Markert. 2015b. Joint graphical models for date selection in timeline summarization. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics*, volume 1, pages 1598–1607. Association for Computational Linguistics.

Giang Binh Tran, Tuan A Tran, Nam-Khanh Tran, Mohammad Alrifai, and Nattiya Kanhabua. 2013. Leveraging learning to rank in an optimization framework for timeline summarization. In *SIGIR 2013 Workshop on Time-aware Information Access (TAIA)*.

Sha Wang, Yuchen Li, Hanhua Xiao, Lambert Deng, and Yanfei Dong. 2023. Web news timeline generation with extended task prompting. *arXiv preprint arXiv:2311.11652*.

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

Zhenhua Wang, Lidan Shou, Ke Chen, Gang Chen, and Sharad Mehrotra. 2014. On summarization and timeline generation for evolutionary tweet streams. *IEEE Transactions on Knowledge and Data Engineering*, 27(5):1301–1315.

Weiqi Wu, Shen Huang, Yong Jiang, Pengjun Xie, Fei Huang, and Hai Zhao. 2025. Unfolding the headline: Iterative self-questioning for news retrieval and timeline summarization. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 4385–4398, Albuquerque, New Mexico. Association for Computational Linguistics.

Rui Yan, Jian-Yun Nie, and Xiaoming Li. 2011a. Summarize what you are interested in: An optimization framework for interactive personalized summarization. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1342–1351, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Rui Yan, Xiaojun Wan, Jahna Otterbacher, Liang Kong, Xiaoming Li, and Yan Zhang. 2011b. Evolutionary timeline summarization: a balanced optimization framework via iterative substitution. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, pages 745–754.

Yi Yu, Adam Jatowt, Antoine Doucet, Kazunari Sugiyama, and Masatoshi Yoshikawa. 2021. Multi-timeline summarization (mtls): Improving timeline summarization by generating multiple summaries. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 377–387.

Chenlong Zhang, Tong Zhou, Pengfei Cao, Zhuoran Jin, Yubo Chen, Kang Liu, and Jun Zhao. 2025a. DTELS: Towards dynamic granularity of timeline summarization. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2682–2703, Albuquerque, New Mexico. Association for Computational Linguistics.

Guibin Zhang, Luyang Niu, Junfeng Fang, Kun Wang, LEI BAI, and Xiang Wang. 2025b. Multi-agent architecture search via agentic supernet. In *Forty-second International Conference on Machine Learning*.

Jusheng Zhang, Zimeng Huang, Yijia Fan, Ningyuan Liu, Mingyan Li, Zhuojie Yang, Jiawei Yao, Jian Wang, and Keze Wang. 2025c. KABB: Knowledge-aware bayesian bandits for dynamic expert coordination in multi-agent systems. In *Forty-second International Conference on Machine Learning*.

Yusen Zhang, Ruoxi Sun, Yanfei Chen, Tomas Pfister, Rui Zhang, and Sercan Arik. 2024. Chain of agents: Large language models collaborating on long-context tasks. *Advances in Neural Information Processing Systems*, 37:132208–132237.

A Baselines Details

We compare the performance of MAS-TLS with several prior works. For non-LLM traditional baselines, we report the numbers as published in the original papers on the same benchmarks and evaluation protocol. For LLM-based baselines and MAS-TLS, we use the same backbone LLM (Qwen3-32b) for all methods and repeat each experiment three times with different random seeds, reporting the average performance.

- TRAN (Tran et al., 2013): applies submodular optimisation over a headline graph to select informative, non-redundant daily events using influence, temporal cues and duplication-aware diversity.
- MAR (Martschat and Markert, 2018): applies sub-modular optimisation for sentence selection, balancing content coverage with both temporal and textual diversity.
- DATE (Gholipour Ghalandari and Ifrim, 2020): uses supervised learning with date-specific features to summarise by day, followed by an unsupervised summary for each date.
- SDF (La Quatra et al., 2021): adopts a "date-first" strategy that summarises dates and then performs summary-driven graph ranking to choose the most important days.
- CLUST (Gholipour Ghalandari and Ifrim, 2020): employs Markov-based event clustering and ranks clusters by how often their dates occur across the entire input set.
- EGC (Li et al., 2021): models an event graph and compresses it into salient sub-graphs via time-aware optimal transport to select key events.
- TimeChunk (Sojitra et al., 2024): Adopts a hierarchical map-reduce strategy to handle long context. It sorts documents chronologically and segments them into chunks, employing an LLM to generate intermediate summaries which are recursively summarized to form the final timeline.
- HM (Zhang et al., 2025a):generates date-level summaries with LLMs and hierarchically merges them to match the required timeline granularity.

- LLM-TLS (Hu et al., 2024): summarize the events of an article using a large model and treats large language models as pseudo-oracles for event clustering in streaming contexts. By turning off the incremental mode, the method in the article reduces to a standard TLS setting and can be directly compared with other methods.
- CHRONOS (Wu et al., 2025): lets an LLM iteratively self-question and rewrite queries to retrieve news, then generates timelines round by round. The output rounds and other hyperparameters can be adjusted according to the complexity of the dataset.

B Framework Comparison Details

B.1 Single Agent Implementation

To isolate the effect of the multi-agent architecture, we construct a *single-controller agent* baseline in which a single agent performs timeline construction over the entire corpus without any subtopic decomposition, multi-agent interaction, or scheduling. This baseline retains only the local reasoning capabilities that each MAS-TLS sub-agent possesses, while removing all forms of inter-agent coordination.

Initialization. The agent receives the full corpus C and initializes its working document pool $C^{(0)} = C$ and an empty timeline $T^{(0)} = \varnothing$. The timeline is generated iteratively for L rounds.

Local Temporal Analysis. At round r, the agent normalizes all timestamp expressions within $C^{(r-1)}$ (using a standard preprocessing tool such as HeidelTime). Let $t(s)$ denote the normalized date assigned to sentence s. It then computes a frequency-based temporal salience distribution:

$$\mathrm{Freq}^{(r)}(t) = \#\{\, s \in C^{(r-1)} \mid t(s) = t \,\},$$

$$P^{(r)}(t) = \frac{\mathrm{Freq}^{(r)}(t)}{\sum_{t'} \mathrm{Freq}^{(r)}(t')}.$$

The next timeline time point is selected greedily:

$$t^{(r)} = \arg\max_{t} P^{(r)}(t).$$

Event Extraction and Summarization. Given the selected time point $t^{(r)}$, the agent gathers all sentences in the pool associated with this time:

$$E^{(r)} = \{\, s \in C^{(r-1)} \mid t(s) = t^{(r)} \,\}.$$

A single LLM call produces a short event description:

$$y^{(r)} = \mathrm{LLM}_{\mathrm{summ}}(E^{(r)}),$$

and the timeline is updated as

$$T^{(r)} = T^{(r-1)} \cup \{(t^{(r)}, y^{(r)})\}.$$

Pool Update and Iteration. To avoid redundant events, the agent removes used sentences:

$$C^{(r)} = C^{(r-1)} \setminus E^{(r)}.$$

The process repeats until L timeline entries are generated.

This single-agent baseline conducts timeline construction in a purely local and single-perspective manner. It uses the same LLM backbone as MAS-TLS but lacks sub-corpus specialization, cross-agent interaction, and Bayesian scheduling. Thus it serves as a strong *single-agent* control for evaluating the architectural benefits of MAS-TLS.

C Supplementary Notes on the Experimental Setup

C.1 Supplementary Notes on Datasets

Table 4 summarizes the statistics of the three benchmarks used in our experiments. We use only publicly available news benchmarks, do not collect new personal data, and do not redistribute raw articles—only derived timeline outputs and aggregated statistics. Below we provide detailed descriptions of their characteristics and the specific challenges they pose.

Table 4: Statistics of the evaluation datasets used in this work.

Metric	T17	Crisis	Entities
# of Topics	9	4	47
# of Timelines	19	22	47
Avg. # Articles	508	2,310	959
Avg. # Pub. Dates	124	307	600
Avg. Duration (days)	212	343	4,437
Avg. Timeline Length (L)	36	29	23

T17 (Timeline 17) (Tran et al., 2013). This is a standard multi-document summarization corpus constructed from major news outlets (e.g., CNN, BBC). It covers general-interest topics ranging from technology (e.g., "Apple iPad") to public health (e.g., "H1N1"). We use T17 to evaluate the baseline capability of MAS-TLS in handling standard journalistic reporting with moderate time spans.

Crisis (Tran et al., 2015a). This dataset consists of articles covering four protracted armed conflicts: Syria, Yemen, Libya, and Egypt. Unlike general news, Crisis timelines are characterized by:

- **High Ambiguity:** The narratives are often chaotic, with overlapping reports of violence, diplomacy, and humanitarian issues.
- **Topic Shifts:** The focus frequently shifts between different warring factions and international interventions.

These properties make Crisis an ideal testbed for our *Non-Stationary Bayesian Scheduling*, as the system must dynamically adapt to evolving sub-narratives (e.g., from protests to civil war).

Entities (Gholipour Ghalandari and Ifrim, 2020). This dataset focuses on the biographies of public figures (e.g., "Arnold Schwarzenegger", "Morgan Tsvangirai") spanning extremely long periods. It exhibits specific challenges:

- **Extreme Duration:** The timeline spans are significantly longer than T17 or Crisis. Some of the topics have a time span of several decades.
- **Distributional Skew:** Reporting is highly bursty, concentrated around specific controversies or achievements, leaving long "dormant" periods.

We utilize Entities to rigorously test the *Efficiency* and *Topical Balance* of MAS-TLS. The high skewness challenges the system to avoid overfitting to high-frequency peaks while maintaining coverage of long-tail life events.

C.2 Computation of AR and Date-F1

Following Martschat and Markert (Martschat and Markert, 2017), we evaluate timeline quality with alignment-based ROUGE (AR-1/AR-2) and Date-F1.

Preliminaries. Let $\mathcal{Y} = \{(t_\ell, s_\ell)\}_{\ell=1}^{L}$ be the system-generated timeline and $\mathcal{R} = \{\mathcal{Y}_k^*\}_{k=1}^{K}$ the set of reference timelines for the same query Q, where $\mathcal{Y}_k^* = \{(t_\ell^{*(k)}, s_\ell^{*(k)})\}_{\ell=1}^{L_k}$. For any timeline $\mathcal{Y}$, we denote by

$$D(\mathcal{Y}) = \{t_\ell : (t_\ell, s_\ell) \in \mathcal{Y}\}$$

the set of distinct dates. For a date $t \in D(\mathcal{Y})$, $s_{\mathcal{Y}}(t)$ is the (possibly concatenated) daily summary at t.

Let

$$D_{\text{ref}} = \bigcup_{k=1}^{K} D(\mathcal{Y}_k^*), \qquad D_{\text{sys}} = D(\mathcal{Y}).$$

For each reference date $d^* \in D_{\text{ref}}$ and system date $d \in D_{\text{sys}}$, we define a temporal weight

$$w_{d^*,d} = \frac{1}{|d^* - d| + 1}, \tag{11}$$

where dates are measured on the normalized day index used in our main formulation.

Alignment. AR first aligns reference dates to system dates using both temporal proximity and content similarity. Following the `align+ m:1` variant of Martschat and Markert (Martschat and Markert, 2017), each reference date $d^* \in D_{\text{ref}}$ is assigned an aligned system date

$$f(d^*) \in D_{\text{sys}},$$

and multiple d^* are allowed to align to the same system date (many-to-one alignment). The alignment is chosen by minimizing a cost that penalizes temporal distance and low ROUGE-1 similarity; we use the authors' official implementation.

Per-date ROUGE. For a reference date d^*, let $s_{\text{ref}}(d^*)$ denote the concatenated reference daily summary at d^*, obtained by concatenating all reference sentences with date d^* across $\{\mathcal{Y}_k^*\}_{k=1}^{K}$. With the aligned system date $f(d^*)$, we define per-date ROUGE-n recall and precision as

$$R_n^{\text{rec}}(d^*) = \text{ROUGE-}n_{\text{rec}}\big(s_{\text{ref}}(d^*), s_{\mathcal{Y}}(f(d^*))\big),$$

$$R_n^{\text{prec}}(d^*) = \text{ROUGE-}n_{\text{prec}}\big(s_{\text{ref}}(d^*), s_{\mathcal{Y}}(f(d^*))\big),$$

where ROUGE-n_{rec} and ROUGE-n_{prec} are the standard sequence-level ROUGE-n recall and precision.

Alignment-based ROUGE (AR-1 / AR-2). AR-n aggregates per-date ROUGE-n scores via temporal weights from Eq. (11). We define

$$\text{AR-}n_{\text{rec}} = \frac{\sum_{d^* \in D_{\text{ref}}} w_{d^*,f(d^*)}\, R_n^{\text{rec}}(d^*)}{\sum_{d^* \in D_{\text{ref}}} w_{d^*,f(d^*)}},$$

$$\text{AR-}n_{\text{prec}} = \frac{\sum_{d^* \in D_{\text{ref}}} w_{d^*,f(d^*)}\, R_n^{\text{prec}}(d^*)}{\sum_{d^* \in D_{\text{ref}}} w_{d^*,f(d^*)}}.$$

The final AR-n score (we report AR-1 and AR-2) is the harmonic mean of AR-n_{rec} and AR-n_{prec}:

$$\text{AR-}n = \frac{2\,\text{AR-}n_{\text{rec}}\,\text{AR-}n_{\text{prec}}}{\text{AR-}n_{\text{rec}} + \text{AR-}n_{\text{prec}}}.$$

Date-F1. While AR-1/AR-2 focus on content quality under soft temporal alignment, Date-F1 directly measures the quality of date selection. We treat a date as "predicted" if it appears at least once in the system timeline, and as "gold" if it appears in any reference timeline:

$$D_{\text{sys}} = D(\mathcal{Y}), \qquad D_{\text{ref}} = \bigcup_{k=1}^{K} D(\mathcal{Y}_k^*).$$

Date-level precision and recall are

$$P_{\text{date}} = \frac{|D_{\text{sys}} \cap D_{\text{ref}}|}{|D_{\text{sys}}|}, R_{\text{date}} = \frac{|D_{\text{sys}} \cap D_{\text{ref}}|}{|D_{\text{ref}}|},$$

and Date-F1 is their harmonic mean:

$$\text{Date-F1} = \frac{2\,P_{\text{date}}\,R_{\text{date}}}{P_{\text{date}} + R_{\text{date}}}.$$

Intuitively, Date-F1 is high when the system fires on the same set of dates as human-written timelines, regardless of the exact wording of the daily summaries.

C.3 Computation of JS Divergence, CV and Gini

This appendix provides the detailed definitions of the Jensen–Shannon (JS) divergence and the Coefficient of Variation (CV) used in our temporal and topical balance analysis. All methods use the same target timeline length L for each topic (dataset default).

C.3.1 Temporal Jensen–Shannon (JS) Divergence

Let the corpus be

$$C = \{d_1, \ldots, d_{|C|}\}, \qquad t(d) \in \mathbb{Z},$$

where $t(d)$ denotes the normalized publication day of document d. The model-generated timeline is

$$T = \{(t_1, s_1), \ldots, (t_L, s_L)\}, \qquad t_i \in \mathbb{Z}.$$

Step 1: Corpus and timeline histograms. We convert both sets of timestamps into daily probability distributions. The corpus histogram is

$$P_C(t) = \frac{|\{\, d \in C \mid t(d) = t \,\}|}{\sum_{t'} |\{\, d \in C \mid t(d) = t' \,\}|},$$

and the timeline histogram is

$$P_T(t) = \frac{|\{\, i \mid t_i = t \,\}|}{\sum_{t'} |\{\, i \mid t_i = t' \,\}|}.$$

Both P_C and P_T satisfy $\sum_t P_C(t) = 1$ and $\sum_t P_T(t) = 1$.

Step 2: Jensen–Shannon divergence. Define the mixture distribution

$$M(t) = \frac{1}{2}\big(P_C(t) + P_T(t)\big).$$

The JS divergence between the corpus and timeline distributions is

$$JS(P_C, P_T) = \tfrac{1}{2}\text{KL}(P_C \parallel M) + \tfrac{1}{2}\text{KL}(P_T \parallel M),$$
$$KL(P \parallel M) = \sum_t P(t)\,\log\frac{P(t)}{M(t)}.$$

A larger corpus–timeline JS suggests that the generated timeline is less proportional to corpus reporting volume (i.e., weaker coupling to bursty document-frequency peaks). We do not treat larger JS as inherently better; it is used as a diagnostic and should be interpreted together with the gold reference and imbalance measures. Following our implementation, we report the JS distance

$$\text{JSDist}(P_C, P_T) = \sqrt{JS(P_C, P_T)}.$$

C.3.2 Coefficient of Variation (CV) of Inter-Event Intervals

For any sequence of event dates $\{t_i\}_{i=1}^{L}$ (from a system/gold timeline, or from corpus documents), we sort them:

$$t_1 \le t_2 \le \cdots \le t_L.$$

Step 1: Adjacent temporal intervals. We compute day-level gaps between consecutive events:

$$\Delta_i = t_{i+1} - t_i, \qquad i = 1, \ldots, L-1.$$

Step 2: CV computation. Let

$$\mu = \frac{1}{L-1}\sum_{i=1}^{L-1}\Delta_i, \sigma = \sqrt{\frac{1}{L-1}\sum_{i=1}^{L-1}(\Delta_i - \mu)^2}.$$

The Coefficient of Variation is defined as

$$\text{CV} = \frac{\sigma}{\mu}.$$

Higher CV indicates irregular, noise-driven temporal clustering (burstiness), whereas a CV closer to the reference timeline (Gold) implies a well-paced narrative structure. Typically, raw corpora exhibit extremely high CV due to media redundancy, and effective summarization should reduce this variance to establish temporal continuity.

C.3.3 Topical JS Distance and Gini in a Corpus-Built Topic Space

We additionally assess *topical* balance by comparing topic distributions of the corpus, the gold timeline, and the model-generated timeline in a shared topic space.

Topic space (dataset-level). For each dataset, we train a topic model (BERTopic with HDBSCAN) on sentence-level corpus units to induce a shared topic space with K topics and a topic assignment function

$$z(\cdot) : \text{text} \to \{-1, 1, \ldots, K\},$$

where $z(x) = -1$ denotes an outlier/noise assignment.

Topic distribution (excluding noise). Given any set of text units $X = \{x_1, \ldots, x_{|X|}\}$ (corpus units or timeline items), we obtain topic labels $\{z(x)\}$ by projecting X into the trained topic model. We compute topic counts over *valid* topics (excluding noise):

$$c_X(k) = |\{\, x \in X \mid z(x) = k \,\}|, k = 1, \ldots, K,$$

and define a smoothed topic distribution

$$P_X(k) = \frac{c_X(k) + \alpha}{\sum_{k'=1}^{K} \big(c_X(k') + \alpha\big)}, k = 1, \ldots, K, \tag{12}$$

where $\alpha > 0$ is a small additive smoothing constant (we use $\alpha = 0.1$). Noise is reported separately as

$$r_X = \frac{|\{\, x \in X \mid z(x) = -1 \,\}|}{|X|}.$$

Topical JS distance (square-root JS). Let P_X and P_{G} be topic distributions computed by Eq. (12) for an object X (a system timeline or the corpus) and the gold timeline, respectively. Define the mixture distribution

$$M(k) = \frac{1}{2}\big(P_X(k) + P_{\mathrm{G}}(k)\big).$$

We first compute the Jensen–Shannon divergence

$$\mathrm{JS}(P_X, P_G) = \tfrac{1}{2}\,\mathrm{KL}(P_X \parallel M) + \tfrac{1}{2}\,\mathrm{KL}(P_G \parallel M)$$

$$\mathrm{KL}(P \parallel M) = \sum_{k=1}^{K} P(k)\,\log\Big(\tfrac{P(k)}{M(k)}\Big).$$

Following our implementation, we *report* the JS *distance* (i.e., the square root of the JS divergence):

$$\mathrm{JSDist}(P_X, P_G) = \sqrt{JS(P_X, P_G)}.$$

A smaller topical JS distance indicates that the system allocates attention across topics more consistently with human references.

Gini coefficient for topical skewness. To quantify topical skew (concentration on a few topics), we compute the Gini coefficient on the topic probability vector P_X. Let $p_{(1)} \le \cdots \le p_{(K)}$ be P_X sorted in non-decreasing order. Then

$$\mathrm{Gini}(P_X) = \frac{2\sum_{i=1}^{K} i\, p_{(i)}}{K \sum_{i=1}^{K} p_{(i)}} - \frac{K+1}{K}.$$

$\mathrm{Gini}(P_X) = 0$ corresponds to a uniform topic distribution, while larger values indicate stronger topical concentration/heavy-tail structure. We treat Gini as a descriptive diagnostic and interpret it together with topical JS distance and the gold reference.

C.4 Implementation Semantic-space Scatter Visualization

We provide a qualitative semantic-space visualization to compare how different methods select content relative to the same corpus structure. For each dataset, we build a sentence-level corpus pool and collect timeline texts from the gold reference and system outputs. All texts are normalized with identical rules (e.g., remove empty/URL/all-symbol fragments, short-text filtering, and exact deduplication after normalization). To avoid visual bias from different output lengths, we use fixed-size subsampling: we draw a background sample from the corpus and an equal-sized sample from each method.

All units are encoded using the same Sentence-Transformer; embeddings are cached by hashing normalized text. We fit UMAP only on corpus embeddings to obtain a corpus-built 2D map, and project gold/system embeddings into the same coordinate system via UMAP transform. Each plot shows the corpus as a light-gray background and overlays one method (Gold/Base/Single/MAS) in color, enabling direct inspection of semantic coverage and selection preference under a shared coordinate frame.

D Supplementary Results

In this section, we provide additional experimental results that were omitted from the main text due to space constraints.

D.1 Detailed Model Transferability

To examine whether the gains of MAS-TLS depend on a specific (proprietary) backbone, we evaluated the framework with three representative backbones of large language models (LLMs). We selected three representative backbones:

- **GPT-4o-mini**: A strong closed-source commercial baseline with favorable cost–quality trade-offs.
- **DeepSeek-R1-32B**: An open-weight model geared toward reasoning-style generation.
- **Qwen-3-32B** (our default): Our primary open-weight backbone balancing instruction following and reasoning.

Table 5 presents results on the **Entities** dataset. Overall performance is consistent across all tested backbones. Notably, the open-weight **Qwen-3-32B** achieves the best scores (AR-1 0.099, Date-F1 0.251), slightly above **GPT-4o-mini** (AR-1 0.095, Date-F1 0.246). The reasoning-oriented **DeepSeek-R1-32B** remains competitive (AR-1 0.092, Date-F1 0.243), with only a small drop.

These results suggest that MAS-TLS transfers robustly across the evaluated backbones and is not tied to a single model family in this setting. We hypothesize that the proposed multi-agent decomposition and coordination components, which provide a structured procedure for evidence organization and candidate consolidation, enabling medium-sized open models to reach comparable performance without relying exclusively on proprietary APIs.

Table 5: **Model Transferability on Entities.** MAS-TLS maintains stable performance across different backbones. The relative AR-1 spread is below 8% .

Backbone	AR-1	AR-2	Date-F1
GPT-4o-mini	0.095	0.036	0.246
DeepSeek-R1-32B	0.092	0.038	0.243
Qwen-3-32B	**0.099**	**0.041**	**0.251**

D.2 Sensitivity Analysis

We analyze the sensitivity of MAS-TLS to four core parameters on the Entities dataset: (1) The number of sub-agents M, which determines the granularity of semantic decomposition; (2) The coordination coefficient β, which controls the influence of cross-agent distinctiveness; (3) The forgetting factor γ, which governs the system's adaptability to non-stationary topic shifts; (4) The fairness weight λ_{fair}, which balances the trade-off between exploiting high-performing agents and exploring dormant ones. Figure 7 illustrates the performance trends (AR-1, red, left axis) and computational cost (Token consumption, blue, right axis) as we vary one parameter at a time while keeping the others fixed at the reference setting.

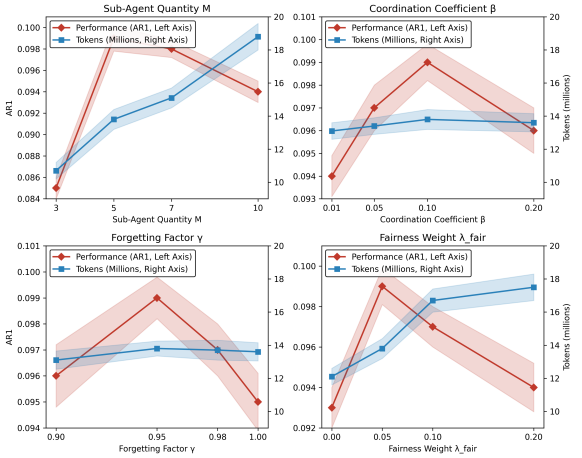

Figure 7: **Hyperparameter Analysis.** We report the impact of four key parameters on generation quality (AR-1; red diamonds; left axis) and computational cost (token consumption in millions; blue squares; right axis). Shaded regions indicate variability across 3 independent runs. We vary one parameter at a time while fixing the others to the selected default ($M = 5, \beta = 0.1, \gamma = 0.95, \lambda_{\text{fair}} = 0.05$).

Impact of Sub-Agent Granularity (M). As shown in Figure 7 (top-left), increasing M from 3 to 5 yields the largest performance gain (AR-1 improves from 0.085 to 0.099), suggesting that moderate semantic decomposition is crucial for covering complex timelines. However, further increasing M to 7 brings a slight drop, and $M = 10$ degrades quality, while token consumption keeps increasing monotonically. Overall, $M = 5$ offers the best efficiency sweet spot, balancing coverage and cost.

Impact of Coordination Strength (β). The coordination coefficient (top-right) exhibits an "inverted U-shape" trend. A moderate penalty ($\beta = 0.1$) outperforms the near-zero setting ($\beta = 0.01$), supporting the value of inter-agent coordination for reduc-

ing redundancy. Meanwhile, "token consumption changes only slightly as β varies, suggesting that β mainly reshapes selection priorities rather than the overall amount of generation. Over-penalizing overlap ($\beta = 0.2$) hurts performance by discouraging the selection of central, high-salience events.

Impact of Forgetting Factor (γ). Figure 7 (bottom-left) shows that γ substantially affects AR-1 on ENTITIES. The system underperforms at $\gamma = 1.0$ (no forgetting), where accumulated historical feedback makes the controller less responsive to recent accept/reject signals. Peak performance at $\gamma = 0.95$ suggests that moderately discounting older feedback helps emphasize recent evidence and improves robustness.

Exploration-Exploitation Trade-off (λ_{fair}). The fairness weight (bottom-right) controls exploration. A small bonus ($\lambda_{\text{fair}} = 0.05$) alleviates agent starvation and improves performance over the greedy baseline ($\lambda_{\text{fair}} = 0$) with a modest increase in cost. In contrast, larger fairness weights (e.g., $\lambda_{\text{fair}} \geq 0.1$) substantially increase token consumption and reduce quality, as low-relevance agents are activated too frequently.

Table 6: **Cross-lingual Evaluation on Chinese DTELS.** Results on Chinese DTELS-Bench. MAS-TLS improves coverage (AR-1/AR-2) and remains competitive on temporal precision (Date-F1).

Method	AR-1	AR-2	Date-F1
HM	0.062	0.035	0.202
LLM-TLS	0.071	0.033	0.215
CHRONOS	0.088	0.038	0.218
MAS-TLS (Ours)	**0.096**	**0.041**	**0.226**

D.3 Cross-Lingual Generalization

To examine whether the proposed agentic architecture generalizes beyond English corpora, we conduct an additional evaluation of MAS-TLS on a Chinese subset derived from DTELS-Bench. The goal is to assess cross-lingual robustness without introducing any changes to the core system design.

In this setting, we keep the overall pipeline intact and limit adaptations strictly to language-dependent components. Specifically, we adjust prompting strategies and incorporate Chinese-specific preprocessing steps, including sentence segmentation and normalization of temporal expressions. Importantly, no modifications are made to the agent topology or the coordination mechanisms, ensuring that the underlying architecture remains fully consistent with the English setup.

Unless otherwise specified, we follow the default target length defined in DTELS-Bench and adopt the same evaluation protocol used in our main experiments. This setup allows for a controlled comparison while providing a more nuanced indication of how the framework performs on non-English data drawn from a portion of the benchmark.

Table 6 reports the results. MAS-TLS achieves an AR-1 of 0.096 and Date-F1 of 0.226, outperforming the strongest baseline CHRONOS on all metrics. We report AR-1/AR-2 as coverage metrics under alignment-based evaluation, and the AR-1 gain suggests that the decomposed workflow remains effective in the Chinese setting. We note that improvements on AR-2 and Date-F1 are smaller, which may reflect differences in temporal expression realizations and preprocessing noise in Chinese corpora.

D.4 Supplementary Temporal Imbalance Analysis on ENTITIES

To quantify the *temporal imbalance* in the input corpus and examine whether MAS-TLS mitigates corpus-induced temporal skew, we analyze the distribution of documents/events over time on ENTITIES. Entity-centric corpora are often temporally bursty: documents may concentrate around a few high-publicity periods, such as scandals, trials, appointments, deaths, or other salient incidents. A timeline system that follows the raw corpus density too closely may therefore over-select events from these dense periods while under-representing quieter but still important stages of the entity's trajectory.

For each topic Q, we construct a discrete time histogram over calendar years. Specifically, let $\mathbf{c} = (c_1, \ldots, c_B)$ denote the yearly counts, where B is the number of years in the topic span and c_b is the number of items falling into the b-th year. For the input corpus, c_b counts documents in year b and $\sum_{b=1}^{B} c_b = N$ denotes the total number of corpus documents for the topic. We construct the same type of yearly histogram for the system output timeline and the gold timeline, where c_b counts selected or reference events rather than corpus documents. All distributions for the same topic are evaluated over the same topic-specific year span, so that the imbalance scores are comparable across corpus, system output, and gold reference.

We report three complementary imbalance metrics, where lower values indicate less temporal concentration under this diagnostic. **(1) Coefficient of Variation (CV)**, defined as

$$\mathrm{CV}(\mathbf{c}) = \frac{\sigma(\mathbf{c})}{\mu(\mathbf{c})},$$

measures the relative dispersion of yearly counts. A high CV indicates that a small number of years contain disproportionately many documents or events, while a lower CV suggests a more even temporal spread. **(2) Gini coefficient** is computed over yearly counts and measures inequality of temporal mass across years. Compared with CV, Gini provides a direct inequality-oriented view of how concentrated the distribution is. **(3) Jensen–Shannon (JS) divergence** is computed between the normalized year distribution and a uniform temporal reference. Let

$$p_b = \frac{c_b}{\sum_{b'=1}^{B} c_{b'}}, \qquad u_b = \frac{1}{B}.$$

The JS divergence between p and u measures how far the observed temporal distribution deviates from uniform coverage over the topic span. We use this value as an information-theoretic diagnostic of temporal concentration.

Figure 9 and Figure 10 show that the input corpus is often highly skewed in time, confirming that temporal imbalance is a prominent property of ENTITIES. MAS-TLS generally reduces this skew relative to the input corpus, yielding lower CV and Gini values for many topics and bringing the output distribution into a range comparable to the gold timeline. These results suggest that the proposed decomposition and temporal coordination mechanisms help the system avoid simply reproducing the densest temporal bursts in the corpus.

Figure 11 further compares MAS-TLS and the gold timeline from an information-theoretic perspective. MAS-TLS often obtains JS-to-uniform values that are comparable to, and in many cases lower than, those of the gold timeline. This indicates that the generated timelines are not overly concentrated in a narrow set of years. However, we do not interpret lower JS divergence as universally better, since human-written gold timelines are not necessarily uniform and may legitimately emphasize periods with more important events.

Importantly, these metrics are intended as supplementary diagnostics rather than standalone measures of timeline quality. A perfectly uniform timeline is not necessarily optimal for every topic. Therefore, lower CV, Gini, or JS divergence should be understood as evidence that the system mitigates excessive corpus-induced temporal skew, not as a replacement for semantic coverage and temporal accuracy metrics such as AR and Date-F1. Under this interpretation, the supplementary analysis supports the claim that MAS-TLS improves temporal balance while maintaining relevance to the gold timeline.

D.5 Evidence Support Audit: Implementation Details

Since TLS outputs are factual event statements (date + description), we conduct a small-scale supplementary audit to check whether each generated event is supported by the evidence that is *available to the system* during generation. This audit is particularly relevant for our agentic workflow, where each sub-agent maintains a running set of high-confidence event candidates and the master performs adjudication across agents.

For each timeline instance, we construct the evidence set from the system's *native intermediate artifacts* produced during inference. Specifically, we use the evidence sentences summarized/retained during the per-agent event maintenance stage (Step 2 in our pipeline), i.e., the set of sentences that the agentic newsroom selects as high-value evidence before master adjudication. This definition ensures that annotators only see information that the system itself had access to when producing the final timeline, avoiding post-hoc evidence augmentation.

For each benchmark, we randomly sample 50–100 event statements from the final generated timelines. To avoid position bias, we draw events across early/middle/late parts of timelines when applicable. For each sampled event, we provide annotators with: (i) the generated event statement (date + description), and (ii) the corresponding system-visible evidence set .

Each item is labeled independently by three annotators in a blind setting (method identifiers removed, randomized order). Annotators label two binary attributes for each event: **(i) Content support**: whether the event description is supported by the provided evidence; **(ii) Time support**: whether the event date is consistent with (or directly supported by) the evidence. When evidence is insufficient to verify an attribute, annotators mark it as *not supported* for a conservative estimate. We

use majority vote as the final label. Inter-annotator agreement is around 80% (average pairwise percent agreement) for both attributes.

We summarize each audited event into one of four mutually exclusive categories based on *content support* and *time support*: **Fully supported** (both supported), **Time-only supported** (time supported but content not), **Content-only supported** (content supported but time not), and **Unsupported** (neither supported). We report the category proportions per dataset and visualize the distribution in Fig. 8.

To assess whether adjudication (LLM_{rank}) affects grounding, we run the audit on two variants under the same protocol: MAS-TLS (full) and MAS-TLS W/O LLM_{rank}. This isolates the effect of adjudication on evidence support while holding the evidence construction and candidate generation stages fixed.

Figure 8 reports the four-way evidence-support breakdown for MAS-TLS with and without LLM_{rank}. Across benchmarks, most events are **Fully supported** by system-visible evidence (about 63–80%), while the **Unsupported** category remains consistently small (about 5–8%). ENTITIES shows a lower fully-supported share and more partially-supported cases (time-only/content-only) than CRISIS and T17, which is expected given its finer-grained event distinctions and higher ambiguity in aligning dates and descriptions.

Comparing w/ and w/o LLM_{rank}, the distributions are nearly unchanged across all datasets: the fully-supported portion increases slightly with adjudication, while the unsupported portion changes marginally (overall $\approx$6.6% w/ vs. $\approx$7.0% w/o). This suggests that adjudication does not introduce additional unsupported statements, and any main-metric gains are not achieved by trading off evidence grounding.

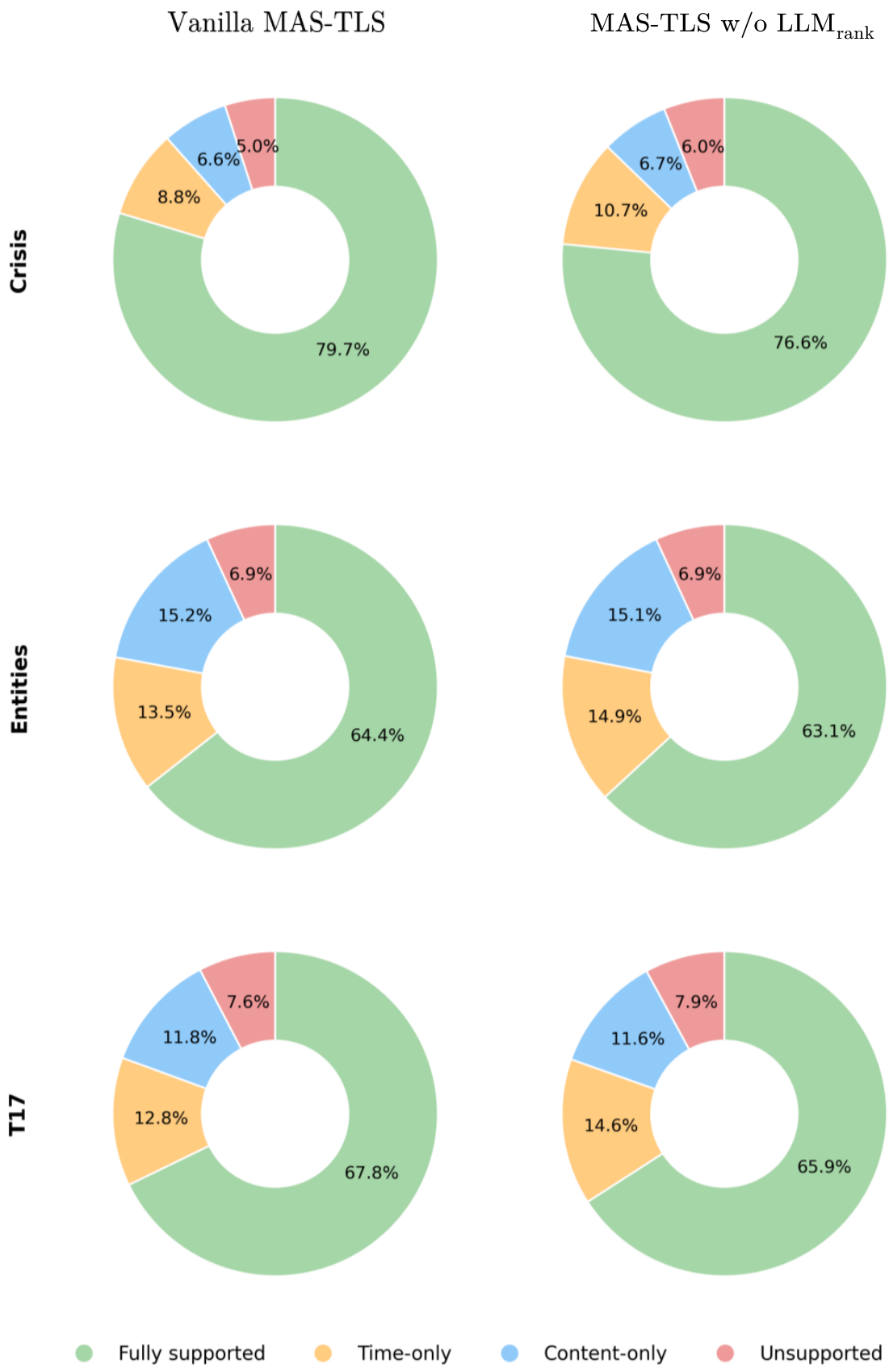

Figure 8: **Evidence support audit (four-way breakdown).** We categorize sampled event statements into **Fully supported**, **Time-only**, **Content-only**, and **Unsupported** based on whether the event content and date are supported by system-visible evidence. Results are shown for CRISIS, ENTITIES, and T17, comparing MAS-TLS with and without LLM_{rank}.

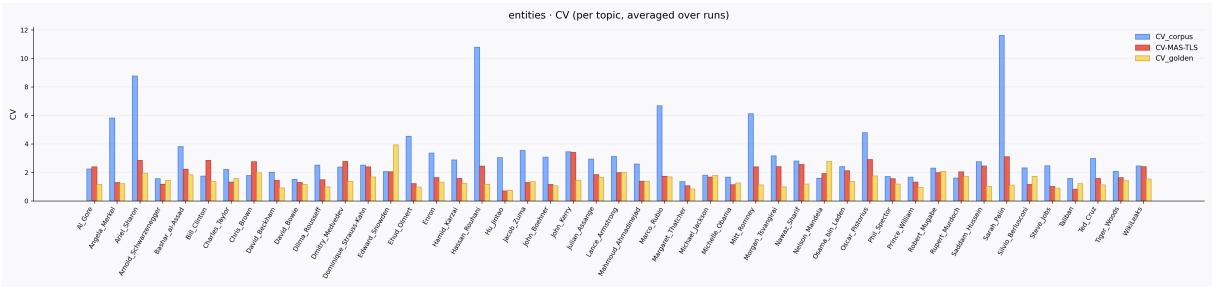

Figure 9: **Temporal imbalance on ENTITIES measured by CV (per topic, averaged over runs).** We compute the coefficient of variation of yearly counts for the input **corpus**, the **MAS-TLS** generated timeline, and the **gold** timeline. Lower values indicate less temporal concentration and a more balanced distribution over years.

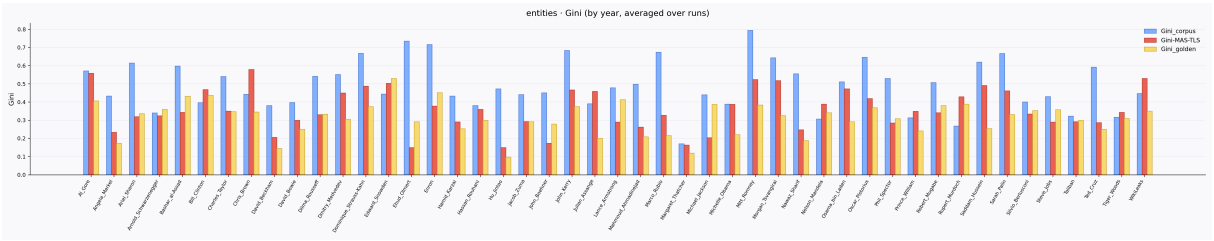

Figure 10: **Temporal imbalance on ENTITIES measured by Gini (by year, averaged over runs).** The Gini coefficient characterizes the inequality of yearly mass: higher values indicate stronger temporal imbalance. MAS-TLS consistently lowers the temporal inequality compared to the input corpus, moving closer to the gold reference.

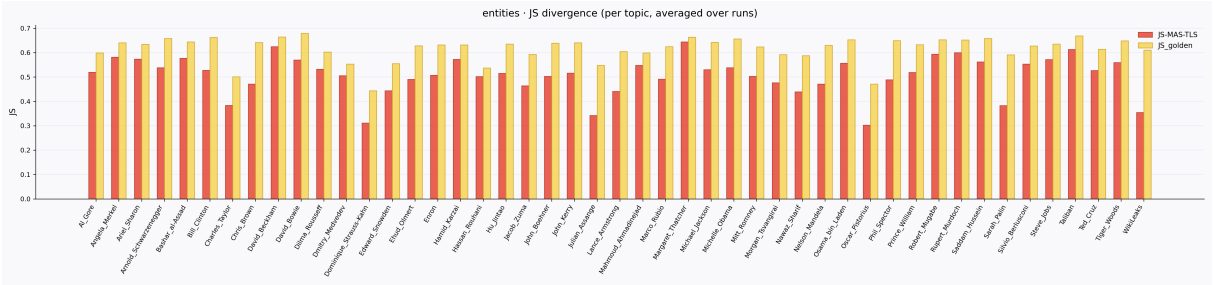

Figure 11: **Temporal imbalance on ENTITIES measured by JS divergence (per topic, averaged over runs).** We report the Jensen–Shannon divergence between the normalized yearly distribution and a uniform reference. Lower values imply a distribution that is less concentrated in a small subset of years.