Beyond Moment: Rethinking Evaluation Paradigms for Timeline Summarization in the Era of LLMs

Anonymous ACL submission

Abstract

Timeline summarization (TLS) has traditionally been evaluated with date-first metrics such as Date-F1 and A-ROUGE, which are well matched to classical date-centric pipelines. Modern LLM-based TLS, however, increasingly relies on event identification, semantic aggregation, and milestone abstraction, making date agreement less suitable as the first matching criterion. This evaluation-order mismatch can understate event-level gains. We address this issue with a semantic-alignment framework that first matches predicted and reference milestones by event equivalence, and then evaluates temporal grounding and summary quality. The framework instantiates four semantic-alignment metrics: **SA Score**, **SA Date-F1**, **SA-ROUGE**, and **STA-ROUGE**. Experiments on **DATEWISE**, **LLM-TLS**, and **FS-LLM-TLS** show that semantic-first evaluation gives a more coherent account of event-level differences than standard date-first metrics. Human judgments on an annotated subset further support the external plausibility of this evaluation order. The results suggest that modern TLS evaluation should complement date-first assessment with semantic-first matching. [1] [2]

1 Introduction

Timeline summarization (TLS) condenses temporally ordered documents into concise narratives of key events (Allan et al., 2001; Chieu and Lee, 2004; Yan et al., 2011). Classical TLS was commonly formulated in date-centric terms: systems identify salient dates and then generate or select summaries for those dates (Steen and Markert, 2019; Martschat and Markert, 2017). This formulation shaped both system design and evaluation.

Date-first metrics such as Date-F1 and A-ROUGE are natural under the classical formulation because they preserve dated timeline structure and compare summary content only after temporal alignment has been established (Martschat and Markert, 2017). However, Table 1 illustrates the core failure mode studied in this paper: date agreement alone can be insufficient for deciding whether a predicted milestone matches the reference milestone. Here, a human reader would treat **LLM-TLS** as event-correct and **DATEWISE** as event-incorrect, yet date-first scoring gives **LLM-TLS** both zero **Date-F1** credit and a lower A-ROUGE-1 score (0.2593 vs. 0.3226), misrepresenting both semantic validity and wording quality.

Table 1: Example from the **Entities** benchmark (Gholipour Ghalandari and Ifrim, 2020) on the topic **Sarah Palin**. From a human event-level reading, **LLM-TLS** (Hu et al., 2024) matches the reference milestone whereas **DATEWISE** (Gholipour Ghalandari and Ifrim, 2020) does not. Yet **Date-F1** gives zero credit to **LLM-TLS** because of a one-day date drift and rewards **DATEWISE** because it matches the reference date.

System	Milestone (Time: Event)
GT	`2008-09-03`: Accepts the Republican Party's nomination for vice president at the Republican National Convention.
LLM-TLS	`2008-09-04`: Republican vice presidential pick Sarah Palin is nominated at the Republican National Convention.
DATEWISE	`2008-09-03`: John McCain (second from right) and Sarah Palin (far right) on stage at the Republican National Convention.

Recent LLM-based TLS systems increasingly operate at the level of events rather than dates: they identify events, aggregate evidence across documents, abstract milestone summaries, and then place those milestones on a timeline (Hu et al., 2024; Qorib et al., 2025; Song et al., 2025). Under this formulation, dates remain necessary temporal attributes, but they are no longer an adequate

[1] Code and data will be released on GitHub upon paper acceptance.

[2] Section 1, 2, and 5 have been lightly refined with LLM assistance.

first criterion for deciding whether two milestones should be matched.

The resulting issue is an evaluation-order mismatch rather than an intrinsic flaw in earlier metrics. Date-first metrics are appropriate when systems are optimized to select salient dates and summarize them. When improvements arise instead from event construction and milestone selection, however, date-first matching can discard semantically correct milestones before temporal fidelity or summary quality is assessed. The evaluation signal then becomes misaligned with the changes a method actually makes.

We test this claim by comparing standard date-first metrics with a compact set of semantic-alignment (**SA**, hereafter) metrics: **SA Score**, **SA Date-F1**, **SA-ROUGE**, and *Semantic & Temporal-Alignment ROUGE* (**STA-ROUGE**). Experiments on **Entities**, **Crisis**, and **T17** compare **DATEWISE** (Gholipour Ghalandari and Ifrim, 2020), **LLM-TLS** (Hu et al., 2024), and a lightweight diagnostic variant, *Full-Stage LLM-TLS* (**FS-LLM-TLS**), and then validate the metric signal against human milestone judgments. Across these settings, date-first metrics expose an evaluation-order mismatch when event-level gains are filtered through temporal matching first, while semantic-alignment metrics provide a more faithful signal of milestone validity, temporal grounding, and summary quality.

In summary, our main contributions are:

- **Semantic-Alignment Evaluation Framework.** We formulate the mismatch between event-centric TLS generation and date-first evaluation, and propose a framework in which event correspondence is established before temporal fidelity and summary quality are assessed.
- **Metric Instantiation.** We instantiate this framework with **SA Score**, **SA Date-F1**, **SA-ROUGE**, and **STA-ROUGE**, which separately measure semantic coverage, temporal correctness after semantic matching, and summary quality over aligned milestones.
- **Diagnostic Testbed and External Validation.** We use **FS-LLM-TLS** as a lightweight diagnostic testbed, not as a standalone pipeline contribution, to test whether semantic-first metrics better reflect event-level gains than date-first metrics. We further use human milestone judgments as an external check on the resulting evaluation signal.

2 Related Work

TLS has historically shown a close relationship between task formulation and evaluation practice. Classical systems typically constructed date-indexed daily summaries, and classical metrics were designed to evaluate date agreement and date-conditioned content overlap. More recent LLM-based systems give a more explicit role to event discovery, semantic aggregation, and milestone abstraction. This section reviews that trajectory as background for the evaluation assumptions analyzed next.

2.1 Date-Centric TLS Methods

Early TLS is commonly described as constructing dated daily summaries. Allan et al. (2001) extracted one sentence for each event in a topic stream. Chieu and Lee (2004) formulated timeline construction as query-based event extraction, where important events are placed along a timeline and represented by extracted sentences. Yan et al. (2011) later modeled trans-temporal dependencies across sentences while still operating in a date-indexed extractive setting.

Later work introduced abstraction while largely preserving the same output structure. Chen et al. (2019) proposed a memory-based abstractive TLS model with explicit time-series modeling. Steen and Markert (2019) likewise defined TLS as identifying key dates and providing short descriptions for them. At the benchmark level, Gholipour Ghalandari and Ifrim (2020) described TLS through subtasks such as date selection and date summarization. Taken together, these works explain why TLS is often treated as selecting salient dates and then producing summaries for those dates.

2.2 Metrics Aligned with Date-Centric TLS

This date-centric formulation also explains the evolution of TLS evaluation. Plain ROUGE (Lin, 2004) ignores the temporal structure that defines timelines. To address this limitation, Martschat and Markert (2017) introduced Date-F1 and alignment-based ROUGE variants, arguing that TLS evaluation should preserve dated structure while avoiding overly brittle exact date matching. Martschat and Markert (2018) further emphasized the importance of date selection and dated daily summaries. In this context, date-first metrics are reasonable fits for the dominant date-centric view of the task.

Beyond timeline-specific metrics, later evalu-

ation work explores stronger semantic signals. BERTScore (Zhang et al., 2020) replaces exact lexical overlap with contextual embedding similarity, and QuestEval (Scialom et al., 2021) evaluates summaries through question generation and answering. LLM-based evaluators such as G-Eval (Liu et al., 2023) and GEMBA (Kocmi and Federmann, 2023) show strong correlation with human judgments in general generation tasks. These approaches increase semantic sensitivity, but they do not specify how structured TLS milestones should be matched while keeping semantic coverage and temporal fidelity distinct. Appendix A expands on why general-purpose semantic evaluators remain insufficient as end-to-end metrics for TLS.

2.3 LLM-based TLS and the Emerging Tension

Recent LLM-based TLS studies give event-level decisions a more explicit role. Hu et al. (2024) proposed an incremental LLM-based pipeline that detects moments, clusters them into events, abstracts milestones, and then constructs a timeline. Zhang et al. (2025) further moved away from a fixed date list by studying dynamic timeline granularity, which requires the system to decide what event units should survive at different levels of abstraction. Qorib et al. (2025) extended this direction to constrained TLS, where timelines must satisfy user-specified relevance constraints and therefore depend on event selection and abstraction rather than a fixed global date list. Song et al. (2025) further showed that temporal reasoning over event sequences can improve timeline summarization in social media.

These studies do not make dates irrelevant, nor do they invalidate date-first evaluation. They do suggest that dates increasingly function as temporal attributes of inferred milestones rather than as the only initial unit of selection. As a result, a mild tension emerges between event-centric modeling decisions and metrics that begin by deciding matchability through date agreement or temporal proximity. The next section analyzes this tension as a date-centric matching assumption.

3 Problem Analysis

3.1 Date-Centric Evaluation Assumption

Section 2 showed that date-first TLS metrics were historically well matched to date-centric systems. Metrics such as **Date-F1** and **A-ROUGE** were designed to preserve timeline structure rather than flatten a timeline into plain text (Martschat and Markert, 2017, 2018). Under that formulation, aligning outputs by date before comparing their summaries is natural. The implementation logic of **Date-F1** and **A-ROUGE** is summarized in Appendix B.

These metrics encode a *date-centric matching assumption*. **Date-F1** requires date agreement before any milestone can receive credit. **A-ROUGE** relaxes exact agreement through date-aware alignment and distance-weighted lexical overlap, but it still lets temporal proximity determine which milestone pairs are compared.

This assumption becomes too strong once generation is organized around events and milestones rather than dates alone. In modern TLS, especially LLM-based pipelines, systems increasingly identify events, aggregate evidence, abstract milestone descriptions, and then attach those milestones to a timeline. Dates remain necessary temporal attributes, but they are not sufficient evidence that two summaries describe the same event. Table 1 illustrates the problem: the semantically correct **LLM-TLS** milestone receives zero **Date-F1** credit because its date drifts, while the date-matching **DATEWISE** sentence receives credit even though it does not express the reference event. Date-first matching can therefore discard a semantically correct milestone before its event content is compared, or compare temporally proximate summaries that are not event-equivalent.

This is the evaluation-order mismatch. Temporal agreement remains important, but using it as the first matching criterion makes event equivalence depend on date proximity. As milestone quality increasingly depends on event identification, abstraction, and selection, date-first matching can systematically understate event-level gains whenever they are not accompanied by exact or near-exact date agreement. This risk is likely to grow: recent advances in preference optimization (Rafailov et al., 2023), self-improvement (Yuan et al., 2024), and distillation (Hsieh et al., 2023) make it increasingly plausible that stronger semantic judgment will be pushed into later milestone-selection decisions. Evaluation in this setting should therefore first establish event equivalence, and only then assess temporal fidelity and summary quality.

3.2 Partial Shift to Event-Centric Generation

Recent TLS pipelines clarify why this mismatch matters for evaluation. Both **LLM-TLS** (Hu et al.,

2024) and **DTELS** (Zhang et al., 2025) first construct event-bearing candidate units and then filter them into a final timeline. For evaluation, this behavior can be reduced to two higher-level processes: *candidate summary construction*, which compresses source documents into event-bearing textual units, and *candidate consolidation and selection*, which clusters, abstracts, and filters those units into the final timeline.

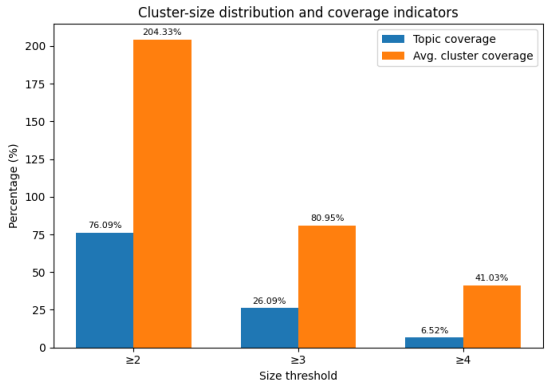

Figure 1: Cluster-size and coverage indicators can be highly imbalanced, illustrating why frequency-based signals alone may not capture milestone salience.

This abstraction shifts the evaluation question from date assignment alone to event-level milestone quality. The shift toward event-centric TLS is visible, but not uniform across stages. Relative to **DATEWISE**, **LLM-TLS** changes the *construction* side most clearly: candidate milestones are produced through LLM-based aggregation and abstraction rather than date-indexed snippet selection alone (Gholipour Ghalandari and Ifrim, 2020). Gains at this stage are therefore gains in event-level construction quality, not merely in date prediction.

The later *consolidation and selection* stages are different. Even in modern pipelines, they can still rely heavily on frequency-based proxies such as cluster size or coverage statistics, with semantic judgment playing a smaller role. Figure 1 shows why this matters: statistical signals are informative, but can be strongly misaligned with true milestone salience. They are useful as auxiliary evidence, but weaker as the core criterion for deciding which milestones survive into the final timeline.

Date-first evaluation is only partially sensitive to either kind of event-level gain. It captures an improved milestone when date agreement is preserved, but can suppress the same gain when the event is grounded in a nearby nonmatching date or when a semantically stronger milestone replaces a more frequent but less salient one.

This observation motivates **FS-LLM-TLS**. Rather than replacing **LLM-TLS**, it serves as a lightweight diagnostic variant that extends semantic judgment more directly into milestone consolidation and selection. This creates a setting in which event-level gains arise not only from better candidate construction, but also from the decision of which milestones deserve inclusion. The next section turns this analysis into a semantic-alignment framework and uses **FS-LLM-TLS** only to probe how different metrics respond to those gains.

4 Method

This section has two parts. Section 4.1 presents the main contribution: a semantic-alignment evaluation framework for event-centric TLS. Section 4.2 describes **FS-LLM-TLS** as a supporting diagnostic testbed derived from **LLM-TLS** (Hu et al., 2024).

4.1 Semantic-Alignment Metrics

4.1.1 Design Rationale

The framework follows the evaluation-order analysis in Section 3. The first question is whether a predicted milestone and a reference milestone describe the same underlying event. Temporal fidelity and summary quality are then assessed on semantically aligned pairs.

LLMs serve here as the implementation mechanism for semantic milestone matching. The evaluator first establishes event equivalence, and the downstream metrics then assess semantic coverage, timestamp accuracy after semantic matching, and summary quality on aligned pairs. Appendix C records the event-equivalence prompt.

4.1.2 Metric Definitions

Semantic-Alignment Score (SA Score). Let $\widehat{\mathcal{M}}$ denote the set of predicted–reference milestone pairs judged event-equivalent after candidate retrieval and LLM verification. The full realization of this alignment procedure, including candidate pruning, prompting, and tie-breaking, is given in Appendix D. The *SA Score* measures semantic coverage over the reference timeline:

$$\text{SA Score} = \frac{1}{N}\sum_{j=1}^{N} \mathbf{1}\left[\exists i : (i,j) \in \widehat{\mathcal{M}}\right],$$

where N is the number of reference milestones.

Semantic-Alignment Date-F1 (SA Date-F1). Among semantically aligned pairs, we further check whether timestamps coincide, yielding a Date-F1-style (Martschat and Markert, 2017) measure over the aligned milestone set. Let $\mathcal{P} = \{(d_i^P, y_i^P)\}_{i=1}^M$ and $\mathcal{R} = \{(d_j^R, y_j^R)\}_{j=1}^N$ denote the predicted and reference timelines, respectively. Here d is the timestamp, y is the milestone text, and superscripts P and R distinguish predicted and reference milestones. Let $\widehat{\mathcal{M}}_{\text{date}} = \{(i,j) \in \widehat{\mathcal{M}} \mid d_i^P = d_j^R\}$. We compute

$$P_{\text{SA-date}} = \frac{|\widehat{\mathcal{M}}_{\text{date}}|}{|\mathcal{P}|}, \quad R_{\text{SA-date}} = \frac{|\widehat{\mathcal{M}}_{\text{date}}|}{|\mathcal{R}|},$$

and define

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

Semantic-Alignment ROUGE (SA-ROUGE). We compute ROUGE-n (Lin, 2004) over semantically aligned pairs:

$$\text{SA-ROUGE}_n = \frac{1}{|\widehat{\mathcal{M}}|} \sum_{(i,j)\in\widehat{\mathcal{M}}} \text{ROUGE}_n(y_i^P, y_j^R).$$

Semantic & Temporal-Alignment ROUGE (STA-ROUGE). To jointly enforce semantic consistency and temporal fidelity, we further introduce STA-ROUGE. Let

$$\widehat{\mathcal{M}}_\epsilon = \{(i,j) \in \widehat{\mathcal{M}} \mid |d_i^P - d_j^R| \le \epsilon\},$$

where ϵ is a temporal window imposed after semantic alignment. We then compute

$$\text{STA-ROUGE}_n = \frac{1}{|\widehat{\mathcal{M}}_\epsilon|} \sum_{(i,j)\in\widehat{\mathcal{M}}_\epsilon} \text{ROUGE}_n(y_i^P, y_j^R).$$

This metric gives a stricter evaluation by requiring both semantic alignment and temporal grounding.

4.2 Diagnostic Testbed: FS-LLM-TLS

Motivated by the partial shift analyzed in Section 3.2, we use **FS-LLM-TLS** as a lightweight diagnostic testbed derived from **LLM-TLS** (Hu et al., 2024). In the later ablations, we denote its three localized modifications as **+S1**, **+S3**, and **+S4**; in this notation, **FS-LLM-TLS** is the full combination **LLM-TLS + S1 + S3 + S4**. Its role is not to introduce a new pipeline contribution, but to extend semantic judgment beyond candidate construction into milestone consolidation and selection. This makes it possible to examine whether the proposed metrics are responsive to event-level gains that date-first matching may only weakly reflect.

Only the components needed for this diagnostic role are summarized here. The comprehensive architecture is provided in Appendix E, and the stage-specific prompts are provided in Appendices F–H.

We make three localized changes. **(S1) Unified snippet extraction:** each document is converted into a grounded event snippet with structured arguments (Appendix F). **(S3) Cluster-level abstractive summarization:** clusters are summarized by the LLM rather than reduced to a single representative sentence (Appendix G). **(S4) Hybrid milestone selection with LLM judgment:** frequency-based proxies remain as shortlist signals, but final milestone choice is made with batched LLM judgments over salience (Appendix H). The third change is the most important one for the evaluation question because it extends semantic judgment into milestone selection.

5 Experiments

We evaluate the framework in three steps: cross-system comparison, external plausibility, and diagnostic utility. The goal is not to establish a new best pipeline, but to test whether semantic-first evaluation changes the evaluation profile of modern TLS.

5.1 Experimental Setup

Following Hu et al. (2024), we evaluate on three established topic-TLS benchmarks: **Entities** (Gholipour Ghalandari and Ifrim, 2020) (47 topics, up to 12 years), **Crisis** (Tran et al., 2015) (4 crisis-related topics), and **T17** (Tran et al., 2013) (9 topics, about 7 months).

We compare **DATEWISE** (Gholipour Ghalandari and Ifrim, 2020), the classical date-centric baseline; **LLM-TLS** (Hu et al., 2024), the upstream LLM-based system; and **FS-LLM-TLS**, our lightweight diagnostic testbed.

Following previous work, we report three standard date-first metrics: **Date-F1**, **AR1** (A-ROUGE-1), and **AR2** (A-ROUGE-2) (Martschat and Markert, 2017). We also report the semantic-alignment metrics from Section 4: **SA Score**, **SA Date-F1**, **SA-R1**, **SA-R2**, and **STA-R1**. Unless otherwise noted, these metrics use the `Codex-5.3` evaluator; direct human validation is reported in Section 5.3.

Unless otherwise specified, we use the `Llama-2-13B` model (Touvron et al., 2023)

Table 2: Cross-system evaluation on Entities, Crisis, and T17 under date-first and semantic-alignment metrics.

Dataset	Method	Standard Date-First Metrics			Semantic-Alignment Metrics (Ours)				
		Date-F1	AR1	AR2	SA Score	SA Date-F1	SA-R1	SA-R2	STA-R1
Entities	DATEWISE	0.205	0.057	0.017	0.194	0.120	0.044	0.013	0.040
Entities	LLM-TLS	0.232	0.091	0.037	0.281	0.152	0.115	0.049	0.086
Entities	FS-LLM-TLS	0.256	0.095	0.040	0.330	0.172	0.130	0.061	0.099
Crisis	DATEWISE	0.286	0.089	0.026	0.350	0.244	0.115	0.029	0.083
Crisis	LLM-TLS	0.329	0.112	0.032	0.380	0.286	0.134	0.039	0.097
Crisis	FS-LLM-TLS	0.348	0.120	0.044	0.395	0.320	0.142	0.045	0.112
T17	DATEWISE	0.528	0.120	0.037	0.247	0.232	0.108	0.029	0.099
T17	LLM-TLS	0.380	0.267	0.121	0.482	0.404	0.352	0.062	0.237
T17	FS-LLM-TLS	0.410	0.280	0.130	0.523	0.458	0.392	0.074	0.269

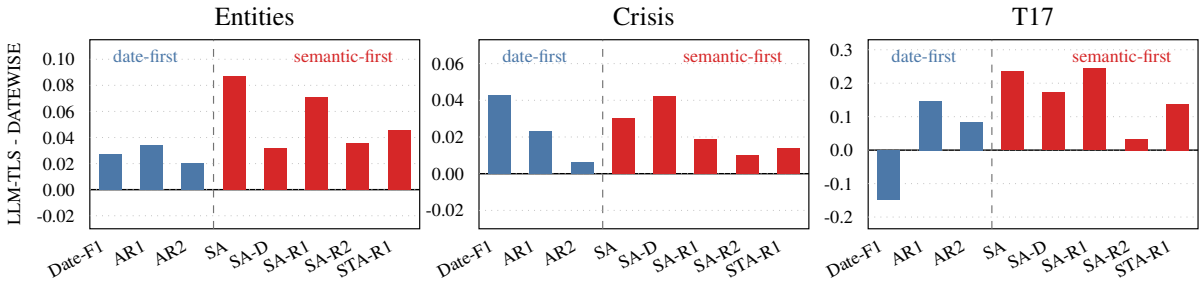

Figure 2: Metric profiles for **LLM-TLS** minus **DATEWISE** on each dataset. Blue bars report **Date-F1**, **AR1**, and **AR2**; red bars report **SA Score**, **SA-D** (**SA Date-F1**), **SA-R1**, **SA-R2**, and **STA-R1**. Positive values favor **LLM-TLS**; negative values favor **DATEWISE**. Each panel uses its own y-axis scale for readability.

throughout all experiments involving LLM-TLS and FS-LLM-TLS. The training environment and hardware details are described in Appendix I.

5.2 Cross-System Evaluation Profile

Table 2 reports the raw scores for **DATEWISE** (Gholipour Ghalandari and Ifrim, 2020), **LLM-TLS** (Hu et al., 2024), and **FS-LLM-TLS** under the same metrics. Figure 2 reorganizes the same evidence into three dataset-specific panels and plots the metric profile of the main cross-paradigm transition, **LLM-TLS** minus **DATEWISE**. The section asks whether date-aware lexical scoring and exact-date matching give a clean account of event-level system changes.

A-ROUGE provides only a mixed signal. Its date-decayed weighting lets lexical overlap contribute more when dates are close, so it partly reflects content similarity but does not isolate semantic coverage from temporal proximity. The semantic-alignment metrics separate these factors: **SA Score** first measures event-equivalent milestone coverage, **SA Date-F1** then checks exact timestamps on that aligned set, and **SA-ROUGE/STA-ROUGE** evaluate wording quality under explicit semantic alignment. On **Entities**, for example, **LLM-TLS** improves over **DATEWISE** under both families, but the decomposed signal is clearer: Δ**AR1**$=$ 0.034, compared with Δ**SA Score**$=$ 0.087 and Δ**SA-R1**$=$ 0.071. On **Crisis**, the profile is comparatively stable (Δ**Date-F1**$=$ 0.043, Δ**SA Date-F1**$=$ 0.042), showing that semantic-first evaluation does not mechanically rewrite every result.

Date-F1 can be actively misleading when date agreement conflicts with event equivalence. The clearest case is **T17**: **Date-F1** favors **DATEWISE** over **LLM-TLS** ($\Delta = -0.148$), while its own date-aware lexical companion moves in the opposite direction (Δ**AR1**$=$ 0.147) and the semantic-alignment metrics move more strongly in that same event-level direction (Δ**SA Score**$=$ 0.235, Δ**SA Date-F1**$=$ 0.172). The same diagnosis holds after adding the diagnostic **FS-LLM-TLS** variant: relative to **LLM-TLS**, **FS-LLM-TLS** raises **SA Score** on **Entities**, **Crisis**, and **T17** by $+0.049$, $+0.015$, and $+0.041$, respectively. On **T17**, **Date-F1** still leaves **FS-LLM-TLS** below **DATEWISE** (0.410 vs. 0.528), whereas **SA Score** and **SA Date-F1** place it above **DATEWISE** (0.523 vs. 0.247, and 0.458 vs. 0.232). This is the mismatch the proposed metrics are designed to expose: event equivalence is established first, and temporal grounding and summary quality are then assessed on that semantically valid set.

Table 3: Cross-evaluator consistency of **SA Score** across **DATEWISE**, **LLM-TLS**, and **FS-LLM-TLS** on **Entities**, **Crisis**, and **T17**. The last column reports row-wise variance across evaluators.

Dataset	Method	Codex-5.3	GPT-5.4	Gemini-2.5	Qwen3-30B-A3B	Var. ($\times 10^{-6}$)
Entities	DATEWISE	0.194	0.195	0.192	0.190	3.69
Entities	LLM-TLS	0.281	0.280	0.280	0.275	5.50
Entities	FS-LLM-TLS	0.330	0.330	0.330	0.330	0.00
Crisis	DATEWISE	0.350	0.350	0.348	0.346	2.75
Crisis	LLM-TLS	0.380	0.380	0.383	0.378	3.19
Crisis	FS-LLM-TLS	0.395	0.395	0.395	0.390	4.69
T17	DATEWISE	0.247	0.245	0.247	0.245	1.00
T17	LLM-TLS	0.482	0.482	0.482	0.479	1.69
T17	FS-LLM-TLS	0.523	0.520	0.520	0.521	1.50

5.3 Implementability and External Plausibility of SA Matching

This subsection tests whether semantic-first matching can be stably implemented and whether its judgments agree with human milestone labels.

5.3.1 Cross-Evaluator Consistency

We first test whether the framework depends on a single evaluator choice.

Table 3 reports **SA Score** on **Entities**, **Crisis**, and **T17** under four evaluators. The same system ordering is preserved throughout: **DATEWISE** (Gholipour Ghalandari and Ifrim, 2020) remains lowest, **LLM-TLS** (Hu et al., 2024) remains intermediate, and **FS-LLM-TLS** remains highest. The variance column is uniformly small, staying on the order of 10^{-6} across all rows. Semantic-first matching is therefore stably realizable at the level relevant to system comparison.

5.3.2 Agreement with Human Judgments

Evaluator consistency alone is not sufficient. We therefore compare a standard date-first matcher with the Codex-5.3 realization of **SA Score** against human labels on the same annotated subset.

TLS milestone correctness is binary, so our primary statistic is the **Matthews Correlation Coefficient (MCC)** (Chicco and Jurman, 2020). Exact confusion-matrix counts, macro-averaging, and MCC computation are given in Appendix J.

Four researchers annotated milestone correctness on 12 sampled **Entities** topics. A dual-blind subset yielded Cohen's Kappa (Cohen, 1960) $\kappa = 0.93$; the exact protocol and kappa computation are also reported in Appendix J. Because this pool is limited to those 12 topics, we use it as external plausibility evidence rather than as a full human benchmark across all datasets.

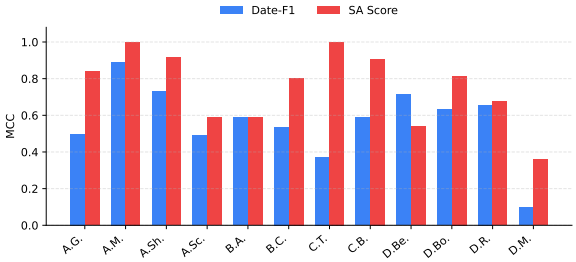

Figure 3: Topic-level MCC against human milestone judgments on the shared 12-topic **Entities** subset. Blue bars denote **Date-F1**; red bars denote **SA Score** instantiated with **Codex-5.3**. Appendix J gives full topic names, abbreviated labels, and raw counts.

Figure 3 compares the topic-level MCC of the binary matching decisions induced by **Date-F1** and **SA Score**. **SA Score** is higher on most topics; one topic is effectively tied, and one favors **Date-F1**. The difference is therefore broad rather than universal. This pattern is consistent with the semantic-first design: event equivalence is established before temporal fidelity is assessed. Appendix J, especially Table 6, reports the corresponding macro-averaged Precision, Recall, F1, and MCC.

5.4 Distinct Utility for Event-Centric TLS

To understand what semantic-first evaluation makes visible beyond the cross-system results above, we examine controlled changes within the same system family. SA metrics yield a more interpretable response profile over event validity, temporal grounding, and aligned summary quality than date-first evaluation.

Table 4 localizes this pattern on the Entities dataset through a stage-wise ablation. Here, **FS-LLM-TLS** is the **LLM-TLS** baseline augmented with **S1**, **S3**, and **S4**.

The first two rows show that **FS-LLM-TLS** im-

Table 4: Stage-wise ablation on the Entities dataset. Rows involving Stage 4 primarily test milestone-selection changes, whereas Stage 1 and Stage 3 mainly affect candidate construction and realization.

Method	Standard Date-First Metrics			Semantic-Alignment Metrics (Ours)				
	Date-F1	AR1	AR2	SA Score	SA Date-F1	SA-R1	SA-R2	STA-R1
LLM-TLS	0.232	0.091	0.037	0.281	0.152	0.115	0.049	0.086
FS-LLM-TLS	**0.256**	**0.095**	**0.040**	0.330	0.172	0.130	**0.061**	**0.099**
LLM-TLS + S1	0.230	0.092	0.037	0.285	0.154	0.116	0.050	0.086
LLM-TLS + S3	0.232	0.094	**0.040**	0.285	0.152	0.121	0.052	0.089
LLM-TLS + S4	0.253	0.093	0.039	0.332	0.175	0.127	0.057	0.095
LLM-TLS + S1+S3	0.236	**0.095**	**0.040**	0.288	0.156	0.123	0.055	0.095
LLM-TLS + S3+S4	0.252	0.094	0.039	**0.346**	**0.177**	**0.132**	0.058	0.098
LLM-TLS + S1+S4	0.249	0.093	0.039	0.332	0.175	0.131	0.058	0.097

proves the event-level signal of **LLM-TLS**: **SA Score** rises from 0.281 to 0.330 (+0.049), with smaller gains in **SA Date-F1** (+0.020) and **SA-R1** (+0.015). The single-stage rows identify where this gain comes from. **S1** and **S3** each raise **SA Score** only to 0.285 (+0.004), whereas **S4** raises it to 0.332 (+0.051). This is consistent with the role of **S4**: adding LLM judgment to milestone selection makes the retained milestones more semantically consistent with the reference timeline.

The combined rows validate the same interpretation. The construction-only combination **S1+S3** remains close to the base system in **SA Score** (0.288), while combinations that include **S4** stay much higher: **S3+S4** reaches 0.346 and **S1+S4** reaches 0.332. The decomposed metric profile is therefore useful diagnostically. **SA Score** captures the semantic effect of selection, **SA Date-F1** shows whether those semantically valid milestones are temporally grounded, and the aligned ROUGE variants reflect the wording quality of the retained summaries. Appendix K analyzes why temporal gains remain more limited, and Appendix L gives a qualitative case study of unconstrained LLM selection.

5.5 Scaling Outlook for Milestone Selection

The broader implication is that LLMs may affect TLS beyond the summarization step. LLMs already provide a natural mechanism for candidate abstraction, but milestone selection still lacks a mature LLM-centered recipe. We do not train such a selector here. However, established directions such as preference optimization, reinforcement-style self-improvement, and distillation (Rafailov et al., 2023; Yuan et al., 2024; Hsieh et al., 2023) make it plausible that stronger semantic judgment will increasingly be pushed into milestone selection.

As a minimal scaling check, Appendix M reports **FS-LLM-TLS** on **Entities** with 13B, 30B, and 70B model configurations. **SA Score** rises from 0.330 to 0.348 and 0.352, and **STA-R1** rises from 0.099 to 0.104 and 0.105, while **Date-F1** changes only from 0.256 to 0.258 and 0.262. This is not a scaling law, but it supports the narrower view that additional model capacity can improve event-level milestone quality in ways that exact-date evaluation only weakly exposes.

6 Conclusion

Date-first metrics remain appropriate for classical date-centric TLS, but they are less well matched to modern settings in which system gains increasingly come from event-level construction and selection.

We therefore propose semantic-first evaluation for TLS. The framework first establishes event equivalence and then assesses temporal grounding and summary quality. Its concrete instantiations are **SA Score**, **SA Date-F1**, **SA-ROUGE**, and **STA-ROUGE**.

The experiments support this evaluation order. Across **DATEWISE**, **LLM-TLS**, and the diagnostic testbed **FS-LLM-TLS**, semantic-first metrics give a more coherent account of event-level differences, and human judgments on the annotated subset support this interpretation. Overall, TLS evaluation should complement date-first assessment with semantic-first matching when the goal is to measure event-centric outputs.

7 Limitations

Our work proposes an evaluation framework for timeline summarization based on semantic alignment. This approach has several limitations. First, it relies on LLM-based semantic similarity, which may be sensitive to prompting strategies and introduces additional computational overhead. Second, sentence-level evaluation may not fully capture complex or long-span events with finer-grained temporal structures. Finally, although semantic alignment is a general principle, our framework is validated only on timeline summarization, and its extension to other structured generation tasks requires further adaptation and validation. A more detailed discussion is provided in Appendix N.

8 Ethics Statement

This work focuses on the evaluation of timeline summarization systems and does not involve the deployment of models in real-world decision-making settings. All experiments are conducted on publicly available datasets, and no personal or sensitive information is collected or generated.

Our human annotation study is limited to assessing the factual correctness of the predicted milestones with respect to existing ground-truth timelines. Annotators evaluate binary correctness decisions and do not assess sensitive attributes or make value-based judgments. All annotations are performed by researchers and used solely for evaluation purposes.

The proposed evaluation framework leverages large language models to assess semantic similarity. While LLM-based evaluation may inherit biases present in the underlying models, our use is restricted to comparative evaluation rather than content generation or decision support. We encourage future work to further examine robustness and fairness considerations when extending semantic-alignment evaluation to other tasks or domains.

References

James Allan, Rahul Gupta, and Vikas Khandelwal. 2001. Temporal summaries of news topics. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 10–18.

Xiuying Chen, Zhangming Chan, Shen Gao, Meng-Hsuan Yu, Dongyan Zhao, and Rui Yan. 2019. Learning towards abstractive timeline summarization. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 4939–4945.

Davide Chicco and Giuseppe Jurman. 2020. The advantages of the matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics*, 21(1):6.

Hai Leong Chieu and Yoong Keok Lee. 2004. Query based event extraction along a timeline. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 425–432.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

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.

Cheng-Yu Hsieh, Chun-Liang Li, Chih-Kuan Yeh, Hootan Nakhost, Yusuke Fujii, Alexander Ratner, Ranjay Krishna, Chen-Yu Lee, and Tomas Pfister. 2023. Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8003–8017, Toronto, Canada. Association for Computational Linguistics.

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, Bangkok, Thailand. Association for Computational Linguistics.

Tom Kocmi and Christian Federmann. 2023. Large language models are state-of-the-art evaluators of translation quality. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 193–203, Tampere, Finland. European Association for Machine Translation.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-Eval: NLG evaluation using GPT-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. 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, Valencia, Spain. Association for Computational Linguistics.

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.

Muhammad Reza Qorib, Qisheng Hu, and Hwee Tou Ng. 2025. Just what you desire: Constrained timeline summarization with self-reflection for enhanced relevance. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(23):25065–25073.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36:53728–53741.

Thomas Scialom, Paul-Alexis Dray, Sylvain Lamprier, Benjamin Piwowarski, and Jacopo Staiano. 2021. QuestEval: Summarization asks for fact-based evaluation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6594–6604, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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, Hong Kong, China. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, and 49 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288.

Giang Binh Tran, Mohammad Alrifai, and Eelco Herder. 2015. Timeline summarization from relevant headlines. In *Advances in Information Retrieval*, volume 9022 of *Lecture Notes in Computer Science*, pages 245–256, Cham. Springer International Publishing.

Giang Binh Tran, Mohammad Alrifai, and Dat Quoc Nguyen. 2013. Predicting relevant news events for timeline summaries. In *Proceedings of the 22nd International Conference on World Wide Web Companion*, pages 91–92.

Rui Yan, Liang Kong, Congrui Huang, Xiaojun Wan, Xiaoming Li, and Yan Zhang. 2011. Timeline generation through evolutionary trans-temporal summarization. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 433–443, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, Xian Li, Sainbayar Sukhbaatar, Jing Xu, and Jason E. Weston. 2024. Self-rewarding language models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 57905–57923. PMLR.

Chenlong Zhang, Tong Zhou, Pengfei Cao, Zhuoran Jin, Yubo Chen, Kang Liu, and Jun Zhao. 2025. 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.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating text generation with BERT. In *International Conference on Learning Representations*.

Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M. Meyer, and Steffen Eger. 2019. MoverScore: Text generation evaluating with contextualized embeddings and earth mover distance. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 563–578, Hong Kong, China. Association for Computational Linguistics.

A Discussion on Alternative Semantic Evaluation Metrics

This appendix expands the discussion in Section 2.2. Several general-purpose semantic evaluation metrics, such as **BERTScore** (Zhang et al., 2020), **MoverScore** (Zhao et al., 2019), and **QuestEval** (Scialom et al., 2021), have been proposed to assess semantic similarity beyond surface-level overlap. While effective for sentence- or document-level generation tasks, these metrics do not provide explicit event-level alignment or account for temporal structure, which are central to timeline summarization. Directly applying such metrics as end-to-end evaluation signals would obscure key error types in TLS, including **missing events**, **spurious milestones**, and **temporal misalignment**. In contrast, our framework leverages semantic similarity as an intermediate signal to establish event correspondence, upon which task-specific evaluation of semantic coverage and temporal fidelity can be performed.

B Date-First Metrics: Date-F1 and A-ROUGE

This appendix supplements the date-first matching discussion in Section 3.1.

Date-F1. This metric measures the overlap between the sets of predicted and reference dates, ignoring event content. Let $D_{\mathcal{P}}$ and $D_{\mathcal{R}}$ denote the sets of dates in the predicted and reference timelines, respectively. Precision (P_{date}) and Recall (R_{date}) are defined as:

$$P_{\text{date}} = \frac{|D_{\mathcal{P}} \cap D_{\mathcal{R}}|}{|D_{\mathcal{P}}|}, \quad R_{\text{date}} = \frac{|D_{\mathcal{P}} \cap D_{\mathcal{R}}|}{|D_{\mathcal{R}}|}.$$

The Date-F1 score is then computed as the harmonic mean:

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

A-ROUGE. We follow the date-aware implementation used in the `tilse` toolkit for the standard date-first TLS metrics (Martschat and Markert, 2017). Unlike **Date-F1**, which gives credit only under exact date agreement, **A-ROUGE** couples temporal proximity with lexical overlap in a single alignment-and-scoring procedure. Let d_i^P and d_j^R denote the dates of the i-th predicted and j-th reference milestones, and let y_i^P, y_j^R denote their texts. The toolkit first constructs a many-to-one alignment

$$\mathcal{M}_{\text{A}} = \text{Align}_{\text{date+content}}(\mathcal{P}, \mathcal{R}),$$

where alignment is determined jointly by date distance and approximate ROUGE-1 overlap. After alignment, each matched pair receives a date-decayed weight

$$w_{ij} = \frac{1}{|d_i^P - d_j^R| + 1}, \qquad (i,j) \in \mathcal{M}_{\text{A}}.$$

Its lexical contribution is then discounted by that weight:

$$c_{ij}^{(n)} = w_{ij} \cdot \text{ROUGE}_n(y_i^P, y_j^R).$$

Exact date matches therefore receive full weight, while a drift of k days contributes only $1/(k+1)$ of the same lexical-overlap credit.

This design uses date distance and lexical overlap as a computable proxy for event correspondence. In other words, event correspondence, temporal agreement, and wording quality are partially mixed in the same score.

Notice: Our semantic-first framework separates these dimensions once event-level alignment is established explicitly: **SA Score** evaluates event correspondence, **SA Date-F1** evaluates temporal correctness on aligned milestones, and the aligned ROUGE variants evaluate summary quality on the aligned set. This yields feedback that is more interpretable than date-aware lexical matching alone.

C Prompt Templates for Semantic Evaluation

This appendix expands the evaluator design in Section 4.1.1. For the main semantic-matching instantiation reported in the paper, we employed the `Codex-5.3` API as the evaluator for the **SA metrics**. The model was provided with the following prompt and decision rules to judge the event equivalence between a predicted sentence and a ground-truth reference sentence.

Event Equivalence Judgment Prompt

Task Instruction:
You are given two sentences:
- **GT sentence:** from the ground-truth timeline.
- **Pred sentence:** from the predicted timeline.

They are roughly aligned by position, but may or may not describe the same event. Your job is to judge whether they refer to the same underlying event,

following the rules below.

Decision Rules:

1. **Same event definition:** Same factual situation, even if wording differs. Look for overlap in: key actors, approximate time period, location, core action, cause, or consequence.
2. **Time tolerance rule (Important):**
   - Small or reasonable discrepancies in time are acceptable (e.g., "early March" vs. "March 5", "late 2023" vs. "December 2023").
   - Major time shifts are NOT acceptable (e.g., different years).
3. **Phrasing:** Synonyms, paraphrasing, different levels of detail, or political framing do not prevent a match.
4. **Different event:** Factual conflict or distinct situations. Substantially different actors, actions, causes, or outcomes → not match (e.g., Resignation vs. Dismissal).
5. **Detail Asymmetry:** If one sentence is more detailed than the other but clearly refers to the same situation, mark as match.

Output Format:
Provide a binary judgment: "yes" or "no", plus a brief explanation.

D Detailed Realization of Semantic-Alignment Metrics

This appendix expands the metric definitions in Section 4.1.2. To avoid leaving the realization underspecified after the main-text compression, we include the fuller **Semantic-Alignment Metrics** subsection from the most complete previous method chapter.

D.1 Semantic-Alignment Metrics

D.1.1 Design Rationale

Our design follows both the event-centric nature of modern TLS and the principle of human judgment. When readers compare two timelines, they first ask whether the same event is being described; only then do they ask whether the date is exactly right or whether the wording is fluent. In long timelines, a modest temporal shift may still refer to the same milestone, while a small semantic mismatch may indicate a completely different event. Thus, a milestone must first be semantically aligned. If not, writing quality or timestamp accuracy should not receive credit. Based on this principle, we calculate SA Score, with SA Date-F1 testing timestamp accuracy after semantic matching and SA-ROUGE testing summary quality on semantically aligned pairs. Semantic & Temporal-Alignment ROUGE (STA-ROUGE) combines both, giving an even stricter evaluation.

D.1.2 Metric Definitions

Semantic-Alignment Score (SA Score). To establish a principled foundation for semantic alignment, we design a candidate-based alignment strategy grounded in semantic similarity. Given a predicted timeline $\mathcal{P} = \{(d_i^P, y_i^P)\}_{i=1}^M$ and a reference timeline $\mathcal{R} = \{(d_j^R, y_j^R)\}_{j=1}^N$. Here d_i^P and d_j^R denote the dates of the i-th predicted and j-th reference events, and y_i^P, y_j^R their associated summaries. We first compute pairwise similarity score of the pair.

$$s_{ij} = \mathrm{sim}(y_i^P, y_j^R) \in [0, 1],$$

where $\mathrm{sim}(\cdot)$ is derived from an encoder $\phi(\cdot)$. These scores serve only as soft weights for candidate retrieval. For each reference milestone y_j^R, we select the top-K_{cand} candidates.

$$C_j = \mathrm{TopK}_{i \in [M]}\, s_{ij}, \quad K_{\text{cand}} = 5,$$

ensuring that only the most semantically relevant predicted events are considered. K_{cand} is the candidate pool size (default 5), and $\phi(\cdot)$ specifies the similarity backbone. This design controls the computational budget.

For each candidate $i \in C_j$, we query the LLM with a yes/no prompt, and this yields a binary judgment. The complete prompt template and decision guidelines are detailed in Appendix C. To ensure a neutral and unbiased evaluation, we employ the `Gemini-2.0-Pro` API as the independent evaluator for all event-equivalence judgments, separate from the generator models (e.g., Llama-2 series).

$$\delta_{ij} = \mathbf{1}[\text{LLM-judge}(y_i^P, y_j^R) = \text{Yes}] \in \{0, 1\}.$$

$$\gamma_j = \max_{i \in C_j} \delta_{ij},$$

which indicates whether y_j^R is successfully matched by at least one predicted candidate. The matched set is then

$$\widehat{\mathcal{M}} = \{(i^*, j) \mid \gamma_j = 1,\ i^* = \arg\max_{i \in C_j : \delta_{ij} = 1} s_{ij}\}.$$

Here $\mathcal{M}$ denotes a generic set of candidate index pairs (i, j) between predicted and reference timelines, while $\widehat{\mathcal{M}}$ is the optimized set actually retained after candidate filtering and LLM judgment. In other words, $\mathcal{M}$ represents the search space of all possible alignments, and $\widehat{\mathcal{M}}$ is the final alignment outcome. Finally, the *SA Score* is computed

as the fraction of reference milestones that are semantically covered:

$$\text{SA Score} = \frac{1}{N}\sum_{j=1}^{N}\gamma_j.$$

Here N is the number of reference milestones, while K_{cand} denotes the candidate pool size.

Semantic-Alignment Date-F1 (SA Date-F1). Among semantically aligned pairs, we further check whether timestamps coincide. Let $D_{\mathcal{P}}^{\text{match}} = \{d_i^P \mid (i,j) \in \widehat{\mathcal{M}}, d_i^P = d_j^R\}$ and $D_{\mathcal{R}}^{\text{match}} = \{d_j^R \mid (i,j) \in \widehat{\mathcal{M}}, d_i^P = d_j^R\}$. We compute

$$P_{\text{SA-date}} = \frac{|D_{\mathcal{P}}^{\text{match}}|}{|\mathcal{P}|}, \quad R_{\text{SA-date}} = \frac{|D_{\mathcal{R}}^{\text{match}}|}{|\mathcal{R}|},$$

and define

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

Note that in our setting $|\mathcal{P}| = |\mathcal{R}|$, so the denominators coincide and precision and recall are always identical; hence, in subsequent tables we report only the unified F1 score for SA Date-F1.

Semantic-Alignment ROUGE (SA-ROUGE). Finally, we compute ROUGE-n between summaries of each $(i,j) \in \widehat{\mathcal{M}}$ and average:

$$\text{SA-ROUGE}_n = \frac{1}{|\widehat{\mathcal{M}}|}\sum_{(i,j)\in\widehat{\mathcal{M}}} \text{ROUGE}_n(y_i^P, y_j^R).$$

Semantic & Temporal-Alignment ROUGE (STA-ROUGE). To jointly enforce semantic consistency and temporal fidelity, we further introduce STA-ROUGE. Specifically, we restrict ROUGE (Lin, 2004) computation to pairs $(i,j) \in \widehat{\mathcal{M}}$ that are not only judged semantically equivalent by the LLM ($\delta_{ij} = 1$) but also have timestamps within a small tolerance window ϵ, i.e., $|d_i^P - d_j^R| \le \epsilon$, consistent with A-ROUGE practice. Formally,

$$\text{STA-ROUGE}_n = \frac{1}{|\widehat{\mathcal{M}}_{\text{time}}|}\sum_{(i,j)\in\widehat{\mathcal{M}}_{\text{time}}} \text{ROUGE}_n,$$

where

$$\text{ROUGE}_n = \widehat{\text{ROUGE}}_n(y_i^P, y_j^R),$$

$$\widehat{\mathcal{M}}_{\text{time}} = \{(i,j) \in \widehat{\mathcal{M}} \mid |d_i^P - d_j^R| \le \epsilon\}.$$

This metric captures both semantic alignment and correct temporal grounding within tolerance, providing a stricter assessment of timeline summarization quality.

E Details of the FS-LLM-TLS Diagnostic Testbed

This appendix expands the diagnostic-testbed description in Section 4.2. We provide the detailed architecture and workflow of **FS-LLM-TLS** as used in this paper. Consistent with the main text, it is included only as a lightweight diagnostic testbed for probing metric behavior.

The testbed extends the four-stage pipeline of LLM-TLS (Hu et al., 2024) in a targeted way so that semantic judgment can be examined more explicitly in later-stage milestone decisions.

As illustrated in Figure 4, the process moves sequentially:

- **Stage 1 (Snippet Extraction):** An LLM generates a timestamped event snippet for each document.
- **Stage 2 (Event Clustering):** Snippets are grouped based on semantic similarity and temporal proximity.
- **Stage 3 (Abstractive Summarization):** Cluster-level summarization produces a single, coherent milestone description.
- **Stage 4 (Milestone Selection):** The final timeline is constructed by selecting and ordering L milestones.

While retaining the overall structure of LLM-TLS, we modify three stages—unified snippet extraction, cluster-level abstractive summarization, and hybrid milestone selection—so that semantic judgment can be traced more clearly across the construction and selection stages discussed in the main text.

F Prompt Templates for Summarization (Stage 1 in FS-LLM-TLS)

This appendix expands the Stage 1 summarization-side modifications summarized in Section 4.2. We provide the full one-shot prompt templates used to generate grounded event snippets. We designed specific prompts for each dataset (Entities, Crisis, and T17), incorporating a domain-relevant example from the training set to guide the model.

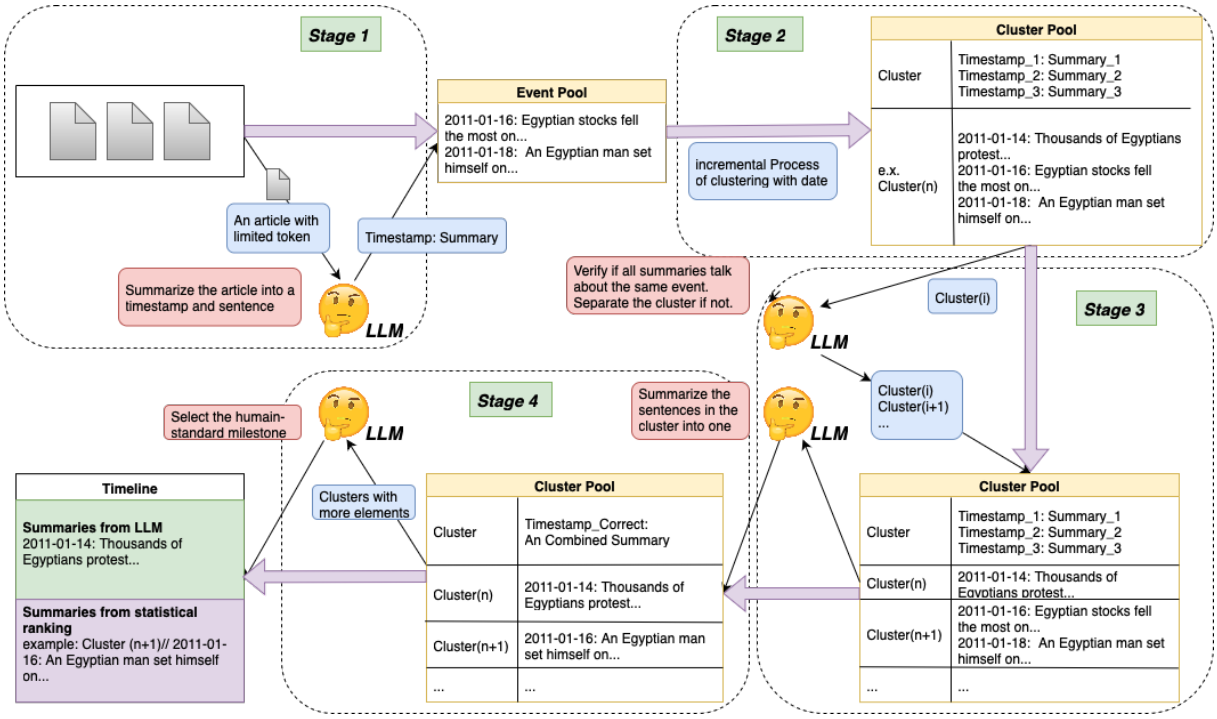

Figure 4: **Overview of the FS-LLM-TLS diagnostic testbed.** The process operates in four stages: (1) *Unified Snippet Extraction*, (2) *Event Clustering*, (3) *Abstractive Summarization*, and (4) *Hybrid Selection*. The LLM is integrated at each stage so that semantic judgment can play a larger role in both candidate construction and milestone selection.

F.1 Prompt for Entities Dataset

One-Shot Prompt (Entities)

Instruction
Review the news article associated with the provided keyword. Identify and summarize the 1-2 most significant event related to the keyword.

Format
YYYY-MM-DD: One-sentence Summary

##################

Keyword 1
Bill Clinton

Content 1
Title: Bush plea tries to rebuild US image
Publish Date: 2005-01-04
Content:
President George W. Bush yesterday appoints two former presidents , George H.W. Bush and Bill Clinton , to launch a joint appeal to the American people for emergency donations to victims of the Indian Ocean tsunami .(2005-01-03) They stood side by side in the White House 's Roosevelt Room as Mr Bush told reporters : " We 're here to ask our fellow citizens to join in a broad humanitarian relief effort . " [...] (Content truncated for brevity)

Event Summary Related to 'Bill Clinton'
2005-01-03: President George W. Bush appoints former Presidents George H.W. Bush and Bill Clinton to head fundraising efforts for humanitarian aid after a devastating tsunami in Indian Ocean.

##################

Keyword 2
{keyword}

Content 2
{content}

Event Summary Related to '{keyword}'

F.2 Prompt for Crisis Dataset

One-Shot Prompt (Crisis)

Instruction
Review the news article associated with the provided keyword. Identify and summarize the 1-2 most significant event related to the keyword.

Format
YYYY-MM-DD: One-sentence Summary

##################

Keyword 1
syria syrian

Content 1
Title: UN condemns Syria after massacre of civilians
Publish Date: 2012-05-28
Content:
UN condemns Syria after massacre of civilians Emergency meeting of UN discusses response to attack that killed 34 children Syrian government " in the strongest possible terms " for heavy - weapon attacks on the town of Houla , where 108 people , and up

While the carefully worded statement stopped short of blaming anyone for the " close - range attacks " [...] (Content truncated for brevity)

Event Summary Related to 'syria syrian'
2012-05-27: The U.N. Security Council condemns Syrian army artillery and tank barrages on a civilian neighborhood where 108 people , most of them women and children , were killed , suggesting in a carefully worded statement that government forces were largely responsible .

##################

Keyword 2
{keyword}

Content 2
{content}

Event Summary Related to '{keyword}'

F.3 Prompt for T17 Dataset

One-Shot Prompt (T17)

Instruction
Review the news article associated with the provided keyword. Identify and summarize the most significant event related to the keyword.

Format
YYYY-MM-DD: One-sentence Summary

##################

Keyword 1
iraq iraqi

Content 1
Publish Date: 2005-12-15
Content:
President accepts full responsibility for invasion Pentagon to plant pro-US stories in global media President Bush conceded yesterday that much of the pre-war intelligence on Iraq " turned out to be wrong " but took sole responsibility for ordering the invasion , which he said was justified by the potential threat posed by Saddam Hussein .(2005-12-14) [...] (Content truncated for brevity)

Event Summary Related to 'iraq iraqi'
2005-12-14: George Bush concedes that much of the pre-war intelligence on Iraq 'turned out to be wrong' but insists the conflict was justified anyway due to the threat posed by Saddam Hussein .

##################

Keyword 2
{keyword}

Content 2
{content}

Event Summary Related to '{keyword}'

G Prompt for Event Synthesis (Stage 3 in FS-LLM-TLS)

This appendix expands the Stage 3 change summarized in Section 4.2. After clustering semantically similar sentences, the next challenge is to synthesize these noisy, redundant clusters into a single, high-quality event summary that mimics the style of ground-truth timelines. We designed the following prompt to instruct the LLM to act as a "Lead Editor," enforcing strict objectivity and conciseness while fusing unique details from multiple sources.

Prompt for Stage 3: Event Synthesis

System Instruction
You are a Lead Editor for a historical timeline database. Your task is to synthesize a **cluster of related sentences** into a **single, high-quality timeline event summary**.
Input Data
You will receive:
1. **Date:** The date of the event cluster.
2. **Cluster Sentences:** A list of sentences extracted from news articles that refer to the same event.

Output Requirements (Ground Truth Style)
1. **Synthesize:** Combine unique details (Who, What, Where, Why) from the cluster into ONE comprehensive sentence.
2. **Objective Tone:** Remove journalistic fluff (e.g., "Breaking news", "Sources report", "It is said that"). State the event directly as a fact.
3. **Conciseness:** The output must be a single, dense sentence (typically 15-30 words).
4. **Resolution:** If sentences conflict, prioritize the most specific and factual details.
5. **Format:** YYYY-MM-DD: [Summary Sentence]

One-Shot Example (Style Guide)
Date: 2005-01-03
Cluster Sentences:
- Bush asks Clinton and his father to help with the tsunami relief.
- Former Presidents George H.W. Bush and Bill Clinton were appointed by President Bush to lead the fundraising.
- A joint appeal for tsunami victims will be launched by two former presidents.
- White House announces Bush and Clinton to head humanitarian aid efforts.

Synthesized Output:
2005-01-03: President George W. Bush appoints former Presidents George H.W. Bush and Bill Clinton to head fundraising efforts for humanitarian aid after a devastating tsunami.

Real Task
Date: {date_str}
Cluster Sentences:
{cluster_text}
Synthesized Output:

H Prompt for Hybrid Milestone Selection (Stage 4 in FS-LLM-TLS)

This appendix expands the Stage 4 selection change summarized in Section 4.2. Stage 4 retains frequency- and coverage-based shortlist signals, but the final milestone choice is made by an LLM

with dataset-specific selection criteria. We therefore use separate prompts for **Entities**, **T17**, and **Crisis**, so that the LLM is asked to judge milestone salience at the appropriate granularity for each benchmark.

H.1 Prompt for Entities Dataset

Stage 4 Prompt for Entities

You are a senior biographical timeline editor for an international news organization. Your job is to compile a concise life timeline that human editors would recognize as the person's important milestones.
Task:
- Topic/person: {topic}
- Select exactly {L} row_ids from the candidate pool.
- Treat this as final biographical milestone selection, not ordinary news ranking.
- Use only the candidate event texts below.

Strict constraints:
- Do not browse the web.
- Do not search the internet.
- Do not use Wikipedia, biography pages, external timelines, news archives, or external files.
- Do not select an event merely because it mentions the person.
- Return exactly {L} unique row_ids.

Selection rules:
1. Prefer events a human editor would keep in a short biography or life timeline for this person.
2. Prioritize events that changed the person's life status, public role, power, legal/health/family identity, body of work, public impact, or long-term legacy.
3. Within the same process, prefer boundary or durable nodes: beginning, formal confirmation, result, status transfer, lasting public impact, or legacy.
4. Preserve coverage across major life processes instead of overfilling one process.
5. Avoid routine commentary, previews, reactions, gossip, minor appearances, ordinary mentions, and duplicates.
6. If fewer than {L} candidates are ideal, fill the remaining slots with the next-best candidates that are most biography-worthy and least redundant.

Return JSON only:
{"selected_row_ids": ["E00001", "..."], "selection_count": {L}}
Candidates:
{candidates}

H.2 Prompt for T17 Dataset

Stage 4 Prompt for T17

You are a senior news timeline editor for an international news organization. Your job is to compile a concise timeline of the development of a news topic, in the style that human timeline editors would recognize.
Task:
- Topic/keyword: {topic}
- Select exactly {L} row_ids from the candidate pool.
- Treat this as topic-process timeline selection, not ordinary relevance ranking.
- Use only the candidate event texts below.

Strict constraints:
- Do not browse the web.
- Do not search the internet.
- Do not use external timelines, news archives, Wikipedia, or external files.
- Do not select an event merely because it mentions the topic keyword.
- Return exactly {L} unique row_ids.

Core criterion:
Select events that mark meaningful state changes in the topic's development. A state change can be political, legal, military, financial, public-health, disaster-response, investigative, operational, institutional, social, or outcome-related. The exact type depends on the topic.
Selection rules:
1. Prefer candidates that help reconstruct the main dated development of the topic.
2. Prefer nodes that start, redirect, escalate, de-escalate, formalize, quantify, resolve, or materially reframe the process.
3. Respect the topic's natural granularity. For dense day-by-day topics, a daily state update may be valid if it changes the story state; for short retrospective topics, use a higher threshold.
4. Preserve coverage across the timeline's stages rather than overfilling one sub-process.
5. If two candidates describe the same node, choose the clearer, more concrete, more consequence-bearing version.
6. Avoid routine comments, weak reactions, background mentions, predictions, and duplicate rows without new process information.
7. If fewer than {L} candidates are ideal, fill the remaining slots with the next-best candidates that add the most process coverage and the least redundancy.

Return JSON only:
{"selected_row_ids": ["E00001", "..."], "selection_count": {L}}
Candidates:
{candidates}

H.3 Prompt for Crisis Dataset

Stage 4 Prompt for Crisis

You are a senior crisis timeline editor for an international news organization. Your job is to compile a concise timeline of how a political or social crisis develops over time, in the style that human timeline editors would recognize.
Task:
- Crisis topic/country: {topic}
- Select exactly {L} row_ids from the candidate pool.
- Treat this as crisis-process timeline selection, not ordinary relevance ranking.
- Use only the candidate event texts below.

Strict constraints:
- Do not browse the web.
- Do not search the internet.
- Do not use external timelines, news archives,

Wikipedia, or external files.
- Do not select an event merely because it mentions the country, leader, opposition, or crisis keyword.
- Return exactly {L} unique row_ids.

Core criterion:
Select events that mark meaningful state changes in the crisis. A state change can involve public mobilization, violence, governance, negotiation, institutional response, territorial or security control, legal status, international involvement, humanitarian consequence, election/transition, or public legitimacy. The exact type depends on the crisis phase.

Selection rules:
1. Prefer candidates that help reconstruct the crisis's main dated development.
2. Prefer nodes that start, redirect, escalate, de-escalate, formalize, internationalize, resolve, or materially reframe the crisis.
3. Preserve coverage across phases. The timeline may include early mobilization, government response, opposition action, violence/security shifts, institutional decisions, negotiations, transition, and aftermath when present in the pool.
4. Multiple candidates from the same broad process may be selected if each marks a distinct stage or consequence.
5. Do not over-collapse a long crisis into only one protest, one crackdown, or one leadership event.
6. Avoid routine diplomatic comments, repeated calls without action, background mentions, rumors, and duplicate rows without new crisis-state information.
7. If fewer than {L} candidates are ideal, fill the remaining slots with the next-best candidates that improve crisis-process coverage and minimize redundancy.

Return JSON only:
{"selected_row_ids": ["E00001", "..."], "selection_count": {L}}

Candidates:
{candidates}

I Computational Infrastructure

This appendix expands the experimental setup in Section 5.1. All experiments reported in this paper were conducted on a local high-performance computing (HPC) server. The infrastructure consists of a node equipped with **8 × NVIDIA A800 GPUs** (80GB VRAM per GPU).

The models were implemented using PyTorch framework within a Python environment. To ensure reproducibility and efficiency, we utilized the distributed data parallel (DDP) strategy across the available GPUs where applicable.

J Detailed Statistics for Human Agreement Study

In Section 5.3.2, we reported macro-averaged agreement metrics (MCC, F1) to compare **SA Score** with **Date-F1** on the annotated subset.

To ensure consistency in the manual evaluation process, we provided the annotators with the following guidelines regarding the timeline comparison task:

Task Overview:
Annotators compare generated milestones against reference milestones and decide whether each candidate pair describes the same concrete timeline milestone. Judgments are based on the event descriptions and dates shown to the annotator; outside knowledge is not used unless explicitly available in the annotation material.

Definition of a Match:
A generated milestone is counted as correct only when it refers to the same core event, status change, outcome, or biographical/public-impact milestone as the reference. A valid match should align on the main topic entity, the entity's role in the event, the central action or outcome, and the relevant process stage. Shared date, actor, topic, or news context alone is not sufficient.

Decision Criteria:
Annotators were instructed to apply the following criteria consistently:

- **Core Milestone Equivalence:** Paraphrases, different sentence structure, or broader/narrower wording are acceptable only when both descriptions uniquely identify the same milestone.
- **Actor and Role Consistency:** The topic entity must be the main actor, target, office-holder, winner, victim, or otherwise directly affected subject in both descriptions.
- **Process-Stage Consistency:** Related stages of the same process are not automatically matches. For example, candidacy, nomination, election victory, inauguration, resignation, arrest, indictment, conviction, sentencing, hospitalization, and death should be treated as separate milestones unless the two descriptions clearly collapse to the same event.
- **Outcome or Status Change:** When the reference milestone is outcome-based, the generated milestone must capture the same outcome or durable status change, not merely background pressure, discussion, prediction, or an earlier procedural step.

Table 5: Expanded confusion-matrix statistics for MCC computation. For each dataset, the candidate event set is fixed. **TP**: true positive, **FP**: false positive, **FN**: false negative, **TN**: true negative. FN is computed as Human − TP and TN as (Total − Human) − FP.

Dataset Topic	Total	Human (Matched)	Date-F1				SA Score (Codex-5.3)			
			TP	FP	FN	TN	TP	FP	FN	TN
Al Gore	25	10	5	1	5	14	8	0	2	15
Angela Merkel	18	8	7	0	1	10	8	0	0	10
Ariel Sharon	25	8	5	0	3	17	8	1	0	16
Arnold Schwarzenegger	17	8	3	0	5	9	4	0	4	9
Bashar al-Assad	24	7	3	0	4	17	3	0	4	17
Bill Clinton	47	9	3	0	6	38	8	2	1	36
Charles Taylor	18	9	4	1	5	8	9	0	0	9
Chris Brown	24	7	3	0	4	17	7	1	0	16
David Beckham	32	10	6	0	4	22	5	1	5	21
David Bowie	24	8	4	0	4	16	6	0	2	16
Dilma Rousseff	24	10	8	2	2	12	9	3	1	11
Dmitry Medvedev	16	10	6	3	4	3	7	2	3	4
Total	**294**	**104**	**57**	**7**	**47**	**183**	**82**	**10**	**22**	**180**

Table 6: Macro-averaged agreement between automatic metrics and human judgments on milestone correctness over the shared 12-topic subset. **SA Score** is instantiated with **Codex-5.3**.

Metric	Precision	Recall	F1	MCC
Date-F1	0.925	0.543	0.666	0.566
SA Score	0.910	**0.789**	**0.826**	**0.753**

- **Date Relation:** Exact date agreement is not required. A date difference is acceptable when the event core is the same and the shift is plausibly due to a reporting date, time zone, official certification date, or multi-day event. A date match alone does not justify a semantic match, and a date shift that signals a different stage should not be counted as correct.

- **Borderline Pairs:** If two milestones are closely related but differ in stage, outcome, actor role, or specificity, annotators mark them as related but not correct under the strict binary scoring used for MCC and F1.

To provide full transparency regarding the annotation pool, we present the raw annotation counts for all 12 sampled topics in Table 5. The table reports the total generated statements, the number of human-validated correct statements, and the breakdown of True Positives (TP) and False Positives (FP) for both metrics.

For ease of comparison with the main-text grouped bar chart, Table 6 reports the macro-averaged Precision, Recall, F1, and MCC computed from the topic-level counts above.

This macro-averaged summary mirrors the main-text topic-level visualization. **SA Score** maintains precision comparable to **Date-F1** while improving recall, F1, and MCC, which is consistent with the view that date-first matching is conservative but misses semantically valid milestones when temporal agreement fails.

Computation of Agreement Metrics. Each predicted milestone is treated as a binary decision indicating whether it correctly matches a ground-truth event, as determined by human annotators. Based on the raw counts reported in Table 5, we compute standard agreement metrics as follows. For each dataset, True Positives (TP) correspond to predicted milestones that are validated as correct by humans, while False Positives (FP) denote incorrect predictions accepted by the metric. False Negatives (FN) are human-validated milestones that are missed by the metric.

Precision, Recall, and F1 are computed using the standard definitions:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}},$$
$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}},$$
$$\text{F1} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}.$$

The Matthews Correlation Coefficient (MCC) is computed as:

$$\text{MCC} = \frac{\text{TPTN} - \text{FPFN}}{\sqrt{D}},$$

Table 7: Milestone-level comparison on the **Al Gore** topic. The ground truth preserves cross-domain coverage, while the naive selector collapses the timeline into campaign-stage detail.

Ground Truth (GT): Balanced & Multi-Domain	**Predicted Timeline (PT): Political Detail Bias**
1993: Inaugurated as Vice President.	**1999:** Announces candidacy for 2000 election.
2000: Concedes election to Bush after Supreme Court ruling.	**2000:** Wins Iowa caucuses.
2003: Joins Board of Directors for **Apple Inc.**.	**2000:** Backed by names in showbiz, including Robert De Niro.
2005: Launches cable channel **Current TV**.	**2000:** Senator Joseph Lieberman is chosen as running mate.
2007: *An Inconvenient Truth* wins an **Oscar**.	**2000:** Debates Bush in the first of three presidential debates.
2007: Co-winner of the **Nobel Peace Prize**.	**2000:** Spars with Bush in the final presidential debate.
2016: Meets President-elect Trump on climate issues.	**2000:** Neck and neck with Bush in the daily polls.
(Politics, Tech, Media, Environment)	*(... and 14 more events strictly about Election 2000 ...)*

Table 8: Selection bias of a naive LLM selector on the **Al Gore** topic. The model over-concentrates on the 2000 election cycle and under-represents the broader biography.

Period / Domain	**GT**	**LLM**	**Bias**
1992–1999 (VP Era)	15%	4%	Under
2000 Campaign	**12%**	**88%**	**Over**
2001–2020 (Post-pol.)	73%	8%	Omitted

where

$$D = (\mathrm{TP} + \mathrm{FP})(\mathrm{TP} + \mathrm{FN}) \times (\mathrm{TN} + \mathrm{FP})(\mathrm{TN} + \mathrm{FN}),$$

and TN denotes true negatives. All metrics reported in Section 5.3.2 are macro-averaged across the 12 sampled topics.

Inter-Annotator Agreement. To assess annotation reliability, a dual-blind subset of the data was independently labeled by two annotators. Cohen's κ (Cohen, 1960) was computed at the sentence level, treating each milestone as a binary correct/incorrect decision. Following the standard definition,

$$\kappa = \frac{P_o - P_e}{1 - P_e},$$

where P_o is the observed agreement rate and P_e is the expected agreement by chance. This procedure yields a Cohen's κ score of $\kappa = 0.93$, indicating near-perfect inter-annotator agreement.

These findings help explain why temporal-grounding metrics remain challenging even after semantic matching: part of the remaining difficulty comes from ambiguity in the underlying data rather than from the matching procedure alone.

K Analysis on Temporal Grounding Limitations

This appendix expands the diagnostic discussion in Section 5.4. While FS-LLM-TLS achieves substantial gains in semantic coverage, the improvement in temporal accuracy (Date-F1) is comparatively modest. Our investigation on the **Entities** dataset reveals two primary factors imposing a strict upper bound on temporal precision:

- **Data Ambiguity:** Approximately 50% of the ground-truth timestamps in the reference timelines cannot be exactly recovered from the source articles due to vague temporal expressions (e.g., "late 2023" vs. "2023-12-20").
- **Role of LLMs:** In our pipeline, LLMs serve primarily as semantic reasoners rather than temporal calibrators. While they excel at identifying salient events, they cannot hallucinate precise dates for vaguely timestamped input without external knowledge bases.

L Qualitative Case Study: Selection Bias in Naive LLM

This appendix expands the diagnostic discussion in Section 5.4. It illustrates the behavior of an unconstrained LLM selector on the topic "Al Gore".

Content comparison. Table 7 shows the same effect at the milestone level. The ground truth spans politics, business, media, and environmental activity, whereas the naive selector collapses the timeline into a sequence of campaign details from 2000.

Observation. The naive selector concentrates most of the timeline on a single high-frequency political period. As shown in Table 8, it assigns **88%**

Table 9: Model-scaling check for **FS-LLM-TLS** on the **Entities** dataset.

Model	Standard Date-First Metrics			Semantic-Alignment Metrics (Ours)				
	Date-F1	AR1	AR2	SA Score	SA Date-F1	SA-R1	SA-R2	STA-R1
Llama-2-13B	0.256	0.095	0.040	0.330	0.172	0.130	0.061	0.099
Qwen3-30B-A3B	0.258	0.096	0.042	0.348	0.176	0.138	0.066	0.104
Llama-2-70B	0.262	0.099	0.044	0.352	0.175	0.140	0.068	0.105

of its timeline capacity to the 2000 election cycle. The result is narrow political over-selection and broad omission of milestones from other domains, including technology, media, and environmental advocacy.

M Model-Scaling Check for FS-LLM-TLS

This appendix supports the forward-looking discussion in Section 5.5. Table 9 reports a narrow **Entities**-only check for **FS-LLM-TLS**; the goal is to test whether larger model configurations can strengthen the event-level signal, not to claim a full scaling law.

The larger configurations improve **SA Score** and the aligned ROUGE variants more clearly than **Date-F1**. This pattern is consistent with the interpretation that additional capacity mainly improves event-level milestone selection and realization, while exact-date agreement remains a stricter and more weakly responsive signal.

N Detailed Limitations

Our work focuses on improving evaluation for timeline summarization by grounding comparison on semantic alignment between predicted and reference milestones. This appendix adds several limitations to those summarized in Section 7.

First, our framework relies on LLM-based semantic similarity to establish event-level alignment. Although we observe strong agreement with human judgments, such alignment may be sensitive to the choice of prompting strategy, and it incurs additional computational cost compared to purely heuristic evaluation metrics.

Second, the proposed evaluation assumes that semantic equivalence between milestones can be reasonably assessed at the sentence level. In scenarios involving highly complex events, overlapping sub-events, or events spanning long temporal intervals, sentence-level alignment may fail to capture finer-grained event structures and nuanced temporal relationships.

Finally, our evaluation framework is developed and validated specifically in the context of timeline summarization. While the underlying principle of semantic alignment is broadly applicable, extending this framework to other structured text generation tasks may require task-specific adaptations and additional empirical validation to ensure reliability and effectiveness.