Neurosymbolic Narrative Consolidation: Grounding Abstractive MDS and LLMs with Temporal Event Graphs

Roger Antonio Finger
Universidade do Vale do Rio dos Sinos (UNISINOS)
São Leopoldo, Brazil
Universidade Federal de Goiás (UFG)
Goiânia, Brazil
rfinger@edu.unisinos.br, rogerfinger@discente.ufg.br

Eduardo Gabriel Cortes
Universidade do Vale do Rio dos Sinos (UNISINOS)
São Leopoldo, Brazil
egcortes@edu.unisinos.br

Sandro Jose Rigo
Universidade do Vale do Rio dos Sinos (UNISINOS)
São Leopoldo, Brazil
rigo@unisinos.br

Gabriel de Oliveira Ramos
Universidade do Vale do Rio dos Sinos (UNISINOS)
São Leopoldo, Brazil
gdoramos@unisinos.br

***Abstract*—Neural summarization models and Large Language Models (LLMs) are effective surface realizers, but *Narrative Consolidation* (NC)—unifying overlapping accounts into a single narrative with high event coverage and strict temporal integrity—demands structural guarantees that standard multi-document summarization objectives do not provide. In practice, end-to-end neural generation over long concatenated inputs exhibits strong compression priors, scope collapse, and temporal drift, yielding fluent but structurally unreliable narratives. We introduce *TAEG-ANC*, a neurosymbolic framework that *grounds* abstractive generation in an explicit temporal knowledge structure. A *Temporal Alignment Event Graph* (TAEG) encodes (i) a canonical event backbone and (ii) cross-document event identity links, acting as a symbolic scaffold that constrains the generation process. Rather than delegating global planning to the model, TAEG-ANC decomposes NC into a sequence of event-level fusion problems: for each canonical event, the method retrieves aligned source spans and applies a neural generator only for local, micro-abstractive fusion, while the final narrative is assembled in canonical order. This separation of concerns provides controllability (chronology is enforced by construction) and interpretability (event-level provenance and alignment are explicit). We instantiate TAEG-ANC with four backbones (BART, PEGASUS, PRIMERA, and Gemma 3) and evaluate on a Gospel consolidation resource for the Passion Week (169 canonical events). Across backbones, the neurosymbolic pipeline produces substantially more complete narratives than global baselines and achieves perfect temporal agreement with the canonical order (Kendall's Tau=1.0). The Gemma 3 variant attains the strongest ROUGE-L and METEOR on this dataset, supporting the view that explicit symbolic scaffolds can robustly complement neural generation in long-form, multi-perspective narrative tasks.**

***Index Terms*—Neurosymbolic AI, Narrative Consolidation, Multidocument Summarization, Event Graphs, Abstractive Summarization, Large Language Models.**

I. Introduction

Multi-Document Summarization (MDS) is typically evaluated as an information compression problem, producing a short text that preserves salient content [11], [12]. In contrast, several application domains require *Narrative Consolidation* (NC): given multiple overlapping accounts, produce a single narrative that is (i) chronologically coherent, (ii) comprehensive in event coverage, and (iii) able to fuse complementary details while reducing redundancy [1], [16]. In settings such as legal witness statements or historical reconstruction, the relevant failure mode is often not lack of fluency, but loss of *event coverage* and violations of temporal consistency.

The canonical Gospels exemplify the challenges of NC, presenting a complex mix of overlapping content and interleaved events. This study centers on the Passion Week segment, a domain defined by high narrative density and cross-document redundancy, where maintaining temporal coherence is essential. This allows for rigorous, controlled evaluation by leveraging existing ground-truth resources, including a canonical timeline [25] and an event-aligned consolidated reference [27].

Recent work formalized NC by introducing the *Temporal Alignment Event Graph* (TAEG), a multi-relational graph that explicitly represents event alignment across documents and enforces a canonical temporal backbone [1]. This work adopts the formal definition of NC proposed by them. As shown in that setting, an explicit temporal structure can guarantee correct ordering when the output is generated by iterating over a canonical event list [25].

Viewed through the lens of neuro-symbolic KG–LLM integration, TAEG provides an explicit symbolic constraint system over event identity and temporal order. Unlike recent graph-constrained reasoning approaches that primarily utilize KGs for localized fact retrieval or entity augmentation [5], [8], our framework employs the graph as a structural backbone to dictate the global narrative flow. While the neural model acts as a local surface realizer, this separation of concerns enables controllable inference: global chronology is enforced by the KG backbone, and generation is grounded in event-aligned evidence retrieved from the KG.

So, this paper addresses a complementary question: how can we exploit abstractive neural generators for NC *by grounding them in explicit* global temporal planning. In long-context, end-to-end settings, neural models must simultaneously select content, decide ordering, and realize text, which amplifies compression priors and often triggers scope collapse. As a result, event coverage degrades and temporal consistency becomes brittle when distant evidence is mixed within a single generation pass.

We propose *TAEG-ANC* (Temporal Alignment Event Graph for Abstractive Narrative Consolidation), a neurosymbolic approach that uses TAEG as a fixed scaffold and applies neural generation only at the event level. Concretely, the model fuses passages linked to the same canonical event and the final narrative is formed by concatenating these event-level fusions in canonical order.

The main contributions of this paper are:

1) A *graph-guided abstractive formulation* of Narrative Consolidation that preserves the symbolic temporal backbone of TAEG while enabling localized abstractive fusion.
2) A *backbone-agnostic instantiation* supporting encoder–decoder summarizers and decoder-only LLMs within the same event-level pipeline.
3) An *empirical study* on a Gospel consolidation resource comparing global (end-to-end) baselines against TAEG-guided generation using BART, PEGASUS, PRIMERA, and Gemma 3.

Open Science. To support reproducibility, the code, derived annotations (event alignments and chronology mappings), and instructions to obtain the source texts are publicly available on Zenodo at https://doi.org/10.5281/zenodo.19262044.

II. Related Work

A. Multi-Document and Temporal Summarization

Neural MDS has progressed from extractive methods to large pretrained abstractive models [29]. These systems are typically optimized for conciseness and salience, which may be misaligned with NC objectives. In particular, models fine-tuned for news-style summarization can exhibit strong length priors [32] and compressive-like behavior [33] when given large heterogeneous inputs.

Timeline summarization and multi-document temporal summarization incorporate time as an organizing principle, often by selecting salient dates or clustering events to form a list or dated overview [17]–[19]. NC differs in target format: a single continuous narrative whose structure follows a canonical event order and aims at unification rather than overview [1]. In addition, the Gospel domain is largely *undated* at the sentence level, making approaches that rely on explicit timestamps less applicable.

B. Narrative Consolidation and Information Fusion

Early work on multi-document *information fusion* emphasized combining partially overlapping sources into a coherent output, rather than only selecting salient sentences [16]. NC extends this direction by making *coverage and ordering* primary objectives. Recent surveys position narrative-oriented extraction and reconstruction as a distinct family of tasks, often requiring structural constraints beyond standard summarization objectives [20].

Notably, common long-input strategies like chunking or hierarchical encoding [31] are ill-suited for NC. Such methods assume sequential documents, whereas NC operates on parallel accounts where the same event appears at different positions. Arbitrary partitioning by token count severs the event-level correspondences essential for fusion [16] and ignores the event semantics and temporal structure that undergird the narrative [30], thus motivating an explicit alignment scaffold like TAEG.

C. Neurosymbolic Grounding with Graph Structure

Neurosymbolic AI combines explicit symbolic structure with neural learning to improve reliability and interpretability [2], [3]. Beyond broad surveys, a representative line of work injects logical structure into differentiable learning via soft constraints, e.g., Logic Tensor Networks [4]. Also, graph-based summarization ranges from similarity graphs (e.g., LexRank) [14] to neural graph encoders [15] and explicit semantic/KG-guided summarization [9]. Most of these approaches use graphs primarily to enhance representations.

A complementary direction is grounding LLMs with structured knowledge sources to improve controllability [5], [6]. Recent work also uses Knowledge Graphs (KG) structures to constrain LLM reasoning, e.g., planning–retrieval–reasoning over relation paths [7] or decoding-time graph constraints [8]. While these methods mainly target Question Answering and entity–relation reasoning, they share the key idea we adopt: using an explicit graph as a scaffold to constrain generation. TAEG-ANC follows this spirit, using an explicit event graph to constrain the generation process rather than only refining embeddings.

III. Methodology: TAEG-ANC

Overview. TAEG-ANC operationalizes Narrative Consolidation as a neurosymbolic pipeline that *separates temporal planning from surface realization*. In KG–LLM terms, TAEG functions as an explicit event KG that *constrains* inference via hard structural constraints (canonical order) and evidence selection (event-level subgraph retrieval). As illustrated in Fig. 1, we first use TAEG as a fixed scaffold that (i) anchors each source passage to a canonical event and (ii) enforces a non-negotiable chronological backbone. For each canonical event e_t, the method retrieves the aligned multi-source cluster S_t (vertical column), then applies a neural generator M_θ only to this *micro-abstractive* context to produce a locally fused paragraph p_t. Finally, the consolidated narrative is assembled by concatenating these event-level fusions in canonical order, guaranteeing global chronology by construction. Algorithm 1 formalizes this procedure: it iterates over the canonical timeline T, extracts each `SAME_EVENT` cluster from G, generates p_t when evidence is available, and appends it to the running narrative N with explicit event markers for traceability and ordering checks.

A. Symbolic Scaffold: Temporal Alignment Event Graph

We adopt the TAEG formulation introduced for extractive NC [1]. Let $T = \langle e_1, \ldots, e_M \rangle$ be a canonical timeline of events, where each e_t corresponds to a canonical event with (possibly empty) references in each document. TAEG is a multi-relational graph $G = (V, E)$ where each node $v \in V$ corresponds to a *version* of a canonical event e_t in a given source document (e.g., a verse span associated with e_t). Edges represent two relations:

- Temporal edges (`BEFORE`): directed edges consistent with the temporal order in T, encoding chronology as an explicit constraint.
- Anchoral edges (`SAME_EVENT`): undirected edges linking all nodes that refer to the same canonical event, forming an event-level cluster for fusion.

In TAEG-ANC, the canonical timeline T is treated as a non-negotiable ordering constraint: the consolidated narrative is produced by iterating $t = 1 \ldots M$.

B. Graph Construction and Node Granularity

Graph instantiation is driven by an external canonical alignment resource (a Gospel harmony [25]) that provides: (i) a single chronological ordering of the Passion Week, (ii) a list of 169 canonical events, and (iii) the set of parallel spans in each of the four Gospels for every event.

In our implementation, the graph builder creates separate nodes per gospel per event (e.g., `event_12_mark`, `event_12_luke`) rather than merging all mentions into a single node. This design keeps provenance explicit at the event level: each node stores (i) the canonical event id, (ii) the document label, (iii) the reference span, and (iv) the extracted verse text used for fusion. This design supports KG-enhanced explainability: each paragraph p_t can be audited against its originating node cluster and aligned spans, enabling event-wise provenance inspection and error localization. The builder also caches the constructed graph for reuse across experimental runs, which is important for multi-backbone comparisons.

C. Event-Level Abstractive Consolidation

A key distinction of TAEG-ANC is that it partitions the input by semantic event identity rather than by fixed token length. Standard long-context strategies, relying on explicit chunking [12] or sliding-window attention [11], assume that documents are sequential and can be split into contiguous segments. However, in multi-document Narrative Consolidation, sources are parallel accounts of the same underlying events. Arbitrary chunking would mix unrelated events within a single segment and destroy the cross-document alignment that is

Algorithm 1 TAEG-guided Abstractive Narrative Consolidation (TAEG-ANC)

Input: Documents D, canonical timeline $T = \langle e_1, \ldots, e_M \rangle$, TAEG G, generator M_θ
Output: Consolidated narrative N
$N \leftarrow$ empty string
for $t \leftarrow 1$ to M **do**
  $S_t \leftarrow$ RetrieveEventEvidence(G, e_t) {texts $s_{t,d}$ from the `SAME_EVENT` cluster V_t}
  if $S_t \neq \emptyset$ **then**
    $p_t \leftarrow M_\theta(S_t)$ {event-level abstractive fusion}
    Append(N, `[Event t]` $\|$ p_t)
  end if
end for
return N

essential for faithful fusion. By contrast, TAEG retrieves all and only the passages that refer to the same canonical event, providing the neural generator with a semantically coherent and temporally focused context.

Specifically, for each event e_t:

1) Cluster retrieval: collect the texts associated with all nodes linked by `SAME_EVENT` edges for e_t: $S_t = \{s_{t,1}, \ldots, s_{t,k}\}$.
2) Fusion: generate a paragraph p_t that merges complementary details while discouraging additions not supported by S_t: $p_t \leftarrow M_\theta(S_t)$.
3) Assembly: append p_t to the output narrative in the fixed order of T.

This event-level decomposition limits the effective context size seen by the generator, and isolates the backbone's role to *local fusion* rather than *global planning*. For evaluation and traceability, each generated paragraph is prefixed with an explicit event marker (e.g., `[Event 12]`), which also enables direct ordering checks.

D. Neural Backbones and Generation Controls

We examine four commonly used architectures for generation:

- BART-Large-CNN [10]: an encoder–decoder model commonly used for abstractive summarization.
- PEGASUS [12]: a summarization model pretrained with gap-sentence generation.
- PRIMERA [11]: a multi-document summarization model with pyramid-style pretraining.
- Gemma 3 (4B) [13]: an open-weights instruction-tuned LLM (decoder-only), used in text-only mode for event-level fusion.

For the pure neural baselines, the same backbones are configured to allow longer outputs (larger maximum generation length), but remain constrained by their input tokenization limits. For Gemma 3, we use an instruction prompt that (i) asks for consolidation rather than summarization, (ii) emphasizes retaining unique details across sources, and (iii) discourages unnecessary shortening.

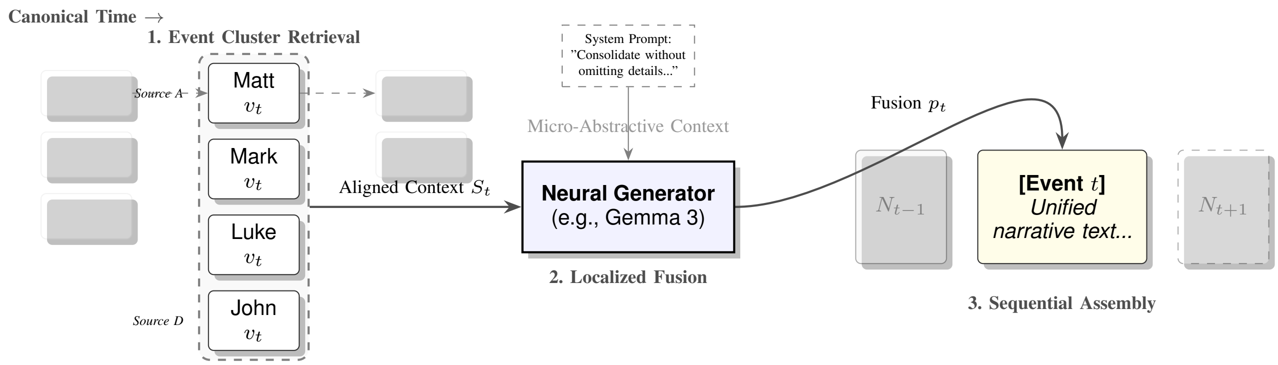

Fig. 1. **Overview of the TAEG-ANC Framework.** The process decouples temporal reasoning from generation. (1) The symbolic graph identifies a cluster of aligned event versions (vertical column) from the canonical timeline. (2) These specific passages are fed into the LLM for localized fusion. (3) The generated segment is appended to the narrative stream, guaranteeing correct chronology.

IV. Experimental Setup

A. Dataset

We use the *Gospel Consolidation Language Resource* introduced with the NC/TAEG formulation [1]. In our implementation, the evaluated corpus corresponds to the Passion Week portion of the four Gospels (Matthew, Mark, Luke, John) [26], aligned to $M = 169$ canonical events via an external chronology file [25]. A manually produced consolidated reference is used as ground truth [27], [28] for automatic evaluation.

B. Baselines

We report results for two groups of generation settings:

1) Pure neural generation: each neural, abstractive backbone is applied end-to-end globally to the distinct source documents, truncated to the backbone's maximum supported input length. We configure each model with the maximum token budgets they support, as reported in Table I. Importantly, several backbones terminate well before reaching the configured decoding budget (by emitting an end-of-sequence, EOS, token), so the very short outputs observed for some models are consistent with early termination and a strong compression prior rather than a binding output-length limit.
2) Neurosymbolic generation: the same neural backbones guided by TAEG, which provides all aligned versions of the same canonical event for event-level (micro-abstractive) fusion. The final narrative is assembled by concatenating the generated event segments in canonical order.

We also report two extractive references for comparison purposes: (i) *TAEG-LexRank*, which performs centrality-based selection within each event cluster, and (ii) a *guided extractive* upper bound (*TAEG-LexRank-TA*) that uses the same gold temporal alignment for selection without generation, providing a reference point under ideal alignment.

C. Metrics

We report metrics aligned with NC objectives:

TABLE I

Token budgets used for the *pure global* baselines. "Input limit" is the truncation length after tokenization. "Decoding budget" is the maximum number of tokens allowed to be generated; models may still stop earlier by emitting an end-of-sequence (EOS) token.

Backbone	Input limit	Decoding budget
BART-Large-CNN	1024 tokens	142 tokens
PEGASUS (cnn_dailymail)	1024 tokens	256 tokens
PRIMERA	4096 tokens	1024 tokens
Gemma 3 (4B; Ollama)	4096 tokens	until EOS

- Kendall's Tau (τ): temporal ordering agreement with the canonical timeline [23]. For TAEG-guided outputs, τ is computed directly from the sequence of emitted event markers; for unanchored baselines, a heuristic matching against the consolidated reference is used to estimate event order.
- ROUGE-1/2/L (F1): token overlap metrics capturing lexical similarity and subsequence structure [22]. ROUGE-L is particularly relevant for Narrative Consolidation as it measures the longest common subsequence, intrinsically rewarding the preservation of sequential structure and narrative flow in long-form texts.
- METEOR: lexical/semantic alignment with synonym handling [21].
- BERTScore (F1): semantic similarity via contextual embeddings [24].

V. Results and Discussion

Our experiments are designed to evaluate the central hypothesis: that a symbolic temporal scaffold is crucial for effective NC. We analyze the results from four perspectives: the failure of global generation, the structural impact of the TAEG scaffold, the behavior of different abstractive backbones, and a qualitative analysis of the generated output.

A. Objective Misalignment in End-to-End Generation

The pure end-to-end baselines, which process a single concatenated document, fundamentally fail at the NC task. As shown in Table II, these models produce outputs that are

TABLE II
PERFORMANCE COMPARISON BETWEEN PURE END-TO-END GENERATION AND TAEG-GUIDED NARRATIVE CONSOLIDATION.

Method	**Strategy**	**Length**	**R-1**	**R-2**	**R-L**	**BERTS**	**METEOR**	τ
Extractive Baselines								
TAEG-LexRank	Extractive	59,408	0.8037	0.6551	0.2065	0.8351	0.3613	0.3048
TAEG-LexRank-TA	Guided Extr.	87,528	0.9805	0.9548	0.9655	0.9863	0.7345	1.0000
Pure Neural (End-to-End)								
Pure-BART	Abstractive	525	0.0113	0.0092	0.0098	0.8240	0.0033	0.5842
Pure-PEGASUS	Abstractive	462	0.0110	0.0093	0.0099	0.8086	0.0032	-0.0454
Pure-PRIMERA	Abstractive	4,352	0.0959	0.0783	0.0678	0.8434	0.0273	0.4471
Pure-Gemma	Abstractive	3,300	0.0451	0.0167	0.0289	0.8094	0.0141	0.3919
TAEG-ANC (Graph-Guided)								
TAEG-BART	Neurosymbolic	42,220	0.6271	0.5163	0.4844	0.8874	0.2633	**1.0000**
TAEG-PEGASUS	Neurosymbolic	33,740	0.5337	0.4286	0.4074	0.8640	0.2161	**1.0000**
TAEG-PRIMERA	Neurosymbolic	83,811	0.8846	0.7085	0.6200	**0.9094**	0.4801	**1.0000**
TAEG-Gemma	**Neurosymbolic**	**101,395**	**0.8883**	**0.7168**	**0.7036**	0.8931	**0.4926**	**1.0000**

drastically shorter than the consolidated reference, reflecting a strong inductive bias towards compression inherited from their pre-training on summarization tasks. This behavior is misaligned with the completeness objective of NC [20].

More critically, these models exhibit a severe *scope collapse*. For instance, the Pure-Gemma model, when given the entire Passion Week narrative, produces a high-level thematic overview rather than a chronological account. It begins its output near the crucifixion and concludes with generic bullet points on *faith* and *forgiveness*, completely omitting the preceding 150+ canonical events. This is reflected in the extremely low temporal correlation scores (e.g., $\tau = -0.0454$ for Pure-PEGASUS), confirming that without an explicit structural guide, even powerful LLMs are unable to perform the long-range planning required for chronological narrative construction.

B. Effect of the Temporal Scaffold

In contrast, all TAEG-ANC variants preserve the canonical event order by construction, achieving a perfect Kendall's τ of 1.0. By decomposing the global planning problem into a sequence of local fusion tasks, the TAEG scaffold effectively decouples temporal structuring from surface realization. This allows the neural backbones to focus solely on their core strength: fusing complementary details from a small, pre-aligned cluster of text passages.

The impact of this decomposition is a dramatic improvement in all content-based metrics. For example, TAEG-Gemma improves the ROUGE-L score from 0.0289 (Pure-Gemma) to 0.7036. This demonstrates that the primary challenge in this NC task is not generation quality but structural integrity, which the symbolic scaffold successfully provides

C. Qualitative Analysis: Fusing Complementary Details

To make the practical difference tangible, Table III presents a case study for a single canonical event: Peter's first denial. The input consists of parallel but non-identical accounts from three Gospels. The Pure-Gemma baseline, suffering from scope collapse, fails to mention the event at all. In contrast, the TAEG-Gemma model successfully *fuses* unique details from each source—the location (courtyard), the specific person (a servant girl), and his direct verbal denial—into a single, coherent paragraph highly faithful to the reference.

TABLE III
QUALITATIVE COMPARISON FOR A SINGLE EVENT (PETER'S FIRST DENIAL)

Input Texts (Aligned by TAEG)
Matthew 26:69-70: Now Peter was sitting outside in the courtyard. And a servant girl came up to him and said, "You also were with Jesus the Galilean." But he denied it before them all, saying, "I do not know what you mean." **Mark 14:66-68:** And as Peter was below in the courtyard, one of the servant girls of the high priest came, and seeing Peter warming himself, she looked at him and said, "You also were with the Nazarene, Jesus." But he denied it... **Luke 22:56-57:** Then a servant girl, seeing him as he sat in the light and looking closely at him, said, "This man also was with him." But he denied it, saying, "Woman, I do not know him."
Pure-Gemma Output (Pure Neural/Abstractive/End-to-End)
... (The model's output discusses only the final events of the crucifixion and provides thematic reflections, completely omitting this earlier event due to scope collapse.) ...
TAEG-Gemma Output (Our Method: Abstractive Graph-Guided)
[Event 135] As Peter was sitting outside in the courtyard, warming himself in the light, a servant girl of the high priest came up to him. Looking closely, she said to him and others that he had also been with Jesus of Nazareth, the Galilean. But Peter denied it before all of them, saying, "Woman, I do not know what you mean. I do not know him."
Consolidated Reference (Ground Truth)
Now Peter was sitting outside in the courtyard. One of the high priest's servant girls came and, seeing Peter warming himself in the firelight, looked closely at him and said, "You also were with Jesus the Galilean, the Nazarene. This man was with him." But he denied it before them all, saying, "Woman, I do not know him or what you are talking about."

D. Backbone Behavior and Metric Interpretation

1) Domain-Specific Adaptation: The PEGASUS Case: Careful checkpoint selection is critical for domain-specific consolidation tasks. PEGASUS provides an instructive example. While the `pegasus-multi_news` variant is theoretically suited for MDS, our preliminary experiments revealed severe domain drift: it excessively hallucinated external news entities (e.g., "The New York Times", "Reuters") when applied to Biblical text. These entities are entirely absent from the source material, representing a form of domain drift where the model's pre-training on news corpora led it to "invent" contextual details inappropriate for the Gospel domain.

To mitigate this, we adopted the `pegasus-cnn_dailymail` variant. Despite its narrower single-document pre-training corpus, its strict source fidelity proved more robust for our event-level fusion task, allowing TAEG-PEGASUS to achieve a ROUGE-L of 0.4074. This underscores a key principle for neurosymbolic NC systems: pre-trained models must be chosen based on domain characteristics and pre-training objectives, not just task similarity.

2) Comparative Performance and Trade-offs: Among the tested backbones, TAEG-Gemma and TAEG-PRIMERA emerge as the top performers. Notably, they excel in different metrics. TAEG-Gemma achieves the highest ROUGE-L (0.7036), suggesting it is superior at preserving n-gram structure and generating lexically similar sentences to the reference, which is crucial for narrative flow. Conversely, TAEG-PRIMERA obtains the highest BERTScore (0.9094), indicating it may be better at capturing semantic meaning even if the surface-level wording differs. This trade-off between lexical fidelity and semantic similarity warrants further investigation, though a formal statistical significance analysis was beyond the scope of this initial study.

A critical observation is that while the scaffold guarantees *global* chronological order, it does not guarantee *local* factuality. We observed instances where backbones, particularly PEGASUS, introduced minor, out-of-domain hallucinations within an otherwise correct event paragraph (e.g., mentioning a "hotel" or "market" setting, which is absent from the source). This highlights a key limitation and an important direction for future work: temporal scaffolds must be complemented by stricter grounding or verification mechanisms to ensure full narrative integrity.

VI. Limitations and Future Work

While our neurosymbolic approach shows significant promise, this study has several limitations that define a clear path for future research.

Generalization and Scalability. While our study relies on a single, small-scale dataset from a niche domain (169 events from the Gospels), it nevertheless captures many of the core challenges inherent to Narrative Consolidation, including event identity, temporal ordering, and cross-source alignment under partial and divergent accounts. To establish the viability of TAEG-ANC as a general framework, future work must validate its performance on larger and more diverse datasets. Potential domains include consolidating multi-source news reports of a developing event, generating unified biographies from different historical accounts, or harmonizing witness testimonies in legal contexts. Such studies would also need to assess the framework's scalability as the number of canonical events grows by orders of magnitude.

Dependence on Canonical Alignments. Our method assumes an a priori canonical timeline and event alignments to isolate the impact of the temporal scaffold. Automatically constructing such perfect graphs from unstructured text introduces significant engineering challenges and computational costs, requiring robust pipelines for cross-document coreference resolution and temporal relation extraction. However, unlike the ancient texts used in our study, modern documents (e.g., news reports, clinical notes) typically feature explicit timestamps and metadata, which can substantially mitigate the cost of automated temporal alignment. While relying on pre-aligned graphs is practical in domains with human-curated timelines (e.g., legal chronologies), a crucial next step is integrating upstream methods for automatic event discovery to build a fully end-to-end system, carefully balancing graph construction costs against narrative quality.

Controlling Verbosity and Factual Grounding. As our results show, the scaffold successfully encourages completeness, but this often leads to outputs that are significantly more verbose than the human-written reference. Future work should explore mechanisms to control verbosity, perhaps by learning to distinguish between complementary details and redundant phrasing. Furthermore, as noted in our analysis, the framework does not prevent local, event-level hallucinations. Integrating the generator with retrieval-based verification, fact-checking modules, or learned confidence scores to gate generation are essential next steps to improve the factual grounding and reliability of the consolidated narrative.

Evaluation Beyond Lexical Overlap. Our evaluation relies on automatic metrics that primarily measure lexical and semantic overlap. These metrics do not fully capture the narrative qualities of cohesion, coherence, and readability. Future work should incorporate human evaluation to assess these higher-level attributes. Additionally, developing new automatic metrics that can evaluate the quality of transitions between events, even when the global order is fixed, would provide a more nuanced understanding of the generated narrative's quality.

VII. Conclusion

This work advances Narrative Consolidation as a neurosymbolic generation problem: rather than relying on purely neural end-to-end decoding to jointly perform planning and realization, we explicitly *represent* and *enforce* the key structural requirements of NC. TAEG-ANC couples a symbolic Temporal Alignment Event Graph—encoding canonical chronology and cross-document event identity—with a neural generator used strictly as a local fusion module. By grounding generation in an explicit scaffold and decomposing consolidation into event-level subproblems, the framework delivers controllability and interpretability: chronology is guaranteed by construction, and event-wise provenance remains explicit through graph-defined alignments.

Empirically, on a Gospel consolidation resource covering 169 Passion Week events, TAEG-ANC consistently outperforms global end-to-end baselines that suffer from compression bias and scope collapse, while preserving strict temporal integrity. Among the tested backbones, the Gemma 3 instantiation achieves the best ROUGE-L and METEOR, indicating that decoder-only LLMs can serve as effective surface realizers when constrained by a symbolic temporal backbone.

Our results also highlight a fundamental limitation of applying generic long-context strategies to multi-document narrative tasks: because sources are parallel rather than sequential, partitioning by token length destroys the event-level alignment that is critical for faithful consolidation. TAEG addresses this by providing a semantically grounded decomposition that keeps related evidence together, regardless of its position in the original documents.

These results position TAEG-ANC as a concrete instance of KG-enhanced inference for LLMs, where a structured event knowledge graph provides hard constraints (canonical order and event identity) and explicit provenance for generation. Moreover, the explicit KG scaffold supports event-level explainability and a natural substrate for validation: each generated paragraph can be traced back to a specific aligned evidence cluster in the graph, offering an auditable alternative to opaque end-to-end long-context decoding.

More broadly, our results suggest that the primary bottleneck in long-form consolidation is not fluency but *structural reliability*. Explicit symbolic scaffolds provide a practical route to neurosymbolic NC systems that are modular, auditable, and robust under long contexts. Future work should integrate automatic event discovery and alignment and further improve local factual grounding, while preserving the core neurosymbolic principle demonstrated here: separating symbolic temporal structure from neural realization to obtain controllable and interpretable consolidation.

Acknowledgment

The authors thank the anonymous reviewers for their constructive feedback.

This work was partially supported by EMBRAPII, AKCIT, CNPq (grant 313845/2023-9) and Positivo Tecnologia S/A.

Generative AI Declaration: The authors acknowledge the use of Google Gemini (Model 3 Pro) to assist in the revision, editing, and linguistic polishing of the text. The authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

References

[1] R. A. Finger, E. G. Cortes, S. J. Rigo, and G. de O. Ramos, "Narrative consolidation: Formulating a new task for unifying multi-perspective accounts," arXiv:2512.18041, 2025. [Online]. Available: https://arxiv.org/abs/2512.18041.

[2] A. d. Garcez and L. C. Lamb, "Neurosymbolic AI: The 3rd wave," *Artificial Intelligence Review*, vol. 56, no. 11, pp. 12387–12406, 2023.

[3] M. K. Sarker, L. Zhou, A. Eberhart, and P. Hitzler, "Neuro-symbolic artificial intelligence: Current trends," arXiv:2105.05330, 2021.

[4] L. Serafini and A. d. Garcez, "Logic tensor networks: Deep learning and logical reasoning from data and knowledge," in *Proc. NeSy*, 2016.

[5] S. Pan *et al.*, "Unifying large language models and knowledge graphs: A roadmap," *IEEE Trans. Knowl. Data Eng.*, vol. 36, no. 7, pp. 3580–3599, 2024.

[6] J. Z. Pan *et al.*, "Large language models and knowledge graphs: Opportunities and challenges," *Trans. Graph Data and Knowledge*, vol. 1, no. 1, pp. 2:1–2:38, 2023.

[7] L. Luo, Y.-F. Li, G. Haffari, and S. Pan, "Reasoning on graphs: Faithful and interpretable large language model reasoning," in *Proc. ICLR*, 2024.

[8] L. Luo *et al.*, "Graph-constrained reasoning: Faithful reasoning on knowledge graphs with large language models," in *Proc. ICML*, pp. 41540–41565, 2025.

[9] X. Ji and W. Zhao, "SKGSUM: Abstractive document summarization with semantic knowledge graphs," in *Proc. IJCNN*, 2021, pp. 1–8.

[10] M. Lewis *et al.*, "BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," in *Proc. ACL*, 2020, pp. 7871–7880.

[11] W. Xiao, I. Beltagy, G. Carenini, and A. Cohan, "PRIMERA: Pyramid-based masked sentence pre-training for multi-document summarization," in *Proc. ACL*, 2022, pp. 5245–5263.

[12] J. Zhang, Y. Zhao, M. Saleh, and P. J. Liu, "PEGASUS: Pre-training with extracted gap-sentences for abstractive summarization," in *Proc. ICML*, pp. 11328–11339, 2020.

[13] Gemma Team, Google DeepMind, "Gemma 3 technical report," arXiv:2503.19786, 2025.

[14] G. Erkan and D. R. Radev, "LexRank: Graph-based lexical centrality as salience in text summarization," *J. Artif. Intell. Res.*, vol. 22, pp. 457–479, 2004.

[15] M. Yasunaga *et al.*, "Graph-based neural multi-document summarization," in *Proc. CoNLL*, 2017, pp. 452–462.

[16] R. Barzilay, K. R. McKeown, and M. Elhadad, "Information fusion in the context of multi-document summarization," in *Proc. ACL*, 1999, pp. 550–557.

[17] Y. Yu, A. Jatowt, A. Doucet, K. Sugiyama, and M. Yoshikawa, "Multi-TimeLine Summarization (MTLS): Improving timeline summarization by generating multiple summaries," in *Proc. ACL-IJCNLP*, 2021, pp. 377–387.

[18] S. Martschat and K. Markert, "Improving ROUGE for timeline summarization," in *Proc. EACL*, 2017, pp. 285–290.

[19] K. K. Mamidala and S. K. Sanampudi, "A novel framework for multi-document temporal summarization (MDTS)," *Emerging Science Journal*, vol. 5, no. 2, pp. 184–190, 2021.

[20] B. Santana *et al.*, "A survey on narrative extraction from textual data," *Artificial Intelligence Review*, vol. 56, no. 8, pp. 8393–8435, 2023.

[21] S. Banerjee and A. Lavie, "METEOR: An automatic metric for MT evaluation with improved correlation with human judgments," in *Proc. ACL Workshop*, 2005, pp. 65–72.

[22] C.-Y. Lin, "ROUGE: A package for automatic evaluation of summaries," in *Text Summarization Branches Out*, 2004, pp. 74–81.

[23] M. G. Kendall, "A new measure of rank correlation," *Biometrika*, vol. 30, pp. 81–93, 1938.

[24] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "BERTScore: Evaluating text generation with BERT," in *Proc. ICLR*, 2020.

[25] R. Aschmann, "Chronology of the four Gospels (harmony of the Gospels)," 2022. [Online]. Available: https://aschmann.net/BibleChronology/ChronologyOfTheFourGospels.pdf.

[26] *The Holy Bible, New International Version*, Biblica, 2011.

[27] A. F. C. Cunha, "Semana da Paixão unificada: Uma harmonização moderna dos Evangelhos," in *Anais do IV Cong. Bras. de Humanismo Solidário na Ciência*, 2025.

[28] A. F. C. Cunha, C. A. Sena, "Sumarização Multi Documentos Agentizada com Ancoragem e Consolidação de Eventos Paralelos: um Experimento com os Quatro Evangelhos," *[E-book, 39p] UFG*, 2026.

[29] C. Ma, W. E. Zhang, M. Guo, H. Wang, and Q. Z. Sheng, "Multi-document summarization via deep learning techniques: A survey," *ACM Computing Surveys*, vol. 55, no. 5, pp. 1–37, 2023.

[30] G. Glavaš and J. Šnajder, "Event graphs for information retrieval and multi-document summarization," *Expert Syst. Appl.*, vol. 41, no. 15, pp. 6904–6916, 2014.

[31] Q. Grail, J. Perez, and E. Gaussier, "Globalizing BERT-based transformer architectures for long document summarization," in *Proc. EACL*, 2021, pp. 1792–1810.

[32] W. Kryscinski, N. S. Keskar, B. McCann, C. Xiong, and R. Socher, "Neural text summarization: A critical evaluation," in *Proc. EMNLP-IJCNLP*, 2019, pp. 540–551.

[33] S. Narayan, S. B. Cohen, and M. Lapata, "Don't give me the details, just the summary! Topic-aware convolutional neural networks for extreme summarization," in *Proc. EMNLP*, 2018, pp. 1797–1807.