ChronoLearn: A GRAG LLM-Based System for Structuring and Exploring Historical Narratives

Mohammad O. ALADDASI, Shahd Abu Hijleh, and Omar Qawasmeh
Princess Sumaya University for Technology
Department of Data Science & Artificial Intelligence, Amman,Jordan
mohammadoaddasi@gmail.com, sha20210504@std.psut.edu.jo, o.alqawasmeh@psut.edu.jo

Abstract

ChronoLearn is a KG–LLM framework to structure and explore user-fed Arabic historical narratives. Traditional history learning methods often rely on fragmented, text-heavy resources that lack contextualisation, interactivity, and clear relational structure. Chronolearn transforms unstructured texts into knowledge graphs KG using an ETL-based Natural Language Processing (NLP) pipeline for entity and relation extraction, followed by schema-guided graph construction. The system integrates KG retrieval with Large Language Model (LLM) generation (GRAG) to produce grounded and explainable narratives. The approach is evaluated in heterogeneous Palestinian and Jordanian sources, including Nakba-related content, using both quantitative metrics and comparative analysis. Additionally, an interactive platform is provided to support the exploration of structured historical knowledge through graph visualisations and narrative storylines, enabling users to explore historical events in a coherent, interactive and interpretable manner. The results demonstrate improved factual grounding and structured reasoning, addressing limitations of text-only approaches in the processing of historical knowledge in Arabic Language

Keywords: Natural Language Processing, Knowledge Graphs, LLMs

1. Introduction

Recent advances in natural language processing (NLP) have brought the importance of treating historical and cultural narratives into attention particularly those related to the Nakba. It facilitated historical sources to be analysed and explored computationally in structured language form. The Nakba, a major historical event with an ongoing impact, meaning catastrophe in Arabic, "refers to the eviction of nearly three-quarters of a million Palestinians from their homes and their transformation into refugees, as well as the destruction of more than four hundred villages and towns in what became the state of Israel, and the erasure of the name Palestine from the map" Interactive Encyclopedia of the Palestine Question (PalQuest) (2022). Historical knowledge consists of interconnected components, including events (e.g., wars and treaties), people and groups (leaders and communities), places (cities and historical sites), and sources (archives, letters, and cultural materials). These elements form a complex relational structure rather than isolated facts.

1.1. Project Background and Motivation

History by its nature is a time-oriented subject with chronology being a major part of its overall structure. It is also a data-rich field, with a number of interconnected sources and references that connect events, people, and places together. However, the teaching of history rarely reflects this, as learning experiences lack strong storytelling, visualisation, and interactivity. This results in a disjointed history experience. Modern learners increasingly expect educational tools to be visual, interactive, and efficient to explore. Despite this, history as a discipline has been slower to adopt technologies such as data visualisation, semantic analysis, and knowledge graphs (KGs).

While artificial intelligence (AI) has introduced new paradigms in learning and information access, history education continues to face a significant technological gap. Interactive teaching methods have been shown to improve engagement; however, most historical content remains presented in static formats. A key trend is the use of Large Language Models (LLMs) through question-answering (Q&A) interfaces. Nevertheless, these systems present several limitations, including lack of source attribution and explainability, susceptibility to hallucinations or inaccurate information, and an over-reliance on text-based interaction. As a result, they provide limited support for structured and visual exploration of historical knowledge.

In this paper, ChronoLearn is proposed as a system to transform user-fed Nakba-related narratives into structured KGs using NLP techniques. A Graph Retrieval-Augmented Generation (GRAG)-based framework is introduced to integrate KGs with LLMs for grounded querying and narrative generation. The system leverages diverse Arabic historical sources to model event-based, cultural, and memory-driven narratives. Furthermore, an interactive platform is provided to support the exploration of structured historical knowledge through visualisation and AI-driven storytelling.

Proceedings of the 2nd International Workshop on Nakba Narratives as Language Resources (Nakba-NLP 2026) @ LREC 2026, pages 43–49

1.2. Problem Statement and Research Objectives

Digital historical information is distributed across multiple sources and formats, making it difficult for users to access structured, reliable, and meaningful knowledge. While the results obtained from traditional search engines are generally broad and vague, the results of new chat-based AI tools are unverified, decontextualised, and carry bias.

The problem is more critical for the Arabic historical narratives, especially those concerning the Palestinian and Jordanian contexts, since the resources for the NLP tools are scarce.

In this research, we seek to answer the research question "How can the integration of KGs with LLMs facilitate the structured representation of Nakba-related historical narratives, compared to traditional approaches?"

2. Related Work

The integration of KGs and LLMs has become an important direction in recent research. Researchers use this combination to build AI systems that are more explainable and interactive. In knowledge-intensive domains such as history, structured representations help organise information and support clear reasoning. KGs represent entities and relationships in a structured form, but LLMs generate free-text responses. KGS and LLMS combination creates systems that connect structured knowledge with semantic understanding.

This section reviews existing research across three main areas: (1) unstructured text-to-KG conversion, (2) KG-enhanced LLM reasoning, and (3) NLP applications for historical and cultural narratives.

2.1. Text-to-Knowledge Graph Conversion

Transforming unstructured text into structured KGs is a key step in building explainable and intelligent systems. This process uses NLP techniques such as Named Entity Recognition (NER) and Relation Extraction (RE). These methods identify important entities in the text and define the relationships between them. Zhang and Soh proposed the Extract-Define-Canonicalise (EDC) framework. Their approach converts textual data into structured KGs through a three-stage pipeline. The stages include entity and relation extraction, schema alignment, and canonicalisation Zhang and Soh (2024). The framework improves scalability and consistency by using retrieval-based schema linking. This allows efficient processing of large and diverse datasets. Similarly, Mohanty introduced EduEmbedd, a framework designed for representing educational content Mohanty (2023). The model captures conceptual entities and pedagogical dependencies. It highlights the value of modelling instructional relationships in addition to factual knowledge.

Wang et al. proposed ChatWeaver, an interactive system that combines human validation (human-in-the-loop) with LLM-based extraction Wang et al. (2025). This design increases reliability by allowing users to review and refine extracted entities and relationships. This level of validation is especially important in history, where contextual accuracy plays a central role. Sun et al. presented a framework that converts historical archives into structured KGs using LLM-based extraction and embedding-based retrieval Sun et al. (2024). Their method includes preprocessing, semantic embedding with FAISS, and LLM-driven relation extraction. The result is a structured and queryable historical knowledge system.

2.2. Knowledge Graph Integration with LLMs

Integrating KGs into LLM pipelines has shown clear improvements in accuracy and explainability. When responses are grounded in structured knowledge, the system produces more reliable and context-aware outputs. This approach also makes the reasoning process easier to trace and understand.

Zhang et al. introduced KnowGPT, a framework that uses KG–based prompting to strengthen reasoning in LLMs Zhang et al. (2024). Their experiments showed better performance compared to baseline models such as GPT-3.5 and GPT-4 across several evaluation benchmarks.

Recent studies in applied domains further highlight the value of combining KGs with retrieval-augmented generation (GRAG). For example, MedSyn integrates KGs with RAG to support question answering over structured medical data. The system achieved stronger factual grounding and reduced hallucination. Akgül et al. explored temporal reasoning over dynamic KGs by incorporating time-series data into graph structures and applying contrastive learning techniques for entity representation Akgül et al. (2025). Their results showed clear improvements in prediction accuracy. However, the framework assumes complete data availability, which may limit its applicability in historical real-world applications.

Yan et al. proposed a KG–guided framework for explainable AI Yan et al. (2025). In their approach, graph relationships validate LLM-generated outputs and assign confidence levels to responses. This design strengthens interpretability and transparency. The evaluation relied mainly on qualitative

feedback, which suggests that future work can include more detailed quantitative validation.

2.3. NLP for Historical and Cultural Narratives

NLP applications in historical and cultural domains focus on structuring and analysing narrative-based data. These approaches support organised representation, interpretation, and exploration of historical content. They also help to create systems that encourage storytelling and interactive learning.

Muralidharan et al. proposed a system that generates multiple-choice questions (MCQs) using KG representations Muralidharan (2024). Their evaluation showed that question generation based on semantic relationships improves learning outcomes and strengthens knowledge retention. This study highlights how structured knowledge can enhance educational applications.

Research Gaps: In summary, after completing the literature review, the following research gaps are identified: First, many systems provide limited support for the Arabic language and regional historical content. Second, current approaches offer few fully automated and scalable pipelines for KG construction. Third, many frameworks still rely on human validation when handling complex or ambiguous information. Fourth, interactive and visual learning tools receive limited attention. Moreover, there is little work focused on transforming Nakba-related narratives into structured and interactive knowledge systems that support computational analysis and exploration. These gaps motivate the development of ChronoLearn. The framework aims to provide an Arabic-focused, interactive, and explainable KG-LLM pipeline that structures and explores Nakba-related narratives in a clear and accessible way.

3. Dataset and Narrative Representation

The dataset is curated to represent diverse forms of Nakba-related narratives. It includes educational texts, cultural heritage descriptions, and scholarly discussions on memory and displacement. Nakba narratives are defined in a broad sense. They cover event-based accounts, cultural artefacts, and memory and post-memory texts. Together, these materials contribute to shaping historical understanding.

The system is evaluated using a heterogeneous dataset that includes three main sources. The first source is an **oral history book**[1], which provides event-based historical content such as treaties, wars, and political developments based on first-hand testimonies Zayed (2013). The second source is an **encyclopedic article**[2], which represents a semi-structured and community-generated text with descriptive cultural content Wikipedia contributors (2026). The third source is a **scholarly study**[3], which presents analytical and narrative-driven perspectives on memory, identity, and lived experience Abu-Lughod (2007).

These sources differ in structure, style, and narrative type. This combination allows evaluation across different forms of historical data. The dataset covers varying levels of structure, from formal historical documentation to narrative and memory-focused texts. In addtion, the dataset reflects three main narrative forms. The first form includes event-based narratives that present structured historical facts such as wars, treaties, and political developments. The second, includes cultural narratives that capture symbolic and identity-based content drawn from songs and folklore. The third narrative includes memory-based, which express lived experiences and inter-generational memory, often found in academic works. This diversity supports the representation of historical knowledge as both factual and interpretive. Such representation is important for modelling different types of narratives.

Processing Arabic historical narratives involves several challenges. Arabic language contains rich structure, varied writing styles, and ambiguity in entity representation. Narrative texts often include implicit relationships and contextual references. These elements make automatic extraction more complex. Nakba-related narratives also carry cultural sensitivity and strong contextual meaning. Much of the meaning appears through memory and interpretation rather than direct factual statements. These challenges support the use of structured knowledge representations that preserve accuracy while maintaining context.

4. Proposed System: ChronoLearn

ChronoLearn is a KG–LLM framework designed to convert unstructured Jordanian and Palestinian historical narratives, especially Nakba-related content, into structured, visual, and interactive knowledge representations. Using a GRAG model, the system combines KGs with LLMs. This integration supports grounded querying, explainable narrative generation, and semantic exploration of historical knowledge. ChronoLearn follows a modular and end-to-end pipeline. Firstly, it ingests user-provided

[1] *Palestinian Refugee Narratives: An Inter-generational Comparison*

[2] *Jafra (song) Arabic Wikipedia Entry*.

[3] *Return to Half-Ruins: Memory, Post-memory, and Living History in Palestine*.

documents. Secondly, it preprocesses and normalises Arabic text. Thirdly, it extracts knowledge triples that reflect key topics and themes. Fourthly, it validates and refines the extracted information. Fifth, it constructs and merges KGs. Finally, it generates multiple outputs, including KG visualisations, RDF exports, and grounded narrative responses.

4.1. System Architecture and KG Construction Pipeline

ChronoLearn follows an ETL-inspired workflow. In the *Extract* phase, documents such as PDFs, webpages, and textual archives are ingested using automated extraction tools, after which the raw content is converted into plain text. The preprocessing stage normalises Arabic orthographic variations, including standardisation of characters (e.g., Alef, Yaa', Ta Marbutah), removal of Tatweel, and optional removal of diacritics (Harakat), improving consistency and downstream NLP performance.

In the *Transform* phase, the system applies LLM-based techniques for semantic analysis. It begins with topic detection, which identifies the main concepts in the document. These topics guide both knowledge extraction and narrative generation by narrowing the semantic focus. The system also performs theme detection. It classifies the narrative type, such as event-based, cultural, or memory-driven. This classification determines schema constraints and shapes the next processing steps.

After semantic analysis, the system extracts knowledge by identifying entities and relationships using LLM-based NER and RE. It structures the extracted information into triples (subject, predicate, object). These triples align with a schema-driven T-Box to maintain consistency in entity types and relationships.

The system then applies a validation and refinement stage to improve reliability. This stage checks schema conformity, removes redundant entries, normalises entity names, and corrects structural inconsistencies in triples. As a result, the KG becomes coherent, accurate, and suitable for further reasoning tasks. During the Load phase, the validated triples form a KG. Entities appear as nodes, and relationships appear as edges. When processing multiple documents, the system merges individual graphs into a unified global KG. It preserves source-level provenance and supports cross-document reasoning and exploration.

On top of the KG, ChronoLearn integrates a GRAG layer. This layer combines graph-based retrieval with LLM-based generation. The system supports two narrative generation strategies. The first strategy follows a topic-guided approach that generates narratives from preprocessed text based on detected topics. The second strategy uses a KG-grounded approach that generates narratives directly from validated triples. This design balances semantic richness with structured factual grounding.

The system also includes an evaluation layer. It assesses generated narratives using both quantitative metrics and qualitative analysis. The evaluation examines semantic similarity, lexical overlap, factual grounding, structural clarity, and completeness. This process enables systematic comparison between generation strategies and ensures reliable outputs.

Finally, ChronoLearn provides an output and interaction layer. It includes KG visualisation for intuitive exploration, AI-generated narratives for storytelling, and RDF export for compatibility with semantic web technologies. Users can explore historical knowledge in a structured, explainable, and visually enriched way. As illustrated in 1, the pipeline moves from ingestion and preprocessing to triple extraction, validation, KG construction, and export. The implementation is available in the project repository[4] upon request, along with all related figures and supplementary materials.

5. Evaluation and Experiments

This section presents the evaluation strategy used to assess system correctness, efficiency, and narrative quality. The evaluation combines quantitative metrics, qualitative analysis, and system-level validation using heterogeneous Arabic historical sources spanning educational, cultural, and scholarly texts. For controlled comparison, each document passes through two processing pipelines:**(i) Method A (LLM + Topics)**; generates narratives guided by extracted themes. **(ii) Method B (KG-based)**; produces outputs grounded in validated KG triples. Both methods use the same inputs to ensure a fair comparison.

The evaluation focuses on three main areas. The first area is data integrity, which includes schema conformity and relational consistency. The second area is performance, which covers latency, scalability, and processing efficiency. The third area is overall system reliability across all integrated components.

The strategy uses multiple semantic, lexical, and structural metrics. These include *Precision@K*, *Recall@K*, and *HITS@K* to measure entity-level accuracy and coverage. *ROUGE-L* evaluates lexical overlap, and *BERTScore* measures semantic similarity. Together, these metrics assess narrative quality from different perspectives and balance factual grounding with semantic alignment.

[4] https://github.com/Mohammad-ALADDASI/ChronoLearn

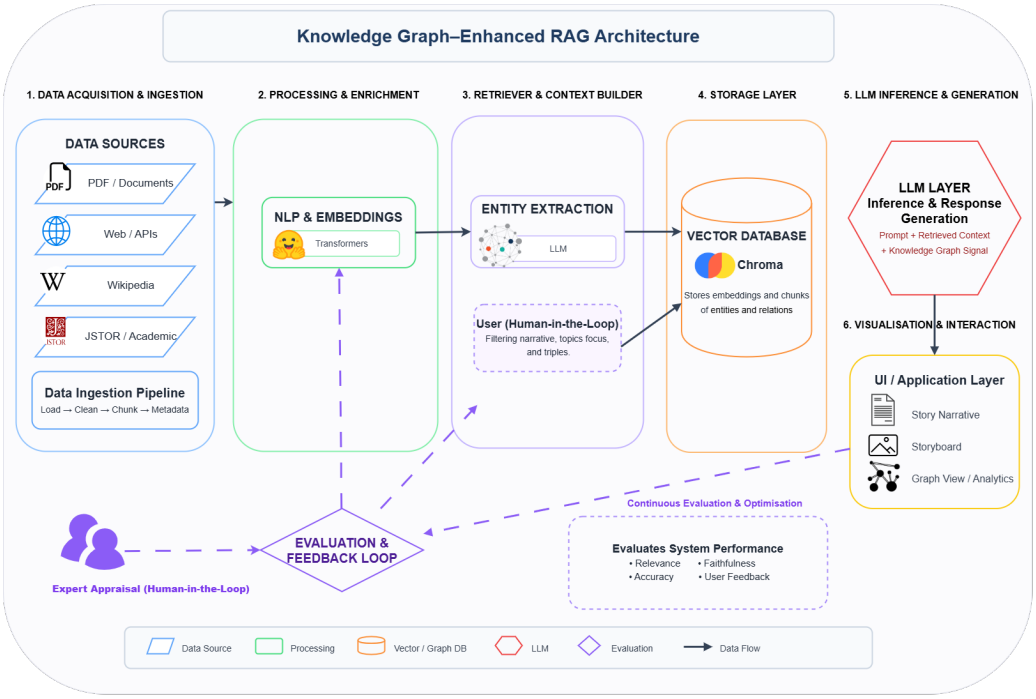

Figure 1: Semantic Knowledge Graph Processing Framework (ChronoLearn).

The evaluation also includes a direct comparison between topic-guided and KG-grounded generation. The system evaluates outputs quantitatively against the source text and qualitatively using an LLM-based evaluator. The qualitative assessment follows four criteria: *topic focus*, *structural clarity*, *factual grounding*, and *completeness*. This comparison highlights the trade-offs between narrative richness and structured accuracy.

User-based and experts' evaluation further complements automated metrics. Participants interact with generated narratives and KG visualisations. They provide feedback on clarity, coherence, factual reliability, and usability. Manual validation is conducted on selected triples and narratives. The team addresses identified issues through iterative refinement of extraction prompts, validation rules, and generation strategies.

5.1. Results and Analysis

Table 1 presents the quantitative comparison between the topic-guided generation approach (Method A) and the KG-based triple grounding approach (Method B).

The results show a consistent and interpretable trade-off between semantic richness and structured factual grounding. Method A achieves higher BERTScore (0.703 vs. 0.691). This indicates stronger contextual and semantic similarity to the reference texts. This suggests that full-text topic conditioning enables the model to generate better outputs that preserve global thematic alignment

Metric	Method A	Method B
BERTScore	0.703	0.691
ROUGE-L	0.097	0.033
Precision@10	0.60	0.80
Precision@5	0.60	0.80
Recall@5	0.130	0.174
HITS@5	1.00	1.00

Table 1: Comparison between Method A (LLM + Topics) and Method B (KG + Triples) generation

coherence. In contrast, Method B outperforms Method A in entity-level precision. It achieves higher Precision@10 (0.80 vs. 0.60) and Precision@5 (0.80 vs. 0.60). This indicates more accurate prioritisation of entities in top-ranked outputs. This reflects the benefits of structured triple grounding in enforcing factual consistency and noise reduction in entity selection.

Method A achieves a higher ROUGE-L score (0.097 vs. 0.033). This indicates stronger lexical overlap and better span-level alignment with the reference content. It suggests that topic-guided generation better preserves surface-level phrasing and narrative continuity, due to the richer contextual input provided to the model.

Method B also demonstrates improved coverage, as reflected in higher Recall@5 (0.174 vs. 0.130). This shows that the triple-based approach retrieves a broader set of relevant entities within the top-K candidates, and overall improving coverage com-

pared to Method A.

Both methods achieve perfect HITS@5 (1.00), confirming that at least one relevant entity consistently appears within the top-ranked outputs. Nevertheless, the relatively low Recall@5 values (0.130 vs. 0.174) suggest that coverage remains limited overall. This indicates that both approaches still miss a significant portion of relevant entities despite improvements.

The pipeline currently reports `linked_named_events: 0`. This component is still under active development and is not yet incorporated as a constraint during generation or evaluation. Importantly, triple validation is conducted independently of event linking. Future integration of event-aware grounding is expected to improve temporal coherence and cross-sentence consistency.

From a qualitative perspective, Method A generates more fluid, coherent, and narratively complete texts, showing smoother transitions and contextual bridging. Method B, on the other hand, produces structurally clearer outputs with stronger factual grounded and fewer theoretical inferences. The comparison highlights a fundamental trade-off between narrative richness and structured accuracy.

Overall, these findings suggest that neither approach dominates across all dimensions. Instead, they expose complementary strengths. The topic-guided generation enhances cohesion and coherence with semantic alignment. Whereas KG-based grounding improves factual precision and entity coverage. These mutual techniques strongly motivates hybrid architectures that combine high-level topic conditioning with structured KG constraints. This enables models to achieve both semantic depth and factual robustness simultaneously.

6. Conclusion

ChronoLearn shows how integrating KGs with LLMs can support structured, interactive, and explainable history learning. By converting unstructured Arabic historical narratives into machine-interpretable knowledge representations, the system supports semantic exploration and grounded narrative generation.

The results indicate that combining graph-based representations with LLMs improves knowledge organisation and strengthens explainability. It also increases user engagement, especially in Arabic historical contexts where structured digital resources remain limited. The comparative evaluation highlights the balance between narrative richness and factual grounding. The findings emphasise the benefit of integrating both approaches within a unified framework.

Future work will extend the system with storyboard generation and advanced narrative construction features to create richer educational experiences. Additional improvements will focus on temporal reasoning, named-event linking, and structured event modelling. The framework will also incorporate human-in-the-loop validation to maintain cultural sensitivity, historical accuracy, and robustness across diverse narrative types.

7. Bibliographical References

Lila Abu-Lughod. 2007. Return to half-ruins: Memory, postmemory, and living history in palestine. In Ahmad H. Sa'di and Lila Abu-Lughod, editors, *Nakba: Palestine, 1948, and the Claims of Memory*, pages 77–104. Columbia University Press, New York, NY, USA.

Ö.F. Akgül et al. 2025. Recipe-tkg: Temporal knowledge graph completion using llms. *arXiv preprint arXiv:2505.17794*.

Interactive Encyclopedia of the Palestine Question (PalQuest). 2022. Meanings of the nakba: Transformations through time and place. Accessed: 2026-02-27.

P. Mohanty. 2023. Eduembedd: An educational knowledge graph embedding framework. In *CEUR Workshop Proceedings*.

D. Muralidharan. 2024. Knowledge graph-based multiple-choice question generation. *arXiv preprint*.

Y. Sun, W. Yang, and Y. Liu. 2024. Knowledge graph construction of oral historical archives using llm-rag. In *ICISDM*.

M. Wang, Y. Gao, M. Fang, and X. Lyu. 2025. Chatweaver: An interactive llm-supported knowledge graph system. In *IEEE AINIT*.

Wikipedia contributors. 2026. Jafra (song). `https://ar.wikipedia.org/wiki/ØňÙĄØśØğ_(ØčØžÙĘÙŁØI)`. Wikipedia, The Free Encyclopedia, Accessed: 27-Feb-2026.

Y. Yan et al. 2025. Knownet: Guided health information seeking via knowledge graph integration. *IEEE TVCG*.

Amal Zayed. 2013. *Palestinian Refugee Narratives: An Inter-generational Comparison*. Ibrahim Abu-Lughod Institute of International Studies, Birzeit University, Birzeit, Palestine.

Q. Zhang, J. Dong, H. Chen, et al. 2024. Knowgpt: Knowledge graph based prompting for large language models. *NeurIPS*.

Z. Zhang and H. Soh. 2024. Extract, define, canonicalize: Towards efficient large-scale knowledge graph construction with llms. *arXiv preprint arXiv:2401.03868*.