國立臺灣大學文學院語言學研究所

碩士論文

Graduate Institute of Linguistics

College of Liberal Arts

National Taiwan University

Master's Thesis

結構化與非結構化輸入於大型語言模型中之比較：以異常事件預測任務中的語意與語用預測能力評估為例

Structured vs. Unstructured Inputs in LLMs: Evaluating the Semantic and Pragmatic Predictive Power in Abnormal Event Forecasting

紀柔安

Jou-An Chi

指導教授：謝舒凱 博士

Advisor: Shu-Kai Hsieh, Ph.D.

中華民國 114 年 7 月

July, 2025

doi:10.6342/NTU202504087

致謝

寫碩論就像是一面照妖鏡，這一年來，它照見了我許多脆弱的時刻，但我終究一步步走過來了。此刻，我想好好感謝這一路上陪伴我的每一個人。

首先，謝謝口試委員張瑜芸老師與謝吉隆老師在百忙之中撥冗參與我的口試，尤其是張老師當天晚上還要趕飛機前往捷克，真的讓我感受到滿滿的支持與情義。謝謝我的指導教授謝舒凱老師，三年來的悉心指導與鼓勵，老師每次在 meeting 中閃閃發光地分享新計畫與研究願景，總讓我重新燃起熱忱。我一直記得老師對我說過：「下定決心就是成功的一半」，雖然我還在學習什麼是全然的決心，但這句話會一直陪我走下去，提醒我勇敢前行。

謝謝實驗室中博士生夥伴品而和 Richard，很開心我們曾一起參與法律 LLM 計畫，那段時光是我碩班最充實也最有成就感的日子。我從你們身上學到了許多研究方法，也感謝品而在我低潮時送上的暖心鼓勵，謝謝 Richard 總是在我遇到技術問題時及時救援，果然是值得依靠的大哥！也謝謝 Amy 在我碩三下提供我實習的機會，雖然那段時間邊寫論文邊實習真的非常忙碌，但我真心喜歡每天去實習上班的感覺。還要謝謝 LOPE 實驗室的其他夥伴們，因為有你們，才讓 LOPE 是一個完整又溫暖的團隊！

接著想感謝我最親愛的朋友們。謝謝「新竹暴飲小窩」的佳妤、曛綺、小馬，寫論文卡關時總會想回小窩一趟，和你們一起喝酒、亂哭、大笑、胡說八道，是我在新竹最溫暖的歸屬。謝謝靜玟、思岑，從高中一路吵吵鬧鬧到現在，雖然常鬥嘴，但我們總能在彼此需要時伸出手。謝謝遠在美國卻還是能分享彼此焦慮的安尼，有尼真好！謝謝快樂小狗好夥伴小曼，和你散步聊天總讓我覺得被療癒；謝謝帆帆，寒假一起在社圖寫論文的日子讓我不孤單。也謝謝國中好朋友小瑤、柔柔、庭瑀、玉梅、賀婷，雖然我很愛搞消失，謝謝你們還是願意包容我，屏東柔柔的婚禮旅行，是我寫論文期間最快樂的一段回憶！謝謝佐佐和馬安兩位好兄

doi:10.6342/NTU202504087

弟，在我寫論文這段期間，總是不厭其煩地聽我 murmur，雖然嘴上不常說好話，但我知道在需要你們的時候，你們都會在。謝謝我的雲端酒友凱為，總是能一語點破我的盲點，是我心中的榜樣——終生張老師！謝謝雖然不擅言詞卻會盡力同理我的建成！謝謝進台大的第一位好朋友凱晴，後會有期！謝謝這一路上曾經幫助、陪伴過我的每一位朋友，我們都要一直快快樂樂地走下去！

最後，謝謝我的家人。謝謝爸爸媽媽，雖然我們常常意見不合、有過不少爭執，但考研前爸爸說過：「失敗不可怕，可怕的是你站不起來。」這句話雖然聽來很硬，但卻在我心裡留下了深刻的力量，也成為我培養韌性的起點。謝謝奶奶、阿公、阿嬤，從小總是偷偷塞零用錢給我，現在我終於要畢業啦！也謝謝我的好夥伴 Momo，曾對我說過：「你不用拍拍你自己，我會拍拍你。」讓我知道，在我累的時候，只要轉頭、就會有人拍拍我。

最後的最後，要謝謝一路走過來的柔安。你真的很棒！未來的路上，也請記得保持謙卑，帶著這一年累積的經驗與收穫，繼續勇敢走得更遠。

doi:10.6342/NTU202504087

摘要

大型語言模型（LLMs）近年被廣泛應用於時間知識圖譜（Temporal Knowledge Graph, TKG）預測任務中，作為傳統圖神經網路方法的泛化替代方案。特別是在異常事件預測這類仰賴時間推理與語用合理性的任務中，模型需要理解一連串行為在時間軸上的邏輯與變化，因此如何提供具備結構化時序資訊的輸入成為關鍵挑戰。然而，目前仍不清楚 LLMs 在此任務中，究竟是依賴結構化的 TKG 輸入效果較佳，或是非結構化的自然語言描述表現更好。本研究比較兩種輸入格式（TKG → Text 與 Text → Text）對於語意一致性與語用推理能力的影響。

TKG 以主詞－關係－受詞的三元組表示事件，並附有明確的時間區間，其結構化格式能提升資訊密度、降低語言歧義，並幫助模型掌握事件順序與語用邏輯，特別適合異常行為偵測等需要精確推理的任務。相對地，非結構化的自然語言輸入雖然更接近 LLM 的預訓練分布，也能提供豐富語境，但容易引入冗詞與語意聯想雜訊。

本研究實驗採用 UCA 資料集，該資料集包含具時間標註的影片說明文字與異常事件標籤，為評估語意生成與語用異常判斷能力提供了合適的實驗資料集。此研究分為兩部分進行評估。首先，使用 BGE 嵌入模型計算 LLM 預測輸出與真實描述之間的餘弦相似度。結果顯示，非結構化輸入的語意對齊表現略高（平均值 0.5978），而結構化輸入亦表現接近（平均值 0.5718），顯示即便上下文減少，TKG 輸入亦能維持語意理解能力。

第二部分實驗中，此研究將不同輸入格式作為上下文，判斷當前幀是否異常，並以 AUC 作為評估指標。結果顯示：使用 TKG 作為時間上下文可提升異常判斷準確性，AUC 分數優於原始文本與摘要文本輸入，突顯其在語用推理上的穩健性。

整體而言，非結構化輸入有利於語意連貫與表達，結構化輸入則在語用推理上

doi:10.6342/NTU202504087

展現出更高穩健性，特別是在異常事件預測這類高度仰賴時間順序與行為合理性的任務中，TKG 輸入可提供更具邏輯性的語用支撐。本研究結果突顯兩者在上下文學習（in-context learning, ICL）任務上的互補潛力，亦為未來混合式提示輸入設計提供了方向，進一步提升模型對時序事件之語用理解能力。

關鍵詞: 大型語言模型、時間知識圖譜、異常事件預測、結構化輸入、非結構化文本、語用推理

doi:10.6342/NTU202504087

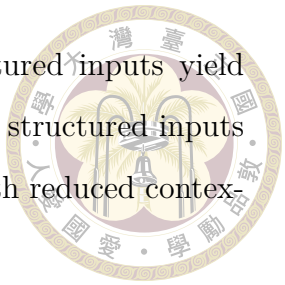
Abstract

Large language models (LLMs) have recently been applied to temporal knowledge graph (TKG) forecasting tasks, offering a generalizable alternative to traditional graph-based approaches. Particularly in abnormal event forecasting, where temporal reasoning and pragmatic coherence are essential, models must interpret the logical sequence and behavioral dynamics of actions over time. This raises a key challenge: how to provide input representations that encode structured temporal information. However, it remains unclear whether LLMs perform better in such tasks when guided by structured TKG inputs or by unstructured natural language descriptions. This thesis compares the effectiveness of structured (TKG $\rightarrow$ Text) and unstructured (Text $\rightarrow$ Text) inputs in forecasting abnormal events, with a focus on both semantic alignment and pragmatic reasoning.

TKGs represent events using subject–relation–object triples with explicit temporal spans. This structured format provides high information density, reduces linguistic ambiguity, and facilitates reasoning over temporal and causal dependencies —features that are particularly beneficial in tasks such as abnormal event detection. In contrast, unstructured textual descriptions offer rich contextual information and better alignment with LLMs' pretraining data, but may introduce interpretive noise due to overgeneralization.

Experiments were conducted on the UCA (UCF Crime Annotation) dataset, which contains temporally annotated video captions and ground-truth anomaly labels, serving as a suitable benchmark for evaluating both semantic generation and pragmatic anomaly detection. A two-part evaluation was performed. First, cosine similarity between the LLM-generated and ground-truth descriptions was computed

doi:10.6342/NTU202504087



using the BGE embedding model. Results show that unstructured inputs yield slightly higher semantic alignment (mean = 0.5978) compared to structured inputs (mean = 0.5718), suggesting that LLMs remain effective even with reduced contextual cues.

In the second experiment, different input formats were used as temporal context to identify abnormal video frames. Using AUC as the evaluation metric, structured TKG inputs outperformed raw and summarized text, demonstrating stronger grounding and robustness in pragmatic reasoning.

Overall, while unstructured inputs promote semantic fluency, structured representations, such as TKGs, enhance logical grounding, particularly in abnormal event forecasting tasks that rely heavily on temporal reasoning and pragmatic coherence. These findings highlight the potential of hybrid input designs—within text-based in-context learning setups—to combine structural clarity with contextual richness, enabling more accurate and pragmatically coherent reasoning in temporal event forecasting.

Keywords: Large Language Models, Temporal Knowledge Graph, Abnormal Event Forecasting, Structured Input, Unstructured Text, Pragmatic Reasoning

doi:10.6342/NTU202504087

Contents

doi:10.6342/NTU202504087

doi:10.6342/NTU202504087

List of Figures

doi:10.6342/NTU202504087

List of Tables

doi:10.6342/NTU202504087

Chapter 1

Introduction

1.1 Background and Research Context

Knowledge Graphs (KGs) have become a foundational component in various natural language understanding and reasoning tasks (Ji et al., 2021). A standard KG represents factual knowledge in the form of triples (h, r, t), where h is the head entity, r the relation, and t the tail entity (Kejriwal, 2019). However, real-world events often unfold over time and require a temporal dimension to be properly represented. To capture this, the Temporal Knowledge Graph (TKG) extends the standard triple with a timestamp τ, forming a quadruple (h, r, t, τ) (Gastinger et al., 2022). This enables the modeling of event dynamics and time-aware reasoning (Goel et al., 2020; Leblay & Chekol, 2018; Trivedi, Faruqui, et al., 2017).

Building on this foundation, Temporal Knowledge Graph Forecasting (TKGF) has emerged as a key approach for modeling and predicting future events in dynamic environments. The task involves extrapolating plausible future relations given a history of time-stamped relational facts, and has been applied in domains such as social dynamics, medicine, and criminal event modeling (Goel et al., 2020; Jin et al., 2020; Lee, Ahrabian, et al., 2023; Trivedi, Faruqui, et al., 2017). Existing approaches for this task fall into two main categories: graph-based and large language model-based (LLM-based) methods. Graph-based approaches typically rely on the underlying graph structure and temporal patterns, using architectures such as recurrent GNNs

doi:10.6342/NTU202504087

(e.g., RE-NET, T-GCN) (Jin et al., 2020; Zhao et al., 2019). While effective, these models often struggle with generalization in sparse or zero-shot scenarios (Han et al., 2021; Ma et al., 2021).

In contrast, LLM-based methods leverage the generalization power of large language models. These can be further divided into embedding-based and text-based approaches. Embedding-based methods transform TKG events into dense vectors for modeling (Goel et al., 2020), while text-based approaches convert structured events into natural language sequences, allowing LLMs to perform in-context learning (Lee, Ahrabian, et al., 2023). The latter is particularly attractive because it requires no additional training and can flexibly interpret input via prompting (Chen et al., 2023). Nevertheless, text-based forecasting with LLMs presents several inherent challenges. Chief among these is the limitation of finite context windows—contemporary LLMs are restricted to processing a bounded number of input tokens per inference step. This constraint becomes particularly problematic in scenarios involving long temporal sequences or tasks requiring context-dependent reasoning, where earlier information may be truncated, underutilized, or entirely lost (Liu et al., 2023).

As prior work has pointed out, both forecasting and anomaly detection involve analyzing time series or timestamped data (Su et al., 2024b). Building on this, this study argues that both TKGF and anomaly detection can be viewed as instances of temporal reasoning tasks, as they require understanding how events unfold and relate over time. While forecasting aims to anticipate future developments based on observed historical patterns, anomaly detection focuses on identifying deviations from expected temporal trajectories. Despite their different objectives—prediction versus deviation classification—both tasks depend on a model's ability to reason over temporally structured inputs and to recognize causality, coherence, and change.

While previous studies have explored LLMs' capabilities in time series forecasting (Gruver et al., 2023; Jin, Wang, et al., 2023), few have examined whether these models truly understand the underlying temporal dynamics. Recent work

doi:10.6342/NTU202504087

by Zhou and Yu (2024) argues that anomaly detection provides a more diagnostic test of temporal reasoning, as it forces models to go beyond average extrapolation and identify structural irregularities. Building on this insight, this study extends the notion of "anomaly" to encompass semantic and pragmatic inconsistencies in human-annotated surveillance captions, and investigate how structured (TKG) and unstructured (caption) inputs influence LLMs' understanding of abnormal events.

1.2 Motivation and Thesis Organization

The core motivation of this thesis is to investigate the comparative effects of structured (e.g., TKG quadruples) and unstructured (e.g., raw captions) temporal inputs on the reasoning and generative capabilities of LLMs. To this end, the present study is situated within the broader framework of abnormal event forecasting, which encompasses both TKG forecasting and anomaly detection. A unified benchmark is established by aligning and merging two complementary datasets: the UCF-Crime (UCF) dataset (Sultani et al., 2018), which offers real-world surveillance videos annotated with anomaly labels, and the UCF Crime Annotation (UCA) dataset (Yuan et al., 2023), which provides human-annotated video-level captions. While each dataset individually lacks certain components essential for a comprehensive evaluation of LLMs' semantic and pragmatic reasoning abilities, their integration—achieved through timestamp alignment—yields a temporally grounded corpus comprising both structured (TKG) and unstructured (caption) representations.

This unified dataset serves as the evaluation benchmark throughout the study. It offers a rich testbed for assessing the temporal reasoning capabilities of LLMs, as it captures unfolding human activities that frequently involve implicit cues related to causality, intention, and behavioral deviation. Rather than relying on synthetic or purely numerical time series data, this study utilizes both human-written captions and their corresponding TKG representations to simulate unstructured and structured temporal contexts, respectively. Two tasks are formulated: (1) forecasting the next-frame caption, and (2) predicting whether the upcoming frame is anomalous

doi:10.6342/NTU202504087

based on prior context. These tasks collectively enable an evaluation not only of predictive accuracy, but also of the semantic coherence and pragmatic depth afforded by different input modalities.

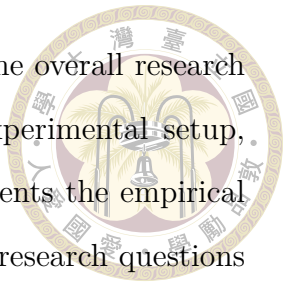
This thesis is guided by the following research questions aimed at assessing the temporal reasoning capabilities of LLMs under different input formats:

- RQ1: Does structured temporal input (e.g., TKGs) provide advantages over unstructured input (e.g., raw captions) for temporal forecasting tasks?
- RQ2: How does temporal context—whether structured or unstructured—impact LLM performance in anomaly detection tasks?

The empirical results reveal a nuanced relationship between input modality and LLM performance across different reasoning tasks. In temporal forecasting tasks, unstructured inputs (e.g., raw captions) yielded slightly higher similarity scores than structured TKGs, though the difference was not statistically significant. In contrast, for anomaly detection tasks, structured inputs demonstrated greater robustness, achieving higher AUC scores than their unstructured counterparts—particularly when prior context was long or uncoherent. Summarization of unstructured inputs provided moderate improvements, suggesting that temporal compression and coherence play a critical role in model performance. These findings highlight a trade-off: while unstructured text may be better suited for generation tasks, structured formats like TKGs offer more consistent support for pragmatic and temporally grounded inference.

This thesis is organized as follows. Chapter 2 provides a literature review structured into three parts: (1)the architecture and pretraining objectives of LLMs, their capabilities in semantic and pragmatic understanding, and their relevance to temporal reasoning; (2) the foundations of knowledge representation with a focus on KGs and TKGs; (3) an overview of temporal reasoning tasks—particularly TKG forecasting and anomaly detection—along with their evaluation frameworks; and (4) prompting strategies for LLMs, contrasting structured inputs (e.g., TKGs) with

doi:10.6342/NTU202504087



unstructured inputs (e.g., raw captions). Chapter 3 introduces the overall research methodology, data preparation procedures, and outlines the experimental setup, including prompt design and evaluation metrics. Chapter 4 presents the empirical results. Chapter 5 discusses findings in relation to the proposed research questions and provides qualitative error analyses for both experiments. Chapter 6 concludes the thesis by summarizing the main contributions and offering suggestions for limitations and future research directions.

By comparing structured and unstructured temporal inputs in the context of abnormal event forecasting, this study aims to empirically examine how different input modalities influence the semantic coherence and pragmatic sensitivity of LLMs.

doi:10.6342/NTU202504087

Chapter 2

Literature Review

Recent advances in LLMs have significantly expanded their applicability across tasks involving language understanding, reasoning, and temporal inference. These models, originally designed for static text processing, are now being evaluated for their ability to handle structured and dynamic information such as KGs and time-sensitive data. In this chapter, we review four key areas of literature that form the foundation for this thesis.

Section 2.1 introduces the architecture and general reasoning capabilities of LLMs, with a particular focus on their emerging potential in temporal reasoning. Section 2.2 discusses knowledge representation using KGs and their temporal extension, TKGs, which provide structured ways to encode evolving events over time. Section 2.3 reviews how LLMs have been applied in temporal reasoning tasks, including forecasting and anomaly detection, and the evaluation benchmarks used to assess model performance. Finally, Section 2.4 examines how different input formats—structured vs. unstructured—affect the performance and interpretability of LLMs in temporally grounded reasoning settings.

2.1 Large Language Models (LLMs)

Large language models (LLMs) refer to neural network architectures with hundreds of millions to billions of parameters, typically based on the transformer architecture

doi:10.6342/NTU202504087

introduced by Vaswani et al. (Vaswani et al., 2017). This architecture leverages self-attention mechanisms to model dependencies between tokens, allowing the model to capture both local and global contextual information. LLMs are generally pretrained on large-scale text corpora using objectives such as masked language modeling (e.g., BERT) or autoregressive language modeling (e.g., GPT), enabling them to learn rich contextual representations of language.

Beyond basic language modeling, LLMs have demonstrated strong performance in a variety of downstream tasks, including question answering, summarization, and commonsense reasoning (Bommasani et al., 2021). These models can generate coherent long-form text, recognize complex linguistic dependencies, and even engage in multi-step inference without task-specific fine-tuning. Their emergent abilities—arising as model size and training data scale—allow them to generalize to tasks for which they were not explicitly trained (Wei et al., 2022).

Recent studies have begun to explore whether LLMs can perform temporal reasoning, such as understanding event order, duration, and causality. Xiong, Payani, et al. (2024) propose a framework where LLMs are fine-tuned using latent temporal graph representations, demonstrating that large models can learn temporal reasoning through structured graph prompts. Furthermore, recent work in temporal question answering shows that LLMs—when guided by well-crafted prompts and auxiliary graph structures—can answer time-sensitive questions and infer event sequences, highlighting their ability to grasp temporal dynamics even in zero-shot or few-shot scenarios (Li et al., 2023; Xiong, Payani, et al., 2024). These findings suggest that LLMs, although primarily designed for static text, possess nascent potential for modeling temporal dependencies when provided with suitably structured inputs.

Given their general-purpose language understanding capabilities and growing evidence of temporal reasoning potential, LLMs offer a promising foundation for exploring how structured and unstructured inputs influence reasoning performance. This thesis adopts LLMs as the core model to investigate how different input modal-

doi:10.6342/NTU202504087

ities—specifically TKGs versus natural language captions—affect the model's ability to understand and reason over time-anchored event data. The goal is not to improve the LLM itself, but to use it as a lens through which to observe the effects of representation format on temporal interpretability.

2.2 Knowledge Representation with Knowledge Graphs (KGs) and Temporal Knowledge Graphs (TKGs)

Knowledge representation serves as a crucial foundation for enabling reasoning in both symbolic and neural models (Liu et al., 2024). Among various structured representations, KGs offer a relational format that is widely adopted in language understanding and inference tasks. However, traditional KGs are static and insufficient for modeling temporally evolving knowledge. To address this, TKGs introduce an explicit temporal dimension, enabling the representation of time-sensitive facts (Cai et al., 2024). TKGs have become particularly useful in forecasting tasks, where the goal is to predict future relational events based on observed temporal patterns. This section introduces the foundations of KGs and TKGs, and contextualizes the role of TKGs in temporal knowledge modeling.

2.2.1 KGs and TKGs

Knowledge Graphs (KGs)

Knowledge Graphs (KGs) are a structured way to represent knowledge by organizing entities and their relationships into directed graphs. Each fact in a KG is encoded as a triple $\langle h, r, t \rangle$, where h denotes the head entity, r represents the relation, and t is the tail entity (Kejriwal, 2019). This triplet structure enables both human interpretability and machine reasoning, and has been widely adopted in applications such as information retrieval, semantic search, and question answering.

KGs excel at capturing semantic relationships among entities in a compact and

doi:10.6342/NTU202504087

structured form. For instance, in the domain of surveillance video analysis, the sentence "a man in black enters the house" can be transformed into a KG triple as follows:

$$\langle \text{man in black}, \text{enters}, \text{house} \rangle$$

Such representations facilitate higher-level reasoning over events, allowing models to infer interactions and detect abnormal behavior by analyzing relational patterns.

By modeling real-world knowledge in a graph structure, KGs support various downstream tasks that require contextual understanding, structured inference, or temporal linking when extended with temporal components (Barrasa & Webber, 2023).

Temporal Knowledge Graphs (TKGs)

While traditional KGs are effective for representing static facts, many real-world scenarios—particularly those involving human behavior and events—unfold over time and require temporal modeling. Temporal Knowledge Graphs (TKGs) extend the standard KG structure by associating each triple with a timestamp, resulting in a quadruple $\langle h, r, t, \tau \rangle$, where τ denotes the time or time interval during which the fact holds true (Jin, Wen, et al., 2023). This additional temporal dimension allows for encoding dynamic relationships that change or evolve over time (Trivedi, Dai, et al., 2017; Trivedi et al., 2018).

In this thesis, the structured input used for LLM prompting is based on TKGs constructed from video-level event annotations in the UCF-Crime Annotation (UCA) dataset. These graphs are used to represent temporally grounded human actions observed in surveillance footage.

For example, the observation "the man in black enters the house between 1.6 and 7.4 seconds" can be encoded as a TKG quadruple:

$$\langle \text{man in black}, \ \text{enters}, \ \text{house}, \ [1.6s, 7.4s] \rangle$$

To model a sequence of evolving actions, consecutive frames can be represented

doi:10.6342/NTU202504087

by a set of temporally ordered TKG quadruples. For instance, the following two observations capture successive interactions by the same subject:

$$\langle \text{man in black},\ \text{approaches},\ \text{house},\ t_1 \rangle$$

$$\langle \text{man in black},\ \text{opens},\ \text{front door},\ t_2 \rangle$$

where $t_1 = [0.0s,\ 3.5s]$ and $t_2 = [3.5s,\ 7.0s]$.

These temporally aligned representations not only preserve the sequential structure of events but also provide a foundation for downstream tasks such as forecasting future actions and detecting anomalies in behavior.

By explicitly modeling time-sensitive event relations, TKGs offer a structured foundation for reasoning over temporally evolving scenes. This enables LLMs not only to interpret isolated facts but also to infer temporal patterns and predict future events based on past behaviors.

2.2.2 TKG Forecasting

Graph-Based Approaches to TKG Forecasting

Graph-based methods for TKG forecasting build upon traditional KG embedding and graph neural network (GNN) techniques to capture temporal dynamics. Early models such as RGCN (Schlichtkrull et al., 2017) and ConvTransE (Jin, Wen, et al., 2023) adapted static KG architectures for timestep-based prediction, but lacked the ability to model temporal continuity. To address this limitation, approaches like RE-Net and Recurrent RGCN introduced recurrent components into GNNs, enabling historical state propagation and temporal aggregation (Chang et al., 2025).

To enhance interpretability, symbolic reasoning models like TLogic (Liu et al., 2022) and Temporal ILP (Xiong, Yang, Fekri, & Kerce, 2024; Xiong, Yang, Payani, et al., 2024) generate temporal logical rules that explain event sequences. These models offer transparency but often require manual rule construction or complex learning schemes. Recognizing the limitations of sparse or noisy graph data, more

doi:10.6342/NTU202504087

recent work integrates textual context. Hybrid models such as SeCoGD (Ma, Ye, Wu, Wang, Cao, & Chua, 2023), LoGo (Ma, Ye, Wu, Wang, Cao, Pang, & Chua, 2023), Glean (Deng et al., 2020), and CMF (Deng et al., 2021) augment TKGs with topic modeling or text-derived embeddings, though they must contend with noise and long-tail distributions in natural language data.

Additionally, models like HisMatch (Li et al., 2022) approach forecasting as a query–candidate matching problem, directly scoring potential object entities for a given query triple. While graph-based models have achieved strong performance on benchmarks like ICEWS (O'brien, 2010) and GDELT (Leetaru & Schrodt, 2013), they often require dataset-specific tuning and lack access to richer contextual cues-motivating the shift toward language model-based approaches that leverage pre-trained knowledge for more generalizable reasoning.

LLM-Based Approaches to Temporal Forecasting

With the advancement of large pre-trained language models (LLMs), researchers have begun exploring their potential for TKG forecasting. Rather than relying on traditional supervised learning, many of these approaches frame TKG prediction as a prompting or language modeling task, enabling inference with minimal or no additional fine-tuning. Broadly, LLM-based methods for TKG forecasting can be categorized into two types: embedding-based approaches and text-based approaches utilizing in-context learning (Chang et al., 2025).

Embedding-Based Methods.

Embedding-based methods integrate pre-trained graph embeddings as an auxiliary modality into the language model's input. For example, KoPA (Zhang, Chen, et al., 2024) proposed an adapter mechanism to prepend KG embeddings to LLM prompts. More recent models like GraphTranslator (Zhang, Sun, et al., 2024) introduce a translation module that converts graph embeddings into token-level representations interpretable by LLMs. Similarly, TEA-LLM (Wang et al., 2024) employs a linear projection to align outputs from graph neural networks (GNNs) with the

doi:10.6342/NTU202504087

LLM's input space. These techniques aim to supplement the model with structured signals that reinforce the understanding of temporal or relational patterns. However, the effectiveness of such integration hinges on the quality of the graph embeddings themselves—poorly trained embeddings may introduce noise or bias, as observed in works like TGL-LLM (Chang et al., 2025).

Text-Based In-Context Learning.

Text-based approaches using in-context learning typically recast TKG forecasting as a completion task, where the LLM is provided with a flattened sequence of historical events in textual or symbolic form and is asked to predict the next plausible fact. Following the retrieval-augmented generation paradigm, models like GPT-NeoX-ICL (Lee, Ahrabian, et al., 2023) dynamically retrieve relevant past facts from the TKG and structure them into prompts formatted as quadruples. Other approaches like GENTKG (Liao et al., 2023) and CoH (Luo et al., 2024) incorporate temporal logical rules or heuristic templates to organize context, guiding the model's predictions through rule-based sampling of historical sequences.

A notable study by Lee, Ahrabian, et al. (2023). showed that even general-purpose LLMs, without fine-tuning, can perform competitively on standard TKG datasets such as ICEWS (O'brien, 2010) and Wiki-TKG (Leblay & Chekol, 2018). Their experiments involved prompting the model with historical triples (e.g., *[Superbowl, Champion, Team]*) and asking it to forecast the next event. Surprisingly, the performance of the LLMs was within a few percentage points (Hits@1) of top-performing graph-based methods. Even more striking, the model maintained its forecasting accuracy when all semantic labels were replaced with arbitrary IDs, implying that LLMs can learn to recognize abstract structural patterns rather than relying solely on semantic knowledge.

These results highlight the capability of LLMs to generalize from structured sequences, even in the absence of real-world semantics. They suggest that both symbolic and structured input formats can serve as effective representations for temporal reasoning, opening new directions for hybrid models that blend symbolic

doi:10.6342/NTU202504087

logic with language understanding.

2.3 LLM Application in Forecasting and Anomaly Detection Tasks

LLMs have been increasingly applied to tasks involving sequential and temporally structured data (Alnegheimish et al., 2024; Jin, Wang, et al., 2023). Among various temporal applications, forecasting and anomaly detection stand out as two essential tasks that benefit from the reasoning capabilities and language understanding of LLMs (Su et al., 2024a). Forecasting involves predicting future observations based on previously observed data, while anomaly detection aims to identify outliers or irregular patterns that deviate from expected temporal behavior. Although these tasks differ in objective, both operate over time-indexed inputs and often rely on models' ability to infer latent temporal dependencies.

An overview of this process is illustrated in Figure 2.1, which depicts the typical pipeline for applying LLMs to forecasting and anomaly detection tasks. The input data—whether in the form of text logs, numerical time series, structured graphs, visual recordings, or speech—is assumed to be temporally anchored. This timestamped information is fed into pre-trained LLMs such as BERT, GPT, LLaMA2, or proprietary models like Claude and Gemini, enabling the model to either project future trends or detect deviations from historical patterns.

Forecasting

Forecasting is a fundamental temporal reasoning task that aims to predict future events or values based on historical data patterns. Traditionally addressed by statistical or deep learning models, forecasting has recently seen the incorporation of LLMs, which bring strong generalization capabilities and flexible prompting mechanisms.

LLM-based forecasting approaches can be grouped into three categories: (1) zero-shot or few-shot prompting, where models like GPT-3 or LLaMA2 are directly

doi:10.6342/NTU202504087

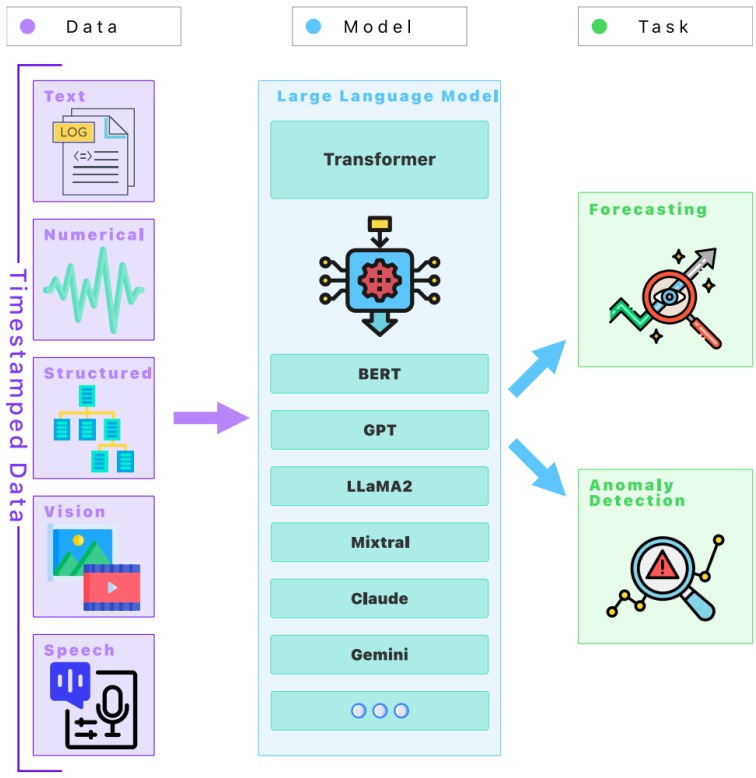
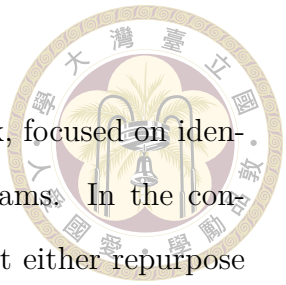


Figure 2.1: Illustration of how LLMs are utilized in forecasting and anomaly detection, with input data typically consisting of time-series or timestamped data.(Su et al., 2024a)

prompted with time series formatted as textual input; (2) fine-tuning, where LLMs such as BERT or RoBERTa are adapted to specific time series datasets; and (3) foundation model usage, where pre-trained models are applied without modification to perform forecasting tasks.

Recent studies exemplify the effectiveness of LLMs in this area. Gruver et al. (2023) employed GPT-3 and LLaMA2 models in zero-shot settings on datasets such as Darts and Informer, reporting competitive results using MAE. Xue and Salim (2023) used GPT-3.5 and BART in prompt-based forecasting tasks across CT and ECL datasets. Xue et al. (2022) fine-tuned BERT variants on SafeGraph data, achieving improved RMSE and MAPE scores. These works demonstrate the versatility of LLMs in modeling temporal dependencies using natural language interfaces and flexible context encoding mechanisms.

doi:10.6342/NTU202504087



Anomaly Detection

Anomaly detection is another critical temporal reasoning task, focused on identifying unexpected or abnormal patterns in temporal data streams. In the context of LLMs, this task is typically addressed using methods that either repurpose pre-trained models for sequence classification or employ prompt-based reasoning to detect deviations from expected patterns.

LLMs have been applied to anomaly detection in several forms: (1) foundation model usage as frozen encoders for log or sensor data, (2) fine-tuning for binary classification of anomalous vs. normal sequences, and (3) prompt-based detection, where LLMs are asked directly whether a given time segment is anomalous.

Multiple studies have demonstrated the potential of LLMs in this context. Dang et al. (2021) fine-tuned BERT on the KPI and Yahoo datasets, achieving strong F1-scores. Lee, Kim, and Kang (2023) explored few-shot and zero-shot detection using BERT and GPT-2 on system logs such as HDFS and BGL. Other works (e.g., Huang et al. (2023) and Zhang et al. (2023)) showed that prompt-based LLMs can capture nuanced behavioral inconsistencies even in noisy or partially labeled settings. These approaches benefit from the ability of LLMs to model complex temporal-textual correlations and implicit anomaly semantics.

Evaluation and Benchmarks

Based on the survey by Su et al. (2024a), the evaluation of LLMs applied to temporal tasks typically relies on task-specific metrics and benchmark datasets that reflect the models' ability to process and reason over time-dependent data.

For forecasting, standard evaluation involves regression-oriented metrics, including Mean Absolute Error (MAE), Root Mean Square Error (RMSE), Mean Absolute Percentage Error (MAPE), Symmetric MAPE (sMAPE), Mean Absolute Scaled Error (MASE), and Overall Weighted Average (OWA). These metrics assess how well a model can project future values from historical patterns, taking into account different levels of scale sensitivity and prediction robustness.

In contrast, anomaly detection is typically evaluated using classification met-

doi:10.6342/NTU202504087

rics, such as Precision, Recall, F1-score, Accuracy, and Area Under the Receiver Operating Characteristic Curve (AUROC). These metrics are especially relevant in imbalanced datasets, where detecting rare anomalies with high recall while minimizing false positives is crucial.

Widely adopted benchmark datasets include:

- **Forecasting:** ETT, Weather, Electricity, Traffic, Darts, Monash, and Amazon Review;
- **Anomaly Detection:** HDFS, BGL, KPI, Yahoo, Thunderbird, OpenStack, and CSIC.

These datasets cover a diverse range of domains, including industrial sensor logs, public infrastructure data, and real-world e-commerce and system activity. As noted by Su et al. (2024a), many LLM-based studies evaluate model performance across different prompting and fine-tuning paradigms using these datasets, facilitating a systematic comparison across input strategies and task formulations (Su et al., 2024a).

2.4 Input Representations and Prompting Strategies for LLMs

Structured inputs (like KG triples, adjacency lists, or learned graph embeddings) are explicit and precise in encoding relationships, whereas unstructured inputs (free-form text, natural language descriptions, or even images) carry rich contextual information but are noisier and harder for models to interpret consistently. Recent studies provide insight into how each type of input influences performance in temporal reasoning and related tasks:

Effectiveness of Structured Data

For tasks requiring relational reasoning (e.g. predicting a missing entity in a future event), structured representations have proven highly effective. The fact

doi:10.6342/NTU202504087

that an LLM can predict events even when entity names are replaced with IDs shows that structure alone (the pattern of connections and timestamps) can be a sufficient signal Lee, Ahrabian, et al. (2023). In forecasting tasks, Chang et al. (2024) observed that LLM prompts formatted with discrete graph facts (triples) were more beneficial than long raw texts describing those facts. The structured format reduces ambiguity –each triple is a concise, atomic fact –which helps the model focus on temporal and relational pattern recognition rather than parsing language. Moreover, in TGL-LLM's benchmarks, the graph-embedding-informed prompts outperformed text-based prompts on all datasets (Chang et al., 2025). This suggests that when an LLM is given information in a structured way (either as triples or encoded embeddings), it can reason more systematically about what comes next, without being distracted by extraneous linguistic details.

Utility of Unstructured Inputs

Unstructured data, such as natural language, can complement structured data by providing background knowledge, explanations, or missing details not captured in the KG. For instance, a news article snippet might mention subtle causal links or entity attributes that are useful for forecasting an outcome. The challenge is that LLMs in zero-shot or few-shot settings might struggle to filter relevant facts from raw text. Chang et al. (2024) found that adding raw news paragraphs to a prompt did not help the model unless it was trained to summarize or attend to them. To effectively use unstructured inputs, techniques like summarization and retrieval are employed: one can first summarize documents into key points (to reduce noise), or retrieve only the most pertinent sentences/events to feed the model. When done properly, this can improve performance –e.g. their work showed that a fine-tuned LLM using summarized news did outperform the same model using only KG triples. Unstructured input shines more in providing semantic nuances: an LLM can leverage its world knowledge and language understanding to infer connections (e.g. knowing that if "X declared Y" appears in text, it implies a relation X, announced, Y). In temporal question answering (TKGQA) (Ong et al., 2023), researchers have combined textual

doi:10.6342/NTU202504087

context with graph structure for better reasoning. For example, GenTKGQA (Gao et al., 2024) uses an LLM to extract a relevant subgraph for a question, then embeds that subgraph to answer the query. Another work, M3TQA (Zha et al., 2024), fuses sentence-level semantic features with entity-level graph features to answer temporal questions. These illustrate that unstructured and structured inputs can be fused to cover each other's blind spots –text offers semantic richness, while graphs offer precision.

While prior studies have explored how different input formats affect LLM performance in forecasting tasks, few have examined how these formats influence a model's ability to interpret and reason over temporally structured information. This thesis focuses on comparing how large language models understand structured inputs, specifically TKGs, versus unstructured inputs such as natural language captions.

To evaluate this, anomaly detection is adopted as a setting in which understanding temporal structure is particularly critical. Detecting anomaly entails more than just completing familiar patterns—it requires a deeper understanding of semantic coherence, pragmatic norms, and contextual irregularities (Zhou & Yu, 2024). In contrast to conventional forecasting, where models may succeed by exploiting statistical regularities, anomaly detection compels the model to recognize subtle shifts, edge cases, and out-of-distribution events. It resists superficial heuristics and instead tests whether the model has internalized a nuanced, contextualized representation of what is normal—and why. As such, anomaly detection offers a more rigorous benchmark for evaluating temporal reasoning, one that prioritizes genuine comprehension over shallow pattern matching.

Accordingly, the abnormal event forecasting task is used as a diagnostic setting to analyze how input modality shapes the model's temporal reasoning. By observing how LLMs respond to anomalous patterns under each input condition, the study assesses their capacity to internalize temporal structures, detect deviations, and encode context-dependent expectations.

doi:10.6342/NTU202504087

Chapter 3

Methodology

3.1 Overviews

This study investigates whether structured or unstructured inputs enable LLMs to more effectively forecast abnormal events, focusing on two key dimensions: forecasting and anomaly detection. To evaluate this, two experiments were designed using the UCF anomaly annotations and the frame-level caption ground-truth from the UCA dataset:

1. The forecasting experiment 3.4 investigates how well an LLM can generate the next-frame semantic description based on prior context. The generated descriptions are then compared against ground-truth captions to assess semantic alignment.

2. The anomaly detection experiment 3.5 evaluates the LLM's ability to identify anomalous frames under different temporal input contexts, including: no context, raw captions, summarized text, and structured TKGs.

3.2 Dataset

This study utilizes both the UCF-Crime dataset (Sultani et al., 2018) and the UCF Crime Annotation (UCA) dataset (Yuan et al., 2023). UCF-Crime is a large-scale benchmark for real-world anomaly detection in surveillance videos, while UCA extends this dataset by providing human-annotated video-level captions. However,

doi:10.6342/NTU202504087

each dataset independently lacks certain components critical for evaluating both semantic and pragmatic capabilities of large language models (LLMs). To address this, a preprocessing step was performed to align and merge the two datasets using shared timestamps, yielding a unified dataset containing video-level captions, temporal spans, and anomaly labels.

The UCF-Crime dataset consists of 1,900 long and untrimmed surveillance videos, totaling over 128 hours. Each video captures real-world scenes from fixed-angle security cameras and contains either normal activities or one of 13 predefined anomalous event types:

Abuse, Arrest, Arson, Assault, Road Accident, Burglary, Explosion, Fighting, Robbery, Shooting, Stealing, Shoplifting, and Vandalism.

Anomalous events in UCF-Crime are broadly defined as criminal or socially deviant behaviors that significantly deviate from routine surveillance footage. However, in its original form, the dataset only provides video-level binary anomaly labels, indicating whether a given video contains any abnormal activity (Sultani et al., 2018). This labeling format supports weakly supervised learning but limits its applicability for tasks requiring fine-grained temporal reasoning.

To enable more precise evaluation, the dataset authors additionally released a supplementary annotation file on the official project website [1]. This file provides segment-level ground truth for a subset of the testing videos, specifying the start and end timestamps of anomalous segments. These temporal spans allow for frame-aligned evaluation of models that predict scalar anomaly scores or identify temporal shifts in regularity.

The annotated segments correspond to observable event sequences involving clearly abnormal activities. For instance, in the Arson category, an anomalous segment may span the full duration of the scene described as "*The man returned to the Christmas tree and continued to light the Christmas tree and successfully lit it.*" This degree of temporal granularity is essential for evaluating the alignment between

[1] Real-world Anomaly Detection in Surveillance Videos : https://www.crcv.ucf.edu/projects/real-world/

doi:10.6342/NTU202504087

predicted anomaly scores and ground-truth event boundaries. While these annotations provide useful temporal segmentation, they do not include natural language descriptions of the visual content, thereby limiting their utility for language-based reasoning tasks.

To complement this limitation, the UCF Crime Annotation (UCA) dataset (Yuan et al., 2023) serves as a multimodal extension of UCF-Crime, offering human-annotated textual descriptions aligned with video segments. UCA includes over 23,000 sentence-level captions covering approximately 110.7 hours of video data. Each caption averages 20 words in length and is temporally grounded with precise start and end timestamps at 0.1-second resolution. In contrast to the weak labels of UCF-Crime, UCA provides fine-grained natural language annotations that describe both normal and anomalous events in detail.

These annotations enhance the dataset with a linguistic modality that enables semantic and pragmatic modeling. The presence of both temporal boundaries and aligned textual descriptions makes UCA particularly suitable for evaluating large language models (LLMs) in temporally grounded reasoning tasks. Additionally, the dataset supports the construction of either structured inputs (e.g., temporal knowledge graphs) or unstructured inputs (e.g., caption sequences), allowing for systematic comparison across input modalities. The dual temporal and semantic structure of UCA is especially valuable for analyzing model performance in forecasting and anomaly detection tasks that require contextual interpretation of evolving scenes. A reference table showcasing aligned captions, TKG triplets, and anomaly labels from both UCF-Crime and UCA can be found in Appendix A.

3.3 Models Used

Two LLMs were employed in this study, each serving a distinct role across the two experiments designed for abnormal forecasting tasks.

GPT-4o-Mini (via OpenAI API).

doi:10.6342/NTU202504087

This model was used exclusively for extracting TKG representations from natural language captions. It was accessed through the OpenAI API and operated using LangChain's `LLMGraphTransformer()` module. A temperature setting of 0.1 was applied to ensure deterministic triple extraction, and no fine-tuning or post-processing was applied beyond temporal alignment. The use of a closed-source model for this preprocessing step was motivated by its demonstrated superior performance in complex zero-shot structural parsing and KG extraction tasks, as evidenced by its robust capabilities in accurately identifying and structuring entities and relations from varied natural language inputs (Carta et al., 2023; Huang et al., 2024). This decision aimed to maximize the quality and reliability of the TKG representations, thereby providing a strong foundation for downstream tasks without introducing confounding errors from subpar extraction.

Mistral-large-latest (via Open Source API).

All downstream LLM inference in both experiments—forecasting and anomaly detection tasks—was conducted using the open-source `mistral-large-latest` model[2], served via API. This model was selected primarily for two key reasons: first, its open-source nature ensures full reproducibility and transparency of our experimental results, a crucial aspect for academic research. Second, Mistral-large-latest is a powerful, instruction-tuned model that has shown remarkable performance across a wide range of natural language processing tasks, including reasoning and generation. Its balance of high performance and accessibility made it an ideal candidate to assess whether such open-source models can effectively perform structured and unstructured reasoning when provided with standardized prompts, without relying on proprietary models for the core forecasting and anomaly detection logic. The same inference parameters were used across all runs: temperature = 0.1, top-p = 1.0, and maximum token length = 128. This consistency ensured a controlled comparison between structured (TKG-based) and unstructured (caption-based) inputs within each task.

[2]Mistral AI:https://docs.mistral.ai/getting-started/models/models_overview/

doi:10.6342/NTU202504087

By separating the TKG extraction phase from the main evaluation model, this setup ensures that the differences observed between input modalities are not confounded by structural encoding quality, allowing for a focused assessment of the LLM's reasoning capabilities.

3.4 Experiment 1: Forecasting

3.4.1 Objective

The forecasting experiment investigates the ability of LLMs to generate semantically plausible next-event descriptions based on prior context. Rather than predicting new knowledge graph triples, the task involves forecasting the natural language caption of a future video frame using various forms of preceding input—including raw frame captions and temporally structured information such as TKGs. The primary focus is on evaluating semantic coherence and contextual appropriateness of the generated output.

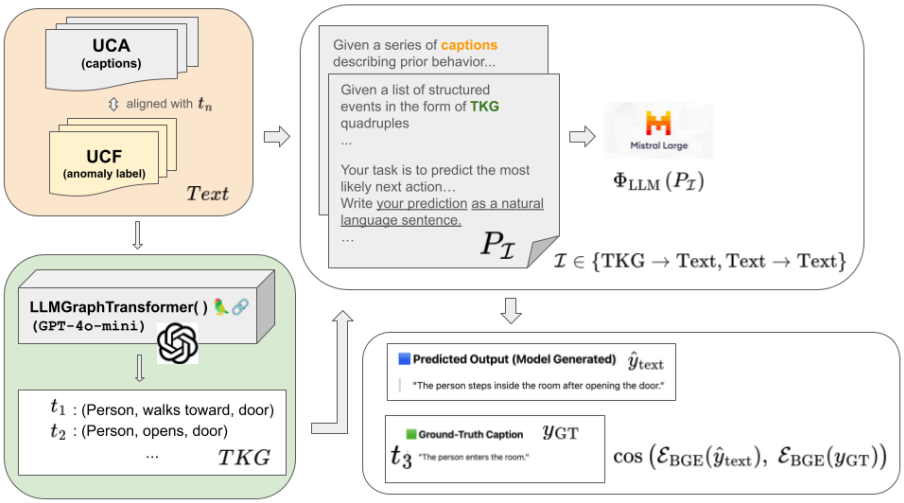


Figure 3.1: Pipeline of the Experiment 1: forecasting task. The model receives temporally ordered input (either structured TKGs or unstructured captions) and generates a next-frame description. The generated output is then compared against ground-truth captions to evaluate semantic alignment.

The central question is: Do structured TKGs or unstructured caption sequences provide more effective input for forecasting task in abnormal event contexts?

doi:10.6342/NTU202504087

To answer this, the forecasting experiment consists of three main stages: (1) constructing a temporal context window from the input, either as a structured sequence of TKG quadruples or as an unstructured sequence of natural language captions; (2) prompting a LLM to generate a predicted sentence describing the most likely next action prior to an abnormal event; and (3) measuring the semantic similarity between the generated prediction and the corresponding ground-truth caption using a sentence embedding model. The overall setup of this experiment is illustrated in Figure 3.1.

3.4.2 Input Settings

Two types of input representations were tested:

Structured (TKG → Text): Temporal Knowledge Graphs (TKGs) were constructed from caption-aligned segments using subject–predicate–object triples with explicit timestamps. These TKGs were then verbalized into prompts describing past events in structured language.

Unstructured (Text → Text): Raw or lightly summarized caption segments were concatenated to form a free-text temporal context, which was then provided as the prompt for generation.

3.4.3 TKG Construction from UCA Captions

To construct the structured input used in the TKG → Text condition, subject – relation – object (S, R, O) triples were first extracted from the original UCA caption data. This was accomplished using the `LLMGraphTransformer()` module from LangChain[3] , which internally leverages a large language model to convert natural language sentences into semantic graph representations. In this study, the underlying language model used for this transformation was `gpt-4o-mini`, accessed via the OpenAI API[4].

[3]LangChain: https://python.langchain.com/docs/introduction/
[4]OpenAI API: https://openai.com/index/openai-api/

doi:10.6342/NTU202504087

Each captioned segment, representing a temporally grounded event, was processed to extract one or more (S, R, O) triples. Instead of using raw numeric timestamps, each event's original start–end time was mapped to an abstract temporal token (e.g., t_1, t_2, t_3, ...), reflecting its chronological order in the video. The final structured format thus took the form:

$$(\text{subject},\ \text{relation},\ \text{object},\ t_n)$$

where tn is a placeholder denoting the n-th event in temporal sequence. These TKG quadruples were then ordered by their original timestamps to form a coherent temporal context, which was embedded into the prompt template for structured prediction.

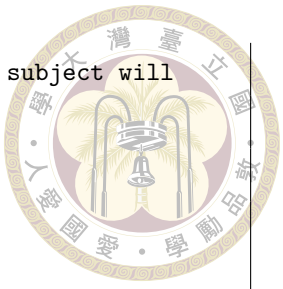
This transformation preserved the relative temporal structure of events while simplifying the representation for the LLM, allowing the model to focus on temporal progression without being distracted by irrelevant or noisy differences in raw timestamp values—such as whether two events are 0.5 seconds apart or 10 seconds apart.

3.4.4 Prompt Design

Each prompt is designed to elicit from the LLM a prediction of what action is most likely to occur immediately before the anomaly. The LLM is constrained to output exactly one sentence as a prediction. Examples of the actual prompt formats used are as follows: **Structured Input Prompt (TKG → Text),** $P_{\text{TKG}\to\text{Text}}$:

```
[Goal]: You are given a list of structured events in the form of temporal
knowledge graph (TKG) quadruples: (subject, relation, object, timestamp).
These represent a subject's past
actions over time.
```


doi:10.6342/NTU202504087



```
Your task is to predict the most likely next action that the subject will
perform **immediately before an abnormal event occurs**.
Write your prediction as a natural language sentence.

[Input - TKG History Before Anomaly]:
{tkg_quadruples}

[Constraint]:
- Predict exactly **one sentence** that describes the next likely action.
- Your output should be **one complete sentence**.

[Output - Predicted Sentence]:
```

Unstructured Input Prompt (Text → Text), $P_{\text{Text}\to\text{Text}}$:

```
[Goal]: The following is a series of natural language captions describing
the subject's behavior leading up to an abnormal event.

Your task is to predict the most likely next action that the subject will
take right before the anomaly occurs. The prediction should be in natural
language.

[Input - Captions Before Anomaly]:
{captions_text}

[Constraint]:
- Predict exactly **one sentence** that describes the subject's next likely
    action.
- Your output should be **one complete sentence**.

[Output - Predicted Caption]:
```

This design ensures that the only difference between the two conditions is the input format, not the task instruction, thus isolating the effect of structured vs. unstruc-

doi:10.6342/NTU202504087

tured information.

3.4.5 Prompted Generation via LLM

In the temporal semantic forecasting experiment, the prediction process was implemented as a forward pass through a LLM, where a context window preceding the anomaly was transformed into a prompt using one of two predefined input-output pathways. The LLM was then asked to generate a single natural language sentence describing the most likely next action.

The generation step can be formally described as:

$$\hat{y}_{\text{text}} = \Phi_{\text{LLM}}\left(P_{\mathcal{I}}\right), \quad \text{where } \mathcal{I} \in \{\text{TKG} \rightarrow \text{Text}, \text{Text} \rightarrow \text{Text}\} \tag{3.1}$$

Here, $\hat{y}_{\text{text}}$ denotes the predicted sentence generated by the LLM Φ_{LLM}, given a prompt $P_{\mathcal{I}}$ constructed from temporal context. The variable $\mathcal{I}$ defines the input-output prompting pathway: either a structured input transformed via $P_{\text{TKG}\rightarrow\text{Text}}$ or an unstructured input passed through $P_{\text{Text}\rightarrow\text{Text}}$. In both cases, the prompt serves as a compact representation of the subject's behavior leading up to the anomaly, and the model is constrained to output exactly one complete sentence.

3.4.6 Metrics

To quantitatively assess how well the predicted sentence $\hat{y}_{\text{text}}$ aligns with the human-annotated ground-truth caption y_{GT}, this experiment adopts a sentence-level semantic similarity metric. Specifically, the similarity is computed using cosine similarity between dense vector representations of the predicted and ground-truth texts, as encoded by the BAAI General Embedding (BGE) model.[5]

Each sentence is passed through the BGE encoder $\mathcal{E}_{\text{BGE}}(\cdot)$ to produce a high-dimensional semantic embedding. The final similarity score is then computed as:

$$\text{Similarity} = \cos\left(\mathcal{E}_{\text{BGE}}(\hat{y}_{\text{text}}),\ \mathcal{E}_{\text{BGE}}(y_{\text{GT}})\right) \tag{3.2}$$

[5] https://huggingface.co/BAAI/bge-m3#bge-m3-paper-code

doi:10.6342/NTU202504087

Cosine similarity measures the angle between the two sentence vectors, where a score closer to 1 indicates stronger semantic alignment. This metric captures paraphrastic similarity without requiring exact lexical overlap, making it particularly suitable for evaluating LLM-generated free-text outputs.

For each input condition (structured vs. unstructured), semantic similarity scores are computed at the segment level and then averaged to report a mean similarity value across the entire evaluation set.

3.5 Experiment 2: Anomaly Detection

3.5.1 Objective

The anomaly detection experiment evaluates how well LLMs can detect abnormal or out-of-context events in surveillance video descriptions under different types of temporal input. Given a series of annotated frames and their descriptions, the model must judge whether a target frame exhibits an anomaly, using prior context in different formats—no context, raw captions, summarized text, or structured TKG. This experiment focuses on the model's ability to reason over event coherence and detect pragmatic inconsistencies.

The central question is: Do structured TKGs or unstructured caption sequences provide more effective input for anomaly detection tasks in surveillance contexts?

To answer this, the experiment consists of three main stages: (1) constructing a temporal context window from the input, either as a structured sequence of TKG quadruples or as an unstructured sequence of natural language captions; (2) prompting a LLM to assign a scalar anomaly score (ranging from 0 to 1) to a predicted the next event; and (3) evaluating how well the predicted scores discriminate between normal and abnormal ground-truth labels using the Area Under the ROC Curve (AUC) metric. An overview of this experiment's design is shown in Figure 3.2.

doi:10.6342/NTU202504087

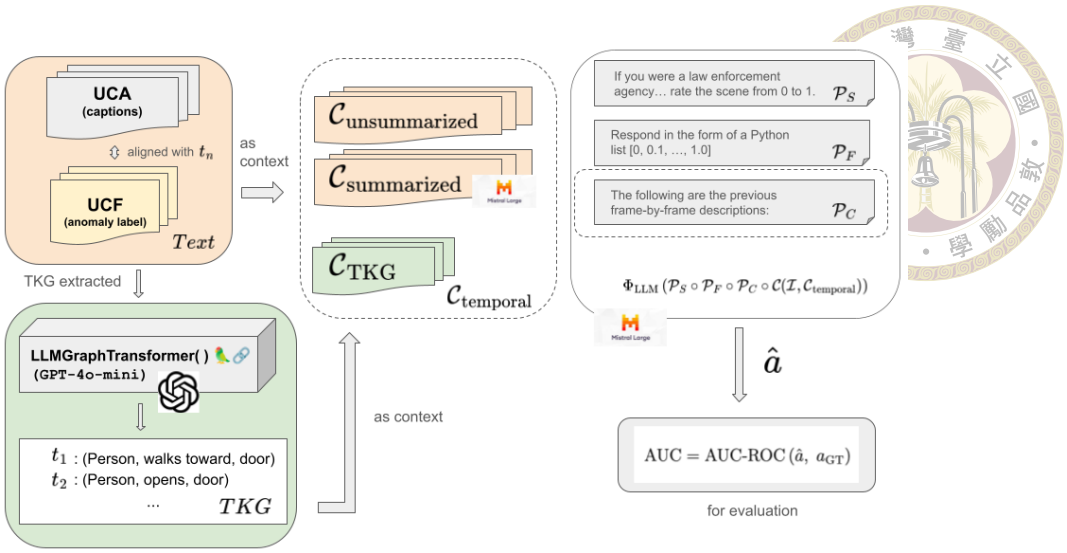


Figure 3.2: Pipeline of the Experiment 2: anomaly detection task. The model receives prior context in one of three forms—raw captions, summarized text, or structured TKGs—and predicts whether the current frame is anomalous. The prediction is compared against the ground-truth anomaly label for evaluation.

3.5.2 Prompt Design

The prompt design in this experiment is inspired by the training-free approach proposed by Zanella et al. (2024), which prompts LLMs to assign anomaly scores to video frames based on natural language descriptions. In our implementation, we adapt their scoring concept and design our own structured prompt schema to explore how different temporal inputs affect anomaly detection performance.

$\mathcal{P}_S$ is a system prompt that provides priors to the LLM regarding anomaly detecting, $\mathcal{P}_F$ instructs the LLM on the desired output format to facilitate automated text parsing, and $\circ$ is the text concatenation operation.

$\mathcal{P}_S$ was constructed as:

```
If you were a law enforcement agency,
how would you rate the scene described on a scale from 0 to 1,
with 0 representing a standard scene and 1 denoting a scene with suspicious
activities?
```

The purpose of $\mathcal{P}_S$ is to process contextual information rather than directly modeling

doi:10.6342/NTU202504087

the specific types of anomalies or providing any prior knowledge about them.

$\mathcal{P}_F$ was formed as:

```
Respond in the form of a Python list, selecting a single number from the
following list:
[0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0].
No textual explanation is required.
The response should begin with '[' and end with ']'.
```

Last but not least, $\mathcal{P}_C$ was designed as:

```
The following are the previous frame-by-frame descriptions: {contexts}.
Current frame description:
```

If the temporal context was excluded, the prompt description, ``The following are the previous frame-by-frame descriptions: contexts,'' would be omitted.

The earlier captions were aggregated as a set and incorporated into $\mathcal{P}_C$, enabling the LLM to reference the temporal context (i.e., $\mathcal{C}_{\text{temporal}}$ in Eq. (3.6)) prior to determining the anomaly in the current frame caption (i.e., $\mathcal{I}$). The temporal context can be categorized into two types: Eq. (3.3) shows that the first type consists of captions from previous frames collected into a list without any summarization by the LLM, resulting in a lack of coherence. Eq. (3.4) shows that the second type involves captions that have been further summarized by the LLM, ensuring greater coherence and consistency:

$$\mathcal{C}_{\text{unsummarized}} = \{\mathcal{C}(\mathcal{I}_{t-1}), \mathcal{C}(\mathcal{I}_{t-2}), \ldots, \mathcal{C}(\mathcal{I}_{t-n})\} \tag{3.3}$$

$$\mathcal{C}_{\text{summarized}} = \Phi_{\text{LLM}}\left(\{\mathcal{C}(\mathcal{I}_{t-1}), \mathcal{C}(\mathcal{I}_{t-2}), \ldots, \mathcal{C}(\mathcal{I}_{t-n})\}\right) \tag{3.4}$$

doi:10.6342/NTU202504087

$$\mathcal{C}_{\text{TKG}} = \{(s_1, r_1, o_1, t_1),\ (s_2, r_2, o_2, t_2),\ \ldots,\ (s_n, r_n, o_n, t_n)\} \tag{3.5}$$

$$\mathcal{C}_{\text{temporal}} = \begin{cases} \mathcal{C}_{\text{unsummarized}} & \text{if using raw captions,} \\ \mathcal{C}_{\text{summarized}} & \text{if using LLM-summarized captions,} \\ \mathcal{C}_{\text{TKG}} & \text{if using TKG quadruples.} \end{cases} \tag{3.6}$$

LLM Φ_{LLM} was asked to select only one score from a list of 11 uniformly sampled values in the interval [0, 1], where 0 means normal and 1 anomalies. The anomaly score was gotten as:

$$\Phi_{\text{LLM}}\left(\mathcal{P}_S \circ \mathcal{P}_F \circ \mathcal{P}_C \circ \mathcal{C}(\mathcal{I}, \mathcal{C}_{\text{temporal}})\right) \tag{3.7}$$

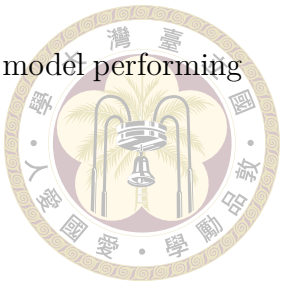
3.5.3 Metrics

To evaluate the model's ability to distinguish between normal and abnormal upcoming events, this experiment adopts the Area Under the Receiver Operating Characteristic Curve (AUC-ROC) as the primary metric. Unlike accuracy-based measures, AUC assesses the model's overall ranking capability across all possible classification thresholds, making it especially suitable for tasks involving probabilistic or scalar anomaly scores.

Each LLM output is a scalar anomaly score $\hat{a} \in [0, 1]$, selected from a fixed set of 11 evenly spaced values. These scores are compared against the binary ground-truth anomaly labels $a_{\text{GT}} \in \{0, 1\}$, as defined in the UCF-Crime dataset.

AUC is computed as the area under the ROC curve, which plots the true positive rate (TPR) against the false positive rate (FPR) across different threshold values. A

doi:10.6342/NTU202504087



model with perfect ranking ability achieves an AUC of 1.0, while a model performing at random chance yields an AUC of 0.5.

Formally, the evaluation is summarized as:

$$\text{AUC} = \text{AUC-ROC}\left(\hat{a},\ a_{\text{GT}}\right) \tag{3.8}$$

where:

- $\hat{a}$: the predicted anomaly score from the LLM
- a_{GT}: the binary ground-truth anomaly label

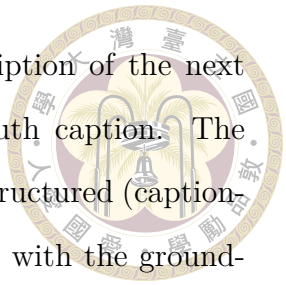
The AUC-ROC metric is threshold-independent, meaning it evaluates the model's ability to rank anomalies above normal events without relying on a fixed decision boundary. This makes it suitable for real-world surveillance applications, where the definition of "abnormal" may vary across operational settings and over time. In the context of this thesis, AUC serves as a robust evaluation criterion for comparing how well LLMs comprehend structured versus unstructured temporal inputs. By quantifying how effectively each input format enables the model to discriminate between normal and abnormal scenes—based on scalar likelihoods rather than binary thresholds—AUC captures not only the model's output behavior, but also its internal understanding of context, coherence, and deviation from normative patterns.

3.6 Summary of Experiments

This section summarizes the two experiments designed to evaluate LLMs under different abnormal events forecasting tasks: forecasting and anomaly detection tasks. While both experiments share a common structure—prompting the LLM with temporally ordered information and comparing the output to human-provided annotations—they differ in output format, evaluation metric, and reasoning objective.

Experiment 1: Forecasting

doi:10.6342/NTU202504087



This experiment assessed whether the LLM-generated description of the next event aligns semantically with the human-annotated ground-truth caption. The model was prompted using either structured (TKG-based) or unstructured (caption-based) temporal context. The generated sentence was compared with the ground-truth sentence using cosine similarity of BGE sentence embeddings. This setup captures the LLM's ability to reproduce fluent and semantically accurate predictions based on preceding context.

Experiment 2: Anomaly Detection

This experiment evaluated whether the LLM could pragmatically judge whether the next event is abnormal, again based on structured or unstructured temporal context. Instead of generating a sentence, the model produced a scalar anomaly score between 0 and 1, selected from a fixed 11-point scale. The predicted score was evaluated using AUC-ROC against ground-truth binary anomaly labels. This pipeline focuses on the model's ability to perform context-sensitive abnormality reasoning.

Together, these two experiments offer a complementary evaluation of how structured and unstructured prompts influence semantic fluency and pragmatic inference in LLMs. The forecasting experiment probes the model's ability to generate semantically coherent event descriptions, reflecting its capacity for surface-level language alignment. In contrast, the anomaly detection experiment evaluates the model's pragmatic reasoning by testing its sensitivity to contextual deviations and behavioral irregularities. By comparing performance across structured (TKG-based) and unstructured (caption-based) inputs in both tasks, this study illuminates how prompt format shapes LLMs' temporal understanding and their effectiveness in different reasoning scenarios.

doi:10.6342/NTU202504087

Chapter 4

Results

4.1 Forecasting Eperiment Results

This section reports the results of Experiment 1, which evaluated the semantic similarity between the LLM-generated sentence and the human-annotated ground-truth caption. The cosine similarity metric was used to compare predictions generated under two different input conditions:

Input Type	Mean Cosine Similarity
Unstructured (Text → Text)	**0.5978**
Structured (TKG → Text)	0.5718

Table 4.1: Mean cosine similarity scores for structured and unstructured input conditions.

The average cosine similarity score for the unstructured input condition (Text → Text) was 0.5978, while the structured input condition (TKG → Text) yielded a slightly lower average of 0.5718. This indicates that, on average, predictions based on unstructured caption input were marginally more semantically similar to the ground-truth captions than those based on structured TKG input.

To complement the aggregate similarity scores presented in Table 4.1, Figure 4.1 illustrates the per-video cosine similarity values across the two input conditions. This figure provides a more fine-grained perspective by displaying the semantic similarity between model-generated and human-annotated captions for each of the 24

doi:10.6342/NTU202504087

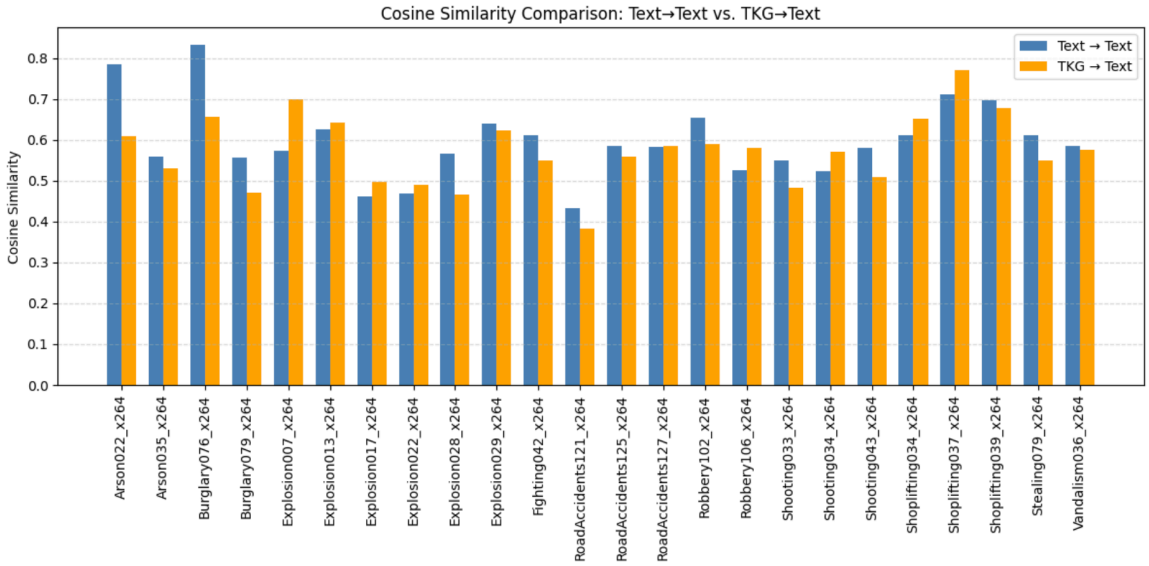


Figure 4.1: Per-video cosine similarity comparison between structured (TKG → Text) and unstructured (Text → Text) input conditions across 24 annotated surveillance clips. The plot highlights the semantic similarity between LLM-generated predictions and ground-truth captions under both input formats.

video instances. While the unstructured input condition (Text → Text) consistently demonstrates slightly higher similarity scores on average, the bar plot reveals non-trivial variation across samples. In certain videos—such as "Explosion017$_x$264" and "Shoplifting039$_x$264"—the structured input (TKG → Text) either matches or exceeds the performance of its unstructured counterpart. These observations suggest that the relative advantage of each input format may be context-dependent, and that relying solely on average metrics may obscure localized performance differences that arise due to event complexity, lexical redundancy, or input sparsity.

To evaluate whether the difference in cosine similarity between the two input conditions was statistically significant, both a parametric and a non-parametric test were employed. The paired t-test was used under the assumption of approximate normality in the distribution of similarity score differences. However, given the relatively small sample size (24 video samples), the assumption of normality may not hold. Therefore, the Wilcoxon signed-rank test was also applied as a distribution-free alternative. This dual-testing approach ensures statistical robustness and mitigates

doi:10.6342/NTU202504087

the risk of drawing unreliable conclusions based on parametric assumptions alone.

Test	Test Statistic	p-value
Paired t-test	$t = 1.7259$	0.0978
Wilcoxon signed-rank test	$W = 90.0$	0.0894

Table 4.2: Statistical test results comparing structured and unstructured input conditions.

While both statistical tests showed a trend favoring unstructured inputs, the results did not reach the conventional threshold for statistical significance ($p < 0.05$). This indicates that although the average similarity score for unstructured inputs was marginally higher, the difference is not strong enough to confirm a meaningful performance advantage. In other words, the structured (TKG $\rightarrow$ Text) and unstructured (Text $\rightarrow$ Text) input formats led to broadly comparable semantic outputs. This finding suggests that unstructured data did not perform substantially better than structured data in terms of semantic alignment.

4.2 Anomaly Detection Experiment Results

This section presents the results of Experiment 2, which evaluated the effectiveness of different temporal input representations in anomaly detection. The AUC-ROC metric was used to assess the model's ability to differentiate between normal and abnormal captions under various input conditions.

Three configurations were tested (see Figure 4.2), with a focus on evaluating how different forms of temporal context affect anomaly detection performance.

The first context condition used unsummarized raw captions from previous frames. This setting resulted in a significantly lower AUC of 56.11, suggesting that unstructured temporal inputs may introduce noise or inconsistencies, overwhelming the model rather than assisting it. In contrast, when these prior captions were summarized by an LLM before being used as input, the model's performance improved to an AUC of 64.84, indicating that temporal coherence contributes positively to anomaly detection.

doi:10.6342/NTU202504087

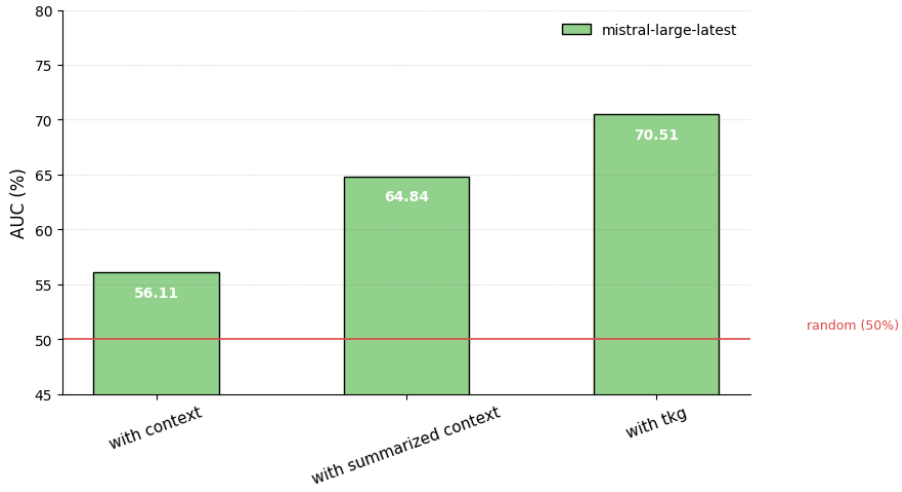


Figure 4.2: AUC performance of the Mistral model under different input configurations.

Finally, the structured input condition, where previous events were encoded as TKG quadruples, yielded the highest AUC of 70.51 among the context-based settings. This result demonstrates that explicitly structured temporal information —organized as (subject, relation, object, time)—can support more effective anomaly reasoning by reducing linguistic ambiguity, enforcing consistency, and anchoring events in time.

These findings highlight the importance of incorporating temporal context into anomaly detection. An event's abnormality often cannot be judged in isolation but must be interpreted in light of prior actions. Structured representations such as TKGs appear to offer a particularly effective way to convey this context, enabling LLMs to perform more robust and temporally informed reasoning.

doi:10.6342/NTU202504087

Chapter 5

Discussion

This chapter addresses the central research questions of this thesis by examining how different temporal input formats—structured versus unstructured—affect large language models (LLMs) in performing event-level reasoning for abnormal event forecasting. Section 5.1 interprets the experimental results in light of the proposed research questions, highlighting the comparative strengths and limitations of each input type. Section 5.2 presents a detailed error analysis to uncover systematic patterns in the model's behavior, with a particular focus on challenges related to semantic inference and pragmatic reasoning in temporal contexts.

5.1 Interpretation of Results and Research Questions

- **RQ1: Does structured temporal input (e.g., TKGs) provide advantages over unstructured input (e.g., raw captions) for forecasting tasks (Experiment 1) ?**

  The unstructured input condition yielded a slightly higher mean similarity score (0.5978) than the structured TKG input (0.5718). However, this difference was not statistically significant, as indicated by both the paired t-test (t=1.7259, p=0.0978) and the non-parametric Wilcoxon signed-rank test

doi:10.6342/NTU202504087

(W=90.0, p=0.0894). These findings suggest that, while unstructured input may produce marginally more semantically aligned outputs on average, the two input formats perform comparably in practice for this forecasting task.

This outcome raises important implications regarding the utility of structured input. TKGs offer a consistent and formal representation that abstracts away surface-level linguistic noise and encourages the model to reason based on event structure and temporal progression. This consistency may be beneficial in downstream tasks that require symbolic manipulation or reasoning. On the other hand, raw captions naturally carry richer lexical and syntactic cues, which may directly benefit tasks emphasizing surface-level semantic similarity.

In this experiment, the lack of a significant advantage for either input type highlights that the LLM is capable of handling both structured and unstructured input formats with similar effectiveness when tasked with generating semantically plausible next-event descriptions. However, the structured format may offer benefits in interpretability, consistency, and extensibility—especially in complex downstream applications where structured input facilitates symbolic reasoning and alignment with other modalities, such as video or sensor data.

Overall, while TKGs did not outperform unstructured captions in terms of raw semantic similarity, their representational strengths suggest potential advantages in more complex, reasoning-intensive applications.

- **RQ2: How does temporal context—whether structured or unstructured—impact LLM performance in anomaly detection (Experiment 2) ?**

  In the anomaly detection task (Experiment 2), the performance of LLMs varied notably depending on how prior context was provided. When previous captions were presented as a raw, unsummarized list, the AUC dropped significantly to 56.11. This indicates that directly accumulating past descriptions

doi:10.6342/NTU202504087

may introduce noise and inconsistencies, which can overwhelm the model and obscure important behavioral patterns.

However, when these prior captions were first summarized into a coherent narrative using an intermediate LLM step, performance improved markedly, reaching an AUC of 64.84. This suggests that coherence and temporal abstraction can mitigate contextual fragmentation and help the model better interpret evolving events.

Structured input in the form of TKGs, which represented past actions as (subject, relation, object, time) quadruples, led to an even stronger performance of AUC 70.51. By converting raw text into symbolic, temporally grounded representations, TKGs reduce linguistic noise, enforce structural consistency, and highlight event progression. These features support the model' s ability to track deviations from expected behavior over time.

Overall, TKGs offered a strong balance between effectiveness and temporal awareness, suggesting that structured representations are more robust for context-dependent anomaly detection task. Their consistency and abstraction appear especially valuable in helping LLMs manage longer temporal contexts and reason over dynamic event sequences.

5.2 Error Analysis

- **Experiment 1: Forecasting**

  To better understand the limitations of TKG-based input in forecasting generation, a qualitative error analysis was conducted by comparing the model's predictions against ground-truth captions. Several recurring failure patterns were observed.

  First, the model often halted at the beginning of an event sequence without successfully predicting the core anomalous action. For example, in the arson

doi:10.6342/NTU202504087

scenario where the ground truth states, *"The man returned to the Christmas tree and continued to light the Christmas tree and successfully lit it,"* the model instead only predicted a preparatory action, such as *"The man will likely try to find another ignition source,"* thereby failing to capture the actual anomalous outcome. This suggests that the model struggles to extend beyond initial event cues toward plausible consequences.

Second, some ground-truth scenarios were highly specific or unexpected, making them difficult to forecast given the available context. In one example, *"The man set the car on fire and caught fire himself,"* the model correctly anticipated the arson attempt but failed to predict that the perpetrator would also catch fire. This type of subtle and low-probability development is especially challenging when contextual signals are sparse or ambiguous.

Third, the TKG-based model frequently lacked narrative progression. It tended to anchor on high-frequency or early-occurring triplets, which led to outputs that were semantically plausible but lacked forward movement in the story. For instance, in the burglary scenario where the ground truth describes *"A total of five people gathered around the door and cooperated to pry it open,"* the model only predicted earlier contextual information such as *"The black car will stop at the house."* This stagnation may reflect a limitation of the symbolic TKG representation, which, despite encoding discrete events, often fragments the causal and temporal flow of actions—especially in scenes involving multiple agents and coordinated behavior.

Lastly, forecasting performance appeared to be task-dependent. The model showed relatively better results in scenarios with richer visual action cues and temporal continuity—such as *arson*, *shoplifting*, and *stealing*. In contrast, it struggled in more abrupt and visually ambiguous categories like *explosion*, *road accidents*, *robbery*, and *shooting*, where the anomalous events are often sudden, less visually grounded, and harder to anticipate without explicit triggers. This disparity underscores the importance of contextual richness and

doi:10.6342/NTU202504087

temporal coherence in structured input representations for effective forecasting.

- **Experiment 2: Anomaly Detection**

  Building upon the previous error analysis of the forecasting experiment, this section turns to the anomaly detection task to assess the model' s ability to distinguish between normal and abnormal segments. In particular, a qualitative analysis of prediction errors is conducted to uncover recurring failure patterns and highlight limitations in the model' s temporal and pragmatic reasoning capabilities.

  The first pattern involves error propagation from prior anomalies. When earlier frames are labeled or perceived as anomalous, the model tends to extend this anomaly judgment to subsequent frames—even if the later frames are independently benign. For example, in a frame describing *"The Christmas tree kept burning and began to emit thick smoke,"* how the model classified it as anomalous, despite the ground truth label being normal. This indicates that anomaly salience is inherited from earlier fire-related events, even when no new escalation is present.

  The second error type reflects over-sensitivity to ambiguous or suspicious scenes, particularly when individuals are described engaging in vague or cautious behavior. In the frame *"The man was pacing nearby, looking around,"* the model predicted an anomaly, although the ground truth label was normal. This suggests that the model may conflate uncertainty with threat, classifying scenes with unclear intentions as anomalous.

  A third error arises from an action-triggered bias, wherein the model tends to label any action as anomalous, especially when verbs such as "try" or "light" are involved—even if the action fails. For instance, the frame *"The man walked down and tried to light a piece of paper but failed to light it,"* was labeled as anomalous by the model, whereas the ground truth label was normal. This

doi:10.6342/NTU202504087

behavior indicates the model's inclination to associate attempted actions with anomalies, without accounting for whether those actions succeeded or posed actual risk.

Lastly, some prediction errors appear to result from annotation misalignment with video timing. That is, the semantic content of the frame may not match the temporal range assigned as anomalous in the annotation file. For example, in the frame *"The man came straight to the door,"* the ground truth label was anomalous, but the content appeared non-threatening, leading the model to classify it as normal. Such discrepancies may reflect temporal label noise introduced during manual annotation.

This error analysis shows that the model demonstrates emerging competence in pragmatic reasoning, successfully leveraging temporal context in many cases. However, it still exhibits systematic biases—such as overgeneralizing from prior anomalies, misinterpreting ambiguous scenes, or conflating attempted actions with completed ones. These patterns suggest that while the model is sensitive to context, its inference mechanisms could benefit from finer-grained distinctions between intent, outcome, and threat level.

doi:10.6342/NTU202504087

Chapter 6

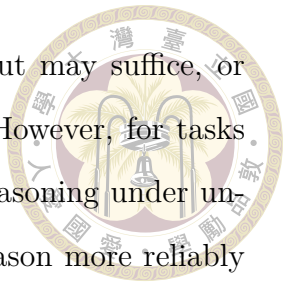
Conclusion

Recent advances in LLMs have enabled promising applications in temporal reasoning, yet questions remain regarding the optimal input format for such models—especially in contexts where data may be noisy, ambiguous, or sparsely annotated. Structured inputs, such as TKGs, offer abstraction and consistency that can mitigate semantic noise, while unstructured inputs preserve linguistic richness and contextual nuance. Understanding how these formats differentially impact LLM performance is critical for downstream applications requiring both semantic fluency and pragmatic reasoning.

This thesis conducted two experiments to examine these differences. In the forecasting task, unstructured textual input yielded slightly higher cosine similarity with ground-truth captions than TKG-based input (0.5978 vs. 0.5718), though the difference was not statistically significant. In contrast, the anomaly detection task demonstrated a clear advantage for structured input, with TKG-based representations achieving the highest AUC score (70.51). These results suggest that while unstructured inputs are expressive and effective for semantic generation, structured inputs confer greater stability, interpretability, and robustness for temporal anomaly reasoning.

The results from the two experiments designed around the abnormal event forecasting tasks in this study reveal critical trade-offs in input format selection for LLMs. For forecasting tasks requiring semantic fluency—such as narrative fore-

doi:10.6342/NTU202504087



casting or descriptive summarization—unstructured textual input may suffice, or even offer slight advantages due to its richer lexical structure. However, for tasks requiring temporal coherence, anomaly detection, or behavior reasoning under uncertainty, structured formats such as TKGs enable models to reason more reliably by abstracting away surface noise and foregrounding event regularities. These observations imply that hybrid prompting strategies—combining structured scaffolds with semantically expressive textual context—may yield the best of both worlds, especially in high-stakes domains like surveillance, legal monitoring, or real-time decision support. Moreover, the robustness of structured input under limited supervision conditions suggests its potential as a stabilizing backbone in future multimodal or low-resource LLM applications.

These findings also carry practical implications for real-world scenarios involving textual criminal case data, such as written testimonies, incident reports, or legal document analysis. In many legal or forensic contexts, information does not arrive in structured form—rather, it is often fragmented across loosely organized paragraphs or ambiguous narratives. Transforming such unstructured content into a structured format like a TKG can provide LLMs with a clear temporal scaffold for reasoning over sequences and causality. The experimental results suggest that while unstructured textual inputs may offer advantages in forecasting, TKGs yield more reliable outcomes in pragmatic inference and anomaly detection—two capabilities that are essential for real-world applications where logical consistency and traceable reasoning are paramount. In particular, TKGs may facilitate applications such as reconstructing timelines across multiple witness statements, identifying contradictory claims, or aligning legal narratives with external knowledge bases. This structured approach moves LLMs beyond surface-level language generation toward deeper, context-aware interpretive tasks.

Despite the controlled experimental design, this study has several limitations. The generalizability of the results is constrained by the choice of models (e.g., Mistral-large-latest) and dataset (UCF-Crime/UCA), which may not reflect per-

doi:10.6342/NTU202504087

formance across domains, modalities, or model scales. The construction of temporal knowledge graphs (TKGs) involved LLM-based extraction and heuristic parsing, introducing potential noise and inconsistencies in structured inputs. Evaluation metrics—cosine similarity for semantic alignment and AUC for anomaly detection—capture only surface-level or threshold-independent performance and do not address deeper pragmatic correctness or real-world constraints. Additionally, this study focused on relatively simplified forms of temporal reasoning, leaving more complex tasks such as causal inference or long-range event tracking for future exploration.

Future work may explore several directions to build upon the current findings. One promising avenue involves incorporating multimodal signals—such as visual and auditory features—into the input space, potentially enriching the temporal grounding of abnormal events. Another area lies in enhancing the reasoning capabilities of LLMs by integrating causal inference or event coreference tracking, which may enable more robust temporal abstraction and narrative coherence. Refining the construction of structured representations like TKGs, especially through improved context-aware extraction methods, could also mitigate input noise and boost reasoning fidelity. Additionally, aligning structured and unstructured inputs through methods such as hybrid prompting or contrastive learning holds potential for more effectively leveraging the complementary strengths of both input types in complex, real-world applications.

In conclusion, the investigation of structured and unstructured inputs for LLM-based temporal event reasoning contributes to a deeper understanding of how language models interpret, generate, and evaluate time-sensitive behavioral data. While challenges remain in capturing temporal nuances and aligning representations with human expectations, this study demonstrates that input format plays a crucial role in shaping semantic and pragmatic model behavior. As research continues to evolve, future LLM applications may benefit from hybrid prompting strategies and enriched temporal abstractions, enabling more robust, interpretable, and generalizable reasoning systems across diverse domains.

doi:10.6342/NTU202504087

Limitations

Despite the structured experimental framework and comparative evaluation conducted in this thesis, several limitations should be acknowledged to provide a realistic understanding of the findings and their scope.

- **Model and Dataset Constraints**

  This study employed GPT-4-turbo-mini for TKG extraction and Mistral-large for generation, which represent only a small subset of LLMs. Results may not generalize to smaller, domain-adapted, or multimodal models. The dataset, based on surveillance captions, also lacks linguistic diversity found in other domains such as legal transcripts.

- **Noisy TKG Construction**

  Temporal knowledge graphs were generated via semi-automatic extraction using LLMs and heuristic rules. This method can introduce noise, including misidentified relations or temporal scopes, potentially weakening the quality of the structured inputs.

- **Limited Evaluation Metrics**

  Semantic evaluation relied solely on cosine similarity of sentence embeddings, which does not capture factual accuracy or pragmatic fit. Anomaly detection used AUC-ROC, which lacks insight into precision, recall, or detection latency —key factors in real-world applications.

- **Simplified Temporal Reasoning**

  Temporal reasoning was approximated by sequential caption aggregation and optional structuring. Advanced reasoning forms—like causal inference or multi-event temporal logic—were not addressed, leaving room for future investigation.

doi:10.6342/NTU202504087

Appendix A

Aligned Samples from UCF-Crime and UCA

video_type	timestamp	caption (text and tkg format)	anomalous
Arson	81.3 - 106	The man walked down and tried to light a piece of paper but failed to light it.	FALSE
		t7 : [Man, WALKED_DOWN, Paper], [Man, TRIED_TO_LIGHT, Paper], [Man, FAILED_TO_LIGHT, Paper]	
	115.8 - 121.2	The man returned to the Christmas tree and continued to light the Christmas tree and successfully lit it.	TRUE
		t8 : [Man, RETURNED_TO, Christmas Tree], [Man, CONTINUED_TO_LIGHT, Christmas Tree], [Man, SUCCESSFULLY_LIT, Christmas Tree]	
Burglary	254.4 - 255.8	Another person opened the trunk, and there were several men in white hiding in the trunk.	FALSE
		t14 : [Another Person, HIDING_IN, Men In White]	
	256.1 - 350.4	A total of five people gathered around the door and cooperated to pry it open.	TRUE

doi:10.6342/NTU202504087

		t15 : [People, GATHERED_AROUND, Door], [People, COOPERATED_TO_PRY_OPEN, Door]	
Explosion	0.0 - 9.0	There are many cars parked on the roadside and many people walking on the roadside.	FALSE
		t1 : [Cars, PARKED_ON, Roadside], [People, WALKING_ON, Roadside]	
	9.0 - 21.3	An explosion occurred in a distant building and produced smoke, and the glass of the building next to it was shaken.	TRUE
		t2 : [Explosion, OCCURRED_IN, Building], [Explosion, PRODUCED, Smoke], [Building, SHAKEN, Glass]	
Fighting	0.0 - 8.2	Security patrolling the door	FALSE
		t1 : [Security, PATROLLING, Door]	
	8.1 - 25.4	A man wearing a hat approaches and fights with the security guard	TRUE
		t2 : [Man Wearing A Hat, FIGHTS, Security Guard]	
RoadAccidents	2.3 - 5.5	A man in pink pants walked by	FALSE
		t1 : [Man, WORE, Pink Pants]	
	7.3 - 13.5	A black car was hit by a white car while passing through the alley, and a woman was also knocked to the ground.	TRUE
		t2 : [Black Car, HIT, White Car], [Woman, KNOCKED_DOWN, Ground], [Black Car, PASSING_THROUGH, Alley]	
Robbery	9.0 - 12.2	A motorcycle with two men parked next to it.	FALSE
		t2 : [Man1, PARKED_NEXT_TO, Motorcycle], [Man2, PARKED_NEXT_TO, Motorcycle]	

doi:10.6342/NTU202504087

	14.1 - 17.9	The man in blue on the back seat of the motorcycle got off the car and took out a gun and pointed it at the man on the phone.	TRUE
		t3 : [Man In Blue, RIDES, Motorcycle], [Man In Blue, USES, Gun], [Man In Blue, POINTS_AT, Man On Phone], [Man In Blue, GETS_OFF, Car]	
Shooting	0.0 - 24.6	There was a black off-road vehicle parked in front of a store with two men in black clothes.	FALSE
		t1 : [Off-Road Vehicle, PARKED_IN_FRONT_OF, Store], [Men, IN, Off-Road Vehicle], [Men, IN_FRONT_OF, Store]	
	24.6 - 36.7	A man in red clothes raised a gun towards the man in black clothes	TRUE
		t2 : [Man In Red Clothes, RAISES_GUN_TOWARDS, Man In Black Clothes]	
Shoplifting	64.9 - 70.8	The man wearing glasses walked out	FALSE
		t5 : [Man Wearing Glasses, WALKED, Out]	
	71.9 - 78.7	The man in short sleeves moved a box of things	TRUE
		t6 : [Man In Short Sleeves, MOVED, Box Of Things]	
Stealing	79.9 - 81.4	A white car passed by on the road	FALSE
		t12 : [White Car, PASSED_BY, Road]	
	81.4 - 99.8	The man in the striped shirt walked to the middle of the black iron gate of the yard and opened it.	TRUE
		t13 : [The Man In The Striped Shirt, WALKED_TO, The Black Iron Gate], [The Man In The Striped Shirt, OPENED, The Black Iron Gate]	

doi:10.6342/NTU202504087

Vandalism	0.0 - 8.6	A naked man walked into the store, walked to the counter, held the counter with both hands and looked around	FALSE
		t1 : [Naked Man, ENTERED, Store], [Naked Man, WALKED_TO, Counter], [Naked Man, HELD, Counter], [Naked Man, LOOKED_AROUND, Store]	
	14.3 - 18.2	The shirtless man glanced inside the counter and suddenly threw an item on the counter to the ground.	TRUE
		t2 : [Shirtless Man, GLANCED_AT, Counter], [Shirtless Man, THREW, Item], [Item, PLACED_ON, Counter]	

doi:10.6342/NTU202504087

References

Alnegheimish, S., Nguyen, L., Berti-Equille, L., & Veeramachaneni, K. (2024). Large language models can be zero-shot anomaly detectors for time series? *arXiv preprint arXiv:2405.14755*.

Barrasa, J., & Webber, J. (2023). *Building knowledge graphs*. " O'Reilly Media, Inc."

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., et al. (2021). On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

Cai, L., Mao, X., Zhou, Y., Long, Z., Wu, C., & Lan, M. (2024). A survey on temporal knowledge graph: representation learning and applications. *arXiv preprint arXiv:2403.04782*.

Carta, S., Giuliani, A., Piano, L., Podda, A. S., Pompianu, L., & Tiddia, S. G. (2023). Iterative zero-shot llm prompting for knowledge graph construction. *arXiv preprint arXiv:2307.01128*.

Chang, H., Wu, J., Tao, Z., Ma, Y., Huang, X., & Chua, T.-S. (2025). Integrate temporal graph learning into llm-based temporal knowledge graph model. https://arxiv.org/abs/2501.11911

Chang, H., Ye, C., Tao, Z., Wu, J., Yang, Z., Ma, Y., Huang, X., & Chua, T.-S. (2024). A comprehensive evaluation of large language models on temporal event forecasting. *arXiv preprint arXiv:2407.11638*.

Chen, X., Wang, Y., Li, L., & Huang, M. (2023). Kg-llm: injecting knowledge graphs into large language models. *ACL*.

Dang, W., Zhou, B., Wei, L., Zhang, W., Yang, Z., & Hu, S. (2021). Ts-bert: time series anomaly detection via pre-training model bert. *Computational Science–*

doi:10.6342/NTU202504087

ICCS 2021: 21st International Conference, Krakow, Poland, June 16–18, 2021, Proceedings, Part II 21, 209–223.

Deng, S., Rangwala, H., & Ning, Y. (2020). Dynamic knowledge graph based multi-event forecasting. *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, 1585–1595.

Deng, S., Rangwala, H., & Ning, Y. (2021). Understanding event predictions via contextualized multilevel feature learning. *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, 342–351.

Gao, Y., Qiao, L., Kan, Z., Wen, Z., He, Y., & Li, D. (2024). Two-stage generative question answering on temporal knowledge graph using large language models. *arXiv preprint arXiv:2402.16568*.

Gastinger, J., Sztyler, T., Sharma, L., & Schuelke, A. (2022). On the evaluation of methods for temporal knowledge graph forecasting. *NeurIPS 2022 Temporal Graph Learning Workshop*.

Goel, R., Kazemi, S. M., Brubaker, M., & Poupart, P. (2020). Diachronic embedding for temporal knowledge graph completion. *Proceedings of the AAAI conference on artificial intelligence*, *34*(04), 3988–3995.

Gruver, N., Finzi, M., Qiu, S., & Wilson, A. G. (2023). Large language models are zero-shot time series forecasters. *Advances in Neural Information Processing Systems*, *36*, 19622–19635.

Han, X., Liu, Z., & Sun, M. (2021). Robustness of temporal knowledge graph forecasting models to sparsity. *Findings of EMNLP*.

Huang, H., Chen, C., He, C., Li, Y., Jiang, J., & Zhang, W. (2024). Can llms be good graph judger for knowledge graph construction? *arXiv preprint arXiv:2411.17388*.

Huang, S., Liu, Y., Fung, C., Wang, H., Yang, H., & Luan, Z. (2023). Improving log-based anomaly detection by pre-training hierarchical transformers. *IEEE Transactions on Computers*, *72*(9), 2656–2667.

doi:10.6342/NTU202504087

Ji, S., Pan, S., Cambria, E., Marttinen, P., & Yu, P. S. (2021). A survey on knowledge graphs: representation, acquisition and applications. *IEEE Transactions on Neural Networks and Learning Systems*.

Jin, D., Balazevic, I., Allen, C., & Hospedales, T. M. (2020). Re-net: reasoning over knowledge graph paths for temporal knowledge base completion. *AAAI*.

Jin, M., Wang, S., Ma, L., Chu, Z., Zhang, J. Y., Shi, X., Chen, P.-Y., Liang, Y., Li, Y.-F., Pan, S., et al. (2023). Time-llm: time series forecasting by reprogramming large language models. *arXiv preprint arXiv:2310.01728*.

Jin, M., Wen, Q., Liang, Y., Zhang, C., Xue, S., Wang, X., Zhang, J., Wang, Y., Chen, H., Li, X., Pan, S., Tseng, V. S., Zheng, Y., Chen, L., & Xiong, H. (2023). Large models for time series and spatio-temporal data: a survey and outlook. https://arxiv.org/abs/2310.10196

Kejriwal, M. (2019). *Domain-specific knowledge graph construction*. Springer.

Leblay, J., & Chekol, M. W. (2018). Deriving validity time in knowledge graph. *Companion proceedings of the the web conference 2018*, 1771–1776.

Lee, D.-H., Ahrabian, K., Jin, W., Morstatter, F., & Pujara, J. (2023). Temporal knowledge graph forecasting without knowledge using in-context learning. *arXiv preprint arXiv:2305.10613*.

Lee, Y., Kim, J., & Kang, P. (2023). Lanobert: system log anomaly detection based on bert masked language model. *Applied Soft Computing*, *146*, 110689.

Leetaru, K., & Schrodt, P. A. (2013). Gdelt: global data on events, location, and tone, 1979–2012. *ISA annual convention*, *2*(4), 1–49.

Li, X., Cheng, L., Tan, Q., Tou Ng, H., Joty, S., & Bing, L. (2023). Unlocking temporal question answering for large language models using code execution. *arXiv e-prints*, arXiv–2305.

Li, Z., Hou, Z., Guan, S., Jin, X., Peng, W., Bai, L., Lyu, Y., Li, W., Guo, J., & Cheng, X. (2022). Hismatch: historical structure matching based temporal knowledge graph reasoning. *arXiv preprint arXiv:2210.09708*.

doi:10.6342/NTU202504087

Liao, R., Jia, X., Li, Y., Ma, Y., & Tresp, V. (2023). Gentkg: generative forecasting on temporal knowledge graph with large language models. *arXiv preprint arXiv:2310.07793*.

Liu, L., Wang, Z., & Tong, H. (2024). Neural-symbolic reasoning over knowledge graphs: a survey from a query perspective. *arXiv preprint arXiv:2412.10390*.

Liu, N., Zheng, K., Dohan, D., et al. (2023). Lost in the middle: how language models use long contexts. *EMNLP*.

Liu, Y., Ma, Y., Hildebrandt, M., Joblin, M., & Tresp, V. (2022). Tlogic: temporal logical rules for explainable link forecasting on temporal knowledge graphs. *Proceedings of the AAAI conference on artificial intelligence*, *36*(4), 4120–4127.

Luo, R., Gu, T., Li, H., Li, J., Lin, Z., Li, J., & Yang, Y. (2024). Chain of history: learning and forecasting with llms for temporal knowledge graph completion. *arXiv preprint arXiv:2401.06072*.

Ma, W., Ding, Y., Yasunaga, M., & Leskovec, J. (2021). Temporal knowledge graph forecasting with sampled historical subgraphs. *EMNLP*.

Ma, Y., Ye, C., Wu, Z., Wang, X., Cao, Y., & Chua, T.-S. (2023). Context-aware event forecasting via graph disentanglement. *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 1643–1652.

Ma, Y., Ye, C., Wu, Z., Wang, X., Cao, Y., Pang, L., & Chua, T.-S. (2023). Structured, complex and time-complete temporal event forecasting. *CoRR*.

O'brien, S. P. (2010). Crisis early warning and decision support: contemporary approaches and thoughts on future research. *International studies review*, *12*(1), 87–104.

Ong, R., Sun, J., Șerban, O., & Guo, Y.-K. (2023). Tkgqa dataset: using question answering to guide and validate the evolution of temporal knowledge graph. *Data*, *8*(3), 61.

Schlichtkrull, M., Kipf, T. N., Bloem, P., van den Berg, R., Titov, I., & Welling, M. (2017). Modeling relational data with graph convolutional networks. https://arxiv.org/abs/1703.06103

doi:10.6342/NTU202504087

Su, J., Jiang, C., Jin, X., Qiao, Y., Xiao, T., Ma, H., Wei, R., Jing, Z., Xu, J., & Lin, J. (2024a). Large language models for forecasting and anomaly detection: a systematic literature review. https://arxiv.org/abs/2402.10350

Su, J., Jiang, C., Jin, X., Qiao, Y., Xiao, T., Ma, H., Wei, R., Jing, Z., Xu, J., & Lin, J. (2024b). Large language models for forecasting and anomaly detection: a systematic literature review. *arXiv preprint arXiv:2402.10350*.

Sultani, W., Chen, C., & Shah, M. (2018). Real-world anomaly detection in surveillance videos. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 6479–6488.

Trivedi, R., Dai, H., Wang, Y., & Song, L. (2017). Know-evolve: deep temporal reasoning for dynamic knowledge graphs. https://arxiv.org/abs/1705.05742

Trivedi, R., Farajtabar, M., Biswal, P., & Zha, H. (2018). Representation learning over dynamic graphs. https://arxiv.org/abs/1803.04051

Trivedi, R., Faruqui, M., Dauphin, Y., & Yogatama, D. (2017). Know-evolve: deep temporal reasoning for dynamic knowledge graphs. *ICML*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.

Wang, D., Zuo, Y., Li, F., & Wu, J. (2024). Llms as zero-shot graph learners: alignment of gnn representations with llm token embeddings. *Advances in Neural Information Processing Systems*, *37*, 5950–5973.

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., et al. (2022). Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*.

Xiong, S., Payani, A., Kompella, R., & Fekri, F. (2024). Large language models can learn temporal reasoning. *arXiv preprint arXiv:2401.06853*.

Xiong, S., Yang, Y., Fekri, F., & Kerce, J. C. (2024). Tilp: differentiable learning of temporal logical rules on knowledge graphs. *arXiv preprint arXiv:2402.12309*.

doi:10.6342/NTU202504087

Xiong, S., Yang, Y., Payani, A., Kerce, J. C., & Fekri, F. (2024). Teilp: time prediction over knowledge graphs via logical reasoning. *Proceedings of the AAAI Conference on Artificial Intelligence*, *38*(14), 16112–16119.

Xue, H., & Salim, F. D. (2023). Promptcast: a new prompt-based learning paradigm for time series forecasting. *IEEE Transactions on Knowledge and Data Engineering*, *36*(11), 6851–6864.

Xue, H., Voutharoja, B. P., & Salim, F. D. (2022). Leveraging language foundation models for human mobility forecasting. *Proceedings of the 30th International Conference on Advances in Geographic Information Systems*, 1–9.

Yuan, T., Zhang, X., Liu, K., Liu, B., Chen, C., Jin, J., & Jiao, Z. (2023). Towards surveillance video-and-language understanding: new dataset, baselines, and challenges. https://arxiv.org/abs/2309.13925

Zanella, L., Menapace, W., Mancini, M., Wang, Y., & Ricci, E. (2024). Harnessing large language models for training-free video anomaly detection. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18527–18536.

Zha, Z., Qi, P., Bao, X., Tian, M., & Qin, B. (2024). M 3 tqa: multi-view, multi-hop and multi-stage reasoning for temporal question answering. *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 10086–10090.

Zhang, M., Sun, M., Wang, P., Fan, S., Mo, Y., Xu, X., Liu, H., Yang, C., & Shi, C. (2024). Graphtranslator: aligning graph model to large language model for open-ended tasks. *Proceedings of the ACM Web Conference 2024*, 1003–1014.

Zhang, T., Huang, X., Zhao, W., Bian, S., & Du, P. (2023). Logprompt: a log-based anomaly detection framework using prompts. *2023 International Joint Conference on Neural Networks (IJCNN)*, 1–8.

Zhang, Y., Chen, Z., Guo, L., Xu, Y., Zhang, W., & Chen, H. (2024). Making large language models perform better in knowledge graph completion. *Proceedings of the 32nd ACM International Conference on Multimedia*, 233–242.

doi:10.6342/NTU202504087

Zhao, L., Song, Y., Zhang, C., Liu, Y., Wang, P., Lin, Y., & Deng, M. (2019). T-gcn: a temporal graph convolutional network for traffic prediction. *AAAI*.

Zhou, Z., & Yu, R. (2024). Can llms understand time series anomalies? *arXiv preprint arXiv:2410.05440*.

doi:10.6342/NTU202504087