World Wide Web (2025) 28:60
https://doi.org/10.1007/s11280-025-01359-7

Context-driven cold-start Web traffic forecasting

Xin Zhou[1] · Weiqing Wang[1] · Wray Buntine[2] · Christoph Bergmeir[1,3]

Received: 9 March 2025 / Revised: 14 June 2025 / Accepted: 16 June 2025

Abstract
Cold-start forecasting is critical in dynamic scenarios where early-stage forecasting drives key decisions, such as content prioritization, resource allocation, and demand estimation before observable trends emerge. In this work, we explore the potential of multimodal forecasting techniques for cold-start forecasting and offer insights into designing more scalable and adaptive models. In particular, we address context-driven cold-start web traffic forecasting that includes textual content and historical web traffic of relevant web pages to generate forecasts when no historical data is available for the target new web page. To advance research in this area, we collect, clean, and align a high-dimensional, multimodal web traffic dataset. We adopt a Retrieval-Augmented Generation framework, and propose the use of large language models (LLMs) for this task. Our experiments demonstrate that the LLM-based strategy consistently outperforms the statistical baseline across multiple forecasting horizons. The best-performing LLM-based model reduces WRMSPE by 0.81% and WAPE by 4.5%, compared with other methods. Furthermore, LLM-based feature extraction enhances contextual understanding, leading to greater stability in long-horizon forecasts.

Keywords Web traffic analysis · Cold-start · Multimodal · Time series forecasting

1 Introduction

Accurate web traffic forecasting is essential for optimizing server performance, managing resources, and improving user experiences [1–4]. It helps businesses anticipate peak periods,

✉ Weiqing Wang
teresa.wang@monash.edu

Xin Zhou
xin.zhou@monash.edu

Wray Buntine
wray.b@vinuni.edu.vn

Christoph Bergmeir
christoph.bergmeir@monash.edu

[1] Faculty of Information Technology, Monash University, Wellington Rd, Melbourne 3800, Victoria, Australia

[2] College of Engineering and Computer Science, VinUniversity, Đa Tôn, Gia Lâm, Hà Nôi, Vietnam

[3] Department of Computer Science and Artificial Intelligence, University of Granada, Av. del Hospicio, 1, Granada 18012, Spain

Published online: 30 September 2025

prevent overloads, and improve resources allocation [5–7]. In addition, reliable forecasts support cost reduction, and enhanced user engagement [8], ensuring a competitive edge in today's digital landscape.

Cold-start web traffic forecasting refers to the challenge of forecasting the web traffic for newly created or previously untracked web pages, where no historical data is available for the target page. This problem is critical in various real-world applications, particularly in digital content management [9], and e-commerce [10], where accurate early-stage forecasting informs decisions on content promotion, and infrastructure scaling [11]. Most of the existing forecasting models [12–14] rely heavily on historical observations of the target time series to capture temporal dependencies such as trends, seasonality, and autocorrelation. Recent advances have explored cross-modality alignment [15, 16], spatio-temporal training [17–19], and LLM-based traffic forecasting [20], yet these methods still rely on partial or full historical observations. In contrast, cold-start forecasting requires alternative sources of information, such as contextual data extracted from textual descriptions and traffic patterns of relevant pages. Addressing this challenge is essential for optimizing search engine indexing strategies, improving recommendation systems [10], and enhancing user engagement by prioritizing new content based on projected demand [21]. As an underexplored task, cold-start web traffic forecasting is fundamentally different not only from general time series forecasting, but also from existing well-studied tasks such as zero-shot forecasting, few-shot forecasting, and few-history forecasting, which are further clarified in Section 4.1. By establishing a methodological foundation for this task, we pave the way for more adaptive and scalable forecasting solutions in dynamically evolving online environments.

The technical solutions of web traffic forecasting build on existing time series forecasting models [22, 23], which range from univariate and multivariate models [24, 25] to more recent multimodal approaches, as shown in Figure 1(a)-(c). However, all these methods cannot be leveraged to solve the cold-start task as this task has fundamentally different training data to be leveraged as shown in Figure 1(d). Traditional time series forecasting methods [12, 14, 26] rely on prior knowledge of time series patterns such as trends, seasonality, and autocorrelation, making them effective when sufficient historical data is available. However, they struggle with scalability and non-linear behavior, limiting their use in complex forecasting tasks. Advances in deep learning [27–31], such as Recurrent Neural Networks [32], Convolutional Neural Networks [33], transformers [28, 34–38], and Large Language Models (LLMs) [39,

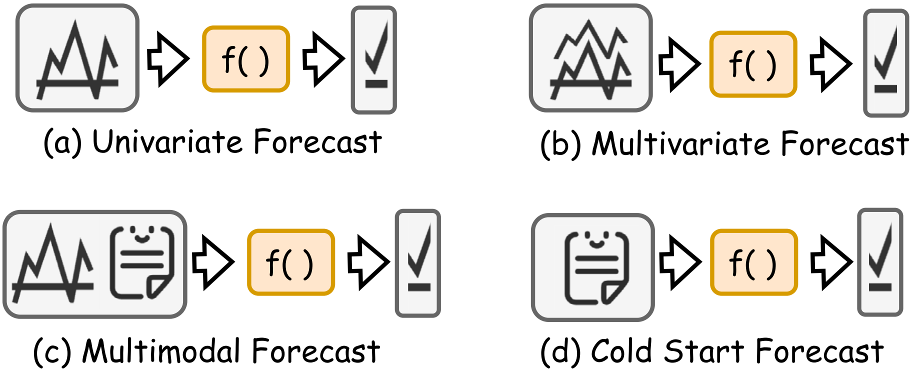


Figure 1 Comparison of tasks in web traffic forecasting. (a)-(c) are the most general tasks that existing works focus on, while (d) is the cold-start task that integrates contextual data

Table 1 Multimodal Web Traffic Datasets

Dataset	Time Information		Channels	Other Modalities	Available
	Time Range	Points			
Kaggle [23]	2015/07/01–2016/12/30	548	145,000	—	✓
GP-Copula [44]	—	365	9,013	—	✓
GPT4MTS [41]	2022/08/17–2023/07/31	349	30	time-relevant text	✗
WikiPopular	2019/07/01–2024/06/30	1,826	728	time-irrelevant text	✓

40], have improved the ability to handle complex, long series. Some multivariate [28, 34] and multimodal models [41, 42] incorporate contextual information, e.g., historical web traffic of relevant web pages or textual news summaries. However, they still depend heavily on historical data from the target series and perform poorly in cold-start scenarios, as shown in Figure 1(d).

In this work, we narrow the gap, as existing models often depend on historical time series data or focus on zero-shot and transfer learning, which are insufficient for true cold-start scenarios. Specifically, we,

- Formalize the task: Cold-Start Web Traffic Forecasting, driven by real-world demands to address scenarios involving new or untracked pages. This task emphasizes the use of contextual data, including textual content and historical traffic of relevant web pages, to enable accurate forecasting without relying on historical data from the target page;
- Publish the first high-dimensional, multimodal web traffic dataset that can be tailored for cold-start forecasting. The existing dataset are not high-dimensional or static multimodal, which make them unsuitable for cold-start forecasting, as shown in Table 1. WikiPopular provides a benchmark for further research and promots the development of new methods in the field[1].
- Develop a Retrieval-Augmented Generation (RAG) framework combining the statistical and the LLM-based strategy to tackle the cold-start challenge. By integrating multimodal contextual data, such as textual content and web traffic of relevant web pages, our framework enhances forecasting and provides standardized resources to support future studies.

Our work addresses cold-start forecasting when no historical data is available, tackling an underexplored problem in web traffic forecasting. Unlike zero-shot learning and few-shot learning, this problem requires a fundamentally different model architecture. Using the RAG framework with contextual data, we demonstrate how forecasting can be improved, providing a scalable and adaptive solution for real-world applications. This research advances the capabilities of web traffic forecasting and paves the way for future exploration in areas such as e-commerce, social media, and news trend prediction, where new content often emerges without historical data.

2 Related works

Cold-start forecasting presents unique challenges distinct from general time series forecasting, as it requires forecasting without any historical data for the target series. Existing

[1] To support further research and benchmarking, we release the codes and data at https://github.com/xinzzzhou/CCWTF.git

time series forecasting methods, including both traditional approaches and deep learning models, rely heavily on past observations, making them unsuitable for cold-start scenarios. While recent advances in multimodal learning and LLMs have improved forecasting capabilities, these approaches still largely depend on historical time series data, limiting their effectiveness in real-world cold-start applications. Other recent works in spatio-temporal forecasting [16–20] similarly assume access to target history, limiting their applicability to cold-start scenarios.

2.1 General time series forecasting

Statistical forecasting methods, such as AutoRegressive Integrated Moving Average (ARIMA) [12], Exponential Smoothing [26], and their variants [13], rely heavily on prior knowledge of time series patterns, such as autocorrelation, trends, and seasonality, to configure model parameters and generate accurate forecasts. Although these methods are effective for stable and linear data patterns, they struggle with non-linear or unexpected behaviors. For multivariate forecasting, models like Vector Auto-Regression [14, 43] extend ARIMA by capturing linear relationships across multiple time series, but their performance deteriorates when inter-series dependencies become nonlinear or complex, limiting their suitability for dynamic or high-dimensional data.

Recent advances [27, 28, 28–34] in deep learning have transformed time series forecasting by capturing complex temporal patterns beyond the limitations of traditional models. These approaches can be broadly categorized into four types. RNN-based models, including LSTM-based architectures [32, 44, 45], are adept at capturing long-term dependencies in sequential data but often struggle with vanishing gradients and processing long sequences efficiently. Convolutional models [33] capture temporal and channel dependencies through local patterns, providing a faster alternative to RNNs and excelling in multivariate forecasting tasks. Attention-based models, such as TST [35], FEDformer [36], Informer [37], and iTransformer [28], leverage self-attention mechanisms to capture long-range dependencies and dynamic temporal relationships, achieving state-of-the-art performance in time series forecasting.

2.2 Large models for time series forecasting

Large Models have shown promise in both multimodal and time series forecasting tasks. Multimodal models, such as GPT4MTS [41], MoAT [42], and TimeCMA [15], leverage LLMs and decomposition methods to incorporate external news into time series forecasting tasks. However, they primarily align textual content with time-stamped events, limiting their applicability to cold-start scenarios, where no historical target data exists.

Recent works, including TEST [39] and TEMPO [40], harness the power of LLMs to capture complex patterns and embed knowledge from extensive datasets. These models, with their enhanced learning capacities, excel at generalizing across diverse time series tasks. Time series foundation models have been used for zero-shot and few-shot forecasting. LagLlama [46] proposes a general univariate probabilistic time series forecasting model trained on a large collection of time series data. However, the use of LLMs for zero-shot time series forecasting is still in its early stages. Recent advancements in zero-shot learning for time series forecasting [47] have primarily tested models on previously unseen data while still utilizing time series inputs. Despite progress, zero-shot forecasting remains underexplored in contexts such as web traffic forecasting, particularly with the increasing availability of publicly accessible, high-dimensional multimodal datasets. Fine-tuning LLMs for specific forecasting tasks is time-consuming and computationally expensive, leaving room for further exploration of lightweight, training-free strategies that can leverage these models for more efficient predictions. UniTime [48] proposes a unified model for cross-domain time series forecasting, which can flexibly adapt to data with varying characteristics. However, these models using time series data only transferring knowledge learned from one domain to another remains difficult due to fundamental differences in underlying data characteristics [49]. This limitation becomes especially apparent in cold-start scenarios, where no historical traffic data exists to provide temporal context.

2.3 Spatio-temporal forecasting

Several recent works propose frameworks for streaming or spatio-temporal forecasting under limited data scenario [16–20]. These methods integrate techniques such as replay-based continual learning [17], federated learning [18], LLM-driven spatial modeling [20], or dataset condensation [19] to enhance generalization under changing or distributed data conditions. However, these approaches typically rely on having at least partial temporal observations for the target instance. For instance, the federated continual learning framework in [18] and uncertainty quantification models in [16] are both designed around the assumption that target traffic series are observable, even if intermittently. This reliance on historical data prevents their direct application to cold-start settings where the temporal signal is entirely absent.

3 Data

Existing research on cold-start forecasting has predominantly focused on domains such as new product forecasting [50], where models predict future demand without historical sales data. However, these studies frequently rely on proprietary datasets, which restricts broad exploration and benchmarking efforts. To address this limitation, we collected a high-dimensional dataset from Wikipedia[2], a rich source of both web traffic data and contextual information. WikiPopular includes daily web traffic time series for pages, accompanied by their contextual data. To ensure relevance and meaningful analysis, we focused on popular Wikipedia pages[3], given the diverse topics available.

3.1 Data Collection

Data collection involves three procedures: time series data collection, textual data collection, and data cleaning. Each procedure is detailed below.

3.1.1 Time series data collection

The time series component consists of daily visit counts for each article, collected through the Wikipedia REST API. This data captures the temporal dynamics of web traffic, revealing trends, seasonality, and irregular patterns. Additionally, relevant series are identified based on semantic similarity between web pages, allowing exploration of contextual relationships

[2] https://www.wikipedia.org/

[3] https://en.wikipedia.org/wiki/Wikipedia:Popular_pages

across multiple pages. These relevant series enhance forecasting accuracy, especially when historical traffic data for the target article is unavailable.

3.1.2 Textual data collection

We supplement the time series data with textual summaries and category labels for each article, collected via Wikipedia's API using web scraping techniques. Summaries provide concise descriptions of each article's content, while the categories reflect its thematic classification. This textual data is crucial for cold-start forecasting, allowing models to forecast web traffic by using content and relationships between web pages, even without historical time series data.

This combination of time series data and contextual data provides a valuable multimodal resource for various forecasting tasks, such as cold-start web traffic forecasting.

3.1.3 Data cleaning

To ensure data quality, we crawled English Wikipedia pages with all-access visit data during the crawling period. We filtered the data by removing pages with ambiguous meanings, and pages not found in the English Wikipedia project. This cleaning process ensures dataset reliability and enhances its applicability in real-world forecasting scenarios.

3.2 Data comparison and analysis

To highlight the uniqueness and value of WikiPopular, we compare it with existing web traffic and multimodal datasets from previous studies, as shown in Table 1. Kaggle provides high-dimensional data but lacks textual information. GP-Copula is also based on Wikipedia data, but lacks a clearly defined dataset structure, e.g., missing time ranges, and doesn't include additional modalities. GPT4MTS contains datasets with 10 events, forecasting NumMentions, NumArticles, and NumSources, so a total of 30 channels.

Although it includes time-relevant text, its limited number of channels and closed-source nature restrict usability. Additionally, its time-relevant text does not support the cold-start problem. WikiPopular stands out by integrating both textual summaries and related time series, while also providing public accessibility upon release. Table 1 provides a summary of the key characteristics of WikiPopular compared to existing ones.

To better illustrate the characteristics of this dataset, we provide descriptive statistics in Table 2. The median page receives 4,633 views daily, while the mean reaches 9,655, reflecting a long-tailed distribution where a small number of pages attract a disproportionate amount of attention, up to over 10 million views daily.

In terms of semantic coverage, WikiPopular dataset covers 728 channels, covering topics of science and nature, e.g., the universe, earth, life, society, and culture, e.g., civilization, people, events, and entertainment, e.g., film, TV, music, sports, video games, books, etc. As shown in Figure 2, certain topics contain more pages due to their large number of subcategories: people-related topics cover singers, actors, athletes, political leaders, and historical figures; entertainment-related topics span sports teams, films, TV shows, music bands, albums, and video games. This rich topical diversity makes the dataset a valuable resource for forecasting web traffic across different domains.

Table 2 Descriptive statistics of the WikiPopular dataset, including traffic distribution and covered semantic categories

median	mean	minimum	maximum	categories
4,633	9,655	1	10,312,178	universe, earth, life, civilization, people, entertainment, event

4 Task distinctions and problem formulation

This section clarifies the distinctions between cold-start forecasting and related tasks and formalizes the problem, emphasizing the unique challenges of cold-start web traffic forecasting.

4.1 Task distinctions

Table 3 lists the comparison of different forecasting tasks. General traffic forecasting relies on past observations of the target series. They focus on modeling temporal patterns such as trend, seasonality, and autocorrelation. This task motivates models to use auto-regressive or series modellings. Cold-start forecasting, in contrast, deals with the scenario where no target history exists. Instead, it uses contextual data. Thus, cold-start forecasting fundamentally differs in its input assumptions and learning objective: it requires cross-entity generalization based on time-irrelevant context, rather than within-entity extrapolation from time series dynamics. This divergence makes general forecasting techniques inapplicable to cold-start forecasting and highlights the need for distinct methodologies.

Beyond above, cold-start forecasting must also be distinguished from several modern forecasting tasks, including zero-shot forecasting [51–54], few-shot forecasting [55–58] , and few-history forecasting. While all these tasks address data-scarce scenarios, they still retain varying degrees of temporal information for the target series. Zero-shot forecasting pertains to the scenario where no prior data is available for a particular instance at inference time, necessitating the use of external knowledge, structural priors, or learned representations to make predictions. Few-shot forecasting provides the model with a limited number of training windows before making predictions on larger unseen windows. Although one could argue that cold-start forecasting eventually transitions into a few-history problem as more data becomes available over time, cold-start forecasting has unique significance compared with few-history forecasting task. For instance, in new product forecasting, solving the cold-start problem allows businesses to make informed decisions before a product is introduced to

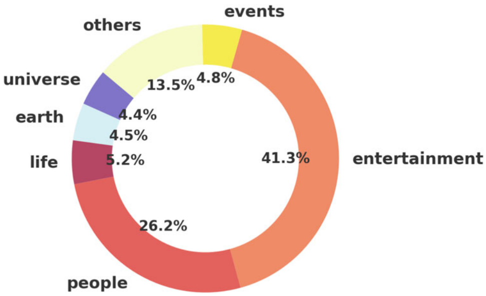


Figure 2 The distribution of the article topics

Table 3 Comparison of different forecasting tasks in terms of data range, number of training windows, and input length per window

Task	Data Range	Number of Training Windows	Input Length per Window
cold-start	Within an individual dataset	No time series input	No time series input
general	Within an individual dataset	Keep a constant proportion of training windows	Keep a constant input length per window
zero-shot	Across multiple datasets	Same as general forecasting tasks	Same as general forecasting tasks
few-shot	Within an individual dataset	Reduced proportion of training windows	Same as general forecasting tasks
few-history	Within an individual dataset	Same as general forecasting tasks	Reduced input length per window

Cold-start forecasting uniquely differs from other tasks as it lacks any time series input, relying entirely on contextual data the market. A robust cold-start forecasting model enables companies to predict demand for products that have not yet been sold, facilitating strategic assortment planning. For an online retailer, such a system could estimate web traffic for potential products, helping optimize inventory decisions. While many recent models have improved performance with multimodal inputs, they generally presuppose access to some historical data for the target instance [15, 19, 20]. These approaches are not directly applicable to cold-start settings, which necessitate learning from context alone. This underscores how cold-start forecasting serves a different function from few-history forecasting, as it provides actionable insights for products that are still in the planning phase.

Given these distinctions, it is imperative to consider cold-start forecasting as an independent research problem, rather than treating it as a subset of few-shot or few-history forecasting. Addressing cold-start forecasting requires fundamentally different methodologies, as it must rely on learned generalization across domains or contextual information, rather than extrapolation from limited in-domain observations. Recognizing and formalizing these differences allows for a more precise characterization of the challenges and potential solutions associated with cold-start forecasting across various applications.

4.2 Problem formulation

Cold-start forecasting, as an underexplored area in web traffic analysis, which focuses on forecasting the web traffic using only contextual data, e.g., textual summaries and categories, of the target page, without relying on any historical web traffic data. Formally, the cold-start forecasting task $f()$ aims to predict the first H time steps of the target new page using its context:

$$f(c_i) = y_{i_{1:T}} \quad \text{or} \quad f(c_i, x_{i_{rel}}) = y_{i_{1:T}}$$

where c_i represents the page summary and categories, providing semantic insights into the content; $x_{i_{rel}} \in R^{K\times T}$ refers to the traffic data in historical T time steps of K relevant web pages; $y_{i_{1:H}} \in R^{1\times H}$ denotes the series values forecasted for the H time steps after T. This setup stands in contrast to both general forecasting and modern forecasting tasks, where $y_{i_{1:T}}$ of the target itself is available for model training. Cold-start forecasting must infer temporal dynamics from time-irrelevant, cross-entity context, making it a distinct methodological challenge.

5 Method

We adopt a RAG architecture [59–61] for time series forecasting, integrating textual information and the series of relevant web pages as supplementary knowledge to enhance the performance of pre-trained LLMs. This approach improves the reasoning capabilities of LLMs, enabling them to generate more accurate forecasts. Specifically, we first design a Relevant Web Pages Retrieval component, employing both the traditional TF-IDF, Word2Vec, SBERT, and LLM-based retrieval to identify the most relevant web pages for the target page. TF-IDF provides keyword-based similarity, while Word2Vec and SBERT offer lightweight semantic matching. Following this, we develop an LLM-based forecasting strategy: a prompt-driven approach utilizing pre-trained language models to infer web traffic patterns.

5.1 Relevant Web pages retrieval

We begin by introducing the method for identifying relevant web pages in scenarios where the target web page lacks historical traffic data. This is based on the assumption that web pages with relevant content are likely to attract comparable patterns of web traffic. Our motivation is to explore contextual relationships between web pages [62] as an alternative to knowledge graphs. While knowledge graphs are powerful, they require significant effort for graph construction and are less suitable for continuous learning with newly introduced web pages. By leveraging a simpler and more flexible framework, we demonstrate the effectiveness of contextual data for cold-start forecasting, a concept validated in prior studies.

By computing the historical traffic of the most relevant web pages, the method provides a simple yet effective solution for forecasting traffic for new or untracked pages. However, the effectiveness of retrieval depends heavily on the quality of the web traffic of relevant web pages. Cosine similarity metric is important in ensuring that the retrieved web pages are closely aligned with the target page, directly influencing the forecast accuracy.

5.1.1 TF-IDF feature extraction

We use TF-IDF, short for Term Frequency-Inverse Document Frequency, to transform the textual web content of each article into numerical feature vectors. The TF-IDF value for a term t in an article a from the corpus A is computed as:

$$\text{TF-IDF}(t, a, A) = \text{TF}(t, a) \times \log\left(\frac{N}{\text{DF}(t)}\right) \tag{1}$$

where $\text{TF}(t, a)$ is the term frequency, e.g., the number of times term t appears in article a, N is the total number of web pages in the corpus, and $\text{DF}(t)$ is the number of documents containing the term t. This formula ensures frequent terms receive lower weights, while unique, informative terms receive higher weights. The TF-IDF vectors are then computed as:

$$\mathbf{v}_a = \left[\text{TF-IDF}(t_1, a, A), \text{TF-IDF}(t_2, a, A), \ldots, \text{TF-IDF}(t_{|V|}, a, A)\right] \tag{2}$$

where $t_1, t_2, \ldots, t_{|V|}$ are the terms in the vocabulary. Once the TF-IDF vectors are generated, we compute the cosine similarity between web pages to measure textual similarity.

5.1.2 Word2Vec and SBERT feature extraction

To capture semantic similarity beyond keyword overlap, we introduce two lightweight embedding-based approaches: Word2Vec [63] and SBERT [64]. For Word2Vec, we use pre-trained word embeddings and compute the representation of each article by averaging the embeddings of its constituent tokens. SBERT, short for Sentence-BERT, on the other hand, directly generates fixed-size sentence-level embeddings from the article's summary using a pre-trained transformer-based encoder. Both methods produce dense vector representations that capture contextual and semantic meaning, enabling more robust similarity comparison even when word choices vary. Cosine similarity is applied in the same way as with TF-IDF to select top-K relevant pages.

5.1.3 LLM-based feature extraction

To improve relevant series identification, we leverage the LLM to extract meaningful text features. Unlike TF-IDF, which depends solely on word frequency, LLM-generated embeddings capture richer semantic relationships by considering the broader context and meaning of the entire text. The LLMs process the textual content of each article to generate dense vector representations, enabling the recognition of similar web pages even when terminology differs, and improved handling of ambiguous or multi-topic content. Such capability proves particularly helpful when dealing with domain-specific terminology, where TF-IDF struggles to identify meaningful connections.

5.1.4 LLM-based feature extraction

To improve relevant series identification, we leverage the LLM to extract meaningful text features. Unlike TF-IDF, which depends solely on word frequency, LLM generated embeddings capture richer semantic relationships by considering the broader context and meaning of the entire text. The LLMs process the textual content of each article to generate dense vector representations, enabling the recognition of similar web pages even when terminology differs, and improved handling of ambiguous or multi-topic content. Such capability proves particularly helpful when dealing with domain-specific terminology, where TF-IDF struggles to identify meaningful connections.

5.1.5 Similarity matrix computation

For each target article, we compute the similarity scores as in (3) using extracted feature vectors, forming a similarity matrix S_{ij}, where i is the target article and j is the relevant article.

$$\text{cosine\_sim}(\mathbf{v}_i, \mathbf{v}_j) = \frac{\mathbf{v}_i \cdot \mathbf{v}_j}{\|\mathbf{v}_i\| \|\mathbf{v}_j\|} \tag{3}$$

where $\mathbf{v}_i \cdot \mathbf{v}_j$ is the dot product of the two vectors, and $\|\mathbf{v}_i\|$ and $\|\mathbf{v}_j\|$ are the Euclidean norms. Then we sort them in descending order, and select the top K most similar web pages, excluding the target article itself:

$$\text{Top-K}(i) = \arg\text{sort}_K \left(S_{ij}\right) \quad \text{for} \quad j \neq i \tag{4}$$

This efficient, interpretable method helps identify relevant series for cold-start forecasting but may overlook deeper semantic connections that go beyond surface-level word similarity.

5.2 LLM-based forecast strategy

In the cold-start forecasting setting, existing time series models struggle due to the absence of historical data for the target web page. To address this, we propose an LLM-based Strategy, which utilizes pre-trained language models to infer web traffic trends from contextual information. By integrating it, we enable a more flexible and adaptive solution that can effectively predict web traffic in cold-start scenarios. Specifically, we employ LLMs as the primary forecasting strategy using prompt-based querying. This eliminates the need for fine-tuning on domain-specific datasets. The pipeline follows: LLMs receive a structured prompt, shown in Table 4 containing both the textual content of the target page and historical traffic data from relevant web pages. Then LLMs generate a forecast based on the prompt, without requiring intermediate feature extraction or transformations. This direct prompting approach leverages LLMs to understand semantic relationships between textual content and user behavior which is reasonable because LLMs have been shown to have the general knowledge because they are trained with enormous data, making it a fast and flexible solution for cold-start forecasting.

Table 4 Prompt designing for the LLM-based strategy, incorporating all the contextual data of the web page

System

You are an expert data analyst specializing in Wikipedia article traffic forecasting. Generate precise 7-day forecasts based on the provided data and article information. For each day, provide a numerical prediction and a brief, insightful explanation considering historical trends, seasonality, recent changes, and the article's context.

User

Generate time series forecasting values for the next {h} days of web traffic from 2019/07/08 to {end_date}.

The title of the article is: {title}

Summary and category of the article: {text}

Please provide your forecast considering the following factors:

1. Overall trends: Any significant rises or drops in traffic.

2. Seasonality: Daily, weekly, or other cyclical patterns.

3. Recent changes: Any notable shifts in the most recent data points.

4. Potential external factors: Consider how the article's topic might influence traffic (e.g., current events, holidays).

5. Historical anomalies: Account for past outliers or unusual patterns.

6. Title of the related article: {related_title}, these articles are sorted according to the cosine similarity of the vectors between the article and the target article. The vectors are generated by the {similar_method}.

7. Web traffic of the related articles during the past 7 days: {related_series}.

For each forecasted day, provide:

- The predicted traffic value

- A brief explanation of why you expect this value, referencing the factors above.

Format your response as:

<begin_predicted_value>[predicted value1, predicted value2, ..., predicted value{h}]<end_predicted_value>,

<begin_explanation>: [Your reasoning]<end_explanation>

The effectiveness of the LLM-based Strategy depends on the selection of the proper parameters: (1) For the model, we use GPT4o-mini, balancing efficiency and accuracy; (2) The temperature is set to 0, ensuring deterministic outputs; (3) The maximum number of tokens is 13,000, to fully cover the prompt without exceeding limits; (4) The batch size is 512, optimizing parallel processing of queries. This configuration ensures optimal performance, minimizing randomness and computational cost.

6 Experiment setup

This section presents experiments on cold-start forecasting, evaluating the proposed methods using appropriate performance metrics and implementation details.

6.1 Evaluation metrics

To assess the performance of our forecasting models, we employ the following error metrics, both of which are scale-independent and suitable for datasets with varying magnitudes:

- Weighted Absolute Percentage Error (WAPE), a widely used forecasting accuracy metric that measures prediction error by weighting absolute errors relative to total actual values. WAPE provides an intuitive and interpretable percentage-based measure, making it effective for comparing forecasting accuracy across datasets and domains.
- Weighted Root Mean Squared Percentage Error (WRMSPE), a variation of Root Mean Squared Percentage Error that accounts for the magnitude of actual values, ensuring that errors on larger values are weighted more heavily. This makes it particularly useful for real-world datasets with diverse scales.

For both WAPE and WRMSPE, lower values indicate better forecasting performance. These metrics ensure robust evaluation by accounting for scaling effects and variability in web traffic data.

6.2 Baseline

To the best of our knowledge, this work is the first to explore cold-start web traffic forecasting. Here we use Statistical Strategy as a forecasting baseline for cold-start forecasting by utilizing the historical traffic of relevant web pages retrieved in Section 5.1. The input data consists of the historical web traffic of relevant web pages. The forecasted traffic for the target traffic is computed as:

$$\hat{y}_i = \frac{1}{K} \sum_{j \in \text{Top-K}(i)} x_{j_{rel}} \tag{5}$$

where $\hat{y}_i$ is the forecasted traffic for the target article i, and $x_{j_{rel}}$ is the historical traffic of the relevant article j. To examine the influence of similarity metrics in the retrieval process, we construct four variants of this Statistical Strategy using different text representations for similarity computation: TF-IDF (*TFStat*), Word2Vec (*W2VStat*), SBERT (*SBertStat*), and LLM-based embeddings (*LLMStat*). These methods differ only in how the Top-K relevant pages are selected. This approach provides a simple, interpretable, and scalable solution, allowing new web pages to be incorporated with minimal computational overhead.

6.3 Implementation details

Our experiments are implemented using PyTorch, HuggingFace's LLaMa[4], and the GPT API. The retrieval-based forecasting approach is conducted as follows:

- Relevant article retrieval is performed using LLaMA 3.2 1B[5], which effectively captures semantic relationships between web pages.
- Forecast generation is handled by GPT-4o-mini[6], which processes contextual information and generates predictions without requiring extensive model fine-tuning.

All experiments are conducted on a single NVIDIA RTX 3090 GPU, ensuring efficient computation and fast inference[7]. Additionally, training-free computation strategies are employed to enable efficient forecasting without the need for costly model training. Models mentioned

[4] https://huggingface.co/

[5] https://www.llama.com/

[6] https://platform.openai.com/

[7] To promote reproducibility, a sample dataset and corresponding codebase can be found in Section 1 for review.

above leverage: pre-trained language models, retrieval-based heuristics, and lightweight statistical techniques. By minimizing computational overhead, our approach is ideal for cold-start scenarios, where historical data is unavailable, and rapid adaptation is crucial.

7 Result and discussions

This section presents our main results, validating the effectiveness of contextual data in cold-start forecasting and providing ablation studies for further insights into our framework. The results, organized into the following subsections, aim to provide insights into various aspects of our approach. Specifically, our experiments address the following key questions:

- Effectiveness of our framework and strategies: How well do the proposed methods perform in cold-start forecasting? Shown in Section 7.1 Main Results;
- Role of contextual data in cold-start forecasting: Does incorporating contextual information enhance forecasting accuracy? Which contextual data contributes to forecasting the most? Shown in Section 7.1 Main Results and Section 7.2.1 Ablation Study 1;
- Impact of relevant web pages: How does the number of relevant web pages, K, influence forecasting performance? Shown in Section 7.2.2 Ablation Study 2.
- Prompt design and article ranking: Can LLM benefit from explicitly ordered relevant articles in prompts? Shown in Section 7.2.3 Ablation Study 3.

7.1 Main Results: cold-start forecasting

Cold-start forecasting evaluates the model's ability to forecast web traffic for newly introduced pages using only contextual data, such as page summaries and the historical traffic of semantically relevant pages, without relying on the target page's own history. To maximize the utility of historical traffic data of relevant web pages, we set the test samples' start date to July 8, 2019, one week after the dataset's initial date. To simulate a real cold-start scenario, we select 50 web pages for testing. To the best of our knowledge, no existing models are specifically designed for this task, so we evaluate in Section 5.2, comparing two forecasting strategies, Statistical and LLM-based, with four types of similarity-based retrieval methods: TF-IDF, Word2Vec, SBERT, and LLM embeddings. These combinations result in eight total methods: *TFStat*, *W2VStat*, *SBertStat*, and *LLMStat* use statistical averaging of relevant time series; *TFLLM*, *W2VLLM*, *SBertLLM*, and *LLM+* use an LLM prompt-based generation strategy. All experiments are conducted on the WikiPopular dataset. The results are summarized in Table 5.

First, we compare the two forecasting strategies, Statistical and LLM-based, across all retrieval methods. We find that LLM-based Strategy outperforms Statistical Strategy, regardless of the underlying retrieval methods. Specifically, *LLM+* achieves the lowest WAPE and WRMSPE across almost all horizons, followed closely by *W2VLLM* and *SBertLLM*. This confirms the advantage of generation-based reasoning over statistical averaging in extracting and utilizing semantic context.

Second, among the Statistical baselines, *LLMStat* shows the best performance, outperforming *TFStat*, *W2VStat*, and *SBertStat*. This suggests that LLM-based retrieval provides semantically richer and more accurate relevant pages than TF-IDF and traditional embedding methods like Word2Vec and SBERT. Although *W2VStat* and *SBertStat* perform reasonably well, their performance lags behind *LLMStat*, especially at longer horizons, indicating limited ability to fully capture contextual alignment using shallow embeddings.

Table 5 Cold-start forecasting results under different combinations of retrieval methods and forecasting strategies

Dataset	WAPE				WRMSPE			
Horizon	1	2	4	7	1	2	4	7
TFStat	1.185	1.137	1.125	1.100	1.865	1.799	1.958	1.893
W2VStat	1.202	1.208	1.212	1.200	1.928	1.999	2.226	2.192
SBertStat	1.272	1.239	1.228	1.084	2.716	2.727	2.824	2.404
LLMStat	0.859	0.884	0.918	0.900	1.352	1.416	1.673	1.599
TFLLM	0.977	0.991	1.035	1.062	1.486	1.609	1.850	1.843
W2VLLM	**0.810**	0.892	0.883	0.930	1.289	1.674	1.695	1.639
SBertLLM	0.888	0.882	0.883	0.932	**1.221**	1.615	1.710	1.582
LLM+	0.817	**0.819**	**0.828**	**0.859**	1.292	**1.361**	**1.616**	**1.564**

TFStat, *W2VStat*, *SBertStat*, and *LLMStat* represent the Statistical Strategy using top-K relevant pages selected by TF-IDF, Word2Vec, SBERT, and LLM-based embeddings, respectively. *TFLLM*, *W2VLLM*, *SBertLLM*, and *LLM+* are their counterparts using the LLM-based forecasting strategy. Metrics reported are WAPE and WRMSPE across horizons 1, 2, 4, and 7. The best results in each column are shown in **bold**

Third, the results verify the **critical role of contextual data and prompt quality** of the LLM-based Strategy. Comparing *TFLLM*, *W2VLLM*, *SBertLLM*, and *LLM+*, we observe that using richer embeddings for retrieval yields better forecasting outcomes. Besides, carefully structured prompts align the model's reasoning with the forecasting objective, optimizing forecasting accuracy. The enhanced ability of LLM, especially *LLM+*, further differentiates them from other methods. These findings underscore the importance of well-crafted prompts, as the effectiveness of LLM depends heavily on how input data is framed. A carefully designed prompt aligns the model's reasoning with the task, improving forecast accuracy.

Lastly, **the forecast horizon stabilized in LLM-based strategy.** The experimental results indicate that forecast accuracy varies with horizon length, with error metrics, WAPE and WRMSPE, increasing slightly as the horizon extends from 1 to 7 days. This trend is expected, as long-term forecasts are more uncertain. While forecasting errors naturally increase over longer horizons, the robustness of *TFLLM* *W2VLLM*, *SBertLLM*, and *LLM+* remains evident, maintaining reliable performance in extended forecast scenarios. Our findings emphasize the importance of leveraging textual data and LLMs' ability to reason over unseen web pages, highlighting new opportunities for cold-start forecasting.

7.2 Ablation study

To further investigate the impact of contextual data and relevant article selection on forecasting performance, we conduct two ablation studies: (1) Effect of contextual data: Exploring how different types of contextual information improve cold-start forecasting; (2) Effect of K-selection: Evaluating how the number of relevant web pages K influences the results; (3) Effect of ranking-aware prompting: Evaluating whether explicitly conveying the order of retrieved articles by similarity in the prompt, e.g., "Most relevant article", improves LLM-based forecasting.

Table 6 Results for the Effect of Contextual Data

Methods	*LLM+*				*TFLLM*			
(a) WAPE Results								
Horizon	1	2	4	7	1	2	4	7
TgtCnt	0.860	0.883	0.878	**0.863**	1.063	1.171	0.879	**0.860**
TCnt+RTit	0.829	0.859	0.872	0.871	1.035	1.140	0.880	0.871
TCnt+RTrf	0.819	**0.841**	**0.870**	0.869	0.992	1.043	**0.878**	0.869
All	**0.818**	0.843	0.887	0.874	**0.977**	**1.024**	0.883	0.877
(b) WRMSPE Results								
TgtCnt	1.449	1.477	1.702	1.646	1.844	1.950	1.651	1.643
TCnt+RTit	1.452	1.479	1.703	1.652	1.851	1.958	1.705	1.702
TCnt+RTrf	1.334	**1.361**	1.620	1.569	1.610	1.678	1.480	1.477
All	**1.292**	1.361	**1.616**	**1.564**	**1.486**	**1.586**	**1.452**	**1.449**

TgtCnt represents the methods using only the web content of the target page as input. *TCnt+RTit* and *TCnt+RTrf* represent the methods that use the content of the target page plus the titles of relevant web pages and the traffic data of relevant web pages, respectively. *All* refers to the methods that use all available contextual data, i.e., web content of the target page, relevant articles' titles, and relevant articles' traffic. The best results are shown in **bold**, and the second-best results are underlined

7.2.1 Effect of contextual data

The ability to incorporate contextual data significantly differentiates LLM-based Strategy from Statistical Strategy, which relies solely on time series web traffic data. In this experiment, we evaluate how various types of contextual information impact forecasting performance, as summarized in Table 6. The results reveal that the quality of contextual data, the choice of relevant article retrieval, and the balance between web traffic and textual data all play a critical role in determining forecast accuracy.

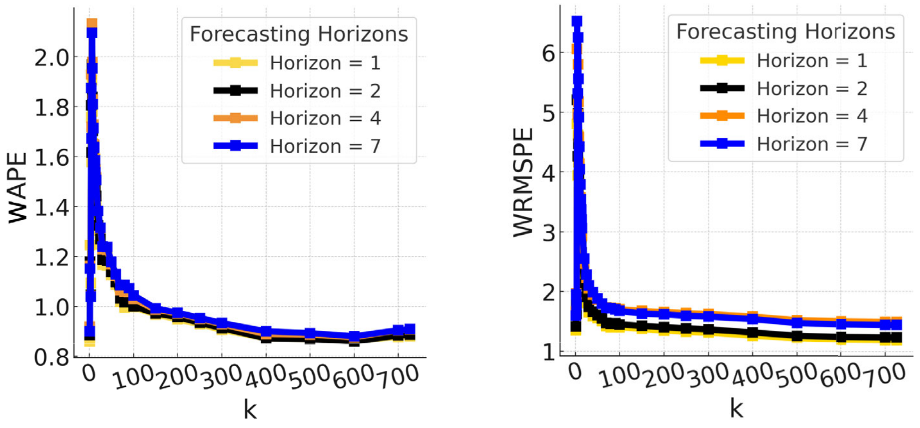
From Table 6, we observe that using only the web content of the target page, *TgtCnt*, yields overall the lowest performance among the contextual methods. However, when we compare this result to *TFStat* in Table 5, the performance of *TgtCnt* is still superior across all metrics. This suggests that even simple content-based forecasting can outperform statistical methods when appropriately leveraging related content. This observation highlights the importance of selecting relevant related series for forecasting. Furthermore, the improvement from LLMs-based Strategy, *TFLLM*, and *LLM+*, demonstrates the effectiveness of using LLMs for relevant web page retrieval, as they provide a richer semantic understanding of relevant articles, leading to better forecasting outcomes. Besides, *TgtCnt* may perform better when the horizon is longer compared to other methods. This may be because content summaries capture high-level semantic information about the web page, which remains stable and relevant over time, especially for long-term forecasts. In contrast, web traffic of relevant web pages, *RTrf*, may be more useful for short-term forecasts but can lose predictive power as the horizon length increases due to the inherent volatility of web traffic.

Furthermore, adding related traffic data *TCnt+RTrf* further boosts performance, and the best results are achieved by combining all contextual data *All*, confirming the benefits of using multimodal information. The results emphasize the need to carefully design retrieval strategies and highlight the potential of LLMs to enhance cold start forecasting by selecting relevant and meaningful related series.

7.2.2 Effect of *K* K-selection

Our experiments examine how the number of relevant web pages, K, affects forecasting in the Statistical Strategy and the LLM-based Strategy, using TF-IDF and LLM for the retrieval of relevant web pages. The results reveal the following main insights. Figure 3 shows the results for the *LLMStat*. Since the Statistical Strategy is relatively lightweight and computationally inexpensive, we run the experiments with K up to the maximum possible, e.g., the total number of available channels minus 1. As shown in Figure 3, the performance in terms of both WAPE and WRMSPE initially improves as K increases, reaching an optimal point before rising slightly and eventually stabilizing at larger values than 200. This suggests that including more relevant articles provides additional relevant information, but beyond a certain point, adding more series introduces irrelevant or noisy data, which begins to impact performance negatively.

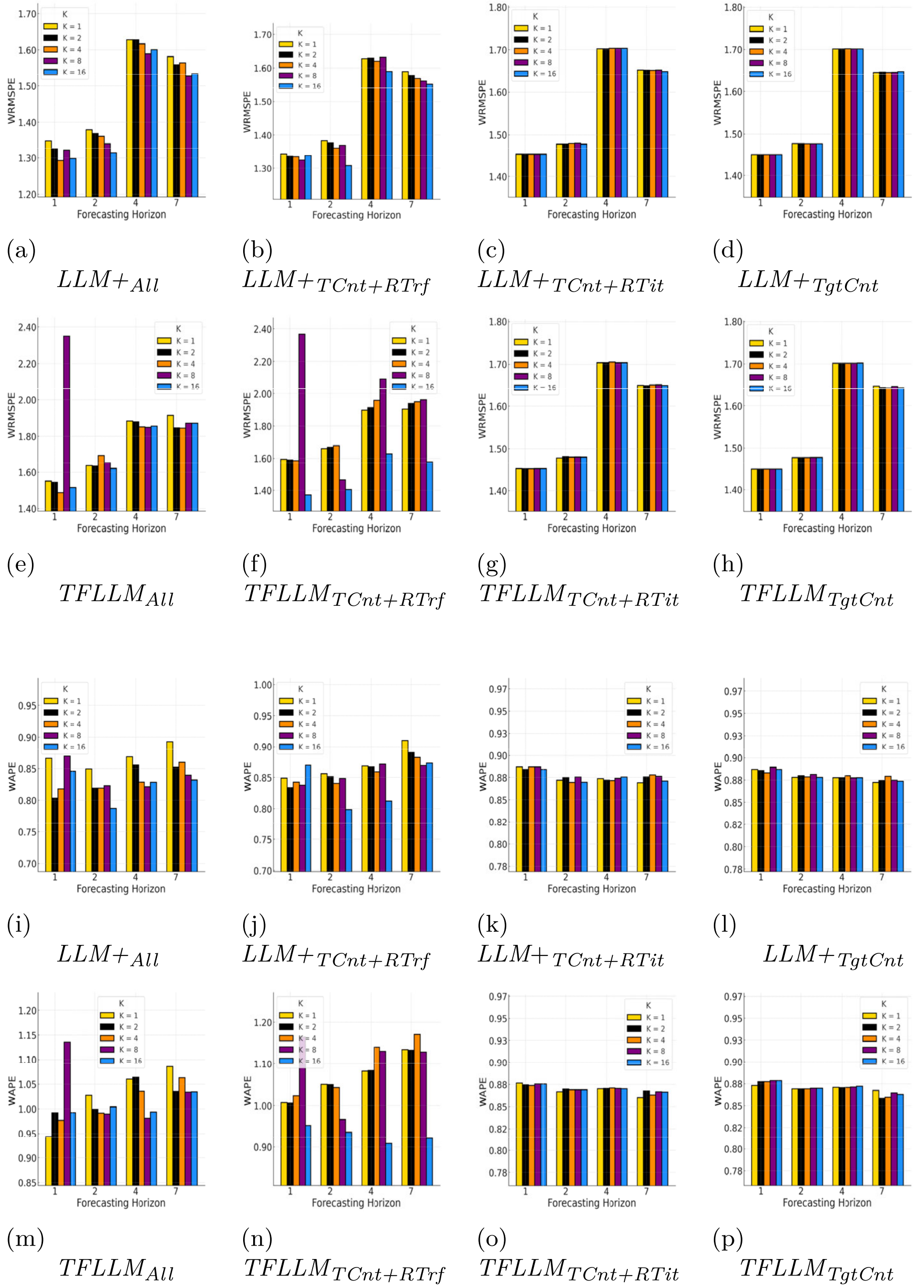
Figure 4 presents the results of *TFLLM* and *LLM+*. In each subfigure, the x-axis represents the forecasting horizon, while the bars indicate performance across different K-values. Using LLM for web page retrieval, shown in Figure 4 (a)–(d) and (i)–(l), we observe that performance initially varies with K but eventually becomes more stable as fewer contextual inputs are used, ranging from all contextual information, e.g., *TgtCnt*, *RTit*, and *RTrf*, to *TgtCnt* only. This indicates that while more contextual data improves performance initially, it also increases the risk of introducing noise, particularly when irrelevant related series are included. From another aspect, the more contextual data included, the higher the chance of achieving better results in Figure 4 (a) and (i). When *All* contextual inputs, the model benefits from richer information, improving overall forecasting accuracy. A similar pattern emerges in the TF-IDF, shown in Figure 4 (e)–(h) and (m)–(p). This confirms the importance of using related series effectively in cold-start forecasting. Another key observation is that different strategies require different optimal K-values. This variability may arise from differences in web page topics or the nature of traffic patterns. Therefore, the selection of K for individual forecasting tasks can further improve prediction accuracy.



(a) WAPE

(b) WRMSPE

Figure 3 Performance for different values of K in *LLMStat*, retrieve relevant articles using LLM and forecasting with Statistical Strategy. (a) and (b) show the forecasting errors measured by WAPE and WRMSPE, respectively



(a) $LLM{+}_{All}$

(b) $LLM{+}_{TCnt+RTrf}$

(c) $LLM{+}_{TCnt+RTit}$

(d) $LLM{+}_{TgtCnt}$

(e) $TFLLM_{All}$

(f) $TFLLM_{TCnt+RTrf}$

(g) $TFLLM_{TCnt+RTit}$

(h) $TFLLM_{TgtCnt}$

(i) $LLM{+}_{All}$

(j) $LLM{+}_{TCnt+RTrf}$

(k) $LLM{+}_{TCnt+RTit}$

(l) $LLM{+}_{TgtCnt}$

(m) $TFLLM_{All}$

(n) $TFLLM_{TCnt+RTrf}$

(o) $TFLLM_{TCnt+RTit}$

(p) $TFLLM_{TgtCnt}$

Figure 4 Performance for different values of K in relevant series selection for *TFLLM*, and *LLM+*, the LLM-based Strategy with TD-IDF and LLM-based relevant article retrieval. (a)–(d) and (i)–(l) represent WAPE and WRMSPE in *LLM+*; (e)–(h) and (m)–(p) represent WAPE and WRMSPE in *TFLLM*. *All*, *TCnt+RTrf*, *TCnt+RTit*, and *TgtCnt* means using all the contextual data, using the content of the web page with the series of relevant web pages, using the content of the web page with the title of relevant web pages, using the content of web page only

Table 7 Cold-start forecasting results under different retrieval and prompting strategies

Dataset	WAPE				WRMSPE			
Horizon	1	2	4	7	1	2	4	7
TFLLM	0.977	0.991	1.035	**1.062**	**1.486**	1.609	**1.850**	**1.843**
TFLLMr	**0.973**	**0.958**	**0.991**	1.129	1.502	**1.574**	1.865	1.904
W2VLLM	0.810	**0.892**	**0.883**	0.930	**1.289**	1.674	**1.695**	**1.639**
W2VLLMr	**0.767**	0.900	0.927	**0.917**	1.311	**1.451**	1.713	1.639
SBertLLM	**0.888**	**0.882**	**0.883**	**0.932**	**1.221**	1.615	1.710	**1.582**
SBertLLMr	0.894	0.908	0.915	0.944	1.287	**1.405**	**1.637**	1.596
LLM+	0.817	0.819	0.828	0.859	1.292	1.361	1.616	1.564
LLM+r	**0.815**	**0.805**	**0.811**	**0.859**	**1.280**	**1.331**	**1.594**	**1.551**

The "r" suffix, e.g., TFLLMr, LLM+r, indicates variants where retrieved articles are not only sorted in descending order of similarity, but this ordering is also explicitly conveyed in the prompt, to help the LLM better prioritize context. Best results are in **bold**

7.2.3 Effect of ranking-aware prompting

We explore whether incorporating the relative importance of retrieved articles into the prompt improves LLM-based forecasting. Instead of explicitly providing numerical similarity scores, which LLMs are often insensitive to, we modify the prompt format to present retrieved articles in descending order of similarity, thereby conveying their relative importance. We refer to these variants with the suffix `r`, such as *TFLLMr*, *W2VLLMr*, *SBertLLMr*, and *LLM+r*. The results are presented in Table 7.

The effectiveness of ranking enhancement varies across retrieval methods. The most consistent and notable improvement is observed with *LLM+*, where *LLM+r* consistently achieves the best performance across all horizons, highlighting that ranking cues are especially effective when paired with semantically rich embeddings. *TFLLMr* also shows clear benefits over *TFLLM*, particularly in short horizons,1 to 2 days, though the advantage diminishes as the horizon lengthens. For *W2VLLM* and *SBertLLM*, the results are more variable and less stable. Some horizons show modest gains with ranking-aware prompting, e.g., WAPE@1 for *W2VLLMr*, while others remain flat or degrade, suggesting that lightweight embeddings may not fully benefit from ranking without deeper semantic grounding. These findings suggest that the utility of ranking-aware prompting depends heavily on the retrieval quality. When the embeddings already encode rich semantic relationships, as in *LLM+*, ranking provides a useful signal for guiding generation.

8 Conclusion

This paper addresses the cold-start problem in web traffic forecasting. We have proposed to use contextual data in the RAG, including textual content and historical web traffic of relevant web pages, through both the Statistical Strategy and the LLM-based Strategy. We further extend the retrieval module with multiple semantic representations, including TF-IDF, Word2Vec, SBERT, and LLM embeddings, enabling a broad and fair comparison across methods. Our experiments confirm that contextual data is important, with models combining web content and relevant series outperforming simpler setups. The LLM-based retrieval strategy consistently delivers superior performance across all settings, demonstrating its strength in extracting semantically aligned reference pages. Moreover, we introduce a ranking-aware prompting study, which improves LLM generation quality by prompting retrieved articles in descending similarity order. Our work advances cold-start forecasting by introducing new methods, releasing multimodal web traffic datasets, and providing insights into the potential of LLMs for complex forecasting tasks without domain-specific training. Future directions include exploring adaptive K selection, fine-grained contextual inputs, enhanced LLM prompting strategies, and fine-tuning the LLM-based Strategy with potential applications in e-commerce, news trends, and other data-scarce domains. In addition, we will investigate category-aware retrieval strategies, which may offer improved contextual alignment by selecting relevant pages within the same semantic class. Although this is not feasible in our current setup due to the absence of explicit category labels in WikiPopular, it represents a valuable extension for future research using datasets with richer structural or categorical annotations.

Acknowledgements The authors acknowledge the use of OpenAI's ChatGPT for writing polishing. They also extend their gratitude to Tuan Anh, a cybersecurity student at VinUniversity, for his valuable contributions to this research. His efforts in collecting the web traffic data used in this study have been instrumental in enabling the analysis and experimentation presented in this manuscript.

Author Contributions Xin Zhou conducted the primary experiments and authored the main manuscript text. Weiqing Teresa Wang supervised and provided expert guidance on the machine learning and deep learning methodologies. Christoph Bergmeir contributed specialist insights on time series forecasting methods. Wray Buntine provided overarching direction, addressed ethical considerations, and assisted significantly in dataset collection. All authors reviewed, critically revised, and approved the final manuscript.

Funding Open Access funding enabled and organized by CAUL and its Member Institutions. The authors sincerely acknowledge the support of Weiqing Wang through the Australian Research Council Discovery Early Career Researcher Award (DECRA) DE250100032, which has provided essential resources for this research. This work contribution is also part of the I+D+i project granted by C-ING-250-UGR23 co-funded by "Consejería de Universidad, Investigación e Innovación" and for the European Union related to the FEDER Andalucía Program 2021-2027". The work is also supported by the Knowledge Generation Project PID2023-149128NB-I00, funded by the Ministry of Science, Innovation and Universities of Spain. Christoph Bergmeir is supported by a María Zambrano Fellowship that is funded by the Spanish Ministry of Universities and Next Generation funds from the European Union.

Data Availability To support further research and benchmarking, we release the codes and data at https://github.com/xinzzzhou/CCWTF.git

Declarations

Ethical Approval Not applicable.

Competing Interests The authors declare no competing interests.

References

1. Shelatkar, T., Tondale, S., Yadav, S., Ahir, S.: Web traffic time series forecasting using arima and lstm rnn. ITM Web Conf. **32**, 03017 (2020). https://doi.org/10.1051/itmconf/20203203017
2. Liu, Z., Yan, Y., Hauskrecht, M.: A flexible forecasting framework for hierarchical time series with seasonal patterns: A case study of web traffic. In: The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval. SIGIR '18, pp. 889–892. Association for Computing Machinery, New York, NY, USA (2018). https://doi.org/10.1145/3209978.3210069
3. Mahanand, A., Prakash, P., Devaraj, A.: Deep learning-based hybrid technique for forecasting web traffic. In: 2023 Fifth International Conference on Electrical, Computer and Communication Technologies (ICECCT), pp. 1–5 (2023). https://doi.org/10.1109/ICECCT56650.2023.10179797
4. Kochetkova, I., Kushchazli, A., Burtseva, S., Gorshenin, A.: Short-term mobile network traffic forecasting using seasonal arima and holt-winters models. Future Internet **15**(9) (2023). https://doi.org/10.3390/fi15090290
5. Ferreira, G.O., Ravazzi, C., Dabbene, F., Calafiore, G.C., Fiore, M.: Forecasting network traffic: A survey and tutorial with open-source comparative evaluation. IEEE Access **11**, 6018–6044 (2023). https://doi.org/10.1109/ACCESS.2023.3236261
6. Zohaib, A., Sheffey, J., Houmansadr, A.: Investigating traffic analysis attacks on apple icloud private relay. In: Proceedings of the 2023 ACM Asia Conference on Computer and Communications Security. ASIA CCS '23, pp. 773–784. Association for Computing Machinery, New York, NY, USA (2023). https://doi.org/10.1145/3579856.3595793
7. Nunnagoppula, H., Katragadda, K., Ramesh, M.: Website traffic forecasting using deep learning techniques. In: 2023 International Conference on Artificial Intelligence and Smart Communication (AISC), pp. 531–536 (2023). https://doi.org/10.1109/AISC56616.2023.10085005
8. He, M., Zhang, H., Zhang, Z., et al.: Invariant representation learning to popularity distribution shift for recommendation. World Wide Web **27**, 10 (2024). https://doi.org/10.1007/s11280-024-01242-x
9. Zhong, H., Zhang, Q., Li, W., et al.: Kpllm-ste: Knowledge-enhanced and prompt-aware large language models for short-text expansion. World Wide Web **28**, 9 (2025). https://doi.org/10.1007/s11280-024-01322-y
10. Zhang, Y., Liao, W., Wang, Y., et al.: Meta-path automatically extracted from heterogeneous information network for recommendation. World Wide Web **27**, 26 (2024). https://doi.org/10.1007/s11280-024-01265-4
11. Ruan, S., Yang, C., Li, D.: Knowledge-enhanced personalized hierarchical attention network for sequential recommendation. World Wide Web **27**, 2 (2024). https://doi.org/10.1007/s11280-024-01236-9
12. Box, G.E.P., Jenkins, G.M.: Time series analysis: Forecasting and control (1970)
13. Zeng, A., Chen, M., Zhang, L., Xu, Q.: Are transformers effective for time series forecasting? In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 37, pp. 11121–11128 (2023)
14. Jouini, N., Medhi, D.: Forecasting contemporaneously aggregated vector arma processes. J. Bus. Econ. Stat. **3**(4), 401–407 (1985)
15. Liu, C., Xu, Q., Miao, H., Yang, S., Zhang, L., Long, C., Li, Z., Zhao, R.: Timecma: Towards llm-empowered multivariate time series forecasting via cross-modality alignment. In: Proceedings of the AAAI Conference on Artificial Intelligence (2025)
16. Qian, W., Zhao, Y., Zhang, D., Chen, B., Zheng, K., Zhou, X.: Towards a unified understanding of uncertainty quantification in traffic flow forecasting. IEEE Trans. Know. Data Eng. (2023)
17. Miao, H., Zhao, Y., Guo, C., Yang, B., Zheng, K., Huang, F., Xie, J., Jensen, C.S.: A unified replay-based continuous learning framework for spatio-temporal prediction on streaming data. In: Proceedings of the 40th IEEE International Conference on Data Engineering (ICDE), pp. 1050–1062 (2024)
18. Miao, H., Zhao, Y., Guo, C., Yang, B., Zheng, K., Huang, F., Xie, J., Jensen, C.S.: Spatio-temporal prediction on streaming data: A unified federated continuous learning framework. IEEE Trans. Know. Data Eng. (2025)
19. Miao, H., Zhao, Y., Guo, C., Yang, B., Zheng, K., Huang, F., Xie, J., Jensen, C.S.: Less is more: Efficient time series dataset condensation via two-fold modal matching. Proceedings of the VLDB Endowment **18**(12) (2025). Extended Version on arXiv:2410.20905
20. Liu, C., Yang, S., Xu, Q., Li, Z., Long, C., Li, Z., Zhao, R.: Spatial-temporal large language model for traffic prediction. In: Proceedings of the 25th IEEE International Conference on Mobile Data Management (MDM) (2024)
21. Chen, J., Zhang, F., Li, H., et al.: Empnet: An extract-map-predict neural network architecture for cross-domain recommendation. World Wide Web **27**, 12 (2024). https://doi.org/10.1007/s11280-024-01240-z

22. Zhou, X., Wang, W., Buntine, W., Qu, S., Sriramulu, A., Tan, W., Bergmeir, C.: Scalable transformer for high dimensional multivariate time series forecasting. In: Proceedings of the 33rd ACM International Conference on Information and Knowledge Management. CIKM '24. Association for Computing Machinery, Boise, ID, USA (2024)
23. Sen, R., Yu, H.-F., Dhillon, I.S.: Think Globally, Act Locally: A Deep Neural Network Approach to High-Dimensional Time Series Forecasting **32** (2019)
24. Obata, K., Kawabata, K., Matsubara, Y., Sakurai, Y.: Dynamic multi-network mining of tensor time series. In: Proceedings of the ACM Web Conference 2024. WWW '24, pp. 4117–4127. Association for Computing Machinery, New York, NY, USA (2024). https://doi.org/10.1145/3589334.3645461
25. Long, Q., Fang, Z., Fang, C., Chen, C., Wang, P., Zhou, Y.: Unveiling delay effects in traffic forecasting: A perspective from spatial-temporal delay differential equations. In: Proceedings of the ACM Web Conference 2024. WWW '24, pp. 1035–1044. Association for Computing Machinery, New York, NY, USA (2024). https://doi.org/10.1145/3589334.3645688
26. Hyndman, R.J., Koehler, A.B., Ord, J.K., Snyder, R.D.: Forecasting with exponential smoothing: The state space approach (2008)
27. Nie, Y., Nguyen, N.H., Sinthong, P., Kalagnanam, J.: A time series is worth 64 words: Long-term forecasting with transformers (2023)
28. Liu, Y., Hu, T., Zhang, H., Wu, H., Wang, S., Ma, L., Long, M.: itransformer: Inverted Transformers Are Effective for Time Series Forecasting (2024)
29. Nam, Y., Yoon, S., Shin, Y., Bae, M., Song, H., Lee, J.-G., Lee, B.S.: Breaking the time-frequency granularity discrepancy in time-series anomaly detection. In: Proceedings of the ACM Web Conference 2024. WWW '24, pp. 4204–4215. Association for Computing Machinery, New York, NY, USA (2024). https://doi.org/10.1145/3589334.3645556
30. Lai, Z., Li, H., Zhang, D., Zhao, Y., Qian, W., Jensen, C.S.: E2usd: Efficient-yet-effective unsupervised state detection for multivariate time series. In: Proceedings of the ACM Web Conference 2024. WWW '24, pp. 3010–3021. Association for Computing Machinery, New York, NY, USA (2024). https://doi.org/10.1145/3589334.3645593
31. Wang, Z., Pei, C., Ma, M., Wang, X., Li, Z., Pei, D., Rajmohan, S., Zhang, D., Lin, Q., Zhang, H., Li, J., Xie, G.: Revisiting vae for unsupervised time series anomaly detection: A frequency perspective. In: Proceedings of the ACM Web Conference 2024. WWW '24, pp. 3096–3105. Association for Computing Machinery, New York, NY, USA (2024). https://doi.org/10.1145/3589334.3645710
32. Xu, D., Cheng, W., Zong, B., Song, D., Ni, J., Yu, W., Liu, Y., Chen, H., Zhang, X.: Tensorized LSTM with Adaptive Shared Memory for Learning Trends in Multivariate Time Series (2020)
33. Cheng, M., Yang, J., Pan, T., Liu, Q., Li, Z.: Convtimenet: A deep hierarchical fully convolutional model for multivariate time series analysis (2024). arXiv preprint arXiv:2403.01493
34. Zhang, Y., Yan, J.: Crossformer: Transformer Utilizing Cross-Dimension Dependency for Multivariate Time Series Forecasting. In: Proceedings of International Conference on Learning Representations (ICLR) (2023)
35. Zerveas, G., Jayaraman, S., Patel, D., Bhamidipaty, A., Eickhoff, C.: A Transformer-based Framework for Multivariate Time Series Representation Learning. In: Proceedings of ACM SIGKDD Conference on Knowledge Discovery & Data Mining (KDD) (2021)
36. Zhou, T., Ma, Z., Wen, Q., Wang, X., Sun, L., Jin, R.: FEDformer: Frequency Enhanced Decomposed Transformer for Long-term Series Forecasting. In: Proceedings of International Conference on Machine Learning (ICML) (2022)
37. Zhou, H., Zhang, S., Peng, J., Zhang, S., Li, J., Xiong, H., Zhang, W.: Informer: Beyond Efficient Transformer for Long Sequence Time-Series Forecasting. In: Proceedings of the AAAI Conference on Artificial Intelligence (AAAI) (2021)
38. Ekambaram, V., Jati, A., Nguyen, N., Sinthong, P., Kalagnanam, J.: Tsmixer: Lightweight MLP-Mixer Model for Multivariate Time Series Forecasting. In: Proceedings of ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD) (2023)
39. Sun, C., Li, H., Li, Y., Hong, S.: TEST: Text prototype aligned embedding to activate LLM's ability for time series. In: The Twelfth International Conference on Learning Representations (2024). https://openreview.net/forum?id=Tuh4nZVb0g
40. Cao, D., Jia, F., Arik, S.O., Pfister, T., Zheng, Y., Ye, W., Liu, Y.: TEMPO: Prompt-based generative pre-trained transformer for time series forecasting. In: The Twelfth International Conference on Learning Representations (2024). https://openreview.net/forum?id=YH5w12OUuU
41. Jia, F., Wang, K., Zheng, Y., Cao, D., Liu, Y.: Gpt4mts: Prompt-based large language model for multimodal time-series forecasting. Proceedings of the AAAI Conference on Artificial Intelligence **38**(21), 23343–23351 (2024). https://doi.org/10.1609/aaai.v38i21.30383

42. Lee, G., Yu, W., Cheng, W., Chen, H.: MoAT: Multi-Modal Augmented Time Series Forecasting. unpublished
43. Sims, C.A.: Macroeconomics and Reality. Econometrica **48** (1980)
44. Salinas, D., Bohlke-Schneider, M., Callot, L., Medico, R., Gasthaus, J.: High-dimensional multivariate forecasting with low-rank Gaussian Copula Processes. In: Proceedings of Advances in Neural Information Processing Systems (NIPS) (2019)
45. Chen, F., Qin, Z., Zhou, M., Zhang, Y., Deng, S., Fan, L., Pang, G., Wen, Q.: Lara: A light and anti-overfitting retraining approach for unsupervised time series anomaly detection. In: Proceedings of the ACM Web Conference 2024. WWW '24, pp. 4138–4149. Association for Computing Machinery, New York, NY, USA (2024). https://doi.org/10.1145/3589334.3645472
46. Rasul, K., Ashok, A., Williams, A.R., Khorasani, A., Adamopoulos, G., Bhagwatkar, R., Biloš, M., Ghonia, H., Hassen, N., Schneider, A., Garg, S., Drouin, A., Chapados, N., Nevmyvaka, Y., Rish, I.: Lag-llama: Towards foundation models for time series forecasting. In: R0-FoMo:Robustness of Few-shot and Zero-shot Learning in Large Foundation Models (2023). https://openreview.net/forum?id=jYluzCLFDM
47. Jin, M., Wang, S., Ma, L., Chu, Z., Zhang, J.Y., Shi, X., Chen, P.-Y., Liang, Y., Li, Y.-F., Pan, S., Wen, Q.: Time-LLM: Time series forecasting by reprogramming large language models. In: Proceedings of International Conference on Learning Representations (ICLR) (2024)
48. Liu, X., Hu, J., Li, Y., Diao, S., Liang, Y., Hooi, B., Zimmermann, R.: Unitime: A language-empowered unified model for cross-domain time series forecasting. In: Proceedings of the ACM Web Conference 2024. WWW '24, pp. 4095–4106. Association for Computing Machinery, New York, NY, USA (2024). https://doi.org/10.1145/3589334.3645434
49. Ye, J., Zhang, W., Yi, K., Yu, Y., Li, Z., Li, J., Tsung, F.: A Survey of Time Series Foundation Models: Generalizing Time Series Representation with Large Language Model (2024). https://arxiv.org/abs/2405.02358
50. Mahajan, V., Wind, Y.: New product forecasting models: Directions for research and implementation. Int. J. Forecast. **4**(3), 341–358 (1988). https://doi.org/10.1016/0169-2070(88)90102-1
51. Amazon Web Services, A.T.: Chronos-Bolt: Pretrained Foundation Models for Fast and Accurate Zero-Shot Time Series Forecasting. https://github.com/amazon-science/chronos-forecasting. Available via AutoGluon-TimeSeries and Hugging Face (2024)
52. Gruver, N., Tschannen, M., Flunkert, V., Gasthaus, J., Januschowski, T.: Large language models are zero-shot time series forecasters (2023). arXiv preprint arXiv:2310.07820
53. Contributors, M.: Mamba4Cast: Zero-Shot Time Series Forecasting. https://github.com/ML4ITS/Mamba4Cast. Open-source project (2024)
54. Team, N.: TimeGPT: Foundation Model for Time Series Forecasting. Available via Nixtla (2023)
55. Lyu, Y., Liu, Y., Wu, H., Ouyang, Y.: Few-shot time-series forecasting with application for vehicular traffic flow. In: 2022 IEEE International Conference on Information Reuse and Integration for Data Science (IRI), pp. 206–213 (2022). https://doi.org/10.1109/IRI54793.2022.00037
56. Banerjee, S., Yellepeddi, K., Banerjee, A.: Few-shot learning for time-series forecasting (2020). arXiv preprint arXiv:2009.14379
57. Muller, R., Ernst, D., Geurts, P.: Zero-shot and few-shot time series forecasting with ordinal regression. In: 28th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN), pp. 529–534 (2020). https://www.esann.org/sites/default/files/proceedings/2020/ES2020-78.pdf
58. Khrabrov, V., Ryzhov, A., Filchenkov, A.: Few-shot time series forecasting in a meta-learning framework. J. Intell. Fuzzy Syst. **46**(3), 3693–3704 (2024). https://doi.org/10.3233/JIFS-231113
59. Su, W., Tang, Y., Ai, Q., Wu, Z., Liu, Y.: Dragin: Dynamic retrieval augmented generation based on the real-time information needs of large language models. In: The 62nd Annual Meeting of the Association for Computational Linguistics (ACL) (2024)
60. Yoran, O., Wolfson, T., Ram, O., Berant, J.: Making retrieval-augmented language models robust to irrelevant context. In: ICLR 2024 Workshop on Large Language Model (LLM) Agents (2024)
61. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-T., Rocktäschel, T., et al.: Retrieval-augmented generation for knowledge-intensive nlp tasks. Adv. Neural. Inf. Process. Syst. **33**, 9459–9474 (2020)
62. Huang, W., Wang, Y., Gan, Z., et al.: An opinion leader mining method based on text contents and network features. World Wide Web **28**, 21 (2025). https://doi.org/10.1007/s11280-025-01331-5
63. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space (2013). arXiv preprint arXiv:1301.3781

64. Reimers, N., Gurevych, I.: Sentence-bert: Sentence embeddings using siamese bert-networks. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing, pp. 3982–3992 (2019)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.