Can LLMs Forecast Internet Traffic from Social Media?

Jonatan Langlet
KTH Royal Institute of Technology & Digital Futures
Stockholm, Sweden

Mariano Scazzariello
RISE Research Institutes of Sweden
Stockholm, Sweden

Flavio Luciani
Namex
Rome, Italy

Marta Burocchi
Namex
Rome, Italy

Dejan Kostić
KTH Royal Institute of Technology
Stockholm, Sweden

Marco Chiesa
KTH Royal Institute of Technology
Stockholm, Sweden

ABSTRACT

Societal events shape the Internet's behavior. The death of a prominent public figure, a software launch, or a major sports match can trigger sudden demand surges that overwhelm peering points and content delivery networks. Although these events fall outside regular traffic patterns, forecasting systems still rely solely on those patterns and therefore miss these critical anomalies.

Thus, we argue for *socio-technical systems* that supplement technical measurements with an active *understanding of the underlying drivers*, including how events and collective behavior shape digital demands. We propose traffic forecasting using signals from public discourse, such as headlines, forums, and social media, as early demand indicators.

To validate our intuition, we present a proof-of-concept system that autonomously scrapes online discussions, infers real-world events, clusters and enriches them semantically, and correlates them with traffic measurements at a major Internet Exchange Point. This prototype predicted between 56-92% of society-driven traffic spikes after scraping a moderate amount of online discussions.

We believe this approach opens new research opportunities in cross-domain forecasting, scheduling, demand anticipation, and society-informed decision making.

1 INTRODUCTION

With the exponential growth of Internet traffic [9, 35] driven by IoT devices, high-resolution streaming, and cloud services, accurate Internet traffic forecasting has become more crucial than ever to prevent costly service outages, degraded user experience, and inefficient resource usage [14].

Today's operators forecast utilization patterns, typically based on *historical trends*, to optimize their infrastructures [3]. However, such forecasting techniques often fall short in anticipating large statistical anomalies, such as extreme sudden surges in utilization caused by impactful non-recurring events like game releases, tournaments, TV series, sports broadcasts, sociopolitical/cultural collective attention, and more [13, 24, 29, 38, 42].

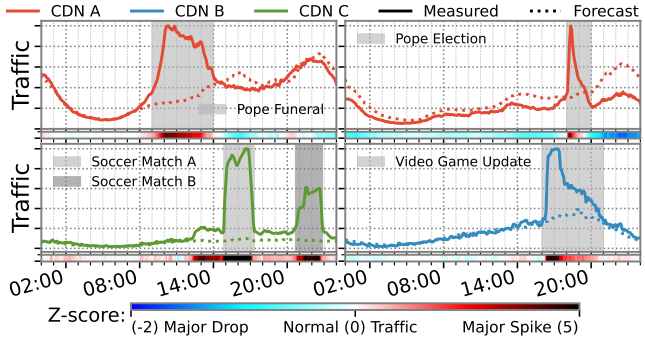

Figure 1: Traffic forecasters cannot predict pattern-breaking event-driven spikes. Each plot shows measured traffic (solid) versus statistical forecasts (dotted) for major CDNs. Z-score color bands indicate the magnitude of deviation. Note how spikes align closely with societal events. Scales and identities omitted per confidentiality policy.

Fig. 1 illustrates four traffic spikes: one video game update (bottom right) and three spikes triggered by the death of the Pope in April 2025: the broadcast of the funeral (top left), the broadcast of the papal election (top right), and two last-minute postponed soccer matches (bottom left). These forecasting failures, stemming from the non-recurring and unexpected nature of the events, can result in tangible operational disruptions that directly impact the end-user experience and system reliability [24, 38].

Operators often resort to *manual* adjustments to prepare their infrastructure for such events. However, *proactively* discovering these events is a time-consuming and complex task. In many cases, operators depend on informal communication channels, such as peer discussions or community chatter, to become aware of potentially impactful events. This reliance on ad hoc knowledge sharing has led to recent initiatives in which operators manually exchange forecasts of anticipated traffic peaks [7]. Even if an event is known, inferring its impact still demands *extensive* manual research into audience size, timing, regionality, and expected data footprint [24].

This entire process is highly resource-intensive, even for hyperscalers, prompting operators to resort to broad overprovisioning as a safeguard [3]. However, overprovisioning imposes increased operational costs and energy consumption. Moreover, as confirmed through conversations with operators, explaining the root societal cause of a spike is often not straightforward, even after a significant spike is verified. This underscores the need for *predictive socio-aware* technologies that understand societal dynamics and their impact on infrastructure.

Although predicting these event-driven spikes is operationally critical, no existing system can reliably extract, structure, and quantify such events, and estimate their digital impact [42]. Today's forecasting systems are largely grounded in low-level, system-centric signals *e.g.,* CPU utilization [22], historical traffic traces [31], and CDN cache hit rates [5].

Bridging society and infrastructure. We argue that *now is the time* for systems to stop treating infrastructure as isolated from society, and begin modelling the reality that drives digital demands. Large Language Models (LLMs) [40] represent a *transformative opportunity*: their unprecedented ability to *interpret* public discourse, *understand* cultural context, and *anticipate* collective behavior makes them uniquely suited to infer relevant events, estimate regional hype, and inform of their infrastructural impact. This shift enables the development of insightful control systems, which we call *socio-technical systems*, that will adapt to societal dynamics, and even go beyond them. For example, socio-technical systems may need to reason both about domain-specific patterns, *e.g.,* whether a soccer team's qualification status affects audience interest in upcoming matches, or reason about synchronized player activities in multiplayer games [30].

Conversely, these systems could allow infrastructure to influence society. For instance, they may identify events competing over the same resources, raising the question: Should impactful events be scheduled to avoid overlap, mitigating the risk of degraded user experience due to network congestion? What about when *event audiences* overlap?

Decoding traffic spikes via public discourse. In this paper, we show how Transformers and LLMs can predict traffic spikes by understanding the underlying societal drivers. We demonstrate the untapped potential in analyzing public discussions, showing that AI can effectively *uncover* and *track* the most relevant upcoming events with sufficient precision to inform traffic forecasting systems. Our findings raise new technical questions and research directions, such as: How can we identify the CDNs implicitly "hosting" an event? And to what extent can we quantify the impact of online engagement by correlating it with shifts in BGP routes?

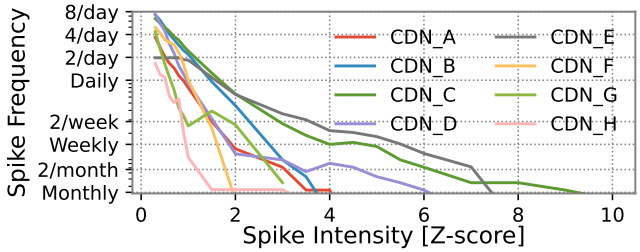

Figure 2: Significant traffic spikes are common, most of which driven by real-world events. We show the spike frequency for a set of major (anonymized) CDNs. The frequency is presented for a range of spike intensities.

Our work, combined with experience in operating a major IXP, has produced insights that we believe will accelerate research into socio-technical systems. We hope these findings will spark broader discussion and inspire future work that builds on and expands beyond the scope of this paper.

Our main contributions include:

- A forecast system for network traffic anomalies using societal signals extracted from public discourse.
- A proof-of-concept AI pipeline that extracts, enriches, deduplicates, clusters, and tracks events to produce abstractions optimized for spike prediction.
- A demonstration that unstructured chatter in aggregate conveys enough information to infer traffic-driving events.
- A public dataset with thousands of traffic-driving events.
- A discussion of opportunities and challenges enabled by socio-technical network systems.

2 MOTIVATION & BACKGROUND

Real-world events frequently trigger sharp, unexpected traffic surges [24–27], leading to degraded user experience [24] or even complete service unavailability [38]. Fig. 1 illustrates this phenomenon using four real-world day-CDN pairs, comparing measured IXP traffic against baseline forecasts for major CDNs serving a European capital. These traffic spikes are driven by societal events, *i.e.,* the papal funeral and election, as well as regionally important soccer matches. Such spikes are directly linked to real-world events, and their characteristics are further shaped by additional societal and geographical factors. For instance, the two soccer matches exhibit distinct traffic patterns: the afternoon match (~3pm) involves a local city team, driving more interest, whereas the evening match (~9pm) is between non-local teams, resulting in still noticeable but less engagement.

Traffic forecasting lacks real-world context. Despite their operational importance, event-driven traffic surges are routinely missed by state-of-the-art forecasting models [14, 32, 33]. These models, trained solely on historical patterns, are fundamentally unaware of one-off societal events. This context blindness has *tangible* consequences for end users. For instance, Netflix experienced a high-profile outage during the Tyson–Paul boxing match [26], resulting in over a million outage reports from 50 countries [38]. These incidents underscore the critical need for a systematic method to bridge the gap between real-world event dynamics and network traffic patterns, reducing the dependency on costly trial-and-error adjustments.

Statistical forecasting often misses critical spikes. Fig. 2 shows that traffic spikes of varying intensities occur regularly at most CDNs. Minor fluctuations (*e.g.,* Z-scores [23] below 2) are common and typically harmless. In contrast, major spikes capable of overwhelming the infrastructure appear multiple times per year, particularly for general-purpose CDNs with limited insight into the content they serve. These are precisely the moments when accurate forecasting matters most and where traditional, history-based models tend to fail. While no public datasets link societal events to traffic surges, our manual investigation of real-world traffic measurements at a major IXP (see Sec. 4.1) suggests that most major spikes are, in fact, *driven by real-world events*.

Predicting traffic requires more than a calendar: it needs cultural context. Previous studies have integrated real-world events into network traffic prediction [1, 4, 29, 41], but they either assume events are already pre-detected and structured [1, 4, 29] or fail to *track* their dynamics [41], limiting effectiveness as events evolve. Yet, anticipating events is not enough, *unexpected changes* in event timing or nature can shift traffic in ways that static models fail to predict (*e.g.,* the postponed matches in Fig. 1). Operators must understand *how much* traffic will flow, *where* and *when*. Take the two matches: although both were part of the same tournament, played on the same day, and delivered through the same CDN, their traffic signatures differ significantly. As noted, this divergence stems from underlying societal factors, *e.g.,* fanbase engagement, that shape demand.

Unfortunately, such contextual metadata is not readily available through traditional event scraping. They are highly dependent on region-specific social knowledge *e.g.,* which sports are popular in a given country or what carries cultural significance. For example, while a soccer match may lead to surges at a European ISP, a US-based ISP is as likely to observe them during NBA games. Capturing this diversity requires not only semantic understanding of events, but also awareness of the broader cultural and societal context in which they occur.

Online chatter holds the missing signal. This is where online discussions become invaluable. Platforms like Reddit, X/Twitter, and news aggregators often hint at traffic-driving events well in advance, signaling timing, scale, sentiment, and availability, either explicitly or implicitly. Unlike static calendars, these dynamic sources reflect what people are *actually planning to do*. In aggregate, they provide rich insight into public interest and emerging trends that, if parsed and structured, could form a live, evolving feed of future digital demands. Even in the papal election case of Fig. 1, where timing is unpredictable due to the conclave's uncertain duration, the system can detect rising interest online and anticipate a spike as soon as the new pope is announced.

LLMs are key enablers. Extracting meaningful information from unstructured and diverse sources is a known complex challenge [37]. In recent years, LLMs have shown exceptional performance across various domains, thanks to their deep semantic and contextual awareness [43]. They excel at filtering out irrelevant or noisy data [17], and can reason through intricate relationships between topics or concepts [6]. While their internal knowledge is limited by a training cutoff, this constraint can be lifted through techniques like tool-augmented training [36], where models dynamically use search tools, and Retrieval-Augmented Generation (RAG) [16], where models access real-time external information. We believe that, when combined with external knowledge sources, LLMs are uniquely equipped to infer relevant events from diverse, unstructured data, placing them in the proper context, both temporally and geographically, while understanding the underlying societal and cultural forces.

We argue for society-aware forecasting. We propose an AI pipeline that extracts, understands, and clusters societal events from online chatter and uses them to predict event-driven traffic. This approach enables proactive allocation, planning, and reduced operational cost. Beyond immediate application, it also opens a new line of research: inferring digital behavior from societal signals at population-scale.[1]

3 PROOF OF CONCEPT

To assess whether online discussions can meaningfully inform traffic forecasts, we built a modular pipeline that parses public discourse, extracts latent societal events, and infers their likely digital footprint. While our current prototype does not yet close the loop with measured traffic data, it aims to approximate how, where, and when such events might shape future load. The overall system is visualized in Fig. 3, where each major component is labeled from ⓐ through ⓓ.

[1] We discuss ethical implications of population-scale "surveillance" in Sec. 5.

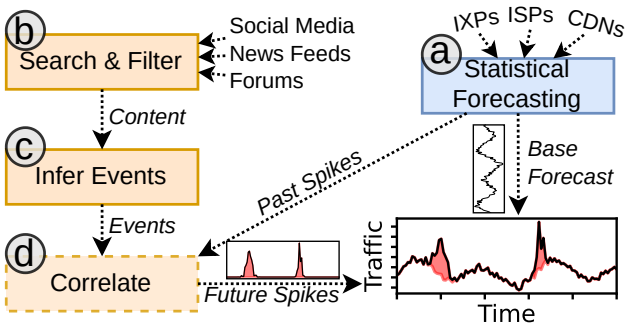

Figure 3: An overview of our approach.

Description	Data Type	Aggregation Method
Date of the event	String	Fixed on creation
Time of the event	String	Fixed on creation
Short description of the event	String	Fixed on creation
Event Category	String	Plurality string
List of relevant entities	String (List)	Entries w/ ≥ 2 votes
Platforms and services	String (List)	Entries w/ ≥ 2 votes
Internet data per user	Integer	Median value
Estimated global audience size	Integer	Median value
Relevance across continents	Float (List)	Per-entry median values
Relevance across nations	Float (List)	Per-entry median values
Duration of the traffic spike	Float	Median value
Likelihood to happen as described	Integer (0–10)	Median value
Semantic categorization vector	Integer (List)	Multi-level clustering

Table 1: Overview of Event Metadata.

3.1 Context-unaware Forecasting (a)

We begin by applying standard time series forecasting techniques to model baseline network traffic. These models capture regular patterns such as weekly cycles and seasonal effects, but remain blind to societal context. While more advanced deep learning forecasters exist [18, 32], they too fail to predict event-driven spikes due to this lack of context. Since our goal is not to optimize time series accuracy but to isolate and understand unexpected load surges, a simple rolling statistical approach suffices as they are known for their reliability [33]. In our solution, by subtracting the predicted baseline from observed traffic, we isolate residual spikes that may reflect external influences or natural variability. Based on manual inspection, nearly all statistically significant spikes (for example, those with $Z \geq 3$) appear to correspond to real-world events (see Sec. 4.1). This decomposition yields a collection of historical anomalies that form the foundation for identifying event-driven patterns and learning how such events may drive future traffic spikes.

3.2 Online Data Collection (b)

To support event inference, we collect and organize public discourse from Reddit, selected for its open API, topical breadth, and active user base. Posts are filtered using combinations of search terms, communities, engagement thresholds, and the presence of outbound links. For each selected post, we retrieve its content, top-level comments, and any linked webpages to capture a rich context. The resulting material is automatically cleaned, preprocessed, and compiled into a single content record per discussion thread.

3.3 Event Inference (c)

The system converts each content record into a streamlined text format suitable for LLM processing. A reasoning model (*i.e.,* 70B DeepSeek-R1 [20]) identifies all upcoming events that might influence network traffic, returning three core attributes: headline, date, and time.

This minimal scope avoids coalescing multiple events within a single discussion, which is common in posts about fixtures, tours, and similar sources with multiple announcements. The extracted events are parsed into structured abstractions and written into the database for downstream processing.

Metadata Inference. Each extracted event is annotated with a rich set of metadata (see Table 1), inferred one at a time by an ensemble of reasoning models (*i.e.,* 14B DeepSeek-R1 [21]). For each field, the system aggregates predictions from three independent LLM runs, repeating if consensus fails.
The core initial metadata, that is, the event category and associated entities (*e.g.,* video game franchises or sport teams), are inferred directly from the extractor output and content record. Later stages use RAG, incorporating relevant Wikipedia articles to provide context and improve inference quality across attributes, such as audience size, geographic reach, and expected data usage.

De-duplication. To handle redundancy, our prototype computes embeddings for all events using an encoder-only Transformer [19] applied to their free-text summaries. Pairwise cosine similarities of the embeddings are used to identify semantically highly similar events occurring on the same date, which are treated as duplicates. When duplicates are found, their content records and metadata inferences are relinked under a single event abstraction, and the redundant entries are removed. All metadata is then re-inferred using the expanded context available after merging.

Semantic Categorization. Since LLM-predicted traffic impacts are inherently noisy and lack ground-truth information, we must correlate events to traffic using historical patterns. Exact matches between past and future events are rare, so we instead learn from semantically similar events. For example, predicting the impact of a UEFA semi-final benefits from prior data on sports matches in general, past UEFA games, and matches involving the same teams (*e.g.,* due to fan bases and regional importance).

Our prototype clusters event embeddings at multiple levels of granularity using k-means, ranging from broad groupings like "sports" at $k = 10$ to "soccer semi-finals involving team X" at $k = 10,000$. Each event is assigned to a cluster at each level, forming a semantic signature that situates it among related events. These dynamically learned categories can be used as features in a prediction model, letting it learn from prior examples even when exact matches do not exist.

3.4 Spike-Event Correlation (d)

Our event abstractions are designed to integrate with context-aware time-series forecasters. While an optimal solution remains an open research challenge, a natural approach is to train models such as the Temporal Fusion Transformer on historical traffic and structured event metadata. Each event contributes a rich set of predictive features, including timing, expected data intensity, duration, geographic scope, relevant platforms, and the multi-level semantic categorization vector. These features help the model learn how different types of events affect different networks in different regions.

By modeling a local time window around each event (*e.g.,* three days), the forecaster can capture both immediate spikes and adjacent effects. Over time, it may also learn negative correlations, such as cases where attention to one event draws users away from other platforms.

3.5 Future Work

While our prototype suffices to demonstrate the feasibility of using online chatter to explain and anticipate traffic spikes, we believe the underlying idea enables more powerful spike prediction systems beyond this initial proof of concept.

Global semantic context. The prototype successfully abstracts events from online content with both direct and implied references. While most relevant context is often captured in surrounding discussions, some broader influences may go unmentioned yet still affect network behavior. For example, an ongoing papal election may significantly alter engagement with a newly released film about the conclave, even if that connection is not made explicit in online chatter.

Incorporating a more global or cultural context, such as prevailing themes or public interests (*e.g.,* a surge in the "zombie" genre), could enhance inference quality. Although our current semantic categorization supports some thematic generalization and spike tuning, it lacks semantic reasoning around broad cultural signals and the current zeitgeist.

Inferring impacted networks. Events rarely specify which networks will serve content, and their delivery paths could dynamically shift. For example, live sports broadcasts may change providers or underlying CDNs [8, 15, 34]. Moreover, routing paths to a given content provider can vary over time due to shifting routing policies or unexpected outages [12]. These route changes can significantly affect the traffic load at different locations depending on the time.

To address this, one solution is to infer likely target networks through learned event-to-traffic patterns and known broadcaster relationships. This is challenging, as commercial agreements are rarely disclosed. One possible approach is to identify events associated with the same inferred service and compare their observed traffic spikes against expected patterns. Consistent deviations can reveal which CDN likely serves the content, thereby tracking these changes. Also, one could incorporate knowledge about BGP updates to infer likely impacted on-path networks (*e.g.,* IXPs).

Filtering massive online content. Reddit receives nearly a million new posts daily [2], whereas X/Twitter sees hundreds of millions [39]. Only a small fraction relates to potentially traffic-driving events, making large-scale analysis costly. To reduce overhead, this stream of content must be filtered before analysis to increase the density of relevant signals and improve the timeliness spike prediction. However, overly aggressive filtering risks missing events with low global visibility but high local impact, such as regional sports matches.

One could design a feedback mechanism that continuously trains a content filtering model based on what content typically leads to high-quality event information. For instance, an encoder-only Transformer can be fine-tuned to quickly estimate contents' relevance before event inference.

4 FEASIBILITY EVALUATION

To the best of our knowledge, *tracking* events, sentiment, and Internet utilization through online chatter is entirely novel. Existing systems do not track the societal context behind anomalies, whereas our entire focus is on doing so. As such, this section does not attempt to compare against prior work but instead examines whether online discussions carry enough signal to support event-driven forecasting.

For our proof-of-concept, we scraped 38,000 webpages, including Reddit, linked websites, and Wikipedia. From this, ~10,000 unique traffic-driving events were inferred and enriched through ~200 A100 [28] compute hours.

Events will be released as a dataset alongside this paper.

4.1 Spike Coverage

Unfortunately, there are no publicly available datasets mapping societal events to their network impact. Therefore, we manually curated a small set of traffic spikes with known causes, identified through public sources such as news articles, release calendars, and operator blog posts. Due to the extensive effort in manually researching individual spikes, we settled with 29 verified event-spike pairs[2].

For each explained spike, we checked whether our system had inferred the same event at that time.

[2]We selected spikes with contiguous $Z \geq 2$ and duration ≥ 20 minutes from between March and July 2025.

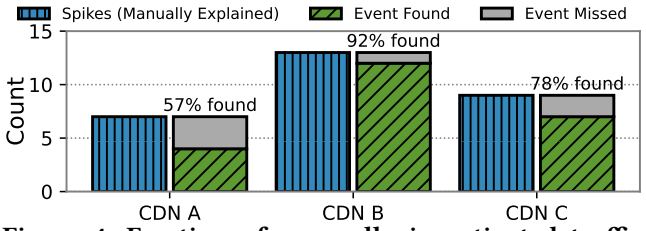

Figure 4: Fraction of manually investigated traffic spikes that were successfully predicted in our prototype, across three CDNs.

As shown in Fig. 4, the system correctly matched a majority of these cases, with coverage rates ranging from 57% to 92% across CDNs.

Although these results already support our idea, some events are missed. These fall into one of two categories: (1) the event *is* discussed in advance, but the scraper had not yet crawled that content, or (2) the event is spontaneous, triggered by breaking news, and therefore appeared with minimal notice (*e.g.,* the death of a public figure).

Takeaway: Public online discourse holds great promise as a source for accurately predicting real-world traffic events, even without manual tuning or ground-truth labels.

4.2 Event-discussion Timelines

We extract upcoming events from general Reddit discussions and measure how far in advance they are mentioned. Fig. 5 shows the cumulative fraction of events first detected at various lead times, grouped by category. Categories with fewer than 1,000 inferred events are grouped as "Others".

TV & Film events typically surface early, and nearly half of them are found in Reddit discussions one month before they occur. Sports exhibit shorter notice periods. While fixtures are typically released in a structured format ahead of time, crucial details, like team matchups, recent performances, and audience anticipation, tend to unfold closer to the event date. Video game events have a mixed behavior, combining early mentions tied to official release dates with last-minute spikes from reviews or content updates.

While this evaluation is based on a prototype pipeline sampling a small subset of online discussions, the trends reflect real structural differences. Some events follow fixed calendars, while others depend on evolving context. For instance, while a movie release might be advertised years in advance, context-dependent information such as actors, budgets, and public interest dynamically evolves up until the release itself. Our results highlight the potential of general-purpose platforms to provide detailed and context-dependent signals for demand-driven events ahead of time.

Takeaway: Online discussions reveal not just that an event will happen, but when public attention begins to build. Different domains show distinct lead time profiles, reflecting how event-information and public interest evolve.

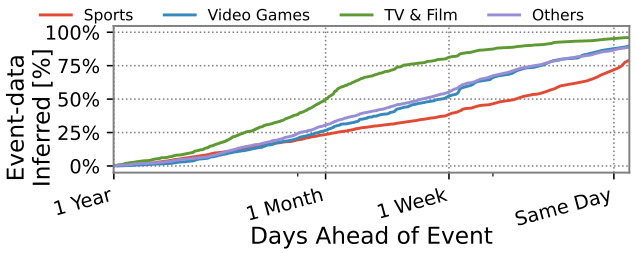

Figure 5: Cumulative fraction of event mentions by time-to-event, grouped by category. Many events are discussed well in advance, while context and engagement often intensify closer to the event date.

5 DISCUSSION

Future application domains. While our primary goal is context-aware traffic forecasting for resource allocation and network management, the event abstraction methodology has applications in many other areas.

Regional event predictions, for example, can guide *proactive CDN cache placement* in areas of anticipated demand or *anticipate mass human movement* for mobile load balancing, GPS routing, and traffic planning during major events.

Socio-technical systems are vulnerable to manipulation via misinformation and "fake news". However, by systematically correlating predicted events with observed utilization data, these systems could identify and downweight sources whose claims repeatedly fail to materialize in empirical data.
This weighting offers a novel, empirically grounded signal of source trustworthiness that could extend beyond forecasting to support misinformation detection, media forensics, or trust-aware content ranking in algorithmic feeds.

Limitations. While this paper highlights the predictive power of online discourse, there are fundamental limitations that constrain the practicality and coverage of this approach.

First, our method assumes reliable access to public online discussions, either through APIs or web scraping. However, access to such data is neither guaranteed nor stable. APIs may impose strict rate limits, require expensive subscriptions, or be deprecated entirely. Web scraping, while more flexible, is increasingly hindered by anti-bot measures [10]. As platforms adopt more aggressive protection against automated access, sustaining large-scale real-time content ingestion becomes more challenging. In short, while the Internet talks about what it will do, *it is increasingly hard to listen.*

Second, certain events occur with little or no advance notice. These include sudden celebrity announcements, geopolitical shocks, and private decision reveals (such as surprise product launches). Because our system relies on textual indicators in public platforms to infer future traffic surges, it will either miss such "spontaneous" events entirely or only detect them as they occur. In such cases, the predictive power of our system collapses to a real-time explainer at best.

How far should LLMs go? As LLMs predict digital events, a key ethical question arises: how far should they go to detect early signals? Some, like traffic spikes from coordinated actions, may stem from private discussions. For example, in games like EVE Online, a 6,000-player assault might only become visible after it happens. Should LLMs "infiltrate" such groups for early insight or would that cross an ethical line?

The situation becomes murkier when game mechanics create fixed, publicly known attack windows. While such structures can make events technically predictable, the actual scale of participation depends on players' strategic choices that are often coordinated in private channels with occasional leaks on public forums. During the largest in-game war [11], 12,000 players coordinated in private channels an attack during a publicly known attack window. Such private channels make potential outages difficult to anticipate.

Our prototype avoids non-public data by design, but as LLMs become more persuasive, the line between monitoring and unethical surveillance blurs. Actors must ask not just *can* we predict an event, but *should* we? This goes beyond gaming. Should LLMs predict DDoS attacks if it means learning from semi-private or encrypted spaces like Telegram?

6 CONCLUSION

Societal dynamics and Internet traffic are tightly coupled. Popular events generate surges in demand, while the way discussions unfold provides crucial metadata for anticipating their impact. From nearly 10,000 unique events extracted from online discourse, our prototype predicted 57–92% of major CDN traffic spikes. This shows that public chatter already encodes much of tomorrow's user behavior, and that with the right abstractions, systems can learn to listen.

We do not offer a complete forecasting solution, but evidence that socio-technical forecasting is both feasible and filled with open questions: how to map events to the delivery networks that serve them, how to capture cultural and regional factors that shape demand, and how to separate fleeting noise from traffic-shaping signals. Addressing these challenges will enable infrastructure that is not only reactive but anticipatory of the society that drives it.

ACKNOWLEDGEMENTS

This work has been partially supported by Knut and Alice Wallenberg Foundation (Wallenberg Scholar Grant for Prof. Dejan Kostić), Vinnova (Sweden's Innovation Agency), the Swedish Research Council (agreement No. 2021-04212), and KTH Digital Futures. The computations were enabled by resources provided by the National Academic Infrastructure for Supercomputing in Sweden (NAISS) through grant agreements no. 2024/22-938 & 2025/22-924.

REFERENCES

[1] Eilaf M.A Babai and Koji Okamura. 2024. A Contextual Approach for Improving Anomalous Network Traffic Flows Prediction . In *2024 IEEE 48th Annual Computers, Software, and Applications Conference (COMPSAC)*. IEEE Computer Society, Los Alamitos, CA, USA, 2203–2208. https://doi.org/10.1109/COMPSAC61105.2024.00353

[2] Backlinko. 2025. Reddit User and Growth Stats. https://backlinko.com/reddit-users#how-many-posts-are-shared-on-reddit-each-year. Accessed: 2025-07-07.

[3] Dario Bega, Marco Gramaglia, Marco Fiore, Albert Banchs, and Xavier Costa-Perez. 2019. DeepCog: Optimizing resource provisioning in network slicing with AI-based capacity forecasting. *IEEE Journal on Selected Areas in Communications* 38, 2 (2019), 361–376.

[4] Juan L. Bejarano-Luque, Matías Toril, Mariano Fernández-Navarro, Carolina Gijón, and Salvador Luna-Ramírez. 2021. A Deep-Learning Model for Estimating the Impact of Social Events on Traffic Demand on a Cell Basis. *IEEE Access* 9 (2021), 71673–71686. https://doi.org/10.1109/ACCESS.2021.3078113

[5] Daniel S Berger. 2018. Towards lightweight and robust machine learning for cdn caching. In *Proceedings of the 17th ACM Workshop on Hot Topics in Networks*. 134–140.

[6] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. 2023. Sparks of Artificial General Intelligence: Early experiments with GPT-4. arXiv:2303.12712 [cs.CL] https://arxiv.org/abs/2303.12712

[7] CDN Alliance. 2025. Traffic Radar Working Group. https://cdnalliance.org/activities/working-groups/traffic-radar/.

[8] Fangfei Chen, Ramesh K. Sitaraman, and Marcelo Torres. 2015. End-User Mapping: Next Generation Request Routing for Content Delivery. In *Proceedings of the 2015 ACM Conference on Special Interest Group on Data Communication* (London, United Kingdom) *(SIGCOMM '15)*. Association for Computing Machinery, New York, NY, USA, 167–181. https://doi.org/10.1145/2785956.2787500

[9] Cisco Systems Inc. n.d.. Cisco Annual Internet Report (2018–2023). https://www.cisco.com/c/en/us/solutions/collateral/executive-perspectives/annual-internet-report/white-paper-c11-741490.html. White Paper. Accessed: 2025-07-07.

[10] Cloudflare. 2025. Cloudflare Just Changed How AI Crawlers Scrape the Internet-at-Large; Permission-Based Approach Makes Way for A New Business Model. https://www.cloudflare.com/en-gb/press-releases/2025/cloudflare-just-changed-how-ai-crawlers-scrape-the-internet-at-large/.

[11] EVE Online. 2021. The Massacre at M2-XFE. https://www.eveonline.com/news/view/the-massacre-of-m2-xfe.

[12] Fast Company. 2024. The Baltic undersea cable cutting highlights the internet's underlying vulnerabilities. https://www.fastcompany.com/91232785/baltic-undersea-cable-cutting-internet-vulnerability.

[13] Anja Feldmann, Oliver Gasser, Franziska Lichtblau, Enric Pujol, Ingmar Poese, Christoph Dietzel, Daniel Wagner, Matthias Wichtlhuber, Juan Tapiador, Narseo Vallina-Rodriguez, et al. 2020. A view of Internet Traffic Shifts at ISP and IXPs during the COVID-19 Pandemic. In *COVID-19 Network Impacts Workshop*. IAB.

[14] Gabriel O Ferreira, Chiara Ravazzi, Fabrizio Dabbene, Giuseppe C Calafiore, and Marco Fiore. 2023. Forecasting network traffic: A survey and tutorial with open-source comparative evaluation. *IEEE Access* 11 (2023), 6018–6044.

[15] Fox Sports Press. 2024. Serie A And Fox Deportes Announce Landmark U.S. Spanish-language Media Rights Agreement.

https://www.foxsports.com/presspass/blog/2024/08/16/serie-a-and-fox-deportes-announce-landmark-u-s-spanish-language-media-rights-agreement/. Accessed: 2025-07-07.

[16] Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yixin Dai, Jiawei Sun, Haofen Wang, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997* 2, 1 (2023).

[17] Izzeddin Gur, Ofir Nachum, Yingjie Miao, Mustafa Safdari, Austin Huang, Aakanksha Chowdhery, Sharan Narang, Noah Fiedel, and Aleksandra Faust. 2023. Understanding HTML with Large Language Models. arXiv:2210.03945 [cs.LG] https://arxiv.org/abs/2210.03945

[18] Chih-Wei Huang, Chiu-Ti Chiang, and Qiuhui Li. 2017. A study of deep learning networks on mobile traffic forecasting. 1–6. https://doi.org/10.1109/PIMRC.2017.8292737

[19] Hugging Face. n.d.. all-MiniLM-L6-v2. https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2. Accessed: 2025-07-07.

[20] Hugging Face. n.d.. DeepSeek-R1-Distill-Llama-70B. https://huggingface.co/deepseek-ai/DeepSeek-R1-Distill-Llama-70B. Accessed: 2025-07-07.

[21] Hugging Face. n.d.. DeepSeek-R1-Distill-Qwen-14B. https://huggingface.co/RedHatAI/DeepSeek-R1-Distill-Qwen-14B-quantized.w4a16. Accessed: 2025-07-07.

[22] Deepak Janardhanan and Enda Barrett. 2017. CPU workload forecasting of machines in data centers using LSTM recurrent neural networks and ARIMA models. In *2017 12th international conference for internet technology and secured transactions (ICITST)*. IEEE, 55–60.

[23] Erwin Kreyszig. 1979. Advanced Engineering Mathematics (Fourth ed.). Wiley. p. 880, eq. 5.

[24] Flavio Luciani. 2022. The Elephant Effect - Considerations on Live Streaming Italy's Serie A Championship. https://labs.ripe.net/author/flavio_luciani_1/the-elephant-effect-considerations-on-live-streaming-italys-serie-a-championship/. Accessed: 2025-07-07.

[25] Doug Madory. 2024. Anatomy of an OTT Traffic Surge: The Fortnite Chapter 2 Remix Update. https://www.kentik.com/blog/anatomy-of-an-ott-traffic-surge-the-fortnite-chapter-2-remix-update/.

[26] Doug Madory. 2024. Anatomy of an OTT Traffic Surge: The Tyson-Paul Fight on Netflix. https://www.kentik.com/blog/anatomy-of-an-ott-traffic-surge-the-tyson-paul-fight-on-netflix/.

[27] Doug Madory. 2025. Anatomy of an OTT Traffic Surge: Netflix Rumbles Into Wrestling. https://www.kentik.com/blog/anatomy-of-an-ott-traffic-surge-netflix-rumbles-into-wrestling/.

[28] NVIDIA. 2025. NVIDIA A100. https://www.nvidia.com/en-us/data-center/a100/.

[29] Andrea Pimpinella, Alessandro EC Redondi, Andrea Pavon, and Luisa Venturini. 2022. Using the (crystal) ball: Forecasting network traffic peaks with football events. In *GLOBECOM 2022-2022 IEEE Global Communications Conference*. IEEE, 4334–4339.

[30] Polygon. 2021. Players in Eve Online broke a world record – and then the game itself. https://www.polygon.com/2021/1/5/22214982/eve-online-world-record-massacre-m2-xfe-ghost-titans.

[31] Nipun Ramakrishnan and Tarun Soni. 2018. Network traffic prediction using recurrent neural networks. In *2018 17th IEEE international conference on machine learning and applications (ICMLA)*. IEEE, 187–193.

[32] Sajal Saha, Anwar Haque, and Greg Sidebottom. 2022. Deep Sequence Modeling for Anomalous ISP Traffic Prediction. In *ICC 2022 - IEEE International Conference on Communications*. 5439–5444. https://doi.org/10.1109/ICC45855.2022.9838262

[33] Sajal Saha, Anwar Haque, and Greg Sidebottom. 2022. An empirical study on internet traffic prediction using statistical rolling model. In *2022 International Wireless Communications and Mobile Computing (IWCMC)*. IEEE, 1058–1063.

[34] Serie A. 2024. CBS Sports And Serie A Announce Renewal Of Media Rights Agreement In The U.S. https://www.legaseriea.it/en/media/serie-a/cbs-sports-and-serie-a-announce-renewal-of-media-rights-agreement-in-the-u-s. Accessed: 2025-07-07.

[35] Amin Shahraki, Mahmoud Abbasi, Md Jalil Piran, and Amir Taherkordi. 2021. A comprehensive survey on 6G networks: Applications, core services, enabling technologies, and future challenges. *arXiv preprint arXiv:2101.12475* (2021).

[36] Zhuocheng Shen. 2024. Llm with tools: A survey. *arXiv preprint arXiv:2409.18807* (2024).

[37] Stefan Stieglitz, Milad Mirbabaie, Björn Ross, and Christoph Neuberger. 2018. Social media analytics – Challenges in topic discovery, data collection, and data preparation. *International Journal of Information Management* 39 (2018), 156–168. https://doi.org/10.1016/j.ijinfomgt.2017.12.002

[38] TechCrunch. 2024. Jake Paul vs. Mike Tyson fight shows Netflix still struggles with live events. https://techcrunch.com/2024/11/16/jake-paul-vs-mike-tyson-fight-shows-netflix-still-struggles-with-live-events/. Accessed: 2025-07-05.

[39] The Social Shepherd. 2025. 21 Essential Twitter (X) Statistics You Need to Know in 2025. https://thesocialshepherd.com/blog/twitter-statistics. Accessed: 2025-07-07.

[40] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).

[41] Xinlei Wang, Maike Feng, Jing Qiu, Jinjin Gu, and Junhua Zhao. 2024. From news to forecast: Integrating event analysis in llm-based time series forecasting with reflection. *Advances in Neural Information Processing Systems* 37 (2024), 58118–58153.

[42] Liang Zhao. 2021. Event prediction in the big data era: A systematic survey. *ACM Computing Surveys (CSUR)* 54, 5 (2021), 1–37.

[43] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. 2025. A Survey of Large Language Models. arXiv:2303.18223 [cs.CL] https://arxiv.org/abs/2303.18223