Measuring GenAI Usage Patterns in a University Campus via Network Traffic Analysis

Minzhao Lyu
University of New South Wales
Sydney, NSW, Australia
minzhao.lyu@unsw.edu.au

Yifan Wang
University of New South Wales
Sydney, NSW, Australia
wangyifan.frank@unsw.edu.au

Vijay Sivaraman
University of New South Wales
Sydney, NSW, Australia
vijay@unsw.edu.au

ABSTRACT

Generative AI platforms backed by large-language models (LLMs) are taking the world by storm. Starting with ChatGPT launched a mere 18 months ago that can generate amazingly human-like text responses to prompts, there are now platforms that can generate code (GitHub Copilot), images (Dall-E), and even video clips (Sora). In this fast evolving world of GenAI, there is huge interest in the community in tracking the usage patterns of these platforms, as well as performance in terms of responsiveness and network load.

Our paper is the first attempt to track usage of emerging GenAI platforms via real-time analysis of network traffic. This can be useful to enterprises seeking to know which GenAI services their employees use most; to Communications Service Providers wanting to know the network loads imposed; and to financial investors needing a pulse on market trends. We begin by explaining the network anatomy of ChatGPT prompt/response interactions in detail, and extend it to six other GenAI platforms supporting text, code, and image generation. We then develop a measurement method to identify and quantify GenAI interactions via real-time analysis of network traffic. We deploy our monitoring system in a University campus over a 5-month period, and reveal interesting insights such as GenAI usage distribution across days of the week and deviations during assessment periods; variation in prompt-to-response-size ratios across the various GenAI platforms; and differences in response times arising from model versions.

CCS CONCEPTS

• **Networks → Network measurement**.

KEYWORDS

Generative AI platform, network traffic analysis

ACM Reference Format:
Minzhao Lyu, Yifan Wang, and Vijay Sivaraman. 2024. Measuring GenAI Usage Patterns in a University Campus via Network Traffic Analysis. In *ASIAN INTERNET ENGINEERING CONFERENCE 2024 (AINTEC '24), August 09, 2024, Sydney, NSW, Australia.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3674213.3674214

1 INTRODUCTION

Generative AI services have surged in prominence over the past 18 months, since the launch of ChatGPT in November 2022. It has revolutionized the way text, code, images, and even videos are created. Tech giants and well-funded unicorns like OpenAI, Google, Microsoft, and Anthropic are accelerating this movement by operating online GenAI platforms that provide users access to well-trained large language models (LLMs), and rapidly launching new versions of LLMs with better response quality and performance. Table 1 lists seven popular GenAI platforms we consider in this paper, including their owner, launch date, estimated market value, and generation capability, that are collected or estimated from various public sources [1, 8, 9, 19, 22–26]. Staggeringly, the GenAI industry is expected to grow at 47.5% CAGR to be worth 1.3 trillion dollars by 2032, overtaking the video streaming and online gaming industry.

Universities and enterprises are keen to understand how their students or workforce use the various GenAI platforms so they can appropriately manage subscriptions and responsible use; telecommunications service providers are anxious about the load GenAI places on their communications infrastructure so they can provision capacity and cloud connectivity; and market watchers are eager to see how GenAI usage and performance trend over time across the various platforms so they can make appropriate investment decisions. As far as we are aware, there are no mechanisms today for these various entities to get timely visibility into GenAI usage and performance in order to make these day-to-day decisions.

This paper is the first study in analysing network traffic in real-time to track the usage and performance of GenAI platforms. By analyzing the network traffic pertaining to

AINTEC '24, August 09, 2024, Sydney, NSW, Australia

ACM ISBN 979-8-4007-0985-2/24/08.
https://doi.org/10.1145/3674213.3674214

Table 1: Seven popular GenAI platforms launched since late 2022, their estimated market values and generative capabilities.

Platform [*Owner*]	**Launched**	**Value**	**Capability**
ChatGPT [*OpenAI*]	Nov 2022	$86bn	text
Bard [*Google*]	Mar 2023	>$100bn	text/image
Claude [*Anthropic*]	Mar 2023	$4bn	text
NewBing [*Microsoft*]	Feb 2023	$154bn	text/image
Dall-E [*OpenAI*]	Aug 2023	$86bn	image
Copilot [*Github*]	Dec 2023	>$3bn	code
Gemini [*Google*]	Feb 2024	>$100bn	text/image

seven popular GenAI platforms (specifically ChatGPT, Bard, Claude, NewBing, Dall-E, Gemini, and Github Copilot), we are able to identify the exchange of prompt and response messages, along with attributes such as message lengths and response times. By deploying this system in a University campus, we draw insights into how often each of these GenAI platforms are used, which days they are used most, how many prompt-response interactions are contained within a session, how response sizes correspond with prompt sizes, and how long they take to respond. Our contributions are summarized as three-fold.

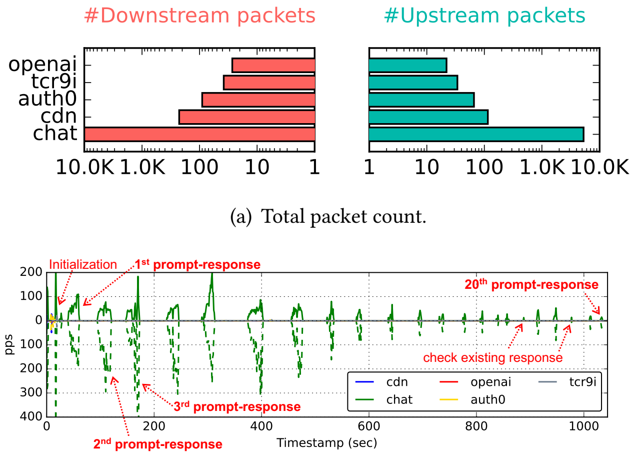
Our **first** contribution (§2) systematically studies the characteristics of network flows and their volumetric patterns when users converse with the seven studied GenAI platforms. Particularly, we dissect the time-series volumetric behaviors of the conversation flow when serving each pair of user prompt and platform response.

Our **second** contribution (§3) develops a real-time network traffic analysis method to detect user sessions and prompt-response pairs of the seven studied generative AI platforms and measure their timing and volumetric usage metrics.

Our **third** contribution (§4) deploys the measurement system in a University campus network and collects data over 5 months (late September 2023 to early March 2024) to reveal interesting insights – ChatGPT is used most heavily, with usage spikes corresponding to news events; Github Copilot usage is aligned with University assessment schedules; NewBing and Dall-E have both large response sizes, but prompts are much shorter in NewBing; several GenAI platforms show multimodal response times, which could be indicative of different model capabilities and versions.

2 TRAFFIC CHARACTERISTICS OF GENAI PLATFORMS

In this section, we discuss network traffic characteristics of seven popular GenAI platforms that have conversations with users by pairs of user prompts and server responses.



(a) Total packet count.

(b) Packet rate.

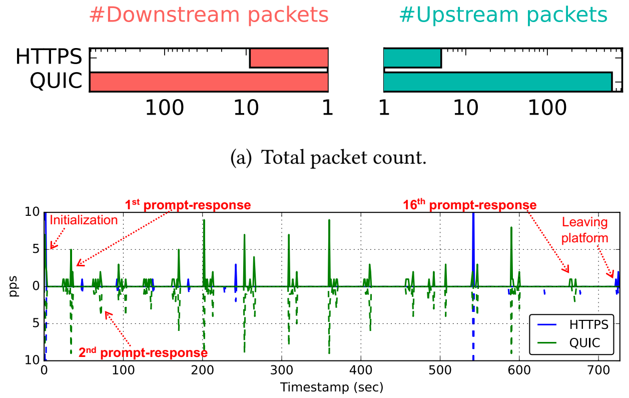
Figure 1: Volumetric profile of service flow labeled by their name prefixes in a ChatGPT session.

By analyzing collected ground-truth traffic capture (PCAP) files of our conversations with the seven platforms (§2.1), we identify important network flows that exhibit distinct volumetric patterns when serving conversations (§2.2) and characterize the delivery of each prompt-to-response pair by the volumetric patterns of service flows (§2.3).

2.1 Lab Experiments

This study covers seven popular GenAI platforms in the market, including OpenAI's ChatGPT, OpenAI's Dall-E, Google's Bard, Google's Gemini, Microsoft's NewBing (now known as Copilot), GitHub's Copilot, and Anthropic's Claude. To have a proper understanding of network behaviors of GenAI platforms when serving user questions (*i.e.,* prompts), we instructed a list of prompts with various word counts from five to the maximum limits (*e.g.,* 500 words for ChatGPT) to be answered by each studied platform. We diversify our user setups by accessing the platforms using three different browsers (Safari, Chrome, and Edge) on both PCs (a MacBook laptop and a Windows desktop) and mobile devices (iOS and android phones connected to wireless hotspot on the Windows desktop).

Network traffic exchanged between our client devices and GenAI platforms on the Internet are captured as PCAP files (one per session) using Wireshark software on PC devices. We analyze the collected PCAP files with our ground-truth prompt-response information to illustrate the network traffic characteristics of GenAI services.



(a) Total packet count.

(b) Packet rate.

Figure 2: Volumetric profile of service flows labeled by their protocols in a Bard session.

2.2 Service Flows and Volumetric Patterns

We first discuss how network flows are utilized to deliver GenAI services using two representative sessions, one for OpenAI's ChatGPT through Chrome browser on an iOS mobile phone and one for Google's Bard through Safari browser on a macOS laptop. In each session, we asked about twenty prompts with several seconds of gap between the adjacent ones and received responses of various word counts, ranging from five words to above one thousand words. Similar insights are obtained for other types of user setups and GenAI platforms with variations only in their service domain names, thus, are not repetitively discussed.

In our packet analysis, we track all network flows (identified by 5-tuples) that support our conversations with GenAI platforms. We observe that they are all TLS encrypted with a Server Name Indication (SNI) field showing the accessed service domain in handshake packets. We therefore label the flows with their contextual service names.

A ChatGPT Session: Network flows that correspond to five unique services are belonging to the domain "*openai.com*" but with unique prefixes, including "*chat*", "*cdn*", "*auth0*", "*tcr9i*" and "*openai*", for user conversation, archived historical content, user authentication, model selection, and platform administration, respectively. In Fig. 1(a), we show the total number of packet counts in both upstream (*i.e.,* client to platform) and downstream (*i.e.,* platform to client) directions, as annotated on the graphs. It is quite clear that the "*chat*" service dominates the total packet counts in both upstream and downstream directions, which is purposed to deliver chat content during the session, followed by "*cdn*" service that is typically for archived content distribution, such as saved chat history and user settings. Other three services that together contribute to less than one third of packet counts are used for administrative purposes, authentication and model selection.

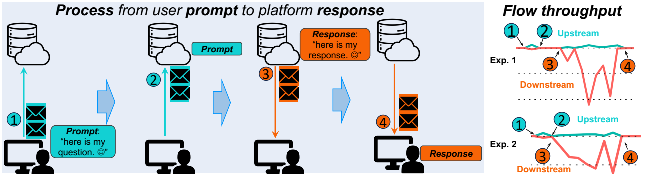

Process from user prompt to platform response
Prompt: "here is my question. ☺"
Prompt
Response: "here is my response. ☺"
Response
1
2
3
4
Flow throughput
Exp. 1
Exp. 2
Upstream
Downstream

Figure 3: Delivery process of one prompt-to-response pair and the underlying flow throughput pattern.

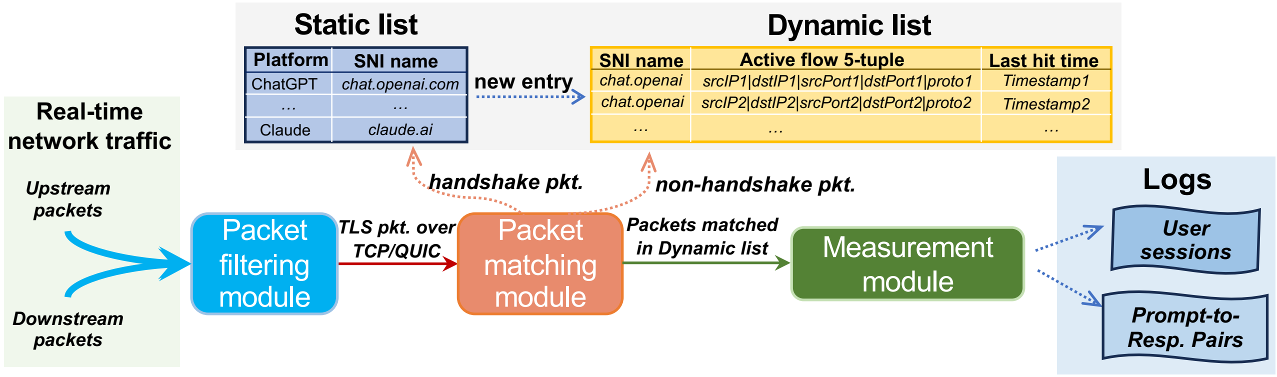
After having a summary-level view of accessed services, we now look at their time-series volumetric patterns as presented in Fig. 1(b). Except for the network flow of "*chat*" service, all other flows are mostly only active during the initialization period (first group of packet spikes before the 20-second timestamp) when we open the browser page of ChatGPT and login using our account. Notably, during this period, "*cdn*", "*openai*", and "*auth0*" have packets being exchanged for page preparation, login authentication and account synchronization, followed by "*tcr9i*" for LLM model selection.

Each bulk of "*chat*" packets after the initialization phase are triggered by our prompt requests. Not surprisingly, the size of each group of upstream or downstream packets is positively related to the size of prompt sent by the user and response sent by the platform. We acknowledge that a precise inference mechanism can be developed to infer word counts in prompts and responses from network traffic as a future work.

A Bard Session: In our representative Bard session, we asked fifteen prompts with response word counts ranging from 1504 to 10. All network flows for Bard session are with the same service name "*bard.google.com*". Unlike ChatGPT and other studied platforms that use multiple services each with a unique purpose, from our repetitive experiments, Google's Bard uses HTTPS flows for administrative and supporting services and use QUIC protocol (developed and promoted by Google) for chat data delivery. Therefore, as shown in Fig. 2(a), for the total packet count belonging to HTTPS and QUIC flows, QUIC dominates by over ten times compared to HTTPS. According to the time-series plot of their packet rates, the QUIC flow becomes active when we send a prompt, while the HTTPS flow is mostly active during the initialization phase and other events such as when we retrieve saved chats from history.

2.3 User Prompts and Platform Responses

After knowing how network service flows are used to deliver a conversation, we now characterize the delivery process



Platform	SNI name
ChatGPT	*chat.openai.com*
…	…
Claude	*claude.ai*

SNI name	Active flow 5-tuple	Last hit time
chat.openai	*srcIP1|dstIP1|srcPort1|dstPort1|proto1*	*Timestamp1*
chat.openai	*srcIP2|dstIP2|srcPort2|dstPort2|proto2*	*Timestamp2*
…	…	…

Figure 4: Real-time system for measuring user conversations with GenAI platforms.

of each pair of client prompt and platform response, *e.g.*, a group of upstream and downstream packets in Fig. 1(b) and Fig. 2(b). We abstract a four-step logical view of each prompt-to-response interaction between a user and a GenAI platform in the left shaded part of Fig. 3. In stage ①, the user types a prompt on the local device and clicks the "send" button so that the input prompt starts being delivered to the GenAI platform over the Internet. The prompt is fully delivered to the platform in stage ②. The platform then uses its large language model to generate a response which starts being sent to the user in stage ③ and is fully received by the user in stage ④. If we annotate the timestamp of each stage occurrence as t_1, t_2, t_3, and t_4, the user waits $\Delta t_{uw} = t_4 - t_1$ (**user waiting time**) to receive a complete response after clicking the "send" button; the genAI server takes $\Delta t_{sr} = t_3 - t_2$ (**server response time**) to select models, provision computing resources and start sending response; and the platform takes $\Delta t_{rg} = t_4 - t_3$ (**response generation time**) to complete response generation.

Two examples of the time-series network flow throughput relating to the four logical steps are shown in the right part of Fig. 3. Example 1 shows a prompt-to-response pair with a long server response time ($t_3 - t_2$) as about 1 second, whereas example 2 has a very short server response time (less than 10ms). It is quite clear that upstream traffic throughput, measured by Byte-per-second (Bps) is higher than downstream throughput when the user is sending prompt as the server replies acknowledgment packets without payload. The downstream throughput is larger than the upstream rate when the server is sending response and the user replies zero-payload acknowledgment packets.

Later on, we will discuss our heuristic measurement techniques for each prompt-to-response pair by capturing the discussed network behaviors.

3 REAL-TIME MEASUREMENT SYSTEM

In this section, we design a real-time network traffic analysis pipeline (§3.1) to detect active user conversations with generative AI platforms and measure their usage patterns using our heuristic measurement techniques (§3.2).

3.1 Network Traffic Processing Pipeline

We design a pipeline that processes real-time network traffic (*i.e.*, packet streams in both upstream and downstream directions) exchanged between users and the Internet to support measurement of user sessions with GenAI platforms. Our pipeline uses virtual network functions (VNF) deployed on a generic server that receives mirrored network traffic from the edge router of a monitored network.

As visually depicted in Fig. 4, our pipeline receives real-time packet streams on two network interfaces (NIC) for upstream and downstream directions, respectively. To avoid unnecessary processing costs on packets that are irrelevant to the conversations, all packets first go through a **packet filtering module** for standard packet header checks. As discussed in §2.2, all conversations with the seven studied platforms are carried by TLS flows over TCP or QUIC. Therefore, all other packets, as identified by their standard header fields, will be discarded without further processing. Note that this stateless module can alternatively be offloaded to programmable data-plane (P4) switches for large-scale deployment.

The second module, namely **packet matching**, extracts server name indication (SNI) fields from TLS handshake packets and checks the SNI records against a static list of service names that carries conversation data of the seven monitored GenAI platforms, such as "*chat.openai.com*" for ChatGPT and "*gemini.google.com*" for Google's Gemini, as discussed in §2.2. Flow 5-tuples of a matched TLS handshake packet will be recorded in a dynamic list to match subsequent packets that belong to the service flow. A recorded 5-tuple

Table 2: Collected dataset for the usage of seven monitored GenAI platforms in our five-month campus deployment.

	chatGPT	**bard**	**claude**	**newbing**	**dall-e**	**copilot**	**gemini**
measurement start date	10/09/2023	10/09/2023	31/10/2023	06/11/2023	06/11/2023	14/11/2023	12/02/2024
#total measurement days	165	165	122	117	119	111	21
#effective measurement days	151	151	118	115	118	110	19
#sessions	39,679	52	82	24	3	331	19
#prompts	54,448	193	83	24	4	663	31

that is not matched by any packet for a certain duration (*i.e.,* 60 seconds in our implementation) will be removed from the dynamic list. All packets that are matched with a service flow of one conversation will be forwarded to our **measurement module**.

3.2 Heuristic Measurement Techniques

The measurement module in our real-time traffic processing pipeline is developed based on our heuristic understanding of network characteristics in §2. Upon arrival of each packet, the module extracts useful information from their metadata fields and headers in the format of <*service name, flow 5-tuple, timestamp, payload size*>. The service name and flow 5-tuple are used to associate a packet with its user session. Timestamp and payload size of the packet are used as metrics for usage measurement.

As visually depicted in Fig. 1(b) and 2(b), a user session with GenAI platforms can have multiple pairs of prompts and the corresponding responses. Each pair is carried by a standalone chunk in the time-series volumetric profile (*i.e.,* packet rate and throughput) of the service flow. Therefore, we employ state-of-the-art time-series chunk detection algorithms [10, 17] with a window-based approach [20, 30] to identify each prompt-to-response pair from each service flow. For each detected prompt-to-response pair, in addition to their volumetric profiles, we also measure their timing metrics including user waiting time Δt_{uw}, server response time Δt_{sr}, and response generation time Δt_{rg} precisely from the time-series volumetric patterns as discussed in §2.3.

Our measurement module produces a record per user session and per prompt-to-response pair to enable insight analysis as will be discussed next. Both record types contain metadata including service name (*e.g.,* "*chat.openai*"), flow 5-tuple, start and finish timestamps, total packet counts and volume in both upstream and downstream directions. In addition, our records of prompt-to-response pairs also include the three timing metrics (Δt_{uw}, Δt_{sr}, and Δt_{rg}).

4 MEASUREMENT INSIGHTS FROM A UNIVERSITY CAMPUS NETWORK

We implement our real-time system as designed in §3 to measure the usage of GenAI platforms in our university campus. Our university IT department has provisioned us a full mirror of network traffic exchanged between our campus network and the Internet. As will be detailed at the end of this paper, ethical clearance is obtained for this measurement study. In this section, we report the insights obtained by analysing system logs (summarised in §4.1) of 151 days from 7pm 10 September 2023 to 3pm 4 March 2024. Insights are discussed around longitudinal daily usage trends of each platforms (§4.2), usage patterns of prompts and responses (§4.3), and timing metrics (§4.4).

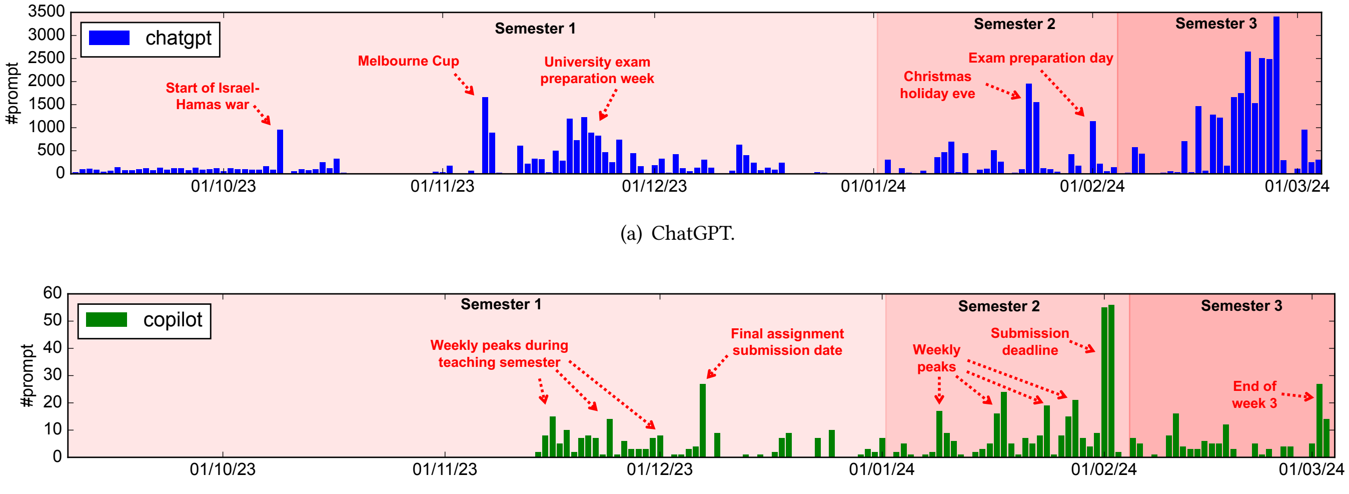
4.1 Dataset Overview

Our collected campus dataset covers usage statistics for all seven GenAI platforms, as listed in Table 2, ordered by the date we start our measurement for each platform. The earliest measurement start date is 10 September 2023 for ChatGPT and Bard, whereas the lastest start date is 12 February 2024 for Gemini. Due to lab maintenance and campus network upgrades during our measurement period, we do not have data for every day. Therefore, we report the number of effective measurement days for each platform in the second row of Table 2. The numbers of captured sessions and prompt-to-response pairs shown in the table indicates that ChatGPT dominates the GenAI usage in our university campus while the other six platforms remain quite minor during the measurement period. GitHub Copilot which is designed for coding supports has also been quite active since we start measuring it on 14 November 2023.

4.2 Temporal Usage Patterns

Our measurement can be roughly divided into three periods, *i.e.,* from 10 September 2023 to 1 January 2024 for the last two months of a semester and the following holiday break; from 2 January 2024 to 5 February 2024 for an entire short semester; and from 6 February 2024 to 4 March 2024 for the beginning of a semester, which are shaded as different regions in Fig. 5.

From Fig. 5(a) for the daily number of prompts with ChatGPT, we can see a clear pattern across the three semesters. In the first semester, the usage level of ChatGPT is the lowest, often below 500 prompts per day except some spikes during certain special events such as 8 October 2023 (the



(a) ChatGPT.

(b) GitHub Copilot.

Figure 5: Daily number of prompts for (a) ChatGPT and (b) GitHub Copilot platforms.

start of Israel–Hamas war), 7–8 November 2023 (Melbourne Cup), and 18–23 November 2023 (University exam preparation week). It is not surprising to see a low usage level on campus during this semester as there was an ongoing debate on whether GenAI should be used by students in education. As for GitHub Copilot which provides coding support, we can see from Fig. 5(b) that there was a spike in daily usage of above 30 prompts on 8 December 2023 (final assignment submission date of this semester), whereas less than 10 prompts are observed on other days.

During the short semester when (a small number of) students are allowed to enroll in at most one university course, the daily usages of ChatGPT are often below 500 prompts except the two days before Christmas holiday and the day before final exams. For GitHub Copilot in Fig. 5(b), we can see weekly usage spikes during four weeks of this short semester, right before the deadlines of weekly assignments for a first-year programming course, particularly during the last two days (1–2 February 2024) of the last week with daily usage exceeding 50 prompts.

In the beginning of the third semester covered in our measurement, a much higher level of daily usage for ChatGPT can be observed as students are more comfortable in using GenAI chatbot for information retrieval. The usage of GitHub Copilot remains at a stable level as the semester just started.

Day of the week: To better capture the weekly usage patterns of GenAI services, we aggregate our daily number of prompts to each platform for each day of the week. A normalized presentation is shown in Fig. 6. As just discussed, Copilot for coding support is mostly used during the end of a week (Thursday to Sunday) as coding assignments in our university are often due on Sunday nights. The usage of ChatGPT is quite evenly distributed across the seven days of a week. Claude and NewBing that provide generic chatbot services have similar usage patterns compared to ChatGPT. Bard has over 60% of its usages occur on Friday and Dall-E is only used on weekdays, which can be quite biased by a heavy user given their minor popularity on campus.

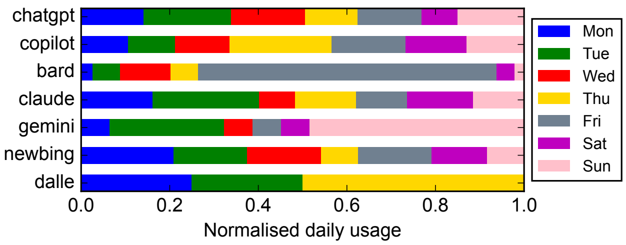


Figure 6: Normalized usage (number of prompts) of GenAI platforms per day of the week.

4.3 Usage Patterns of Prompts and Their Corresponding Responses

After discussing temporal usage patterns of each platform, we now look at the aggregated patterns of their prompts and responses, including the number of prompt-to-response pair per session, response sizes and the ratio between responses with their paired prompts.

Number of prompt-to-response per session: We now look at the number of questions (*i.e.,* prompt-to-response pairs) users typically ask on a generative AI platform per

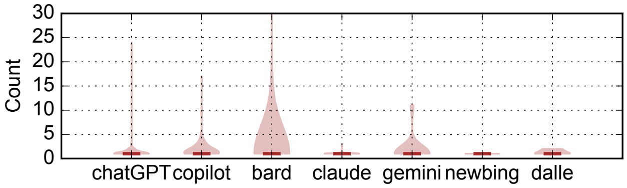


Figure 7: Number of questions (*i.e.,* prompt-to-response pairs) per session.

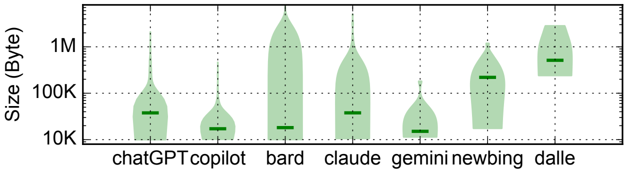


Figure 8: Response size.

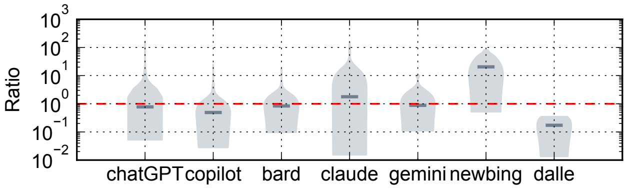


Figure 9: Response-to-prompt size ratio.

session. In Fig. 7, we show the distribution of numbers of prompts per session for each platform considered in this paper as violin plots. The changing horizontal width of each violin along the y-axis indicates the popularity (or probability) at that value. The median value of each plot is highlighted by the horizontal blue line. First of all, although the distributions vary for each platform, it is obvious that the median values for all GenAI platform equal to one. This indicates that users of all GenAI platforms often ask one question then either remain inactive for a long period before asking another question or simply closing the browser/app. Our second observation is that Google's Bard and Gemini have more sessions with a large number of questions (*e.g.,* more than 10) compared to their counterparts. This is likely due to the relatively tedious initialization processes required for a user to start a session on the two platforms. Users may want to ask more questions as compensation. In addition, ChatGPT and GitHub Copilot have a very small fraction of sessions with many prompt-to-response pairs, such as over 10. All Claude and DALL-E sessions have less than 5 questions, and NewBing has exactly one question per session due to its unique implementation.

Response sizes: Similar to the just discussed metric, in Fig. 8, we show the size of responses generated by the seven measured platforms as violin plots. Almost all (99%) of responses from Copilot and Gemini are less than 100KB as both of them can only produce texts for public users. ChatGPT, Bard, and Claude that produce both texts and graphs have some of their response sizes above 1MB. Among the four platforms, ChatGPT is still dominated (more than 85%) by small-sized (less than 100KB) responses, whereas Bard has about 20% of responses larger than 1MB. NewBing and Dall-E platforms have much higher median response sizes (over 300KB) than the others, it can be attributed to that, NewBing is integrated with Microsoft's Windows and Office ecosystem and can generate a wide variety of outputs like graphs and pages of searching results; and Dall-E is specifically made for image generation.

Response-to-prompt ratio: We further investigate into the sizes of prompts sent to each platform with respect to their received responses. Fig. 9 shows the distribution of size ratios for each pair of prompt and response. ChatGPT, Bard, Claude and Gemini have their median ratio around 1 (the dashed red line), showing roughly equivalence in their prompt and response sizes. Among the five platforms, Claude has a wider range of ratios (from 0.01 to 100) distribution as it can produce various types of outputs given text-based prompts. Copilot have most (about 85%) of its prompts larger than the corresponding responses as it takes a block of code and provides debugging and augmenting suggestions that are often smaller than the original codes in size. NewBing and Dall-E exhibit quite extreme patterns compared to others: the former platform has most of its responses more than ten times larger than the corresponding prompts, and the latter one has all its generated images smaller than the input ones. We guess that all prompts to Dall-E are images that request for augmentation and editing, instead of asking for images from text prompts.

4.4 Timing Metrics

Timing metrics (in §2.3) for each response-to-prompt pair, *i.e.,* user waiting time, server response time, and response generation time exhibit different patterns among platforms.

User waiting time: Recalling that this metrics indicates the period of time between when a user starts sending a prompt (① in Fig. 3) until the user receives a complete response (④ in Fig. 3), we show the distribution of Δt_{uw} for each prompt-to-response pair of the seven platforms in Fig. 10. Over 90% of user prompts to four platforms (ChatGPT, Bard, Gemini and Dall-E) have their full responses within two seconds. Such performance for Dall-E is quite surprising as it produce graphs which are expected to take longer times. Claude and Copilot have their median user waiting

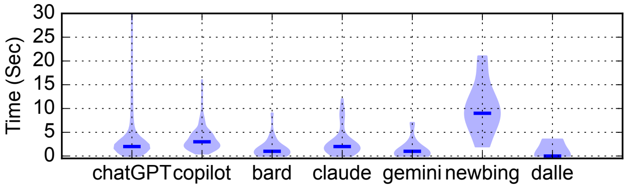


Figure 10: User waiting time Δt_{uw}.

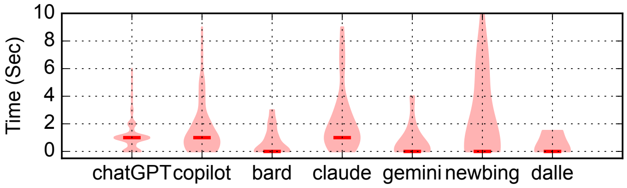


Figure 11: Server response time Δt_{sr}.

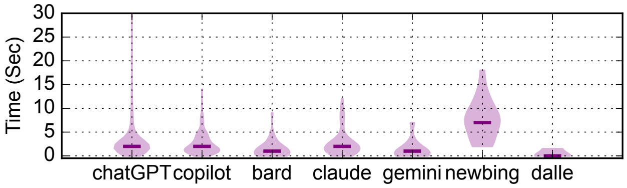


Figure 12: Response generation time Δt_{rg}.

time around 4 seconds, and NewBing has the highest timing performance with a median value near 10 seconds. We observe long tails in the distributions of ChatGPT, Copilot, Bard, Claude and Gemini, which is consistent with their response size distributions. Therefore, we believe that those prompts are with heavy responses.

Server response time: After received a prompt from an user, GenAI platforms select and call appropriate LLMs. This metric Δt_{sr} measures the time taken for this administrative process. The distribution of Δt_{sr} is presented in Fig. 11. ChatGPT exhibits a multimodal pattern in its distribution, which differs from the response size pattern shown in Fig. 8. Apparently, larger responses may not always result in longer server response times. This discrepancy could be attributed to the different times required by the server to call various versions of GPT models (*e.g.,* GPT-3.5 vs GPT-4) for different user prompts and subscription plans. Claude, Gemini, and NewBing also have multimodal distribution patterns. We also observe that, Copilot and Claude have slightly larger median server response time (about 3 seconds) than other platforms, and NewBing exhibits the largest range from nearly 0 second to 14 seconds.

As for the time taken by the platform to generate a complete answer, *i.e.,* ***Response generation time*** Δt_{rg}, shown in Fig. 12, it exhibits similar patterns to user waiting time and is therefore not explicitly discussed.

5 RELATED WORK

GenAI platforms has been evaluated for their response quality in fulfilling specific tasks [12] such as code generation [31], news fact-checking [4], educational assessment [21], and human emotion interpretation [7]. Some prior works measure their social impact [2], ethical trustworthiness [16], and potential risks in economics [28] and medicinal modelling [29]. The works in [6, 27] survey user experience with GenAI services through questionnaires. In this paper, we analyze the network characteristics and measure usage patterns of seven popular GenAI platforms in a large university campus, which has not been achieved by prior works [3, 5, 11, 13–15, 18] on network traffic analysis for different types of user applications in operational networks.

6 CONCLUSION

This paper presents our measurement study on the usage patterns of seven generative AI (GenAI) platforms in our university campus network. We first analyze the characteristics of network flows serving user conversations with GenAI platforms through lab experiments. A real-time network traffic analysis system is then designed and implemented in our university network to measure usage metrics of GenAI services. We report deployment insights for over five months, including temporal patterns, distribution of prompts and responses, and timing performance of each measured platform. The insights into GenAI usage obtained by our system can provide references for various stakeholders, such as enterprises for partnership planning, communications service providers in network optimization, and financial institutions for investment decisions.

ETHICS

University ethical clearance is obtained for this study (UNSW Human Research Ethics Advisory Panel approval number HC211007), which allows us to analyze campus network traffic to infer application usage behaviors including GenAI. Note that user identities remain anonymous. We made no attempt to extract or reveal any personal user information. All results presented are aggregated.

ACKNOWLEDGMENTS

We thank the anonymous reviewers for their feedback. Yifan Wang is supported by the Australian Government's National Industry PhD Program award reference number 35063.

REFERENCES

[1] Anthropic. Claude. https://www.anthropic.com/claude, 2024. Accessed: 2024-03-09.

[2] Baldassarre, M. T., Caivano, D., Fernandez Nieto, B., Gigante, D., and Ragone, A. The Social Impact of Generative AI: An Analysis on ChatGPT. In *Proc. ACM GoodIT* (Lisbon, Portugal, 2023).

[3] Bhuyan, S., Zhao, S., Ying, Z., Kandemir, M. T., and Das, C. R. End-to-end characterization of game streaming applications on mobile platforms. *Proc. ACM Meas. Anal. Comput. Syst.* (Feb 2022).

[4] Caramancion, K. M. News Verifiers Showdown: A Comparative Performance Evaluation of ChatGPT 3.5, ChatGPT 4.0, Bing AI, and Bard in News Fact-Checking. *arXiv* (6 2023).

[5] Chang, H., Varvello, M., Hao, F., and Mukherjee, S. Can You See Me Now? A Measurement Study of Zoom, Webex, and Meet. In *Proc. ACM IMC* (Virtual Event, Nov 2021), IMC '21.

[6] Choi, W., Zhang, Y., and Stvilia, B. Exploring Applications and User Experience with Generative AI Tools: A Content Analysis of Reddit Posts on ChatGPT. *Proceedings of the Association for Information Science and Technology 60*, 1 (2023), 543–546.

[7] Elyoseph, Z., Refoua, E., Asraf, K., Lvovsky, M., Shimoni, Y., and Hadar-Shoval, D. Capacity of Generative AI to Interpret Human Emotions From Visual and Textual Data: Pilot Evaluation Study. *JMIR Ment Health* (Feb 2024).

[8] Fox, M. Microsoft adds $154 billion in market value after it announces $30 per month AI subscription. https://markets.businessinsider.com/news/stocks/microsoft-stock-price-copilot-ai-office-365-monthly-subscription-msft-2023-7, Jul 2023. Accessed: 2024-03-12.

[9] Github. GitHub Copilot - Your AI Pair Programmer. https://github.com/features/copilot, 2024. Accessed: 2024-03-09.

[10] Gutterman, C., Guo, K., Arora, S., Gilliland, T., Wang, X., Wu, L., Katz-Bassett, E., and Zussman, G. Requet: Real-Time QoE Metric Detection for Encrypted YouTube Traffic. *ACM Trans. Multimedia Comput. Commun. Appl. 16*, 2s (jul 2020).

[11] Habibi Gharakheili, H., Lyu, M., Wang, Y., Kumar, H., and Sivaraman, V. iTeleScope: Softwarized Network Middle-Box for Real-Time Video Telemetry and Classification. *IEEE Transactions on Network and Service Management* (Sep 2019).

[12] Hadi, M. U., tashi, q. a., Qureshi, R., Shah, A., muneer, a., Irfan, M., Zafar, A., Shaikh, M. B., Akhtar, N., Wu, J., and Mirjalili, S. Large language models: A comprehensive survey of its applications, challenges, limitations, and future prospects. *TechRxiv* (Nov. 2023).

[13] Li, Z., and Huang, Y. A first look into the third-party web dependencies in china. In *Proc. AINTEC* (Hanoi, Vietnam, Dec 2023).

[14] Lyu, M., Madanapalli, S. C., Vishwanath, A., and Sivaraman, V. Network Anatomy and Real-Time Measurement of Nvidia GeForce NOW Cloud Gaming. In *Proc. PAM* (Virtual Event, Mar 2024).

[15] Lyu, M., Tripathi, R. D., and Sivaraman, V. MetaVRadar: Measuring Metaverse Virtual Reality Network Activity. *Proc. ACM Meas. Anal. Comput. Syst. 7*, 3 (dec 2023).

[16] Magooda, A., Helyar, A., Jackson, K., Sullivan, D., Atalla, C., Sheng, E., Vann, D., Edgar, R., Palangi, H., Lutz, R., Kong, H., Yun, V., Kamal, E., Zarfati, F., Wallach, H., Bird, S., and Chen, M. A Framework for Automated Measurement of Responsible AI Harms in Generative AI Applications. *arXiv* (Oct. 2023).

[17] Mangla, T., Halepovic, E., Ammar, M., and Zegura, E. eMIMIC: Estimating HTTP-Based Video QoE Metrics from Encrypted Network Traffic. In *Proc. Network Traffic Measurement and Analysis Conference* (June 2018), pp. 1–8.

[18] Michel, O., Sengupta, S., Kim, H., Netravali, R., and Rexford, J. Enabling passive measurement of zoom performance in production networks. In *Proc. ACM IMC* (Nice, France, Oct 2022).

[19] Microsoft. Your Everyday AI Companion. https://www.microsoft.com/en-us/bing?ep=278&form=MA13LT&es=31, 2024. Accessed: 2024-03-09.

[20] Mustafa, R. U., Islam, M. T., Rothenberg, C., and Gomes, P. H. EFFECTOR: DASH QoE and QoS Evaluation Framework For EnCrypTed videO tRaffic. In *Proc. IEEE/IFIP Network Operations and Management Symposium* (May 2023), pp. 1–8.

[21] Nguyen Thanh, B., Vo, D. T. H., Nguyen Nhat, M., Pham, T. T. T., Thai Trung, H., and Ha Xuan, S. Race with the Machines: Assessing the Capability of Generative AI in Solving Authentic Assessments. *Australasian Journal of Educational Technology 39*, 5 (Dec 2023), 59–81.

[22] OpenAI. Introducing ChatGPT. https://openai.com/blog/chatgpt, Nov 2022. Accessed: 2024-03-09.

[23] Paris, M. Year In AI: ChatGPT Valued At $86 Billion As Google Rides On. https://www.forbes.com/sites/martineparis/2023/11/30/year-in-ai-chatgpt-now-valued-at-86-billion-google-takes-epic-ride/?sh=49d802b23219, Nov 2023. Accessed: 2024-03-12.

[24] Pichai, S. An Important Next Step on Our AI Journey. https://blog.google/technology/ai/bard-google-ai-search-updates/, Feb 2023. Accessed: 2024-03-09.

[25] Ponciano, J. Alphabet Stock Plunge Erases $100 Billion After New AI Chatbot Gives Wrong Answer In Ad. https://www.forbes.com/sites/jonathanponciano/2023/02/08/alphabet-google-stock-plunge-erases-100-billion-after-new-ai-chatbot-gives-wrong-answer-in-ad/?sh=5f0f6dfd55ce, Feb 2023. Accessed: 2024-03-12.

[26] Ramesh, A., Pavlov, M., Goh, G., Gray, S., Chen, M., Child, R., Misra, V., Mishkin, P., Krueger, G., Agarwal, S., and Sutskever, I. DALL·E: Creating Images from Text. https://openai.com/research/dall-e, Jan 2021. Accessed: 2024-03-09.

[27] Skjuve, M., Følstad, A., and Brandtzaeg, P. B. The User Experience of ChatGPT: Findings from a Questionnaire Study of Early Users. In *Proc. ACM Conference on Conversational User Interfaces* (July 2023).

[28] Wach, K., Duong, C. D., Ejdys, J., Kazlauskaitė, R., Korzynski, P., Mazurek, G., Paliszkiewicz, J., and Ziemba, E. The Dark Side of Generative Artificial Intelligence: A Critical Analysis of Controversies and Risks of ChatGPT. *Entrepreneurial Business and Economics Review 11*, 2 (June 2023), 7–30.

[29] Walters, W. P., and Murcko, M. Assessing the Impact of Generative AI on Medicinal Chemistry. *Nature Biotechnology 38* (01 2020), 143–145.

[30] Wassermann, S., Seufert, M., Casas, P., Gang, L., and Li, K. ViCrypt to the Rescue: Real-Time, Machine-Learning-Driven Video-QoE Monitoring for Encrypted Streaming Traffic. *IEEE Transactions on Network and Service Management 17*, 4 (Dec. 2020), 2007–2023.

[31] Ziegler, A., Kalliamvakou, E., Li, X. A., Rice, A., Rifkin, D., Simister, S., Sittampalam, G., and Aftandilian, E. Measuring GitHub Copilot's Impact on Productivity. *Commun. ACM 67*, 3 (feb 2024), 54–63.