Conversations Reimagined: Human-AI Collaboration in Analyzing How Creators Explain Security and Privacy Tools

Hsieh, Tsai-Hsuan Georgia Institute of Technology, USA | thsieh61@gatech.edu
Lin, Yu-Jie National Taiwan University, Taiwan | r13126008@ntu.edu.tw
Huang, Yu-Wen National Taiwan University, Taiwan | f11126004@ntu.edu.tw
Chou, Kai-Hsiang National Taiwan University, Taiwan | r13922067@ntu.edu.tw
Li, An-Jie Eidgenössische Technische Hochschule (ETH), Zürich | lianli@student.ethz.ch
Kung, Li-Fei National Institute of Cyber Security, Taiwan | lfkung@nics.nat.gov.tw
Jeng, Wei National Taiwan University and National Institute of Cyber Security, Taiwan | wjeng@ntu.edu.tw

ABSTRACT
This study experimentally reproduces the methodology of Akgul et al. (2022) on influencer VPN advertisements, adapting it to a Taiwanese context while integrating extensive generative AI (GAI) support. We developed a three-stage GAI assisted pipeline: transcription using OpenAI's Whisper, segment identification via prompt-based querying with Google's Gemini API, and performance assessment against human annotations. This hybrid human-AI workflow reduces the manual screening burden while preserving analytical rigor, allowing researchers to concentrate on interpretive analysis. The study demonstrates how qualitative research can evolve through experimental integration of AI tools, augmenting human expertise rather than replacing it—particularly in culturally specific domains such as cybersecurity communication in Taiwan.

KEYWORDS
Information security; VPN advertisements; qualitative analysis; Generative Artificial Intelligence

INTRODUCTION
The processing of large-scale digital content and the task-shifting between data collection, processing and content analysis phases often create challenges where researchers must repeatedly communicate and refine to achieve the consistency of the collection standard, processing details, and codebook (Fazeli, Sabetti, & Ferrari, 2023; Bingham, 2023), which not only leads to technical barriers for social science researchers who lacks programming expertise or resources to conduct a comprehensive automated data collection (Palys & Atchison, 2012), but also limits the flexibility of adjusting or refining the dataset during the research process.

To address these methodological questions, we applied our generative AI-assisted framework to cybersecurity knowledge dissemination. Cybersecurity content features dense technical terminology (Reeder et al., 2018), presenting users with comprehension and communication challenges (Das et al., 2018), while rapidly evolving threats necessitate ongoing methodological adaptations (Krombholz et al., 2015). This domain ideally showcases our framework's strengths and addresses the terminology-intensive, dispersed content that challenges traditional qualitative research approaches.

With the growing importance of online media platforms as influential sources of information communication (Kross, Hargittai, & Redmiles, 2021), the Virtual Private Network (VPN) industry actively collaborates with content creators to leverage their connections with audiences, using phrases like "This video is sponsored by (brand) VPN" to signal its sponsorship on popular YouTube videos and podcasts. Additionally, the narratives and rhetorical strategies applied in VPN advertisements have great potential to mislead or confuse the non tech-savvy audiences, influencing their understanding of security and privacy protection practices (Akgul et al., 2024). Another previous research by Akgul et al. (2022) also conducted a random sampling of English-language YouTube videos to examine how content creators collaborated with VPN companies in advertising, inspiring our exploration of its implications in Taiwan, where public cybersecurity education remains an emerging concern.

Given the complexity and scale of video content available on Taiwanese platforms and the challenge of manual annotation, we introduced generative AI-assisted methods across data collection, processing, and analysis. By applying this innovative methodology to culturally specific material, we aim to analyze how the public discusses cybersecurity knowledge across various online media platforms, and how content creators incorporate sponsored VPN advertisements into their content, focusing on their use of security terminology, metaphors, and persuasive strategies they use to explain these privacy tools to general audiences.

AUTOMATING DATA COLLECTION WITH GENERATIVE AI
Our research started with a small-scale manual annotation, we then applied NLP tools and large language models to bridge video collection and content analysis. We first constructed a dataset centered on the detection of VPN

sponsorships. From the top 100 YouTube channels in Taiwan (as listed on Wikipedia), ten were randomly selected to ensure genre diversity. All videos from these channels published over a 180-day period (150 in total) were collected and annotated for (1) any sponsorship and (2) VPN-specific sponsorships. This annotated dataset served as the foundation for downstream model development and evaluation.

Building on the manually annotated dataset, we introduced a three-stage pipeline to support scalable content retrieval and iterative evaluation. The process involves (1) transcription, (2) segment identification via prompting, and (3) performance assessment against human annotations. The complete workflow is depicted in Figure 1.

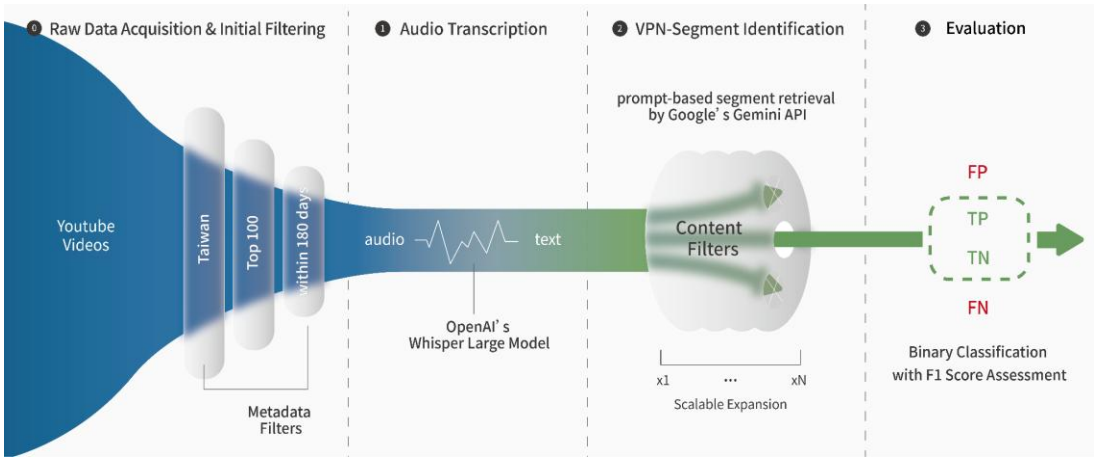

Figure 1. AI-assisted workflow for transcribing, identifying, and evaluating VPN sponsor segments in videos

We improved on Akgul et al. (2022) by using LLMs (Whisper, Gemini API) to enhance data screening and streamline audio-to-text conversion—especially for Mandarin videos, which often lack accurate subtitles. This reduced dependence on platform transcripts and improved screening accuracy.

YouTube Audio Transcription. To convert video content into inputs suitable for large language model (LLM) processing, we extracted audio tracks using youtube-dlp and transcribed them into full-text transcripts. Transcription was performed using OpenAI's Whisper Large, a multilingual speech-to-text model chosen for its adequate accuracy on mixed-language content and acceptable runtime efficiency.

VPN-Segment Identification. We applied a prompt-based segment retrieval using Google's Gemini API to identify transcript segments containing VPN sponsorships. The goal was to construct a scalable pre-screening layer—reducing the manual burden while maintaining sensitivity to subtle or variably placed sponsorship content. Incorporating AI-assisted segmentation helps shift human effort from routine detection to interpretive analysis and saves time spent reviewing entire transcripts to locate short sponsored segments during manual annotation.

Evaluation Method & Tool. To evaluate performance and guide prompt refinement, we framed VPN sponsorship identification as a binary classification task operationalized via prompting. The model returned a transcript segment if VPN sponsorship was detected; otherwise, it produced no output. Model outputs were evaluated against human annotations using confusion matrices (TP, FP, FN, TN) to compute precision, recall, and F1-score.

To support prompt iteration, we built a lightweight web tool that lets researchers test custom prompts, view model outputs, and instantly retrieve evaluation metrics (e.g., F1-score). The interface enables fast, reproducible prompt tuning and pattern discovery. We compared two prompting strategies: (1) End-to-End, directly querying for VPN content, and (2) Two-Phase, first detecting any sponsorships, then narrowing to VPNs. This comparison aimed to test whether staged prompting improved accuracy.

FUTURE RESEARCH PLAN

Based on the established methodology, we hope to explore a more complete method to process qualitative data and analyze the content carrying information security knowledge, in order to develop an evidence-based research solution to public cybersecurity education. With the AI-assisted workflows and prompt engineering demonstrated in this study, social science researchers can retain control over core analytical decisions while expanding their technical capabilities through dialogue with generative AI. This collaboration enables automation of tasks like identifying and extracting relevant media segments—previously reliant on intensive manual effort. Rather than replacing human labor, this approach exemplifies a human-centered model where GAI augments researcher expertise. Through iterative interaction, researchers and AI systems evolve together, showing that human-centered AI is achieved not through control, but through conversation.

DATA AVAILABILITY

The dataset, annotated transcripts, and code used in this study are openly available via an anonymous GitHub to ensure double-blind peer review. All scripts for transcription, prompt-based segment retrieval, and evaluation are included, along with a lightweight web interface for prompt testing and reproducibility. The repository can be accessed at: https://github.com/s3131212/youtube-vpn-ads-analysis

GENERATIVE AI USE

We utilized Gemini 1.5 Flash during the research phases to assist in identifying and extracting transcript segments that potentially contained VPN sponsorships. The AI's involvement was limited to this experimental setup, supporting prompt-based segment retrieval and preliminary classification. In the manuscript preparation phase, large language models including OpenAI's ChatGPT (GPT-4o and GPT-4.5) and Anthropic's Claude Sonnet 3.5 and 3.7 were employed for language refinement and stylistic polishing. Apart from these uses in data processing and editorial assistance, no generative AI tools were involved in the conceptual analysis or substantive writing of this paper—all analytical interpretations and core writing were conducted by the human authors.

AUTHOR ATTRIBUTION

Hsieh, Tsai-Hsuan: conceptualization; investigation; data curation; validation; writing – original draft; writing – review and editing. Lin, Yu-Jie: investigation; data curation; validation; writing – original draft. Huang, Yu-Wen: investigation; project administration; visualization; writing – original draft; writing – review and editing. Chou, Kai-Hsiang: data curation; methodology; software; validation. Li, An-Jie: data curation; methodology; software; validation. Kung, Li-Fei: investigation; project administration; writing – review and editing. Jeng, Wei: conceptualization; resources; funding acquisition; supervision; writing – review and editing.

ACKNOWLEDGMENTS

This work was financially supported by the National Science and Technology Council (NSTC) in Taiwan, under NSTC 113-2627-M-002-021-, and the Center for Research in Econometric Theory and Applications (Grant no. 114L900202 & 114L910509) which is under the Featured Areas Research Center Program by Higher Education Sprout Project of Ministry of Education (MOE) in Taiwan.

REFERENCES

Akgul, O., Roberts, R., Namara, M., Levin, D., & Mazurek, M. L. (2022). Investigating influencer VPN ads on YouTube. 2022 IEEE Symposium on Security and Privacy (SP), 876–892. https://doi.org/10.1109/SP46214.2022.9833633

Fazeli, S., Sabetti, J., & Ferrari, M. (2023). Performing Qualitative Content Analysis of Video Data in Social Sciences and Medicine: The Visual-Verbal Video Analysis Method. *International Journal of Qualitative Methods, 22*. https://doi.org/10.1177/16094069231185452

Palys, T., & Atchison, C. (2012). Qualitative Research in the Digital Era: Obstacles and Opportunities. *International Journal of Qualitative Methods, 11(4)*, 352-367. https://doi.org/10.1177/160940691201100404

Bingham, A. J. (2023). From data management to actionable findings: A five-phase process of qualitative data analysis. *International Journal of Qualitative Methods, 22*. https://doi.org/10.1177/16094069231183620

Reeder, R. W., Felt, A. P., Consolvo, S., Malkin, N., Thompson, C., & Egelman, S. (2018). An experience sampling study of user reactions to browser warnings in the field. Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems.

Das, S., Dingman, A., & Camp, L. J. (2018). Why Johnny doesn't use two factor a two-phase usability study of the FIDO U2F security key. Financial Cryptography and Data Security, 2018.

Krombholz, K., Hobel, H., Huber, M., & Weippl, E. (2015). Advanced social engineering attacks. *Journal of Information Security and Applications, 22*, 113-122.

Kross, S., Hargittai, E., & Redmiles, E. M. (2021). Characterizing the online learning landscape: What and how people learn online. Proceedings of the ACM on Human-Computer Interaction (CSCW).

Akgul, O., Roberts, R., Shroyer, E., Levin, D., & Mazurek, M. L. (2024). As advertised? Understanding the impact of influencer VPN ads. arXiv. https://doi.org/10.48550/arXiv.2406.13017