Artificial Intelligence and Social Computing, Vol. 122, 2024, 237–246
https://doi.org/10.54941/ahfe1004661

Using ChatGPT to Support Criminal Investigations: A Comparative Study of AI and Human Query

Ahad Alotaibi[1,2] **and Chris Baber**[2]

[1]Department of Information Systems, King Faisal University, Al-Hasa, 31982, Saudi Arabia
[2]School of Computer Science, University of Birmingham, B15 2TT, UK

ABSTRACT

This paper examines the role of advanced Artificial Intelligence (AI), particularly Large Language Models (LLMs) like ChatGPT, in supporting and enhancing criminal investigations. We focus on the integration of AI in query generation, intelligence analysis, and the interpretation of vast datasets to identify patterns and connections within criminal activities. Through a comparative study involving human participants and ChatGPT, we investigate the effectiveness of AI-generated queries in the 'North by Southwest' scenario, a simulated criminal case involving drug trafficking and money laundering. The ChatGPT study evaluates the AI's ability to generate a coherent investigation strategy and sequence investigative questions effectively. The human study, involving eight female Ph.D. candidates, assesses the strategies individuals employ when reasoning and developing hypotheses from ambiguous information, specifically focusing on three analytical approaches: following money, crimes, and people. Our findings highlight the complementary nature of AI and human analytical approaches. While ChatGPT provides a structured framework for sifting through evidence, human participants offer detailed, situational insights, particularly in connecting financial, criminal, and interpersonal elements. The study underlines the necessity of evaluating the accuracy and reliability of LLMs, considering the ethical implications and potential biases inherent in AI technologies. We conclude that a collaborative approach, utilizing both AI and human intelligence, can lead to more thorough and efficient investigations, ensuring that AI serves as an augmentative tool rather than a substitute for human expertise in the pursuit of justice.

Keywords: Criminal investigations, Data/frame model, Large language models (LLMS), ChatGPT

INTRODUCTION

The development of generative Artificial Intelligence (AI), particularly Large Language Models (LLMs), offers a revolution in criminal investigations (Brahan et al., 1998; Stepanenko et al., 2020). Advanced AI systems can support investigative queries of evidence, analyze data and uncover connections that may evade human investigators, and provide novel insights.

AI systems' role in evidence analysis and decision support, particularly through artificial neural networks, raises 'fair trial' concerns (Blount, 2021;

© 2024. Published by AHFE Open Access. All rights reserved.

Costantini et al., 2019) and ethical concerns about AI misuse in criminal activities (Khan et al., 2021), stressing the need for accountability and regulation. Hepenstal et al. (2021a) emphasize the importance of transparency and adaptability in AI systems, particularly in developing conversational agents for IA.

In cybercrime, ChatGPT helps mine criminal networks from chat logs, summarize conversations, and visualize information (Iqbal et al., 2012), as well as analyse topics and authors in chat logs to segregate crime-relevant logs and identify hidden criminal topics (Basher and Fung, 2014). ChatGPT, as an advanced AI solution and Language Model (LLM), enhances criminal investigations by assisting in structured query writing, summarizing electronic communications, and analysing search results (Henseler and Beek, 2023). However, its application in digital forensics must be cautiously considered due to potential inaccuracies (Scanlon et al., 2023).

AI systems, including multiagent systems and conversational agents, facilitate data analysis and correlation, supporting human examiners in tasks like identification, prediction, and classification (Hepenstal et al., 2021a; Hoelz et al., 2009; Stepanenko et al., 2020). Combining AI with mobile computing for enhanced crime intelligence through crowdsourcing is also proposed (Khanwalkar, 2016).

In this paper, we explore the use of ChatGPT to generate investigative queries to extract relevant information from evidence from a (simulated) crime. Targeted queries could help investigators narrow their search, focus on potential leads, save time and resources, and identify patterns that may not be immediately apparent, helping to uncover hidden links between individuals or events that could be crucial in solving a crime.

Intelligence Analysis

Intelligence Analysis (IA) involves iterative processes of collecting, processing, and sharing information to support decision-making (Clark, 2013). Kang and Stasko (2011) describe the 'intelligence cycle' as constructing conceptual models of collecting information, analysing, and reporting key findings. This cycle is not linear but involves complex feedback loops where new information helps verify, revise, or reject hypotheses.

Initially, IA involves 'creative, generative, tentative' sensemaking leading to a systematic and less uncertain understanding of the problem (Wong, 2014; Wong and Kodagoda, 2015). IA is characterized by convergence and divergence processes, including anchoring on specific hypotheses, associating evidence with hypotheses, and laddering to explore further explanations (Baber et al., 2016; Elm et al., 2005; William Wong and Kodagoda, 2016). IA requires seeking, cross-checking, and evaluating new information. Hepenstal et al. (2021b) detail questions used in IA, such as associations between individuals and organizations, and communication patterns.

Query Generation: Human vs. AI

In criminal investigations, human query generation requires a human-centered approach that integrates domain knowledge (Qazi and Wong, 2019; Hepenstal et al., 2021a), large datasets van Banerveld et al., 2014; Barrett, 2009), and interactive querying (Coppi et al., 2011; Hepenstal et al., 2021b; Hong et al., 2021).

ChatGPT shows potential in structuring queries and generating Boolean queries for criminal investigations, with its natural language processing and human-like response generation capabilities (Henseler and Beek, 2023; Wang et al., 2023; Goar et al., 2023; Hariri, 2023).

This paper aims to evaluate the accuracy and reliability of Large Language Models (LLMs) in generating relevant queries for criminal investigations. The NxSW scenario (Baber et al., 2016), a criminal case involving drug trafficking and money laundering, was selected for this study. This scenario includes a wide array of evidence sources, such as newspaper reports, telephone logs, bank accounts, and police interviews.

METHODOLOGY

The study presented in this paper was conducted in three stages: employing ChatGPT to generate investigative steps and questions, a simulated Intelligent Analysis (IA) activity with human participants, and comparing the outcomes of these two stages.

In the first stage, ChatGPT was tasked with creating an investigation strategy based on the NxSW scenario. This stage assessed ChatpGPT's capacity to propose investigative queries. For the second stage, Non-expert human participants were recruited to engage in a simulated IA activity with the NxSW evidence. Given that ChatGPT is not a specialist in its analysis, we felt it sensible to compare its questions with those of naive participants (rather than experienced criminal investigators) to determine if similar questions are being created. The third stage involved comparing the steps and questions formulated by the participants to reach the solution with the responses generated by ChatGPT. This comparison allowed for the evaluation of AI's effectiveness in relation to human analysis.

THE CHATGPT STUDY

In this stage, we utilized ChatGPT as an LLM tool to develop an investigation strategy for the NBSW scenario. We accessed ChatGPT.openai using the Google Chrome browser, and we wrote several messages to investigate the NBSW case as follows.

The question posed to ChatGPT was: "*You are an intelligent analyst. I want you to provide me with a top-level view series of questions in the best sequence to search evidence that includes the following: I have phone calls for 8 people, 4 criminal records, account records, some news articles, police statements, and interviews. We also have van rental records and harbour logs that show arrival and departure details for the boats in the marina. Additionally, we have information on 9 suspects, including their names and addresses. A map shows locations of these addresses and other locations of interest. There is a seating plan of the Marina Club Valentine's Day Gala Dinner with names of the people who attendeds.*" As ChatGPT provided different answers each time we altered the sequence of evidence types in the message for each query and added follow-up questions (as form of 'prompt engineering') to confirm the sequence of the steps by asking: *"Can you please arrange the questions you provided earlier based on the best sequence?"*, *"What do you mean by 'patterns'*?", "What patterns are evident in the phone calls of the 8 individuals?"

THE HUMAN STUDY

Participants, Materials and Procedure

Eight female Ph.D. candidates (Working in pairs) with no prior knowledge or expertise in IA were observed performing tasks related to the NxSW exercise. The activity, designed with multiple potential solutions including one correct solution and other distractors. The study was approved by University of Birmingham Ethics (ERN_1408-Jul2023). Following the study's explanation and consent acquisition, each pair was briefed and equipped with notepads, pens, and the dataset, which comprised nine suspect cards featuring photos and addresses, phone and harbour master logs, maps, financial and witness statements, newspaper reports, and more, all printed individually. The participants' objective was to select suspects, corroborate with evidence, and specify arrest locations in sessions lasting around two hours.

Data Collection, Preparation and Analysis

Sessions were recorded with an iPhone 13 Pro and transcribed using the Transkriptor application[1]. During the exercise, we employed a verbal protocol (Ericsson and Simon, 1993). Instead of requesting a concurrent protocol, which could disrupt participants' reasoning processes, we implemented an interruptive protocol. Every 10 minutes, participants were asked to pause their investigation and describe their current analysis. We noted participants' actions, questions, and discussions. Post-exercise, participants were interviewed about the reasoning behind their final decisions and the evidence supporting these.

The experimenter's and participants' notes from each session were compiled and underwent content analysis (using nVivo). In this, we identified specific conclusions reached within each 10-minute interval, such as 'arrest person X' or 'make an arrest in location Y', and the associated information noted at that time. Additionally, we catalogued responses to the specific questions posed during interruptions, e.g., who to arrest, where, and what information led to these conclusions. Conclusions, along with the corresponding information, were transcribed onto post-it notes and arranged in a timeline, elucidating the progression of their investigation and the logical sequence of fact evaluation.

RESULTS AND DISCUSSION

A summary of the questions asked by ChatGPT and the participants is provided in Table 1.

[1] https://transkriptor.com

Table 1. Comparative analysis of ChatGPT and human study strategies.

Human analysis strategies	ChatGPT category	ChatGPT questions	Participants questions
Follow the Money	Preliminary Overview	Are there any unusual transactions in the account records?	Is there a significant sum of money (greater than £10000)? Is there a recurring transaction?
	Communications Analysis	Do any phone calls coincide with significant events or transactions?	Are there any phone calls on the corresponding dates?
	Communications Analysis	Do the account transactions correlate with other activities or events involving the suspects?	Why is a transaction suspicious? Does this transaction have an obvious purpose?
	Official Reports Correlation	Are there inconsistencies or confirmations with the collected evidence and official reports?	
Follow the Crimes	Official Reports Correlation	What information do police statements and interviews reveal about the suspects or events?	Is there any statement or interview?
	News Articles Context	Do news articles provide additional context or corroborate other pieces of evidence?	Is there any crime mentioned in the news articles?
	Physical Evidence Cross-Reference	What insights can be derived from the van rental records and harbor logs regarding the suspects' activities?	Did any of the persons involved in the crime rent a van or own a yacht? What records (renting a car or leaving a yacht) occurred in the same month as the crimes?
Follow the people	Preliminary Overview	Who are the 9 suspects, including their backgrounds and addresses?	
	Preliminary Overview	Are there matches between the 4 criminal records and any of the 9 suspects?	What are the crimes and who committed them?
	Preliminary Overview	What patterns are evident in the phone calls for the 8 individuals?	Do the people involved have any relationships?
	Communications Analysis	Are there frequent communications between certain suspects in the phone calls?	Do the people involved have any relationships?
	Event-Specific Evidence	Who attended the Marina Club Valentine's Day Gala Dinner, and what interactions occurred?	

(Continued)

Table 1. Continued

Human analysis strategies	ChatGPT category	ChatGPT questions	Participants questions
	Event-Specific Evidence	Do attendees' movements correlate with any other evidence, such as phone calls or financial transactions on that day?	What calls occurred on the same date as the crime? And who's involved?
	Synthesis and Hypothesis	Are there any emerging patterns or inconsistencies when cross-referencing all evidence?	
	Synthesis and Hypothesis	How do all pieces of evidence connect when mapped together?	
	Synthesis and Hypothesis	What hypotheses can be formed about the suspects' activities based on the evidence and observed patterns?	

The ChatGPT Study: Analysing Investigation Strategies

ChatGPT categorized the questions into steps deemed most suitable for investigators working on the NBSW scenario. These steps included: Preliminary Suspect and Evidence Overview; Detailed Analysis of Communications and Financial Transactions; Cross-Referencing with Physical and Event-Specific Evidence; Correlating with Official Reports and News Articles; Comprehensive Synthesis and Hypothesis Formation. These steps relate to the descriptions of IA process in the Introduction.

The Human Study: Exploring Evidence Through Three Approaches

Participants in the study explored the evidence using three distinct strategies:

1. Follow the money
2. Follow the crimes
3. Follow the people

(note: approaches 1 and 2 inherently included approach 3).

Each pair identified a starting point for their investigation, e.g., 'follow the money' was used by pairs 2 and 3 (and to some extent 4). They narrowed the frames by looking at each available accounting record. Pairs 1 and 4 began with the criminal records, the newspaper articles and statements. We termed this strategy 'follow the crime'. Although pairs 3 and 4 followed different frames, they came to the same conclusion. As the investigation developed, participants sought to elaborate their frames using additional information which could include adding information from their own knowledge, which we call 'filling the gaps'.

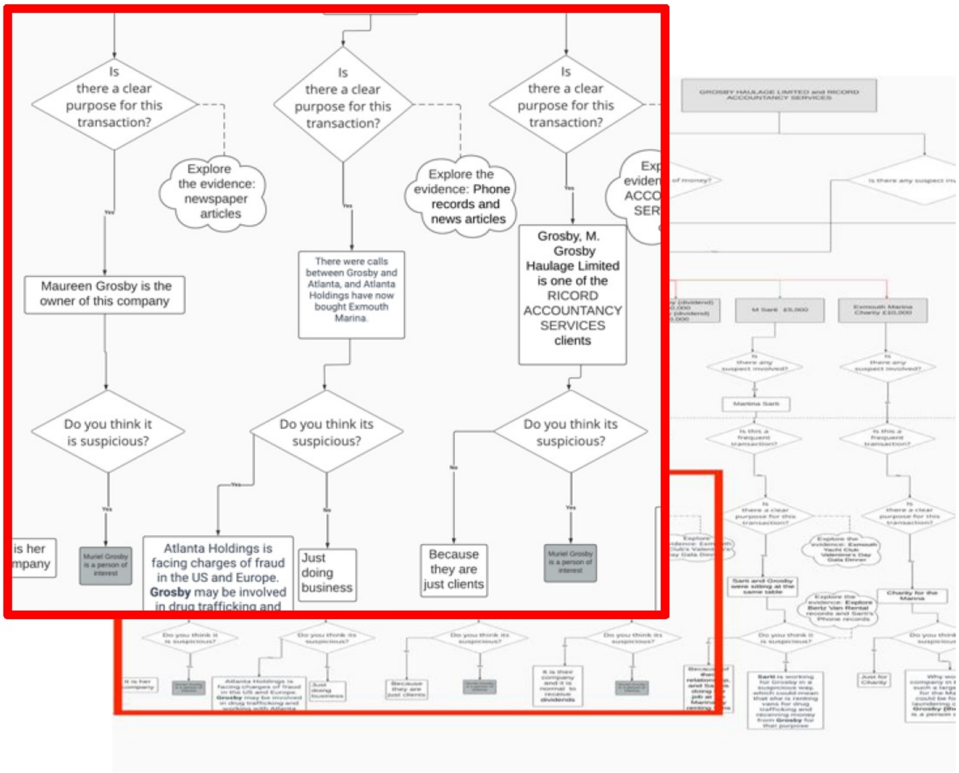
Is there a clear purpose for this transaction?
Explore the evidence: newspaper articles
Yes
Maureen Grosby is the owner of this company
Do you think it is suspicious?
Yes
Muriel Grosby is a person of interest
is her company
Is there a clear purpose for this transaction?
Explore the evidence: Phone records and news articles
Yes
There were calls between Grosby and Atlanta, and Atlanta Holdings have now bought Exmouth Marina.
Do you think its suspicious?
Yes
No
Atlanta Holdings is facing charges of fraud in the US and Europe. Grosby may be involved in drug trafficking and
Just doing business
Is there a clear purpose for this transaction?
Yes
Grosby, M. Grosby Haulage Limited is one of the RICORD ACCOUNTANCY SERVICES clients
Do you think its suspicious?
No
Yes
Because they are just clients
Muriel Grosby is a person of interest
GROSBY HAULAGE LIMITED and RICORD ACCOUNTANCY SERVICES
M Sarli £5,000
Exmouth Marina Charity £10,000
Is there any suspect involved?
Is this a frequent transaction?
Is there a clear purpose for this transaction?
Do you think it is suspicious?
Charity for the Marina
Just for Charity

Figure 1: Following the financial transactions flowchart (with section zoomed in).

Figure 1 gives an example of the timeline constructed from the analysis for a pair applying the ‘follow the money’ strategy. It shows that this pair of participants initially examined the accounting records: one for Grosby Haulage Limited and the other for Ricord Financial Services. Each set of accounts has transactions that are either payable to or received from individuals or organisations. Having identified evidence that mentioned finance, participants then asked questions for each piece of evidence. Questions that we identified or were reported by participants following both strategies in the following (see Table 1).

It should be noted that some of these questions can be answered directly from the evidence, e.g., the sum of money or the recurrence of a transaction. But some of the questions rely on assumptions that the participants were making, e.g., what constitutes a ‘significant sum of money’, or what makes a transaction ‘suspicious’? In this, the evidence does not provide a complete picture and participants need to draw on their own experience, expectations, and assumptions.

The Comparison: Analytical Approaches in ChatGPT and Human Study

We categorized questions according to the strategies that the human participants employed. This alignment helps in understanding how each method was applied in both ChatpGPT and human-driven investigations (see Table 1).

Both approaches involve examining phone call records to establish communication patterns and connections between suspects, and emphasize uncovering relationships among individuals through financial transactions, criminal activities, or direct communication. They also utilize a diverse range of evidence, including financial records, criminal histories, and physical evidence like van rental records and harbour logs. However, there are notable differences.

The participants' questions are more specific and situational, focusing on particular events or transactions, such as crimes mentioned in news articles or calls made on crime dates. Conversely, ChatGPT's questions are broader, providing a foundational framework for the investigation. ChatGPT starts with a general overview of suspects, moves to a detailed analysis of communications and financial transactions, and then to broader synthesis and hypothesis formation. The participants' approach is more fluid and less linear, intertwining elements of money, crimes, and people in the same line of inquiry. ChatGPT delves into suspects' backgrounds and profiles from the outset, while participants focus more on suspects' actions and relationships as revealed through specific activities or transactions. Additionally, ChatGPT suggests a sequential integration of different evidence types, from individual analysis to combined synthesis, whereas participants integrate various evidence types more concurrently.

Limitations and Challenges of the Study

Comparing ChatGPT and human studies present challenges due to the nature of information provided to each. ChatGPT received a summarized version of the case, including key elements and evidence types, while human participants had access to the full evidence bundle. Consequently, its analysis focused on broader patterns and connections due to the summarized nature of the information.

As we deliberately chose to use naïve participants, we expect some difference from professional analysts because of a lack of experience. Interestingly, the types of question we identified are similar to those recorded by Hepenstal et al. (2021b) who conducted their studies with experienced analysts. This suggests that our participants were responding to the available information in a manner similar to more experienced analysts (and, it should be noted, produced similar conclusions to those observed in previous studies using this exercise).

CONCLUSION

The exploration of AI, particularly ChatGPT as an LLM tool, in criminal investigations highlights the potential of integrating advanced technology with human analytical skills. ChatGPT provides a broad framework for investigating suspects, their financial activities, and communication patterns, while participants' questions delve into specifics, focusing on the interconnectedness of financial activities, criminal actions, and personal relationships.

This comparison demonstrates how different investigative approaches can complement each other, forming a comprehensive analysis.

Our study demonstrates that participants adopt strategies combining analytical approaches, focusing on money, crimes, and people. Hepenstal et al., (2021b) note that investigative analysis includes questions developed from evidence, with assumptions used to fill gaps in reasoning. In conclusion, the synergy of AI and human investigation creates a dynamic, informed, and comprehensive approach to criminal justice, suggesting a future where technology and human expertise collaborate closely for enhanced outcomes.

REFERENCES

Baber, C., Attfield, S., Conway, G., Rooney, C., Kodagoda, N., 2016. Collaborative sense-making during simulated Intelligence Analysis Exercises. International Journal of Human Computer Studies 86, 94–108. https://doi.org/10.1016/j.ijhcs.2015.10.001

Barrett, E., 2009. The interpretation and exploitation of information in criminal investigations.

Basher, A. R. M. A., Fung, B. C. M., 2014. Analyzing topics and authors in chat logs for crime investigation. Knowl Inf Syst 39, 351–381. https://doi.org/10.1007/S10115-013-0617-Y

Clark, R. M., 2013. Intelligence Collection. CQ Press.

Coppi, D., Calderara, S., Cucchiara, R., 2011. Iterative active querying for surveillance data retrieval in crime detection and forensics. International Conferences on Imaging for Crime Detection and Prevention. https://doi.org/10.1049/IC.2011.0133

Elm, W., Potter, S., Tittle, J., Woods, D., Grossman, J., Patterson, E., 2005. Finding Decision Support Requirements for Effective Intelligence Analysis Tools. Proceedings of the Human Factors and Ergonomics Society 49th Annual Meeting 297–301.

Ericsson, K. A., Simon, H. A., 1993. Protocol Analysis: Verbal Reports as Data Cambridge. EUA Massachusetts Institute of Technology.

Goar, V., Yadav, N. S., Yadav, P. S., 2023. Conversational AI for Natural Language Processing: An Review of ChatGPT. International Journal on Recent and Innovation Trends in Computing and Communication 11, 109–117. https://doi.org/10.17762/IJRITCC. V11I3S.6161

Hariri, W., 2023. Unlocking the Potential of ChatGPT: A Comprehensive Exploration of its Applications, Advantages, Limitations, and Future Directions in Natural Language Processing. arXiv.org. https://doi.org/10.48550/ARXIV.2304.02017

Henseler, H., Beek, H. V., 2023. ChatGPT as a Copilot for Investigating Digital Evidence. LegalAIIA@ICAIL.

Hepenstal, S., Zhang, L., Kodagoda, N., Wong, B. l. william, 2021a. Developing Conversational Agents for Use in Criminal Investigations. ACM Trans Interact Intell Syst 11. https://doi.org/10.1145/3444369

Hepenstal, S., Zhang, L., Wong, B. L. W., 2021b. Automated Identification of Insight Seeking Behaviours, Strategies and Rules: a Preliminary Study. Proceedings of the Human Factors and Ergonomics Society Annual Meeting 65, 1269–1273. https://doi.org/10.1177/1071181321651348

Hoelz, B. W. P., Ralha, C. G., Geeverghese, R., 2009. Artificial intelligence applied to computer forensics. ACM Symposium on Applied Computing 883–888. https://doi.org/10.1145/1529282.1529471

Hong, J., Voss, C., Manning, C., 2021. Challenges for Information Extraction from Dialogue in Criminal Law 71–81. https://doi.org/10.18653/V1/2021. NLP4POSIMPACT-1.8

Iqbal, F., Fung, B. C. M., Debbabi, M., 2012. Mining Criminal Networks from Chat Log. 2012 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology 332–337. https://doi.org/10.1109/WI-IAT.2012.68

Kang, Y. A., Stasko, J., 2011. Characterizing the Intelligence Analysis Process: Informing Visual Analytics Design Through a Longitudinal Field Study, in: VAST 2011 - IEEE Conference on Visual Analytics Science and Technology 2011, Proceedings. pp. 21–30. https://doi.org/10.1109/VAST.2011.6102438

Khan, K. F., Ali, A., Khan, Z. F., Siddiqua, H., 2021. Artificial Intelligence and Criminal Culpability. International Conference on Intelligent Computing. https://doi.org/10.1109/ICIC53490.2021.9692954

Khanwalkar, S., 2016. Crime Intelligence 2.0: Reinforcing Crowdsourcing using Artificial Intelligence and Mobile Computing.

Qazi, N., Wong, B. L. W., 2019. An interactive human centered data science approach towards crime pattern analysis. Inf Process Manag 56. https://doi.org/10.1016/J. IPM.2019.102066

Scanlon, M., Breitinger, F., Hargreaves, C., Hilgert, J.-N., Sheppard, J. W., 2023. ChatGPT for Digital Forensic Investigation: The Good, The Bad, and The Unknown. ArXiv. https://doi.org/10.48550/ARXIV.2307.10195

Stepanenko, D., Bakhteev, D. V., Evstratova, Y. A., 2020. The use of artificial intelligence systems in law enforcement. Russian journal of criminology 14, 206–214. https://doi.org/10.17150/2500-4255.2020.14(2).206–214

van Banerveld, M., Le-Khac, N. A., Kechadi, M. T., 2014. Performance Evaluation of a Natural Language Processing Approach Applied in White Collar Crime Investigation. International Conference on Future Data and Security Engineering 8860, 29–43. https://doi.org/10.1007/978-3-319-12778-1_3

Wang, S., Scells, H., Koopman, B., Zuccon, G., 2023. Can ChatGPT Write a Good Boolean Query for Systematic Review Literature Search? Annual International ACM SIGIR Conference on Research and Development in Information Retrieval 1426–1436. https://doi.org/10.1145/3539618.3591703

William Wong, B. L., Kodagoda, N., 2016. How Analysts Think: Anchoring, Laddering and Associations, in: Proceedings of the Human Factors and Ergonomics Society. Human Factors an Ergonomics Society Inc., pp. 178–182. https://doi.org/10.1177/1541931213601040

Wong, B. L. W., 2014. How Analysts Think (): Early Observations, in: Proceedings - 2014 IEEE Joint Intelligence and Security Informatics Conference, JISIC 2014. Institute of Electrical and Electronics Engineers Inc., pp. 296–299. https://doi.org/10.1109/JISIC.2014.59

Wong, B. L. W., Kodagoda, N., 2015. How Analysts Think: Inference Making Strategies, in: Proceedings of the Human Factors and Ergonomics Society. Human Factors and Ergonomics Society Inc., pp. 269–273. https://doi.org/10.1177/1541931215591055