



OPEN

Large scale summarization using ensemble prompts and in context learning approaches

Andrés Leiva-Araos^{1,2,9}, Bady Gana³, Héctor Allende-Cid^{3,4,5}, José García⁶ & Manob Jyoti Saikia^{7,8,9}✉

The field of Information Assurance (IA) and Cybersecurity has seen substantial evolution, driven by advancements in technology and the increasing sophistication of threats in the digital age. This study employs Large Language Models (LLMs), as well as other advanced NLP techniques, to conduct a comprehensive analysis of literature from 1967 to 2024. By analyzing a corpus of more than 62,000 documents extracted from Scopus, our approach involves a comprehensive methodology that includes two main phases: topic detection using BERTopic and automatic summarization with LLMs across various periods (annual and decades). By designing targeted queries to extract relevant papers, analyzing textual data, and applying advanced prompting techniques for summarization, we integrate computational models to handle large volumes of data. Our results demonstrate that an ensemble of methods (Ev2) outperforms traditional summarization and density-based approaches, with improvements ranging from 16.7% to 29.6% in keyword definition tasks. It generates summaries that outperform in 5 out of the 7 tested metrics while maintaining the logical integrity of bibliographic references. Our results illuminate the shifts in focus within Information Assurance across decades, revealing key breakthroughs and forecasting emerging areas of significance.

Keywords Information assurance, Cybersecurity trends, Systematic topic review, Large language models, Natural language processing (NLP), Automatic summarization

The landscape of Information Assurance (IA) has experienced profound evolution over the past decades, reflecting both the advancements in technology and the sophistication of threats that characterize the digital age. To systematically trace and analyze these developments, this paper seeks to delve into a comprehensive examination of the literature spanning from 1967 to 2024. Our study leverages state-of-the-art techniques in prompt engineering, LLM (Large Language Model) management, and advanced processing methods to perform a detailed analysis and summarization of topic evolution over time, using robust machine learning approaches for comprehensive document synthesis. This research is inspired by a gap identified in the scholarly discourse: the need for a panoramic view of the IA domain that not only identifies key thematic shifts, but also elucidates emerging trends and domains within the landscape of cybersecurity and information security. The main contribution of this work lies in the application of advanced natural language processing (NLP) techniques to a specific use case, achieving significant improvements in the overall task through preliminary results obtained using in-context learning techniques. The results are promising and demonstrate the potential of these methodologies. However, it is important to highlight that there is room for improvement in the validation stage, due to the lack of standardized metrics applicable at this scale of summarization tasks. This underscores, as we suggest in our conclusions, the need to develop new evaluation approaches and tools that enable rigorous and consistent validation of results in domains with large volumes of information.

In recent years, automatic summarization techniques have been widely used to handle large text corpora, especially in domains like cybersecurity. Traditional methods often rely on semantic similarity and clustering to

¹Department of Computing, University of North Florida, Jacksonville, FL 32224, USA. ²Data Science Institute, Universidad del Desarrollo, Santiago 7610658, Chile. ³Escuela de Ingeniería Informática, Pontificia Universidad Católica de Valparaíso, Valparaíso 2340000, Chile. ⁴Knowledge Discovery, Fraunhofer-Institut für Intelligente Analyse- und Informationssysteme IAIS, 53757 Sankt Augustin, Germany. ⁵Lamarr Institute for Machine Learning and Artificial Intelligence, 53115 Dortmund, Germany. ⁶Escuela de Ingeniería en Construcción y Transporte, Pontificia Universidad Católica de Valparaíso, Valparaíso 2362804, Chile. ⁷Department of Electrical Engineering, University of North Florida, Jacksonville, FL 32224, USA. ⁸Electrical and Computer Engineering Department, The University of Memphis, Memphis, TN 38152, USA. ⁹Biomedical Sensors & Systems Lab, The University of Memphis, Memphis, TN 38152, USA. ✉email: msaikia@memphis.edu

generate summaries. However, these approaches typically face limitations in retaining bibliographic structure and contextual consistency across extensive data. This is a significant challenge in studies that aim to preserve the historical progression and thematic integrity of scientific literature.

Our approach introduces the “Ensemble Prompts” (Ev2) method, which combines prompt engineering with advanced techniques like Chain of Density (CoD) and Few-Shot Learning, among others. This method is distinct in its ability to balance relevance, conciseness, and thematic richness across periods, ensuring summaries that maintain key bibliographic references and the logical flow of topics. Unlike prior methods, which often focus solely on reducing redundancy or clustering themes, our approach generates targeted summaries for each decade, accurately reflecting historical shifts and emerging trends.

This methodology provides a novel solution to the challenges of summarizing large-scale literature while preserving essential bibliographic structure. By enabling decade-wise analysis and tailored prompt structures, our method addresses gaps identified in earlier approaches, particularly in the areas of topic retention, structural coherence, and contextual accuracy. This enhances both the precision and depth of our findings, offering valuable insights into the evolving landscape of cybersecurity research.

Our study leverages the capabilities of BERTopic¹, a machine-learning approach for topic modeling, alongside other NLP methods to provide a detailed analysis of topic evolution over time. Using a corpus of 62,344 documents sourced from Scopus, this study uses a modified systematic topic review methodology, enhanced by the computational capabilities of the “*Mistral-7B-Instruct-v0.1*”² Large Language Model (LLM) from the Mistral-7B series. Our approach is distinguished by its semi-automated analytical framework, which integrates text analysis with the latest developments in NLP to conduct a thorough examination of the literature. By transforming texts into numerical data through FlagEmbedding and leveraging dense vector mapping with the “*BAAI/bge-large-en-v1.5*” model³, our methodology facilitates a nuanced classification, and semantic analysis of themes. At the time of conducting the experiments, the BAAI/bge-large-en-v1.5 model ranked among the top 10 models on Hugging Face’s MTEB leaderboard, demonstrating competitive performance across a broad range of tasks, despite utilizing fewer parameters compared to other leading models⁴. This process is foundational to the effective integration of LLM vector databases, significantly enhancing the precision and depth of our literature review.

At the core of our analysis is the utilization of dimensionality reduction techniques such as Uniform Manifold Approximation and Projection (UMAP)⁵ for efficient topic clustering, which preserves critical structural details while enabling a refined delineation of topics through Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN)⁶. This approach is augmented by the application of c-TF-IDF for precise topic representation, improving thematic accuracy in document clusters. Through this innovative combination of BERTopic and “*Mistral-7B-Instruct-v0.1*”, our study presents a dynamic, scalable framework for analyzing extensive text data, streamlining the review process, and ensuring that our findings accurately reflect the burgeoning knowledge base in Information Assurance.

The central focus of our study is a comprehensive summarization process aimed at generating coherent and insightful summaries across various periods, including annual and decadal intervals. This process incorporates targeted query design for extracting relevant papers, data analysis and visualization, calculation of a normalized importance score based on citations and publication age, grouping papers by year, and using “*gpt-4-turbo-2024-04-0*”⁷ as the LLM for summarization. To implement this, we utilized libraries from the OpenAI API and LangChain⁸. By dividing the corpus into manageable chunks and employing MapReduce techniques, we efficiently manage the data volume while maintaining coherence and accuracy in the summaries.

This paper sets out to not only chronicle the progression of themes within IA but also to offer valuable insights for researchers, practitioners, and policymakers. By providing a decade-wise analysis, we aim to illuminate the shifts in focus, identify pivotal breakthroughs, and forecast emerging areas of importance within the fields of information security and cybersecurity. This endeavor seeks to contribute a novel perspective to the academic community, fostering a deeper understanding of the evolutionary trajectory of Information Assurance and setting a foundation for future research and innovation in the domain.

The rest of this paper is structured as follows. In “Related Work”, offers a comprehensive literature review, focusing specifically on surveys within the field of Information Assurance. This review sets the stage by contextualizing our research within the existing body of knowledge. Section “Proposed Approach”, details the experimental methodology employed in our study, providing a thorough explanation of the processes and technologies used. In “Results and Findings”, we present our outcomes, which are categorized and analyzed by decade, starting from 1967 up to the present day. This chronological arrangement helps to highlight trends and shifts in the domain over time. Finally, Section “Conclusion and Future Work”, concludes our study and discusses potential future challenges and improvements both in our framework and in the applied use cases.

Related work

Recent advances in the field of automatic document summarization have aimed to address the challenges associated with summarizing large volumes of text, which is essential for faster understanding and real-life applications. Semantic similarity and clustering techniques are crucial for generating effective summaries of extensive text collections, though these processes are computationally intensive and time-consuming. To tackle these challenges, frameworks based on MapReduce technology⁹ have been proposed, leveraging its proven capabilities in handling Big Data. For instance, a novel framework utilizes semantic similarity-based clustering and Latent Dirichlet Allocation (LDA) for summarizing large text collections¹⁰, demonstrating scalability and effectiveness in terms of compression ratio, retention ratio, ROUGE¹¹, and Pyramid scores¹². Additionally, aspect-based summarization systems, which play a significant role in analyzing web reviews, have also benefited from MapReduce frameworks. These systems employ node optimization algorithms to enhance the accuracy of generated summaries, proving more effective than traditional methods¹³. Furthermore, with the proliferation

of digital information, the need for efficient automatic text summarization has become critical. Techniques involving Density-Based Spatial Clustering of Applications with Noise¹⁴ algorithms for clustering and Hidden Markov Models¹⁵ for summarization within a MapReduce framework have shown promise in handling large document collections. These methods utilize preprocessing steps and machine learning techniques to improve the accuracy and relevance of summaries, particularly in topic-based abstract summarization tasks¹⁰. While summarization techniques have evolved significantly, the MapReduce framework remains a robust alternative for managing the volume factor, effectively addressing the challenges posed by large-scale text data.

The scholarly exploration of Information Assurance has emerged over recent years, manifesting in a rich tapestry of research that spans various facets of cybersecurity and information security. This growing interest is exemplified by works in Wursch et al.¹⁶, who critically examined the efficacy of LLMs in extracting cybersecurity concepts, highlighting the challenges and opportunities within this domain. Similarly, in Yao et al.¹⁷, the authors provided a comprehensive survey on the security and privacy implications of LLMs, shedding light on the intricate balance between technological advancements and the emerging vulnerabilities in cybersecurity frameworks. In a more focused vein, Jamal et al.¹⁸ introduced an improved transformer-based model for detecting phishing, spam, and ham emails, leveraging the power of LLMs to enhance detection accuracy. These studies underscore the dynamic interplay between AI technologies and cybersecurity, reflecting a growing trend toward leveraging advanced computational techniques to fortify information security.

Within the expansive domain of cybersecurity research, several studies stand out for their innovative approaches and contributions to understanding and mitigating prevalent threats. In Jamal et al.¹⁸, the authors introduced the IPSDM model, an advanced solution predicated on refining the BERT family of models, specifically to enhance the detection of phishing and spam emails. This model represents a significant step towards leveraging NLP and machine learning (ML) technologies to address the continuously evolving tactics employed by cyber adversaries. In addition, the work in Elluri et al.¹⁹, embarked on a comprehensive examination of more than 150 research articles, meticulously selecting and analyzing the 50 most pertinent and recent studies to shed light on the sophisticated methodologies utilized by cybercriminals. Their work provides invaluable insights into the cyber threat landscape, offering a detailed overview of the technological and tactical advancements in cybercrime. Concurrently, Wang et al.²⁰ research contributes to the scholarly discourse through a classic systematic literature review (SLR) focused on Cyber Threat Hunting. This SLR serves as a critical resource for academics and practitioners alike, offering a consolidated view of current methodologies and strategies to identify and neutralize cyber threats, reinforcing the foundational knowledge necessary to develop more effective cybersecurity measures. Together, these studies underscore the importance of continuous research and innovation in the fight against cyber threats, highlighting the potential of NLP and ML in crafting sophisticated defense mechanisms.

Furthermore, the literature significantly emphasizes systematic literature reviews (SLRs) that employ NLP techniques to dissect and analyze specific cybersecurity threats. For instance, the work in Salloum et al.²¹ conducted a systematic review of phishing email detection using NLP, providing valuable insights into the effectiveness of various feature extraction and classification algorithms. The application of NLP has predominantly been oriented towards enhancing detection mechanisms, classification accuracy, and the overall efficacy of cybersecurity measures. This trend is exemplified in Salloum et al.²¹, where the authors undertake a classic systematic review of research focusing on the utilization of NLP for the detection of phishing emails—a pervasive threat with significant financial repercussions. Through an exhaustive analysis of 100 articles spanning from 2006 to 2022, this study delineates the critical domains of feature extraction and classification algorithms, particularly spotlighting support vector machines. The research underscores the widespread adoption of techniques such as Term Frequency-Inverse Document Frequency (TF-IDF), word embeddings, and the utilization of the Nazario phishing corpus. This study, along with others focusing on AI-driven cybersecurity²² and intrusion detection systems in wireless networks²³, illustrates the pivotal role of NLP and AI in advancing cybersecurity research. Notably, the work by Yang et al.²⁴ on information security in chat-bots introduces a novel perspective, addressing key security challenges and proposing solutions like blockchain and end-to-end encryption.

Upon an initial examination of the literature, it becomes clear that while a significant volume of surveys exists within this domain, they predominantly focus on specialized areas such as cellular networks, security intelligence modeling, FinTech, steganography, and cybercrime in social networks. It is noteworthy that the application of NLP techniques in SLRs remains sparse, with a primary emphasis on Phishing Email Detection. In our analysis of 125 pre-prints, we encountered a limited array of reviews; however, we have identified several tools, like SecureBERT²⁵, that promise to be exceptionally beneficial for our research. Based on this review, our project emerges as a pioneering endeavor aimed at comprehensively summarizing decades of research within the Information Assurance domain. This initiative uniquely details the evolution of these scholarly efforts over time, facilitated by an innovative amalgamation of techniques for topic extraction and text summarization.

Our research builds on these foundational works, endeavoring to bridge the gap between individual studies and a holistic understanding of the IA domain. By employing a unique combination of BERTopic analysis and advanced NLP techniques, our study aims to synthesize decades of research, providing a panoramic view of the evolving trends and themes in information assurance. This endeavor not only complements existing research but also aims to catalyze further exploration and innovation in the fields of information security and cybersecurity, highlighting the critical role of NLP and Artificial intelligence (AI) in shaping the future of digital security.

To the best of our knowledge, our proposal represents a unique effort to summarize decades of research in the domain of Information Assurance and describe how these efforts have evolved over time.

(TITLE-ABS-KEY ("information assurance")
OR TITLE-ABS-KEY ("information security")
OR TITLE-ABS-KEY (cybersecurity))

Table 1. Scopus query statement.

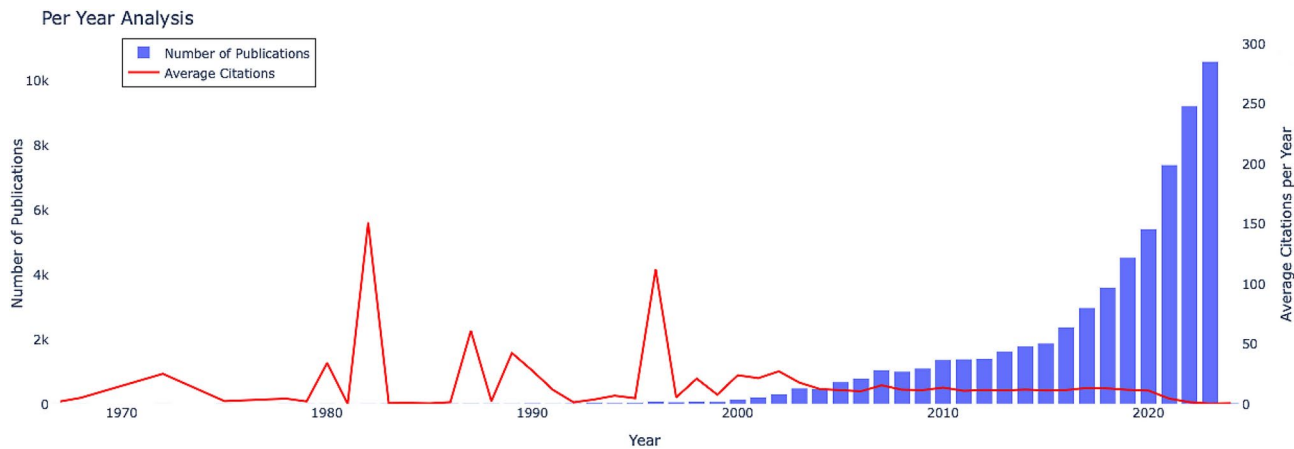


Fig. 1. The red line shows the average citations per year. The peaks in citations per year observed correspond to fundamental editions of books on the subject.

Methods

In our approach to exploring the evolution of AI through the lens of NLP, we have employed a series of advanced techniques and methodologies to process and analyze an extensive corpus of literature spanning from 1967 to 2024.

In-context learning (ICL)²⁶ forms the cornerstone of our research, allowing the model to dynamically leverage contextual information for more accurate and contextually aware outputs. The remain foundation of our approach is built upon the BERTopic algorithm, enhanced by the capabilities of LangChain, Chain of Density (CoD)²⁷, and few-shot learning techniques²⁸, alongside various summarization methods. This strategy enables us to systematically segment the collected documents by decade, facilitating an in-depth exploration of thematic evolution and emerging trends within the IA domain.

Document extraction

The initial query (see Table 1), based on three specific keywords, retrieved a total of 62,344 documents. This dataset included 4,422 book chapters, 1,771 reviews (1,400 published since 2017), 122 pre-prints, and 25 additional reviews identified using NLP techniques. From this initial set, 1,551 documents were excluded for lacking abstracts, and an additional 2,297 were removed due to missing author information. This filtering process resulted in a final working dataset of 58,496 documents.

Our analysis pays particular attention to the temporal distribution of documents and citation metrics over the years, presented in Fig. 1. The red line illustrates the average yearly citations and the blue bars marks the amount of documents each year. This visualization underscores significant peaks in academic attention, potentially corresponding to groundbreaking publications or shifts in the field's focus. Since 2000, there has been an exponential growth in research on security issues globally, with notable increases from 2020 onward.

Topic modeling

The next step in our workflow involves Topic Modeling to analyze the corpus of documents collected earlier. We use BERTopic, an established method that applies advanced NLP techniques for topic modeling in large datasets. By integrating BERTopic with the Mistral-Instruct-v0.1 model, a 7B pre-trained model, we generate contextual topic representations efficiently without processing individual documents (see Fig. 2).

For document embedding, we use BAAI/bge-large-en-v1.5 to generate contextual embeddings that capture the semantic details of the dataset. Next, we apply UMAP for dimensionality reduction, converting high-dimensional embeddings into a lower-dimensional space to improve computational efficiency and clustering performance in topic modeling. Clustering is performed using HDBSCAN, which groups the reduced embeddings into distinct topics. This process allows for the extraction of meaningful information from large datasets, facilitating further analysis as discussed in our paper.

The hyperparameters used in UMAP and HDBSCAN, such as *n_neighbors*, or *min-samples* and *min-cluster-size* respectively, were selected based on prior studies and the original hyperparameter configurations outlined in BERTopic¹. The settings were further validated through their application in document clustering within scientific domains, as seen in García et al.²⁹, and were chosen for their demonstrated effectiveness in producing

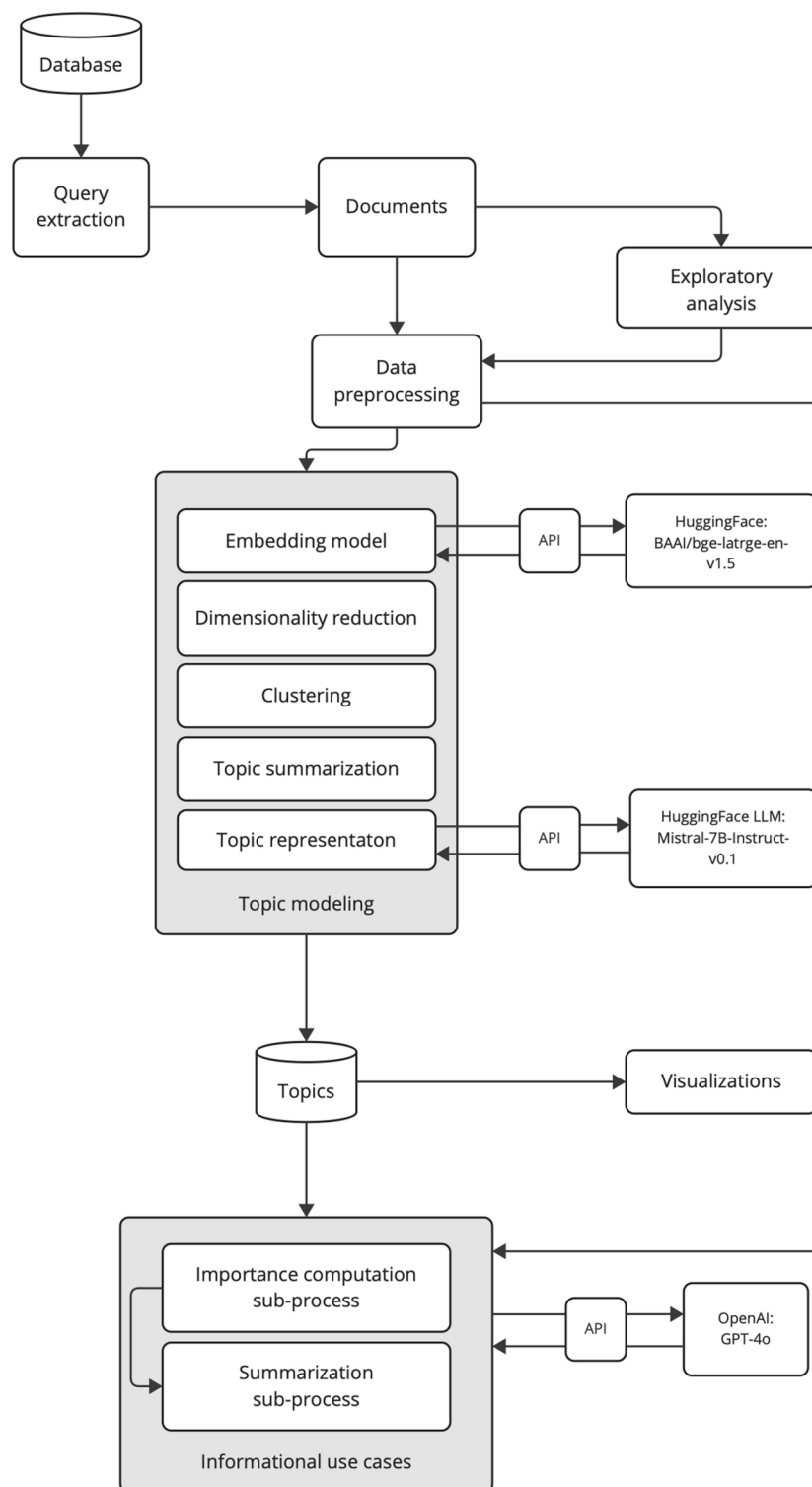


Fig. 2. Workflow used in the project..

stable and interpretable clusters, as highlighted in studies like Gana et al.³⁰. While exhaustive hyperparameter tuning was not conducted, the chosen settings yielded consistent clustering results that aligned with our goal of identifying coherent thematic groups within the document set.

To enhance the accuracy of topic representation, we use c-TF-IDF¹. This variant of the classic TF-IDF refines how topics are represented by assessing the distinctiveness of documents within one group relative to others. In the BERTopic framework, c-TF-IDF shifts the focus from individual documents to clusters. Documents within each cluster are merged into a single document, and the term frequency within this cluster is assessed. This is

```

Article: "{text}"
You will generate increasingly concise, entity-dense summaries of the above Article.
{
  "System": "You are a helpful, respectful and honest assistant for labeling topics.",
  "Example": {
    "Documents": [
      "Enhancing Cybersecurity Measures: A Comparative Study of Automated vs. Manual Threat Detection.", "The Role of Natural Language Processing in Improving Intrusion Detection Systems.", "Machine Learning Approaches to Identify Potential Threats in Cybersecurity."
    ],
    "Keywords": [
      "cybersecurity", "threat detection", "automated threat detection", "natural language processing (NLP)", "machine learning", "intrusion detection systems"
    ],
    "Label": "Advancements in Automated Cybersecurity Threat Detection"
  },
  "Main": {
    "Instructions":
      "I have a topic that contains the following documents: [DOCUMENTS]
      The topic is described by the following keywords: '[KEYWORDS]'. Based on the information about the topic above, please create a short label of this topic. Make sure you to only return the label and nothing more."
  }
}

```

Fig. 3. Prompt structure in topic representation.

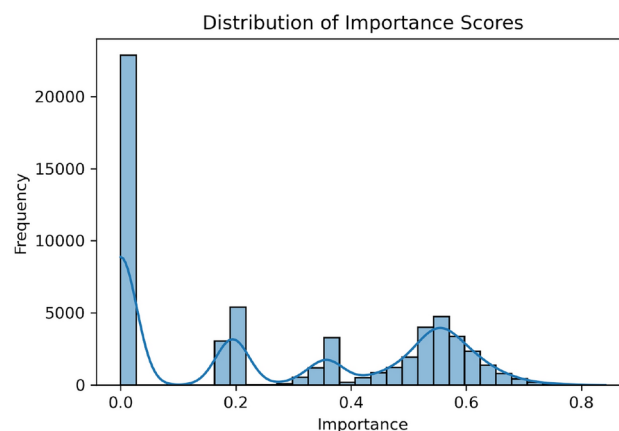


Fig. 4. Histogram of document importance scores with a KDE curve, showing the multimodal distribution and variability in document impact..

combined with the inverse document frequency, calculated as the logarithm of the ratio of the average number of words per cluster to the frequency of the term across all clusters. This adjustment accentuates the significance of terms within clusters, facilitating the generation of distinct word distributions for each document cluster.

Prompt engineering plays a key role in our methodology, focusing on generating custom names for the clusters identified in the BERTopic process. This approach combines CoD and Few-shot Learning techniques²⁸ to enhance effectiveness. By integrating these techniques, we create descriptive names for each cluster, leveraging the keywords identified through the c-TF-IDF process to capture the thematic content of each cluster. This ensures that the names reflect the core topics in our analysis of Information Assurance. The prompt used is shown in Fig. 3.

Document importance factor

We propose a new metric called Impact Factor (I). This metric is used to determine the relative significance of a set of documents, allowing us to assess the importance of yearly summaries in relation to one another. This ensures that when generating summaries for extended periods (e.g., a decade, as shown in prompts in Figs. 11 and 12), more weight is given to years with higher relative importance.

The distribution of importance scores shows a multimodal nature, suggesting the existence of multiple peaks in the data; see Fig. 4. This highlights the need to develop a robust indicator that can adequately handle this variability in document importance. The boxplots in Fig. 5 show that including papers with zero citations significantly affects the distribution, underscoring the importance of excluding these documents when calculating the annual average importance to more accurately reflect the impact of key works in the field. To calculate the importance of each document, several new columns were created. The citation counts and the number of

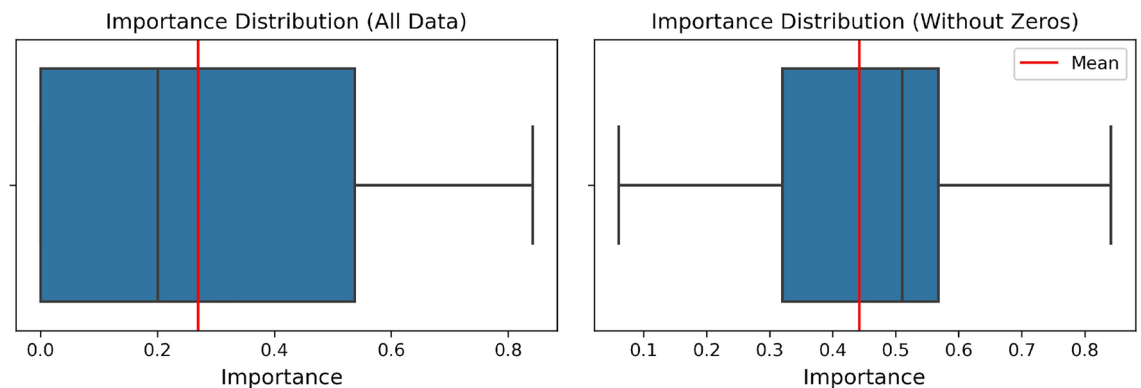


Fig. 5. Comparison of importance scores with and without zero-citation papers, highlighting the impact of zero-citation papers on the overall distribution..

years since publication were normalized (starting from 2025), to ensure consistency in the data analysis. The *Importance Factor*, I , of each document was then computed using the formula in Eq. 1:

$$I = (0.5 \times C + 0.5 \times (1 - Y)) \times (1 - C_p) \quad (1)$$

where:

- I is the computed importance.
- C is the normalized citation count.
- Y is the normalized number of years since publication.
- C_p is the citation penalty.

The *Citation Penalty* (C_p) is defined in Eq. 2:

$$C_p = \left(\frac{\min(C_t) - C}{\min(C_t)} \right) \times \left(1 + 0.5 \times \left(\frac{Y}{\max(Y)} \right) \right) \quad (2)$$

where $\min(C_t)$ is defined with expert judgement as 3. This penalty factor ensures that older papers with many citations are weighted appropriately compared to very recent papers with few citations. One specific year could have a very important paper whose contribution might be undervalued because it shares the same year with many zero-citation papers. That year is certainly important. The final calculation of the annual average importance excludes papers with zero citations to accurately reflect the contribution of pivotal works in the field over time. The method for calculating importance can be adapted based on the specific needs of the research project. As will be demonstrated later, the significance of this calculation lies in its ability to determine which papers are selected for summarization based on their quartile rankings.

Volume considerations

Given the rapid evolution of the discipline and the extensive corpus of over 62,000 papers, it is crucial to address the increasing volume of publications as we approach the year 2024, see Fig. 6. To address the challenge of processing large volumes of documents, particularly for years with an overwhelming number of publications, we implemented a MapReduce strategy⁹. This approach is essential due to the inherent limitations of even the most advanced LLMs like GPT-4 in handling extensive text inputs. Given this constraint of the model's input window, we must carefully manage the batch size for processing. For example, when working with a model that has a 128K token input window, we must allocate fewer documents per batch in order to reduce the input tokens. This approach allows for the multiple summarization steps and instructions required by techniques like CoD and our Ensemble Prompts, which process the data iteratively, thereby increasing the overall token usage per input-output operation. This results in a practical limit of around 18.29K tokens per document, which translates to approximately 14,076 words and 101,728 characters. Utilizing the *chunk_size* as a parameter in both LangChain and OpenAI API, we can effectively divide the text into manageable portions, ensuring efficient and coherent summarization despite the volume of data.

Initially, we divide each document into smaller, manageable chunks to ensure that each segment fits within the model's processing window. This segmentation is crucial as it allows the model to process text efficiently without exceeding its token limit. Once the documents are split, we process each chunk individually. If a chunk exceeds the processing capacity or causes errors, we further divide it around a central point, typically identified by specific markers or patterns in the text. This recursive division continues until all sub-chunks are within a manageable size for the model. After processing each chunk to generate intermediate summaries, we then combine these summaries into a coherent final summary. If the combined summary itself exceeds the LLM's processing window, we recursively apply the MapReduce process again, dividing the combined summary into smaller parts and summarizing those until the final summary is within the token limit. This aggregation step

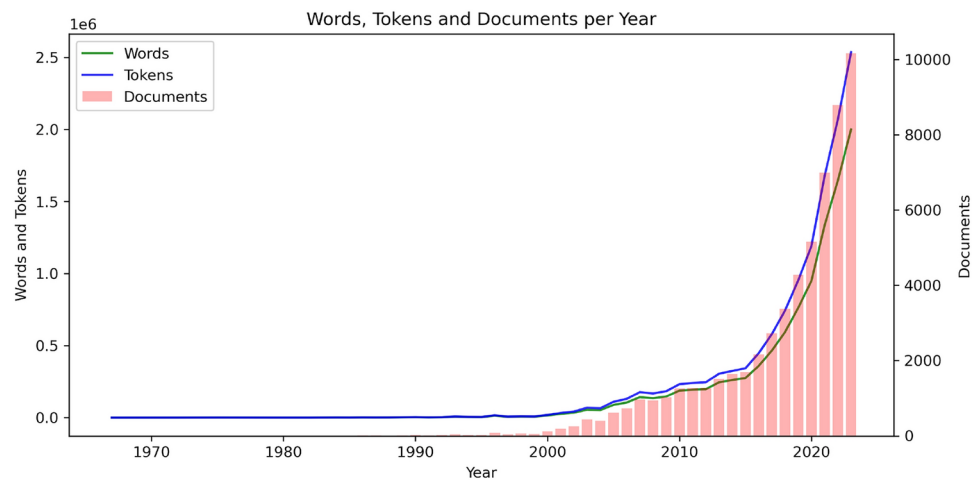


Fig. 6. Distribution of documents over the years.

ensures that the final output retains the context and relevance of the original documents while fitting within the model's constraints. By effectively managing the chunk size and using sophisticated techniques for handling large text volumes, our MapReduce strategy ensures efficient and coherent summarization of extensive document sets, facilitating robust and scalable analysis of large-scale data. See Algorithm 1.

Require: *documents, prompt_map, prompt_reduce*
Initialize *temperature, frequency_penalty, presence_penalty*
Input: *documents, prompt_map, prompt_reduce*
Output: Final summarized document
Ensure: Final summarized document

function SUMMARIZATION(*documents, prompt*)
 Initialize *intermediate_summaries* as an empty list
 for *document* in *documents* **do**
 chunks $\leftarrow$ *split(document)*
 for *chunk* in *chunks* **do**
 try
 intermediateSummaries $\leftarrow$ *summarize(chunk)*
 except
 Summarization(chunk, promptMap)
 end for
 end for
 combinedSummary $\leftarrow$ *combine(intermediateSummaries)*
 return *combined_summary*
end function
Call *Summarization(documents, prompt_map)*
Call *Summarization(combined_summary, prompt_reduce)*

Algorithm 1. Recursive MapReduce summarization strategy

Summarization use case

The final step in our work extends to extracting information from the literature based on the specific themes identified in the previous steps. Among this information, we include referenced summarizations and trends in entity recognition. This comprehensive analysis aims to illuminate underresearched areas, highlight interdisciplinary connections, and guide future research directions.

Our project aims to automate the process of searching and summarizing primary literature on a given topic, providing a novel tool for researchers and practitioners to gain insights into the evolving trends and challenges in the cybersecurity domain, among others.

To ensure quality and coherence in the generated summaries, we employ three distinct prompting techniques: Vanilla (V), Chain of Density (CoD), and our novel proposal, Ensemble Prompts (Ev2). The Ev2 prompts, in particular, were designed with a focus on detailed language to ensure comprehensive topic representation. For

Write a concise summary of the following abstracts:

"{text}"

CONCISE SUMMARY:

Combine the concise summaries into a comprehensive overview:

"{text}"

COMPREHENSIVE OVERVIEW:

Fig. 7. Vanilla prompt templates for Map and Reduce processes.

Article: "{text}"

You will generate increasingly concise, entity-dense summaries of the above Article.

Repeat the following 2 steps 5 times.

Step 1. Identify 1-3 informative Entities (";" delimited) from the Article which are missing from the previously generated summary.

Step 2. Write a new, OOdenser summary of identical length which covers every entity and detail from the previous summary plus the Missing Entities.

A Missing Entity is:

- Relevant: to the main story.
- Specific: descriptive yet concise (5 words or fewer).
- Novel: not in the previous summary.
- Faithful: present in the Article.
- Anywhere: can be located anywhere in the Article.

Guidelines:

- The first summary should be long (4-5 sentences, ~80 words) yet highly non-specific, containing little information beyond the entities marked as missing. Use overly verbose language and fillers (e.g., "this article discusses") to reach ~80 words.
- Make every word count: re-write the previous summary to improve flow and make space for additional entities.
- Make space with fusion, compression, and removal of uninformative phrases like "the article discusses".
- The summaries should become highly dense and concise yet self-contained, e.g., easily understood without the Article.
- Missing entities can appear anywhere in the newsummary.
- Never drop entities from the previous summary. If space cannot be made, add fewer new entities.

Remember, use the exact same number of words for each summary.

Answer in JSON. The JSON should be a list (length 5) of dictionaries whose keys are "Missing_Entities" and "Denser_Summary".

Fig. 8. CoD prompt template for Map and Reduce processes.

our work, the use of verbose prompts was essential to capture nuanced themes accurately, enhancing the clarity and depth of the generated summaries and ensuring a rigorous reflection of the analyzed content.

Both CoD and Ev2 techniques are utilized for the Map and Reduce components, handling initial document chunk summarization and the consolidation of these summaries into a coherent final output.

Prompts

In this subsection, we provide the detailed prompts used for the summarization process in our study. The prompts are tailored to leverage different techniques to enhance the quality and coherence of the generated summaries. By presenting these prompts, we aim to offer a comprehensive understanding of the methodologies employed and facilitate the reproducibility of our results.

Vanilla (V) prompts

The Vanilla prompting technique represents the most straightforward approach, see Fig. 7. This method involves a basic template where the model is provided with the text to be summarized, without additional context or guiding instructions. The simplicity of this approach makes it a useful baseline for comparing the effectiveness of more sophisticated prompting methods. These prompts direct the model to generate summaries based solely on the given input text, ensuring clear and unbiased baseline outputs for further comparison. They are utilized for both the Map and Reduce components, facilitating the initial summarization of document chunks and the subsequent consolidation of these summaries into a coherent final output.

Chain of density (CoD) prompts

The Chain of Density (CoD) prompting technique represents a more advanced approach to summarization, see Fig. 8. This method guides the model through a structured process, emphasizing the density and relevance of the information extracted from the text. The CoD prompts are designed to iteratively refine the summaries, ensuring that the most critical and informative content is retained. The CoD prompt is utilized for both the Map and Reduce components.

In the Map phase, the model summarizes individual document chunks, focusing on extracting dense and relevant information. In the Reduce phase, these intermediate summaries are further refined and consolidated into a final coherent summary.

Ensemble (Ev2) prompts

The Ensemble Prompts (Ev2) technique combines multiple prompting strategies to enhance the summarization process. This method integrates context, role, tone, few-shots, procedural structure, and fool's trap shots to guide the model in generating high-quality summaries. The Ensemble Prompts are designed to provide a comprehensive and nuanced approach, ensuring that the summaries are not only accurate but also contextually rich and coherent. The fool's trap shots is a technique that introduces evidently false information into the one-shot to ensure that the examples are not incorporated into the model's response. Similarly to the concept of fool's trap, the work on Query-Based Adversarial Prompt Generation³¹ discusses the creation of adversarial prompts that challenge the robustness of model responses. While this study primarily focuses on adversarial prompts aimed at identifying and preventing harmful outcomes, the idea of introducing information to test the model's accuracy is highly applicable in our context.

The Ev2 prompts are utilized for both the Map and Reduce components. In the Map phase, see Fig. 9, the model generates initial summaries for document chunks with a multi-faceted approach. In the Reduce phase, see Fig. 10 these summaries are further refined and combined into a final coherent summary.

Summarization LLM

In our study, we use the "GPT-4-Turbo-2024-04-09" model³², an advanced LLM developed by OpenAI, known for its high performance in natural language processing tasks. Key parameters influencing the model's output include temperature, presence penalty, and frequency penalty. Temperature controls the randomness of predictions; higher values produce diverse outputs, while lower values yield more deterministic responses. Presence and frequency penalties reduce word and token repetition, promoting variety in the output. For our summarization process, all parameters are set to zero, ensuring concise and consistent summaries free from unnecessary repetition, which is critical for clarity and coherence in our literature review.

Decade summaries prompts

In our study, we used a specific prompt and one-shot example to summarize the documents by decades. The primary prompt, see Fig. 11, instructed the model to act as an Expert Academic Advisor in summarizing literature reviews with a professional, helpful, and strictly academic tone. The task involved summarizing the text while considering the average importance *I* of each input year's summary, indicated by a scale from 0 to 1. The model

CONTEXT: This is a scientific survey paper about Information Assurance.

ROLE: You are an Expert Academic Advisor in summarizing literature reviews for Information Assurance.

TONE: Professional, helpful, and strictly academic. Use scholarly language, maintain objectivity, and focus on evidence-based insights.

PROCEDURE-- Articles: "{text}"

Step 1: Create an initial topic-densesummary using bibliographic references as citations; e.g., '[Juan Pérez, 1972]'

Step 2: Identify 1-3 new informative topics missing from the previous summary.

Step 3: Add the new informative topics to the previous ones.

Step 4: Write a new dense summary covering all topics and details from the previous summary plus the new topics. Include author references.

Step 5: Repeat Steps 2 to 4, four times.

Step 6: Rephrase the final summary so that it has no more than 2,000 tokens; consider to maintain the bibliographic references.

A missing informative topics is:

- relevant to the main story,
- specific yet concise (5 words or fewer),
- novel (not in the previous summary),
- faithful (present in the article),
- anywhere (can be located anywhere in the articles).

A bibliographic reference is:

- a name and a year surrounded in square brackets (e.g., '[Lulú, 2020]').

GUIDELINES-- Follow these Guidelines:

- The first summary should be short (20 sentences, less than 100 words) yet highly non-specific, containing little information beyond the topics marked as missing.
- The last summary MUST HAVE LESS THAN 2,000 tokens.
- Use overly verbose language and fillers (e.g., "this article discusses...", "In [Lulú, 1979] the authors...").
- Do not use flamboyant language.
- Make every word count: rewrite the previous summary to improve flow and make space for additional topics.
- The summaries should become highly dense and concise yet self-contained, i.e., easily understood without the articles.
- Missing topics can appear anywhere in the new summary.
- Never drop topics from the previous summary.
- IMPORTANT: Ensure to include relevant academic citations in the classic bracketed format, this way: 'In [Anderson J.P., 1972], the authors present the findings of a planning study that addresses the computer security requirements of the USAF, recommending urgent research and development to secure information processing systems for command, control, and support within the Air Force.'
- IMPORTANT: Do NOT use the examples given in the instructions.
- IMPORTANT: some paragraph of your summary should reference to more than one paper if the topic of the sentence in the summary is referenced in them; DO NOT include more than three references. Here you have an example: 'In [Juan Pérez J.P., 1972][Pepito Perez, 1978], the authors use the concept of...'
- IMPORTANT: Return only the last summary, nothing else.
- Eliminate redundant concluding paragraphs.
- Mark the output with "SUMMARY: " for clarity.

Fig. 9. Ev2 prompt template for Map process.

Generate a concise and dense summary of the following text.

"{text}"

Follow these steps:

Step 1: Count the number of words in the text accurately. Use the split() method on the text as if it were a Python string, using white spaces as separators.

Step 2: Generate a summary containing fewer than 2000 words.

Step 3: Count the words in the summary from Step 2 using the same split() method logic from Step 1.

Step 4: If the word count from Step 3 is less than 2000, return the summary from Step 2 and stop the process. Otherwise, generate a new concise and dense summary using the summary from Step 2.

Step 5: Repeat Steps 3 and 4 until the summary meets the maximum word count of 2000 words.

GUIDELINES:

- Aggressively group and condense similar content and topics.
- Use concise language to combine related points from different summaries.
- Include the author and the publication year reference in the summary, this way: '[Author, Year]' (author and year inside the brackets).
- Where possible, cite multiple sources within a single sentence to minimize repetition.
- Do not include more than three references consecutively. Example: 'In [Juan Pérez J.P., 1972][Pepito Perez, 1978] and [Lulú, 2000], the authors use the concept of...'
- Ensure the final summary captures the essence of the original content in a highly condensed form.
- IMPORTANT: Return only the final summary, nothing else. Eliminate all the intermediate steps, any conclusion, summary or similar paragraphs included at the end of your summary; e.g., 'In summary, ...'.
- Mark the output with "SUMMARY: " for clarity.

Fig. 10. Ev2 prompt template for Reduce process.

ROLE: You are an Expert Academic Advisor in summarizing literature reviews.

TONE: Professional, helpful, and strictly academic. Use scholarly language, maintain objectivity, and focus on evidence-based insights.

TASK: Summarize the text below, considering the importance of each paragraph. The importance is measured on a scale from 0 to 1, where 0 is not important and 1 is of maximum importance. The importance is indicated at the beginning of each paragraph in square brackets, as follows: "[Importance=0.5]"

PROCEDURE: Articles: "{text}"

1. Read all the articles and identify the IMPORTANCE at the beginning of each paragraph, surrounded by square brackets. This is the structure [IMPORTANCE=value], and this is an example [IMPORTANCE=0.5].

2. Create a list of the IMPORTANCE values.

3. Determine the quartiles Q1, Q2 and Q3 of the previous list. To do this do the following:

3.1. Sort the list: Arrange the IMPORTANCE numbers in ascending order.

3.2. Find the Median:

- If the number of data points (n) is odd, the median is the middle number.
- If n is even, the median is the average of the two middle numbers.

3.3. Calculate Q1 (First Quartile):

- Q1 is the median of the lower half of the dataset (excluding the median if n is odd).

3.4. Calculate Q3 (Third Quartile):

- Q3 is the median of the upper half of the dataset (excluding the median if n is odd).

3.5. Identify Quartiles:

- Q1: First quartile (25th percentile)
- Q2: Second quartile (50th percentile, the median)
- Q3: Third quartile (75th percentile)

4. Read again all the articles from the beginin and develop a large and detailed summary considering ONLY the paragraphs contained in the articles that have the IMPORTANCE greater than the quartile Q2. Exclude from the summary all the paragraphs with IMPORTANCE less than Q2.

5. Your answer must be in a JSON format with the following keys: "SUMMARY", "INCLUDED", and "NOT_INCLUDED", where:

• "SUMMARY": Contains the summarized text and only the summary. Include the author reference in the summary. Use square brackets to reference the authors. Use the following format: '[Author, Year]', and this is an example: [Lulú, 2025].

• "INCLUDED": An ordered list of tuples with the importance, author and year values USED in the summary; this way [(Importance, "Authors, year")]. Here an example: [(0.765, "Lulú, 2025"), (0.543, "Charly, 2024")].

• "NOT_INCLUDED": An ordered list of tuples with the importance, author and year values NOT USED in the summary; this way [(Importance, "Authors, year")].

6. Ensure that your response includes only the final summary. Do not include any intermediate steps (2, 3 and 4). Carefully review your output to ensure that the summary is coherent and well-structured, and that the lists of included and not included articles each contain approximately half of the articles.

IMPORTANT: This is an example of the desired output: {one_shot}

Fig. 11. Prompt template for decade summarization. For readability, the one-shot is designed in a separate prompt.

was directed to focus on paragraphs over quartiles 2, and 3 of importance, extracting more entities from the most important paragraphs and fewer from the less important ones. The output was provided in JSON format with keys for "SUMMARY", "INCLUDED", and "NOT_INCLUDED" paragraphs. In this prompt, it is clear to see the use of roles, context, few-shots, Chain of Thoughts³³, and Fool's Trap techniques in its construction, forming an ensemble of techniques. The one-shot example prompt, see Fig. 12, demonstrated the desired output format and included a fool's trap, a technique that introduces evidently false information to ensure that the examples are not incorporated into the model's response. In this case, the fool's trap consisted of text generated from LoremIpsum.com³⁴, a common placeholder text used to demonstrate graphic elements or simulate real text. Our experiments demonstrated that this technique is extremely useful for high-volume outputs, as we frequently observed during the development of the prompt that examples would inadvertently appear in the results.

```
{
  "SUMMARY": "Lorem ipsum dolor sit amet [Anderson J.P., 1972], consectetur adipiscing elit. Proin maximus lacus eget odio mattis finibus. Pellentesque eu cursus massa. Fusce a arcu gravida, venenatis lorem vel, fringilla urna [Walter K.G., 1975]. Praesent consequat massa ex, vitae congue ante rhoncus nec. Nunc congue dui ut mauris elementum aliquet. Vivamus at interdum odio, non dapibus odio. Nunc placerat aliquet arcu eu lobortis. Integer lacinia vitae turpis vel finibus. Vestibulum in eros ut lorem dignissim lobortis. Ut sed lectus nec lectus ornare scelerisque ut et leo. Aliquam tempor, sapien sit amet aliquet viverra, enim tortor placerat arcu, et suscipit enim enim luctus est. [Lennon Richard E., 1978].",
  "INCLUDED": [
    {
      "Importance": 0.824,
      "Author and Year": "Anderson J.P., 1972"
    },
    {
      "Importance": 0.697,
      "Author and Year": "Walter K.G., 1975"
    },
    {
      "Importance": 0.582,
      "Author and Year": "Lennon Richard E., 1978"
    }
  ],
  "NOT_INCLUDED": [
    {
      "Importance": 0.155,
      "Author and Year": "Reitman R.P., 1979"
    }
  ]
}
```

Fig. 12. Prompt template for one-shot with fool’s trap based on LoremIpsum. This prompt is concatenated with the one in Fig. 11.

Decade	# docs.	Avg. Tokens	Avg. Words	Ratio Words/Tokens
1970s	12	147.16	115.16	0.77
1980s	78	127.67	100.92	0.78
1990s	412	169.85	135.92	0.79
2000s	5,972	160.31	140.36	0.88
2010s	22,050	190.27	167.03	0.88
2020s	32,268	218.50	190.56	0.87

Table 2. Comparative abstract analysis per decade.

Results

In this section, we describe our main findings; for each decade please see Appendix B of the full version of this work on GitHub³⁵. Table 2 shows a comprehensive summary of the distribution and characteristics of academic documents across several decades, from the 1970s to the present day. It details the total number of documents per decade as well as the average number of tokens and words found in the abstracts of these documents. Additionally, a ratio of words to tokens for each decade is provided, reflecting the density and compactness of the language used in the abstracts over time. Notably, there is a gradual increase in both the number of documents and the number of tokens and words per abstract from the 1970s to the 2020s, indicating a trend towards longer and possibly more detailed abstracts. This evolution may reflect changes in academic standards and the complexity of the topics addressed in publications over the years.

Initial findings and analysis

The results section of this study commences with an analysis of the literature from the 1960s, a decade which, surprisingly, yielded only a single paper relevant to our query on Cybersecurity and Information Assurance. This solitary paper underscores the nascent stage of the information service industry and its foundational challenges³⁶. It highlights the need for a structured separation of costs across computing, communications, and information services, the adoption of multi-access system concepts to economize on infrastructure, and the imperative for developing public, message-switched communications services that incorporate adequate information security measures. During the aggregated decades of the 1960s and 1970s, a total of 60 keywords were identified, with 13 author keywords and 47 index keywords. In these early decades, there were no repeated keywords within the total keywords (column c in Table 3). As the decades progressed, the proportion of unique keywords within the total keywords never exceeded the threshold of 30%. This trend indicates two significant developments: an increase in the variety of studied topics, reflecting the growing complexity of the discipline, and an increase in the number of works on particular themes, resulting in more repeated keywords.

Decade	# Author Keywords (a)	# Index Keywords (b)	# Total Keywords (a+b=c)	Unique Keywords in (c)	%
1960s	0	5	5	0	0.0%
1970s	13	42	55	0	0.0%
1980s	95	318	413	69	16.7%
1990s	383	1,634	2,017	322	16.0%
2000s	7,207	17,936	25,143	6,420	25.5%
2010s	32,137	44,874	77,011	22,798	29.6%
2020s	26,764	39,490	66,254	18,917	28.5%

Table 3. Comparative analysis for Author and Index Keywords lemmatized by decade.

Top 30 Keyword Frequency Distribution in 1970 (no similarity)

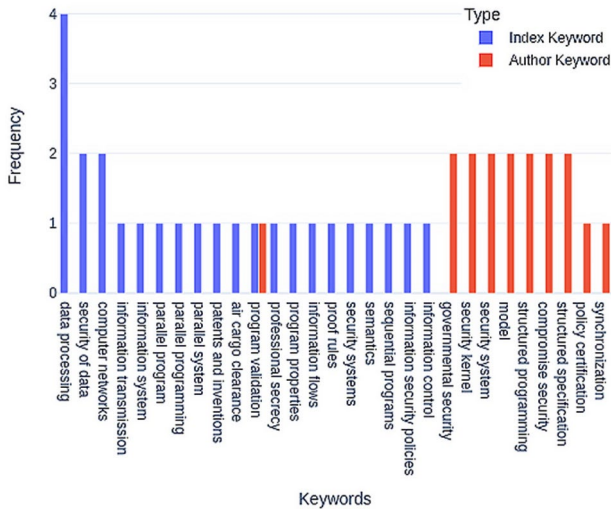


Fig. 13. Top 30 keywords no similarity frequency distribution in 1970s. No match between authors and indexes (journals).

If we analyze the author and index keywords, we observe that their frequencies do not match. This discrepancy can be interpreted as authors claiming to write about certain topics, while readers (journals) are focusing on different subjects. Fig. 13 presents an analysis of the top 30 keywords, where only the keyword *program validation* appears as a common term between the two sets.

Summaries by year

We applied the three methods (V, CoD, and Ev2) to each year of our dataset. We observed different behaviors for V and CoD depending on the need to apply MapReduce, which varies based on the volume of the documents. The *temperature*, *presence_penalty*, and *frequency_penalty* parameters were all initialized to 0. To illustrate the annual summarization techniques, we take the year 2000 as an example. This year included a total of 166 documents. Given the relatively low volume, none of the three templates required the application of the MapReduce strategy. The Vanilla (V) technique produced a comprehensive summary by individually summarizing 96 of the 166 documents from this year, resulting in a total of 2676 words. This method provided a broad overview of the content, capturing a wide array of details from the documents. The CoD technique, known for its focus on extracting dense and relevant information, generated an extremely concise summary of just 46 words; the references are lost. This method distilled the documents down to their most critical points, offering a highly concentrated view of the year’s significant themes. Our novel proposal, the Ensemble Prompts (Ev2) technique, produced an intermediate summary of 275 words. This method maintained the original citation structure, ensuring that the summary was not only coherent but also contextually rich and well-organized. The balance achieved by Ev2 highlights its potential to offer detailed yet concise summaries that respect the integrity of the original documents. If we move to a year with a document volume that necessitates the use of MapReduce, such as 2007 with 728 documents, we observe different behaviors in the models. Vanilla produces a 160 word document but loses the bibliographic references from the original text. CoD again generates a dense summary of just 56 words. Ev2 produces a 179 word text with duplicate references for similar topics. You can find all these results in Appendix B of the full version of this work on GitHub³⁵.

Summaries by decades

In our study, we extended the summarization process to create summaries for each decade, including all documents published within that period. This approach provides a broader perspective on the evolving trends and significant developments over time. However, summarizing documents over a decade poses unique challenges and introduces stress on the evaluation metrics. This stress is similar to the “vanishing gradient” problem³⁷ observed in neural networks with many hidden layers; here, the disparity between the size of the document set and the length of the summary becomes increasingly pronounced. As the size difference grows, the metrics that rely on the original document tend to yield values close to zero; see Fig. 15. This phenomenon intensifies as the volume of documents increases, causing the metrics to lose their effectiveness in analyzing the summary itself. However, they remain useful for comparing the performance of different summarization methods.

We used the outputs from the Ensemble Prompts (Ev2) technique as the source for our decade summaries. Additionally, we adjusted the *temperature* and *presence_penalty* parameters to 0.5. Increasing the temperature allows for more diverse and creative outputs, while a higher presence_penalty discourages the model from repeating the same tokens and encourages the use of new ones, thereby enhancing the richness of the summary. This adjustment helps ensure that the summaries are not only concise and informative but also engaging and varied, capturing the essence of the documents in a more natural and readable manner. For the summarization process, we utilized the average importance factor *I* for the decade. The average was calculated without considering papers with zero citations. We used a specific prompt (Fig. 11) and its corresponding one-shot example (Fig. 12). In the one-shot example, we employed the Fool’s Trap technique to prevent the model from using the example in its output; in this case, we used a text generated from [Ipsium.com](https://www.lipsum.com)³⁴.

Metrics analysis

Currently, metrics for evaluating large-scale summarization tasks are still under development. While recent efforts, such as the combination of metrics proposed in SUSWIR³⁸, or the methodology detailed by Seilt et al.³⁹, which introduces the creation of needles and the calculation of the MINEA score (Multiple Infused Needle Extraction Accuracy), provide advancements, there are still no widely accepted standards that enable objective and consistent evaluation at this scale. This situation underscores the need to further develop specific approaches and metrics for complex automated extraction and summarization tasks. In Table 4, the average values for each metric by decade are presented, offering insight into the trends and effectiveness of different security frameworks over time. The metrics analyzed include SSF, RLF, RDF-CRTD, BAA, SUSWIR, and NIC. These metrics are evaluated using three prompts: CoD, Ev2, and V.

The Semantic Similarity Factor (SSF) measures the semantic alignment between original texts and their summaries, with values ranging from 0 (poor similarity) to 1 (high similarity)³⁸. In Table 4, SSF is evaluated across different decades. The results indicate that in the 1970s, the SSF values were relatively high, with SSF-CoD at 0.6644, SSF-Ev2 at 0.9483, and SSF-V at 0.8686. This suggests that during this period, the summaries maintained a strong semantic alignment with the original texts. However, there is a noticeable decline in SSF values by the 2000s, with SSF-CoD dropping to 0.3639 and SSF-Ev2 to 0.6522. This decrease could indicate challenges in maintaining semantic consistency between original texts and their summaries over time. The 2020s show a slight recovery in SSF values, with SSF-CoD at 0.4507 and SSF-Ev2 at 0.6995, suggesting some improvement in semantic alignment in recent years. Overall, the SSF metric highlights the evolving effectiveness of summarization techniques in preserving the original semantic content.

The Relevance Factor (RLF) measures how well the summary captures the key points of the original content³⁸, utilizing the METEOR (Metric for Evaluation of Translation with Explicit ORDERing) score⁴⁰. METEOR is an advanced metric that considers exact word matches, stemmed matches, synonym matches, and the ordering of matched words, providing a comprehensive evaluation of textual similarity. In Table 4, the RLF is evaluated

Decade	SSF			RLF			RDF-CRTD		
	CoD	Ev2	V	CoD	Ev2	V	CoD	Ev2	V
1970	0.6644	0.9483	0.8686	0.0778	0.3411	0.1424	0.1497	0.6334	0.2529
1980	0.6971	0.9151	0.8499	0.0608	0.2743	0.1316	0.0943	0.4193	0.2263
1990	0.4903	0.8652	0.7667	0.0091	0.0511	0.0537	0.0151	0.0809	0.0861
2000	0.3639	0.6522	0.7812	0.0008	0.0041	0.0232	0.0012	0.0056	0.0378
2010	0.3224	0.6858	0.6993	0.0001	0.0007	0.0002	0.0002	0.0010	0.0004
2020	0.4507	0.6995	0.7461	0.0010	0.0046	0.0046	0.0018	0.0068	0.0100
Decade	BAA			SUSWIR			NIC		
	CoD	Ev2	V	CoD	Ev2	V	CoD	Ev2	V
1970	0.1968	0.7215	0.2953	0.3951	0.7066	0.5200	0.5476	0.6058	0.9786
1980	0.3615	0.6144	0.5535	0.3615	0.6144	0.5535	0.5345	0.8803	0.9340
1990	0.2845	0.4657	0.3847	0.2845	0.4657	0.3847	0.5000	0.9308	0.9340
2000	0.2390	0.3460	0.4017	0.2390	0.3460	0.4017	0.4348	0.8232	0.8239
2010	0.2292	0.3600	0.3951	0.2292	0.3600	0.3951	0.4694	0.9308	0.9340
2020	0.2686	0.3694	0.4111	0.2686	0.3694	0.4111	0.5000	0.9308	0.9340

Table 4. Descriptive Statistics by Decade, Metric, and Prompt (Mean Values).

across different decades. The results show that in the 1970s, the RLF values are relatively low, with RLF-CoD at 0.0778, RLF-Ev2 at 0.3411, and RLF-V at 0.1424. This suggests that the summaries from this period were not very effective in capturing the key points of the original content. Moving into the 1980s, we see a slight improvement, with RLF-CoD at 0.0608, RLF-Ev2 at 0.2743, and RLF-V at 0.1316, though the overall relevance remains low. The 1990s continue this trend with further reductions in relevance, particularly notable in RLF-CoD at 0.0091 and RLF-V at 0.0537, indicating challenges in maintaining the core content in summaries. By the 2000s, the values for RLF show a significant drop, with RLF-CoD almost negligible at 0.0008 and RLF-Ev2 at 0.0041. This drastic decline may reflect either a shift in summarization techniques or a need for improved methods to ensure that the summaries are capturing the essential points. In the 2010s and 2020s, the values remain consistently low, indicating that despite advancements, the relevance of summaries compared to their original texts has not substantially improved. For instance, in the 2020s, RLF-CoD is at 0.0010, and RLF-Ev2 is at 0.0046. It is important to note that the RLF metric is significantly impacted by the volume of text, especially in more recent years. As the volume of content has increased, maintaining relevance in summaries has become more challenging. This factor should be considered when interpreting the consistently low RLF values in recent decades. Overall, the RLF metric highlights ongoing challenges in creating summaries that effectively capture the key points of the original texts, suggesting that more sophisticated or targeted summarization methods may be required to improve relevance.

The Redundancy Factor (RDF) measures the amount of redundant information in a summary³⁸. A high redundancy score indicates that the summary contains less repetitive information, making it concise and more useful. This metric involves comparing each sentence in the summary with every other sentence to calculate their pairwise semantic similarity using Cosine Similarity. Sentences with high similarity scores are considered redundant. The RDF is calculated by determining the proportion of sentence pairs with low redundancy, providing an overall score that ranges from 0 to 1, where higher scores indicate lower redundancy. For this analysis, we use the Cosine Threshold Redundancy Detector (CTRD), as it has shown to be more sensitive to massive texts. In Table 4, the RDF is evaluated across different decades. The results indicate that in the 1970s, RDF values are moderately low, with RDF-CoD at 0.1497, RDF-Ev2 at 0.6334, and RDF-V at 0.2529. These values suggest that summaries from this period contained a fair amount of redundant information. As we move into the 1980s, we observe a decrease in redundancy, with RDF-CoD at 0.0943, RDF-Ev2 at 0.4193, and RDF-V at 0.2263, indicating improved conciseness in summaries. The 1990s show a further reduction in redundancy, particularly in RDF-CoD at 0.0151 and RDF-V at 0.0861, reflecting efforts to produce more concise summaries. By the 2000s, RDF values continue to decrease, with RDF-CoD at 0.0012 and RDF-Ev2 at 0.0056, suggesting significant strides towards reducing redundancy in summaries. However, the values for RDF-V in the 2000s are relatively low at 0.0378, indicating some challenges in maintaining this trend across all prompts. In the 2010s and 2020s, the RDF values show slight fluctuations but remain relatively low overall. For instance, in the 2020s, RDF-CoD is at 0.0018 and RDF-Ev2 at 0.0068, reflecting a consistent effort to minimize redundancy. The increasing volume of text in recent decades has made it more challenging to maintain low redundancy, yet the use of CTRD has proven effective in identifying and reducing repetitive information.

The Bias Avoidance Analysis (BAA) metric checks whether a summary introduces subjective opinions or biases that are not present in the original content. This is achieved by comparing the named entities present in both the original text and the summary using Jaccard Similarity⁴¹, which measures the similarity between two sets. Named Entity Recognition (NER)⁴² identifies and categorizes entities such as names, dates, and locations in the text, allowing for a precise evaluation of how well the summary retains key information from the original document. The BAA values range from 0 (indicating significant introduction of bias) to 1 (indicating perfect overlap and no bias). In Table 4, the BAA is evaluated across different decades using three prompts: CoD, Ev2, and V. The results indicate that in the 1970s, the BAA values were relatively moderate, with BAA-CoD at 0.1968, BAA-Ev2 at 0.7215, and BAA-V at 0.2953. This suggests that while there was some introduction of bias, the overlap of named entities between the original texts and summaries was reasonably high. Moving into the 1980s, we observe an improvement, with BAA-CoD at 0.3615, BAA-Ev2 at 0.6144, and BAA-V at 0.5535, indicating a reduction in the introduction of bias in summaries. The 1990s show a slight decrease in BAA values, particularly in BAA-CoD at 0.2845 and BAA-V at 0.3847, suggesting some challenges in maintaining unbiased summaries. However, the values for BAA-Ev2 remain relatively high at 0.4657, indicating that this prompt still maintained a good overlap of named entities. By the 2000s, BAA values further decrease, with BAA-CoD at 0.2390 and BAA-Ev2 at 0.3460, suggesting an increase in bias introduction. Despite these challenges, BAA-V remains stable at 0.4017. In the 2010s, the BAA values show a slight improvement, with BAA-CoD at 0.2292, BAA-Ev2 at 0.3600, and BAA-V at 0.3951. This trend continues into the 2020s, with BAA-CoD at 0.2686, BAA-Ev2 at 0.3694, and BAA-V at 0.4111, reflecting efforts to reduce bias and improve the overlap of named entities between original texts and summaries.

The Summary Score without Reference (SUSWIR)³⁸ is a comprehensive metric that evaluates the quality of summaries without requiring human-generated reference summaries. SUSWIR combines four individual metrics: SSF, RLF, RDF, and BAA. Each of these metrics evaluates different aspects of the summary, and their weighted combination provides a holistic assessment.

$$SUSWIR(X, Y) = w_1(SSF(X, Y)) + w_2(RLF(X, Y)) + w_3(RDF(Y)) + w_4(BAA(X, Y)) \quad (3)$$

In this sense, $SUSWIR(X, Y)$ is the weighted combination of four factors that assess the quality of the generated summary Y in relation to the original document X . Each factor is assigned a weight, denoted by $w_i \in [0, 1]$, which indicates the relative importance of each factor in the overall assessment. The sum of these weights is constrained such that $\sum_i w_i = 1$. In our study, we utilized specific weights as follow:

$$\{ 'SSF' : 0.4, 'RLF' : 0.3, 'RDF' : 0.1, 'BAA' : 0.2 \}$$

The Semantic Similarity Factor (SSF) is assigned the highest weight of 0.4 because it measures the similarity between the summary and the original text, ensuring that the essential meaning is preserved. The Relevance Factor (RLF) is given a weight of 0.3 as it assesses how well the summary captures the critical points of the original text, making it a crucial aspect of a good summary. The Redundancy Factor (RDF) has the lowest weight of 0.1, reflecting the observation that the summaries are very dense, leaving little room for duplicated content or concepts. Finally, the Bias Avoidance Analysis (BAA) is weighted at 0.2, acknowledging that while it is important to ensure the summary does not introduce biases, the small size of the summary relative to the original text reduces the likelihood of significant bias, thus necessitating a relatively lower weight compared to SSF and RLF. In Table 4, the SUSWIR is evaluated across different decades using three prompts: CoD, Ev2, and V. The results indicate that in the 1970s, the SUSWIR values were relatively high, with SUSWIR-CoD at 0.3951, SUSWIR-Ev2 at 0.7066, and SUSWIR-V at 0.5200. This suggests that summaries from this period were generally effective, balancing semantic similarity, relevance, low redundancy, and minimal bias. Moving into the 1980s, we see a slight improvement in SUSWIR values, with SUSWIR-CoD at 0.3615, SUSWIR-Ev2 at 0.6144, and SUSWIR-V at 0.4785, indicating overall good summary quality. The 1990s show a moderate decline in SUSWIR values, particularly in SUSWIR-CoD at 0.2845 and SUSWIR-V at 0.4229. This decline suggests that summaries during this period faced challenges in maintaining quality across all evaluated aspects. By the 2000s, SUSWIR values continue to decrease, with SUSWIR-CoD at 0.2390 and SUSWIR-Ev2 at 0.3460, reflecting difficulties in producing high-quality summaries during this decade. In the 2010s, the SUSWIR values begin to show improvement, with SUSWIR-CoD at 0.2292, SUSWIR-Ev2 at 0.3600, and SUSWIR-V at 0.3951. This trend continues into the 2020s, where SUSWIR-CoD reaches 0.2686, SUSWIR-Ev2 0.3694, and SUSWIR-V 0.4111. These improvements suggest that more recent summaries have become better at balancing the four aspects evaluated by SUSWIR. The Normalized Inverted Conciseness (NIC) metric is designed to evaluate the conciseness of summaries relative to their original texts, particularly in contexts where multiple documents are concatenated. This metric adjusts for the number of documents and ensures that the conciseness score is scaled between 0 and 1, where a value closer to 1 indicates better conciseness. In Table 4, the NIC is evaluated across different decades using three prompts: CoD, Ev2, and V. The results indicate that in the 1970s, the NIC values were relatively high, with NIC-CoD at 0.5476, NIC-Ev2 at 0.9786, and NIC-V at 0.6058. This suggests that summaries from this period were generally concise relative to their original texts. Moving into the 1980s, NIC values remained strong, with NIC-CoD at 0.5345, NIC-Ev2 at 0.8803, and NIC-V at 0.9340, indicating consistent conciseness in the summaries. The 1990s show a slight decline in NIC values, particularly in NIC-CoD at 0.5000 and NIC-V at 0.4183, suggesting some challenges in maintaining conciseness. However, NIC-Ev2 remained high at 0.9308, indicating that this prompt continued to produce concise summaries. By the 2000s, NIC values remained relatively stable, with NIC-CoD at 0.4348, NIC-Ev2 at 0.8232, and NIC-V at 0.8239, reflecting consistent efforts to maintain conciseness in summaries. In the 2010s, the NIC values show further improvement, with NIC-CoD at 0.4694, NIC-Ev2 at 0.9308, and NIC-V at 0.9340. This positive trend continues into the 2020s, with NIC-CoD at 0.5000, NIC-Ev2 at 0.9308, and NIC-V at 0.9340, demonstrating significant improvements in producing concise summaries. Overall, the analysis of the metrics across different decades reveals several trends in the quality of summaries. The SSF metric highlights the evolving effectiveness of summarization techniques in preserving the original semantic content. The RLF metric underscores ongoing challenges in creating summaries that effectively capture the key points of the original texts, especially as the volume of content has increased. The RDF metric shows notable improvements in reducing redundancy over time, while the BAA metric reflects efforts to minimize bias introduction. The NIC metric demonstrates consistent efforts to maintain conciseness in summaries, with significant improvements in recent decades. Among the three prompts evaluated (CoD, Ev2, and V), the Ev2 prompt generally produced the best results across most metrics. This indicates that the Ev2 prompt is particularly effective in generating high-quality summaries, balancing semantic similarity, relevance, low redundancy, and minimal bias. The overall trend suggests ongoing advancements in summarization techniques, contributing to the production of more effective and high-quality summaries over time.

Methodological evaluation of summarization metrics

In this section, evaluations are conducted using various statistical techniques to analyze the performance and variability of different summarization metrics over time. The attached figures illustrate examples of these analyses, including radar plots to evaluate the average performance of different techniques, box plots to observe the variability of the metrics, and line graphs to analyze trends over the years. Additionally, significance tests are performed to assess the normality of the data points of summaries generated by year. Since the data does not meet the normality condition, the non-parametric Kruskal-Wallis⁴³ test is applied to determine the significance of the differences among the three compared prompts (CoD, Ev2, and V). This approach helps identify significant variations in the quality of the summaries generated by each technique over different periods, providing a deeper understanding of the performance and evolution of summarization techniques. In Fig. 14, the radar plot illustrates the average performance across various metrics for the three summarization techniques (CoD, Ev2, and V). This visualization provides a clear comparison of how each technique performs relative to the others on multiple criteria, highlighting strengths and weaknesses in different areas. It can be observed that Ev2 consistently performs better across all metrics compared to the other two prompts, CoD and V, indicating its overall superiority in summarization performance.

Additionally, line graphs have been added to illustrate the BAA and RLF metrics over different years. As shown in Fig. 15, it is observed that Ev2 consistently outperforms the other techniques across various years in both metrics. However, it is also noted that all metrics show a trend of decreasing towards zero over time. This trend can be primarily attributed to the increasing volume of articles in recent years, which tends to cause a

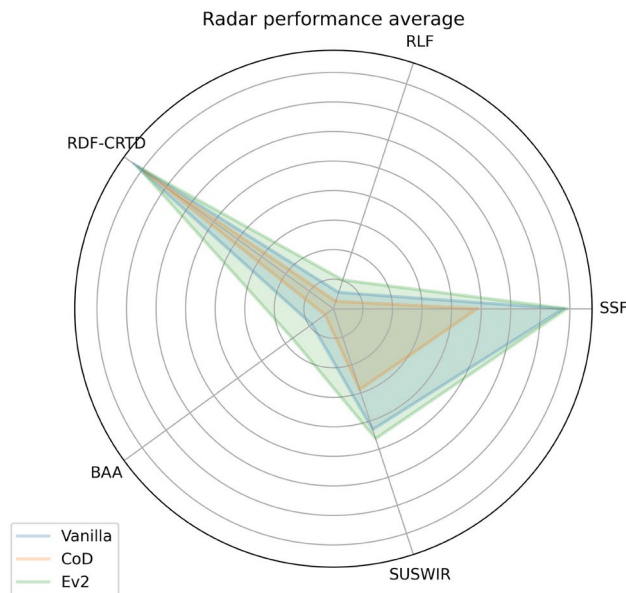


Fig. 14. Metrics average performance.

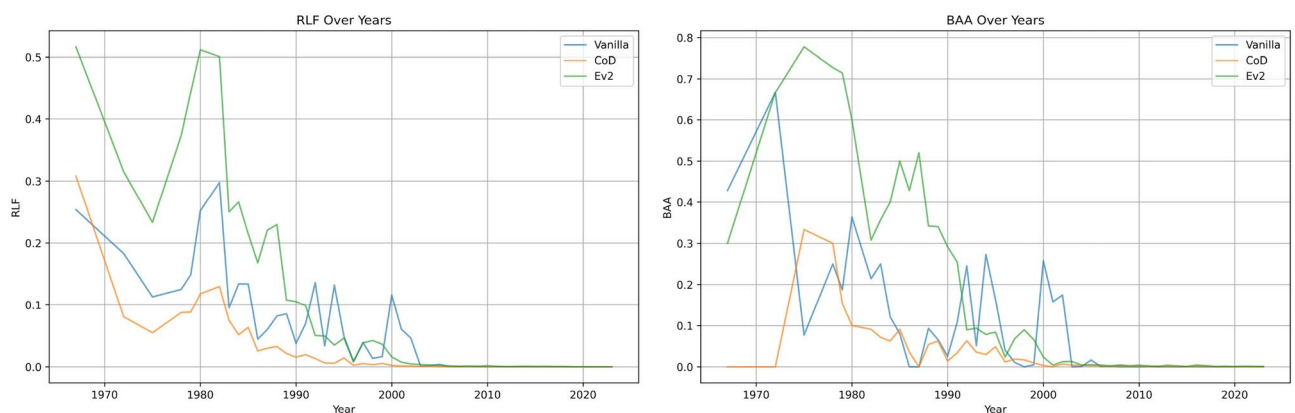


Fig. 15. RLF and BAA metrics over years.

greater deterioration in the metrics. In Fig. 16, the box plot shows the variability of the SSF metric for the three prompts. It is observed that Vanilla and Ev2 have similar results, although Ev2 exhibits greater dispersion. CoD, on the other hand, shows inferior results compared to Vanilla and Ev2.

Finally, to ensure that the observed differences in the various graphs are significant, the results of the Shapiro-Wilk test and the Kruskal-Wallis test are presented in Tables 5 and 6, respectively. The results of the Shapiro-Wilk test in Table 5 confirm that the populations do not follow a normal distribution, as indicated by the p-values being below the typical significance level of 0.05 for most metrics and prompts. This non-normality justifies the use of the non-parametric Kruskal-Wallis test for further analysis. In Table 6, the Kruskal-Wallis test results are displayed. This test is a non-parametric method used to determine if there are statistically significant differences between the distributions of multiple groups. The results indicate that for all metrics analyzed (SSF, RLF, Cons, RDF-CRTD, BAA, SUSWIR, and NIC), there are significant differences among the three prompts (Ev2, CoD, and Vanilla). Specifically, the p-values for each metric are well below the typical significance level of 0.05, suggesting that the variations observed in the visualizations are indeed statistically significant. These findings validate the superiority of the Ev2 prompt across most metrics and confirm the presence of meaningful differences in performance and variability between the prompts. Additionally, it is important to note that this statistical test was applied using 49 data points from the year-by-year analysis of the metrics, providing a robust basis for these conclusions.

While our methodology demonstrates the capability of handling extensive corpora with advanced NLP techniques, we recognize that further efforts are required to solidify the robustness of our results. Specifically, we acknowledge the need for more rigorous validation frameworks to evaluate the outputs generated by large-scale summarizations. At present, there is a lack of standardized metrics applicable to such large-scale tasks, which

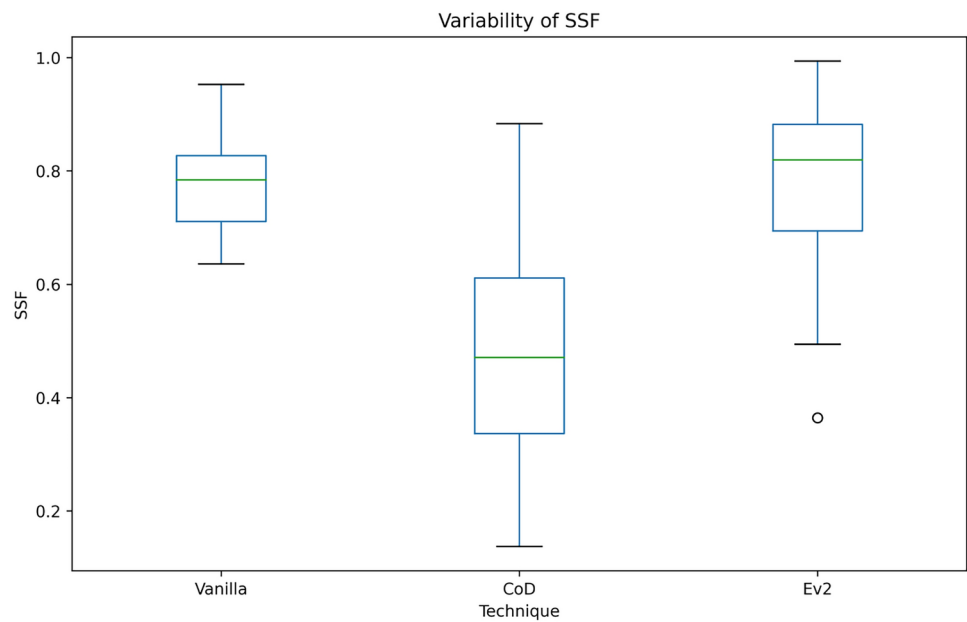


Fig. 16. Variability of SSF metric.

Metric	Prompt	W	p-value
SSF	Vanilla	0.9791	0.05279
	CoD	0.9790	0.05243
	Ev2	0.9548	0.0579
RLF	Vanilla	0.7753	3.04×10^{-7}
	CoD	0.5440	4.13×10^{-11}
	Ev2	0.6992	1.01×10^{-8}
Cons	Vanilla	0.7808	3.99×10^{-7}
	CoD	0.5649	7.97×10^{-11}
	Ev2	0.6840	5.48×10^{-9}
RDF-CRTD	Vanilla	0.8765	1.02×10^{-4}
	CoD	0.6396	1.02×10^{-9}
	Ev2	0.9423	0.0181
BAA	Vanilla	0.7064	1.37×10^{-8}
	CoD	0.5542	5.68×10^{-11}
	Ev2	0.7420	6.35×10^{-8}
SUSWIR	Vanilla	0.9104	0.0012
	CoD	0.9535	0.0513
	Ev2	0.9088	0.0011
NIC	Vanilla	0.4116	9.68×10^{-13}
	CoD	0.9541	0.0544
	Ev2	0.9370	0.0114

Table 5. Results of the Shapiro-Wilk test for normality for each metric and prompt. The W statistic measures how well the data follows a normal distribution, with a lower p-value indicating a departure from normality.

poses challenges in objectively assessing the quality and accuracy of the summaries. Addressing this gap presents a significant opportunity for future research, as developing specialized metrics tailored to these applications would enable more reliable and consistent evaluations.

Conclusion

Automatic summarization techniques are crucial for managing the overwhelming volume of scientific literature, as they enable researchers to quickly extract key information and stay up to date with the latest developments in their fields.

Metric	Statistic	p-value
SSF	62.92	2.18×10^{-14}
RLF	8.62	0.0135
RDF-CRTD	8.76	0.0125
BAA	12.02	0.0025
SUSWIR	59.81	1.03×10^{-13}
NIC	65.15	7.14×10^{-15}

Table 6. Results of the Kruskal-Wallis test for each metric. The statistic indicates the test statistic, and the p-value indicates the significance level.

Our study successfully addresses this pivotal necessity by demonstrating the capabilities of our framework based on Systematic Topic Review and advanced LLM prompting techniques to extract concise and coherent information from large document corpora. The use of in-context techniques presents its own set of challenges, necessitating successive experiments of refinement and optimization to meet human reader expectations. Based on the systematic experiments performed, we have identified that managing the length and density of summaries through entity recognition is crucial, as it directly correlates with the corpus length and significantly impacts the efficacy of information synthesis. Excessively large corpora have a significant tendency to generate summaries that require further prompts to achieve the desired levels of length and density. Our analysis highlights the challenges posed by LLMs, which tend to generate redundant outputs, especially in extensive corpora. This issue is exacerbated as the size of the corpora often exceeds the model’s input token capacity (300,000 tokens in the versions used in our study), necessitating the division of processing into smaller chunks and sacrificing the overall coherence of the summary. In this regard, our conclusion points to the need to develop more nuanced prompting strategies that can handle extensive data inputs without compromising the quality, density, and precision of the output.

In terms of Information Assurance as a discipline, our research contributes to a deeper understanding of the dynamic and evolving nature of cybersecurity over the years. Our work allows for an understanding of how various computational techniques and methodologies have been developed and integrated to strengthen security frameworks that protect digital assets in an increasingly interconnected world. Our work significantly contributes to the field of summarization by addressing the challenges of processing and coherently presenting large volumes of data, particularly in the rapidly evolving domain of Information Assurance. The methodologies and techniques developed here not only enhance the efficiency of summarization processes but also ensure that crucial information is accessible and understandable. This has profound implications for research, academia, and systematic content dissemination, facilitating informed decision making and knowledge dissemination in areas critical to technological and security advancements. The potential uses of our approach extend beyond academic inquiry, impacting the ways industries and governments handle large-scale information, thus playing a pivotal role in shaping informed and secure digital environments.

Our Future Work will address improvement aspects in processing techniques and information extraction from large corpora of documents. The use of MapReduce techniques to process large corpora that exceed the context window size of LLM models can help manage volume but still have room for improvement in the added coherence of the summaries. Similarly, although the assembly of prompting techniques, such as CoD and Few-Shot, generates a sequence of steps in the summary process, it is still necessary to improve the technique to control the length, density, repeatability, and coherence of the summaries. Regarding metrics, we project our work towards the inclusion of automatic and analogous (human) review mechanisms and methods that allow for an objective evaluation of the quality and accuracy of the results obtained in large summarizations. Additionally, we plan to validate the proposed methodology by applying it across diverse topics to assess its robustness and generalizability. By testing the framework with datasets from varied domains, we aim to identify potential limitations and further refine the methodology, ensuring its applicability beyond the initial scope. This cross-domain validation will not only enhance the reliability of our approach but also provide insights into its adaptability and effectiveness in different research contexts, contributing to its broader utility in the field.

Key contributions

Our study’s primary contribution lies in providing a good insight into the evolution of Information Assurance (IA) by presenting a transparent and scalable methodology adaptable to other domains. While the innovative use of in-context learning (ICL) and advanced techniques like BERTopic, UMAP, HDBSCAN, Chain of Density (CoD), and few-shot learning offer a promising foundation for enhancing contextual accuracy in topic modeling and summarization, these technical approaches require further in-depth analysis to fully validate their effectiveness. Nevertheless, they serve as a strong starting point for exploring their applicability in large-scale datasets. The integration of practical strategies, such as the MapReduce framework and pre-configured hyperparameters, ensures the study’s reliability and scalability, making it a valuable resource for understanding IA’s thematic evolution.

Constraints and limitations

One limitation of this study is the absence of exhaustive hyperparameter fine-tuning for the algorithms employed, such as UMAP and HDBSCAN. While we used hyperparameters validated in prior studies and the original

BERTopic configurations, more targeted fine-tuning could potentially enhance clustering results. Certain methods and techniques employed in this research require the reader to refer to their original publications for full methodological details. However, this does not impact the reproducibility of the experiments, which are thoroughly documented and available on GitHub¹. The study highlights the limitations of current automatic metrics when applied to summarizing large documents, as they tend to become less effective with increasing input size, showing a “dilution” effect similar to the “vanishing gradient problem.” While this study does not include human validation, it seeks to bypass this step due to the impracticality imposed by the sheer volume of data processed.

Data availability

All the results from the prompt executions, the obtained summaries, as well as the data and code used, are available on GitHub <https://github.com/Puertarra/InformationAssurance.git>.

Received: 27 August 2024; Accepted: 14 March 2025

Published online: 25 March 2025

References

- Grootendorst, M. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794* (2022).
- Jiang, A. Q. *et al.* Mistral 7b. *arXiv preprint arXiv:2310.06825* (2023).
- Beijing Academy of Artificial Intelligence (BAAI). BAAI/bge-large-en-v1.5: A Pre-trained Language Model. <https://huggingface.co/BAAI/bge-large-en-v1.5> (2023). Accessed: 2024-03-01.
- Muennighoff, N., Tazi, N., Magne, L. & Reimers, N. Mteb: Massive text embedding benchmark. *arXiv preprint arXiv:2210.07316* (2022).
- McInnes, L. & Healy, J. Umap: Uniform manifold approximation and projection for dimension reduction. *J. Open Source Softw.* (2018).
- Campello, R., Moulavi, D. & Sander, J. Density-based clustering based on hierarchical density estimates. *Adv. Knowl. Discov. Data Mining* **7819**, 160–172. https://doi.org/10.1007/978-3-642-37456-2_14 (2013).
- OpenAI. Gpt-4 technical report. *ArXiv abs/2303.08774* (2023).
- Contributors, L. Langchain. GitHub repository (2023).
- Dean, J. & Ghemawat, S. Mapreduce: A flexible data processing tool. *Commun. ACM* **53**, 72–77 (2010).
- Nagwani, N. K. Summarizing large text collection using topic modeling and clustering based on mapreduce framework. *J. Big Data* **2**, 6 (2015).
- Lin, C.-Y. Rouge: A package for automatic evaluation of summaries. In *Text Summarization Branches Out* 74–81 (2004).
- Harnly, A., Nenkova, A., Passonneau, R. & Rambow, O. Automation of summary evaluation by the pyramid method. In *Recent Advances in Natural Language Processing (RANLP)* 226–232 (2005).
- Priya, V. & Umamaheswari, K. Aspect-based text summarization using mapreduce optimization. *Computational Intelligence and Sustainable Systems: Intelligence and Sustainable Computing* 131–139 (2019).
- Ester, M., Kriegel, H.-P., Sander, J. & Xu, X. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Knowledge Discovery and Data Mining* (1996).
- Belerao, K. T. & Chaudhari, S. B. Summarization using mapreduce framework based big data and hybrid algorithm (hmm and dbscan). In *2017 IEEE International Conference on Power, Control, Signals and Instrumentation Engineering (ICPCSI)* 377–380. <https://doi.org/10.1109/ICPCSI.2017.8392320> (2017).
- Würsch, M., Kucharavy, A., David, D. P. & Mermoud, A. Llm perform poorly at concept extraction in cyber-security research literature. *arXiv preprint arXiv:2312.07110* (2023).
- Yao, Y. *et al.* A survey on large language model (llm) security and privacy: The good, the bad, and the ugly. *High-Confidence Computing* 100211 (2024).
- Jamal, S. & Wimmer, H. An improved transformer-based model for detecting phishing, spam, and ham: A large language model approach. *arXiv preprint arXiv:2311.04913* (2023).
- Elluri, L., Mandalapu, V., Vyas, P. & Roy, N. Recent advancements in machine learning for cybercrime prediction. *J. Comput. Inf. Syst.* 1–15 (2023).
- Wang, Z. A systematic literature review on cyber threat hunting. *arXiv preprint arXiv:2212.05310* (2022).
- Salloum, S., Gaber, T., Vadera, S. & Shaalan, K. A systematic literature review on phishing email detection using natural language processing techniques. *IEEE Access* **10**, 65703–65727 (2022).
- Sarker, I. H., Furhad, M. H. & Nowrozy, R. Ai-driven cybersecurity: An overview, security intelligence modeling and research directions. *SN Comput. Sci.* **2**, 173 (2021).
- Kumar, Y. & Kumar, V. A systematic review on intrusion detection system in wireless networks: Variants, attacks, and applications. *Wirel. Pers. Commun.* 1–58 (2023).
- Yang, J., Chen, Y.-L., Por, L. Y. & Ku, C. S. A systematic literature review of information security in chatbots. *Appl. Sci.* **13**, 6355 (2023).
- Aghaei, E., Niu, X., Shadid, W. & Al-Shaer, E. Securebert: A domain-specific language model for cybersecurity. In *International Conference on Security and Privacy in Communication Systems* 39–56 (Springer, 2022).
- Cahyawijaya, S., Lovenia, H. & Fung, P. Llm are few-shot in-context low-resource language learners. *arXiv preprint arXiv:2403.16512* (2024).
- Adams, G., Fabbri, A., Ladhak, F., Lehman, E. & Elhadad, N. From sparse to dense: Gpt-4 summarization with chain of density prompting. *arXiv preprint arXiv:2309.04269* (2023).
- Brown, T. *et al.* Language models are few-shot learners. *Adv. Neural Inf. Process. Syst.* **33**, 1877–1901 (2020).
- García, J. *et al.* Relevance of machine learning techniques in water infrastructure integrity and quality: A review powered by natural language processing. *Appl. Sci.* **13**, 12497 (2023).
- Gana, B., Leiva-Araos, A., Allende-Cid, H. & García, J. Leveraging llms for efficient topic reviews. *Appl. Sci.* **14**, 7675 (2024).
- Hayase, J., Borevkovic, E., Carlini, N., Tramèr, F. & Nasr, M. Query-based adversarial prompt generation (2024). [arXiv:2402.12329](https://arxiv.org/abs/2402.12329).
- Achiam, J. *et al.* Gpt-4 technical report. *arXiv preprint arXiv:2303.08774* (2023).
- Wei, J. *et al.* Chain-of-thought prompting elicits reasoning in large language models. *Adv. Neural Inf. Process. Syst.* **35**, 24824–24837 (2022).
- Lorem Ipsum. Lorem ipsum. <https://www.lipsum.com/> (s.f). Texto de relleno utilizado en la industria de la impresión y diseño gráfico.

¹<https://github.com/Puertarra/InformationAssurance.git>

35. Puertarra. Information assurance. GitHub repository (2024). <https://github.com/Puertarra/InformationAssurance.git>.
36. Dennis, J. B. A position paper on computing and communications. In *Proceedings of the first ACM symposium on Operating System Principles* 6–1 (1967).
37. Hochreiter, S. The vanishing gradient problem during learning recurrent neural nets and problem solutions. *Int. J. Uncertain. Fuzz. Knowl.-Based Syst.* **6**, 107–116 (1998).
38. Foysal, A. A. & Böck, R. Who needs external references?-text summarization evaluation using original documents. *AI* **4**, 970–995 (2023).
39. Seidl, F. *et al.* Assessing the quality of information extraction. *arXiv preprint arXiv:2404.04068* (2024).
40. Banerjee, S. & Lavie, A. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization* 65–72 (2005).
41. Niwattanakul, S., Singthongchai, J., Naenudorn, E. & Wanapu, S. Using of jaccard coefficient for keywords similarity. In *Proceedings of the International Multiconference of Engineers and Computer Scientists* **1**, 380–384 (2013).
42. Roy, A. Recent trends in named entity recognition (ner). *arXiv preprint arXiv:2101.11420* (2021).
43. Breslow, N. A generalized Kruskal–Wallis test for comparing k samples subject to unequal patterns of censorship. *Biometrika* **57**, 579–594 (1970).

Acknowledgements

The authors thank the Biomedical Sensors & Systems Lab for the research support and the article processing charges. The authors also thank the University of North Florida for the research support. Bady Gana is supported by the National Agency for Research and Development (ANID) Scholarship Program/DOCTORADO NACIONAL/2024-21240115. Bady Gana is supported by Beca INF-PUCV. José García is supported by VIN-CI-DI:039.463/2024.

Author contributions

Andrés Leiva-Araos: concept design, data analysis, experimental design and execution, writing and review. Bady Gana: experimental design and execution, and review. José García: experiment validation, writing, and review. Héctor Allende-Cid: review and platform support. Manob Jyoti Saikia: investigation, validation, project management and revising manuscript. All authors have read and approved the manuscript.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to M.J.S.

Reprints and permissions information is available at www.nature.com/reprints.

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

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2025