A Comparative Analysis of Generative AI Systems for Document Summarisation

Matteo Rinco[1], Enrico Scarso[1], Nima Taraghi[1] and Kathrin Kirchner[2]
[1]Department of Management and Engineering, University of Padua, Vicenza, Italy
[2]Department of Engineering Technology, Technical University of Denmark, Ballerup, Denmark

matteo.rinco@gmail.com
enrico.scarso@unipd.it
nima.taraghi@phd.unipd.it
kakir@dtu.dk

Abstract: Generative AI (GenAI) systems can support knowledge workers in managing their knowledge, for example, by effectively processing explicit knowledge, such as document location, classification, integration, and summarisation. Document summarisation is especially useful in many cases since it allows users to quickly identify and understand the key information in a written document (a technical report, an academic paper, a user manual, etc.). In other words, effective summation facilitates distilling the essential meaning, ideas, or information from documents. At present, the main used GenAI tools allow document summarisation. However, they provide different performances since they are based on different Large Language models. In this paper, we compare the summarising performance of a sample of the most popular GenAI tools, i.e., ChatGPT, Copilot, Gemini, and Claude. Our analysis compares the summaries of six documents (academic and nonacademic, in English and Italian) provided by the four tools. These summaries were obtained through a prompt engineering process in which we specified the requirements for the summaries. These summaries were then analysed using quantitative metrics, such as ROUGE and BERTScore, and qualitative criteria. By integrating both types of analysis, we achieved a comprehensive evaluation, reducing subjectivity and analysing the summaries across multiple aspects. Our analysis results do not allow us to conclude that there is a "best-in-class" tool regarding the summarisation function. However, we find that the different summaries have specific characteristics, such as the length of the sentences or the number of synonyms used connected with the GenAI model on which each is based. Therefore, our results confirm that a correct understanding of how GenAI tools work is needed to use them consciously, exploit their potential, and reduce their limitations. The study has some limitations. In particular, we compared only a limited number of tools based on a likewise limited number of documents. Moreover, as these tools are constantly evolving, their performance continues to improve over time.

Keywords: Generative artificial intelligence, Document summarisation, Knowledge management, Comparative analysis

1. Introduction

Generative Artificial Intelligence (GenAI) systems are valuable tools for enhancing the efficiency and effectiveness of various knowledge management (KM) processes (Alavi et al., 2024; Jarrahi et al., 2023). These systems rapidly process explicit knowledge, facilitating tasks such as document location, classification, integration, and summarisation. Document summarisation is especially beneficial in numerous scenarios, enabling users to quickly identify, comprehend, and use key information in written documents (e.g., technical reports, academic papers, and user manuals). Effective summarisation distils the core meaning, ideas, and information from documents, streamlining knowledge acquisition and utilisation.

Currently, all widely used GenAI systems offer document summarisation capabilities. However, as these systems are built on different Large Language Models (LLMs), they exhibit varying performance regarding the relevance and significance of the summaries they produce and, consequently, the knowledge that can be extracted from them. In light of this, our paper compares the summarisation performance of various popular GenAI systems, namely ChatGPT, Copilot, Gemini, and Claude. In particular, we evaluated summaries of six documents (both academic and non-academic, in English and Italian) generated by these four systems. These summaries were obtained through a prompt engineering process in which the specific requirements of the summaries were outlined. The summaries were then analysed using quantitative metrics, such as ROUGE and BERTScore, and qualitative criteria. By integrating these two types of analysis, we achieved a comprehensive evaluation that minimised subjectivity and examined summaries in multiple dimensions.

The results of the analysis do not allow us to conclude that there is a "best-in-class" system regarding the summarisation function. However, we observed that summaries produced by different tools exhibit specific characteristics, such as variations in sentence length or frequency of synonym usage, which are linked to the underlying LLM used by each system. Consequently, our findings underscore the importance of fully understanding how GenAI tools work. Furthermore, since many of these systems are offered as a component of

a specific technological ecosystem, we also reflected on the ecosystems to which the investigated tools belong. This knowledge is crucial for users to employ these tools consciously, maximise their potential, and mitigate their limitations.

The paper is structured as follows: Section 2 provides the background of the study, while Section 3 outlines the research objectives and methodology. Section 4 presents the results of the comparative analysis, which are then discussed in Section 5. The final section also includes conclusions and addresses the limitations of the study.

2. Background

Generative AI is a branch of AI that can create new content (text, images, audio, video) and has the potential to transform domains that rely on creativity, innovation, and knowledge processing (Feuerriegel et al., 2024; Brynjolfsson and Raymond, 2025). In terms of knowledge creation, GenAI can especially support the combination and internationalisation of explicit knowledge (Alavi et al., 2024). Based on massive amounts of data, GenAI can merge, categorise, aggregate, and summarise knowledge from different data sources and, in this way, generate new content and identify patterns. Compared to traditional IT systems, GenAI is not a passive provider of information but an active co-creator.

GenAI systems are evolving rapidly, particularly in natural language processing. Although early summarisation systems were mostly extractive, selecting and copying essential sentences from the source, GenAI models create novel sentences to capture the essence of the original source, which is challenging since it often struggles to match human-level quality (El-Kassas et al., 2021).

The recent developments in deep learning, particularly the introduction of transformer architecture such as BERT, have significantly enhanced models' ability to understand language context. More recently, the evolution of large language models has demonstrated the rapid progress of GenAI tools in summarisation (Balagopal, 2024).

The rapid development of GenAI has raised questions about the reliability, performance, and overall performance of different tools, directly affecting how practitioners select among them. Every generative model has strengths and weaknesses. Therefore, performance can vary across data sets and evaluation criteria (Betzalel et al., 2024; Tewari et al., 2024). Several authors have already compared different GenAI tools along with several criteria and the background of the application. For instance, a comparative analysis of BART, T5, and PEGASUS revealed that each performance differs depending on the quantitative metric used, highlighting that no single model consistently outperforms others in all aspects (Dharrao et al., 2024). Rane et al. (2024) compared Gemini and ChatGPT with respect to their effectiveness and performance in finance and accounting. Their literature and tool documentation analysis showed that performance differences exist, while both GenAI models contribute to higher efficiency in finance and accounting tasks. Shukla et al. (2024) compared ChatGPT, Gemini and Perplexity regarding their performance in text generation, information retrieval, transparency, and context understanding. They found different strengths and weaknesses, and no overall winning tool could be identified. Sobo et al. (2025) conducted an experimental study, evaluating the generated software code for robots with experiments on whether the intended robot task was completed successfully based on the generated code. The authors found that Claude was the most successful GenAI tool compared to Gemini and ChatGPT.

In our study, we focus specifically on document summarization as an organisational KM activity. Our study differs from existing research in that it considers four different GenAI tools, has a specific focus on document summarisation, and compares the various tools with both quantitative and qualitative measures to achieve a thorough evaluation.

3. Study Goals and Methodology

Our study aimed to evaluate the performance of some GenAI tools in summarising documents. To accomplish this goal, we implemented a four-step methodology. First, we selected the GenAI tools considering each tool's degree of diffusion, its underlying mathematical model, and especially the ability to process documents. Then, we selected the documents to be summarised. Afterward, the interaction methods with the tools were established: appropriate prompts were defined to generate the summaries. Lastly, we established the method for evaluating the produced summaries. In this regard, BERTScore and ROUGE as quantitative metrics were used, together with a qualitative evaluation. In the following, we will give only a brief description of each step of the methodology. Readers interested in more details about the individual steps of the methodology can contact the authors. The comparison was made between September and November 2024.

3.1 Selection of GenAI Tools

We selected four of the most known and diffused GenAI tools. The first was ChatGPT, which is the most widely used tool. It can accept images, documents, and texts as input and provide texts and images as output. At the time of the study, three versions of GPT-4 were available: GPT-4 Turbo, GPT-4o and GPT-4o mini, which had rather comparable performance (OpenAI, 2024). We used GPT-4o, as it was available, with limited access, in the free version: after reaching the allowed usage limit, access to this model was temporarily unavailable. However, we could still use the tool with the GPT-4o mini model.

The second selected tool was Google Gemini. It was available in a free version, but this version did not allow document processing. There were four versions of Gemini: Ultra, Pro, Google, and Nano (Google DeepMind, 2024). We used Gemini Pro 1.5, available with the paid version of the tool, because it was the most advanced model with superior performance (Gemini Team, 2024).

The third choice fell on Microsoft Copilot, which is based on Prometheus, a proprietary technology from Microsoft that uses Bing and GPT (at that time ChatGPT-4 Turbo) to generate responses based on real-time data. We had access to the free version through the University account. A key feature of Copilot is to be integrated into all Microsoft applications (Microsoft, 2024).

Lastly, we chose Claude by Anthropic, specifically the paid version 3.5, which was much more powerful than the free one. This version was based on the Sonnet model (Anthropic PBC, 2024). Table 1 provides a summary of the compared tools.

Table 1: Characteristics of the compared tools

	ChatGPT	**Gemini**	**Copilot**	**Claude**
Documents uploading	Yes	Yes	Yes	Yes
Access	Free	Paid	Free	Paid
Model	GPT-4o	Gemini Pro 1.5	GTP.4 Turbo	Sonnet 3.5
Open source	No	No	No	No
Integration	OpenAI platform	Google services	Microsoft Office	Anthropic platform
Languages	Multilingual	Multilingual	Multilingual	Multilingual

It is worth recalling that the type of access (free or paid) did not affect the quality of the content produced since all the tools used the most advanced model provided by each company. For Copilot and ChatGPT, the difference compared to the paid versions was only in the limitations of the use of the model.

3.2 Document Selection

Given that we wanted to use a mix of quantitative and qualitative metrics, which required significant time to examine each summary, the number of selected documents could not be particularly high. Consequently, we decided to choose six documents for twenty-four summaries to examine. Furthermore, we selected academic and non-academic documents to avoid bias induced by a specific topic. Furthermore, to identify possible due to the language, we used documents in English and Italian. Table 2 gives a brief description of the selected documents.

Table 2: List of selected documents

Document Type	**Topic**	**Language**
Academic	Marketing	Italian
	Innovation management	English
Non-academic	Supply chain management	English
	Circular economy	
	Economy	Italian
	Business information management	

3.3 Structuring Prompts

As is widely known, the quality of responses generated by a GenAI tool does not depend solely on the characteristics of the neural network (depth, parameters) or the data used for training, but also on the formulation of the prompt. Therefore, a properly structured prompt is needed to obtain effective and informative responses (Lo, 2023). In this regard, we referred to the new prompt engineering discipline, which concerns "the process of constructing a prompt for LLMs to obtain precise, coherent, and relevant responses" (Lo, 2023). In particular, we used the CLEAR (Concise, Logical, Explicit, Adaptive Reflective) framework and the Prompt Pattern Catalogue to create effective prompts. The catalogue provides a set of prompt patterns applied to solve common problems and enhances both the tool's responses and the prompts themselves (White et al., 2023). In the study, only some of these patterns have been selected and adapted to the specific context, specifically: Persona, Template, Question Refinement, Cognitive Verifier, and Context Manager.

3.4 Setting the Evaluation Criteria

As mentioned above, we employed a mix of qualitative and quantitative evaluation criteria.

For quantitative analysis, we used the two most common metrics, ROUGE and BERTScore, which complement each other in assessing summary quality. BERTScore is ideal for measuring semantic similarity, whereas ROUGE is better suited for assessing surface-level, word-based overlap (Özbolat, 2023; Dharrao et al., 2024).

In principle, the two metrics can be used to compare the summary generated by AI with the original document or the AI-generated summary with a quality summary produced by a human. Since the second option presents several criticalities, for example, the strict dependence on the quality of the human-generated summary, in this study, we chose the first option, thus compering the AI-generated summary with the original document. It is important to specify, therefore, that the values of ROUGE and BERTScore do not represent a measure of the summary quality, but only whether it manages to capture information from the original text. For the mathematical formulas of the two metrics, please refer to Lin (2004) and Zhang et al. (2019), respectively.

The ROUGE factor is an automatic evaluation metric that allows for similarity between two texts by calculating precision and recall (Lin, 2004). Precision identifies whether information not present in the original text has been added to the summary, while recall identifies parts of the AI-generated summary that are also present in the original text. In our study, the two factors are combined to obtain the F-score.

The BERTScore is a language metric that identifies the similarity between two texts and is based on using Bidirectional Encoder Representations from Transformers (BERT). This model uses a Transformer architecture to process the language (Zhang et al., 2019).

Qualitative analysis was based on subjective evaluation of the authors on five criteria, each representing a specific aspect of the summary. We used a scoring system in which each criterion was assigned a score of one to five (Likert scale). The closer the score assigned to a criterion was to five, the better the summary was considered. The criteria considered were:

- **Alignment** (between the request and the response): indicates the tool's ability to understand the prompt and generate a summary that is consistent with the request. It considers the time needed to define the prompt and how well the response aligns with the user's requests.
- **Comprehension**: indicates the AI tool's ability to understand the content of the original document and identify logical connections. It considers the ability to identify key points of the document (Zipitria et al., 2008).
- **Coherence**: evaluates the structure and logical connections between sentences in the summary. It identifies whether the logical progression of sentences allows the reader to understand the summary and evaluates the presence of redundant or unnecessary sentences for comprehension (Zipitria et al., 2008).
- **Cohesion**: evaluates the fluency of the summary and the use of cohesive devices such as connectives, ellipsis, lexical substitution, etc. The more fluid the summary and the more correctly it connects sentences, the higher the score assigned to cohesion (Zipitria et al., 2008).
- **Lexical competence**: evaluates grammar, lexical accuracy, the ability to rephrase the text, and the style used. A good summary should not copy sentences from the original document but rephrase them to make them appropriate to the context.

This analysis took a long time since each original document had to be read and compared with the summaries produced by the four tools.

4. Findings of the Comparative Analysis

4.1 Prompt Definition

The methodology used to define the prompts employed in the study consists of three macro-phases. The first phase defines a basic architecture using specific frameworks and rules. In the second phase, the AI tool's capabilities are leveraged to improve the architecture defined in the previous macro-phase. The improved prompt is tested in the third phase, verifying that the tool provides an adequate response. The prompt is approved if the verification yields a positive result; otherwise, the process is repeated. Fig. 1 shows the flow chart of the process.

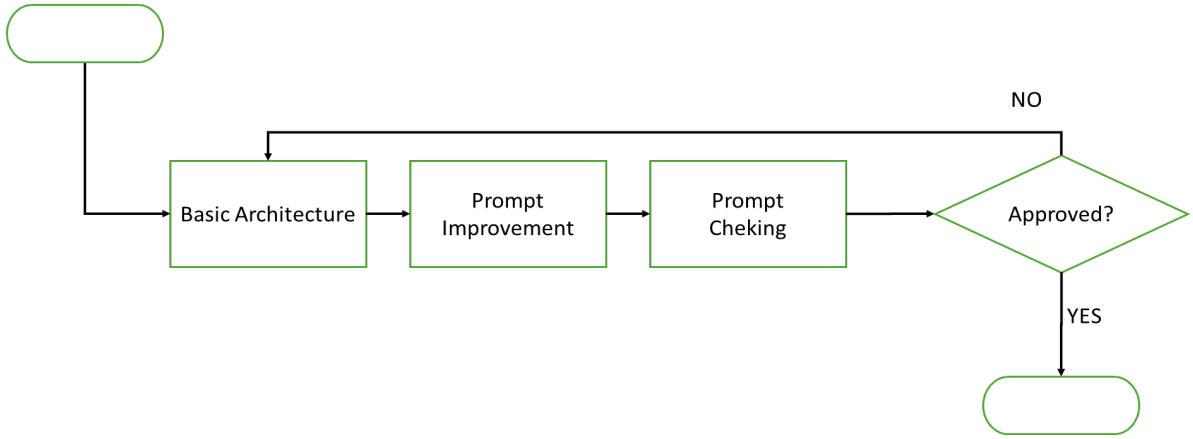

Figure 1: Prompt Engineering process

An example of a final prompt that was created in several iterations for a marketing document with ChatGPT-4 is:

Act as if you were a marketing expert. Use the following structure to generate an academic summary [Introduction, Literature Review, Methodology, Results, and Discussion]. Focus on the logical steps necessary to reach the conclusion and the results obtained, ignoring the bibliography. Maintain an academic tone suitable for a management engineer seeking an informative overview. The keywords to incorporate are: new technologies, fashion, luxuryfication, omnichannel, and retailing. The text must be approximately 15% of the original length, which corresponds to about 1500 words. Strictly adhere to this requirement.

4.2 Validation of Summaries

Two criteria were used to select the summaries to be submitted for comparisons: structure and length. The structure was included in the final prompts and had a binding function during summary generation. When a tool did not meet this requirement, the summary was regenerated until the desired structure was obtained. Using the same structure for all summaries (of the same document) facilitated the comparison of the tools. We met no particular difficulties with this criterion, except for a single case with Google Gemini. As concerns the length of the summaries, first of all, we had to establish. a) how to measure length; b) how to define the length of summaries.

We used the number of words to measure length as it is the most precise method. In particular, we decided that the length of the summaries should be about 15% (only for a document full of graphs and tables, such a value was increased to 20%) of that of the original text. This percentage ensured a balance between summary length and content. Furthermore, a longer length could have flattened the differences between the tools, making evaluation more difficult.

It is worth noting that the length is relevant for calculating quantitative metrics. To make a fair comparison, summaries relating to a specific document had to have the same length. This criterion can also influence qualitative evaluation; a longer summary can more easily include the key points of the original document and insert more details.

4.3 Results

The analysis consisted of calculating individual quantitative and qualitative scores for each summary provided by each tool. Consequently, a total of 6 * 4 * (5 + 5) = 240 individual scores were assigned. After that, scores

were aggregated in various ways to have an overall evaluation of the different tools. For instance, as concerns the qualitative evaluation of the ChatGPT-generated summaries, we assigned the scores shown in Table 3, where the average score is the weighted sum of the individual scores. Additionally, as not all of the five criteria have the same importance, each criteria got a weight (the sum of all weights was 30). The highest weights of 9 were assigned to comprehension and coherence, as a document summary that does not include the content of the original document or reports incorrect information or is logically disconnected, is considered incorrect. The other criteria got a lower weight of 4. Such weights reflect a subjective evaluation of the relative importance of the five qualitative criteria.

Table 3: Scores obtained by the summaries provided by ChatGPT

Summaries	Alignment	Comprehension	Coherence	Cohesion	Lexical competence	Average
Scientific (IT)	5	4	4	3	5	4.13
Scientific (EN)	3	3	4	5	4	3.70
Not scientific (EN)	3	2	3	4	4	2.97
Not scientific (EN)	4	3	4	3	5	3.70
Not scientific (IT)	3	3	4	4	4	3.57
Not scientific (IT)	4	4	4	4	5	4.13

Finally, the scores relating to the four tools were calculated for each criterion. Figure 2 shows the scores obtained in relation to coherence.

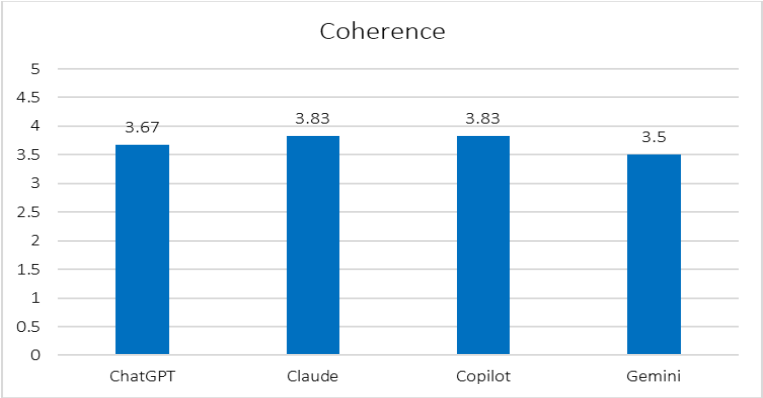

Figure 2: Coherence score

Having calculated the score obtained by each system in each criterion, we calculated the total score of the qualitative analysis as the weighted average score on the five criteria (Fig. 3).

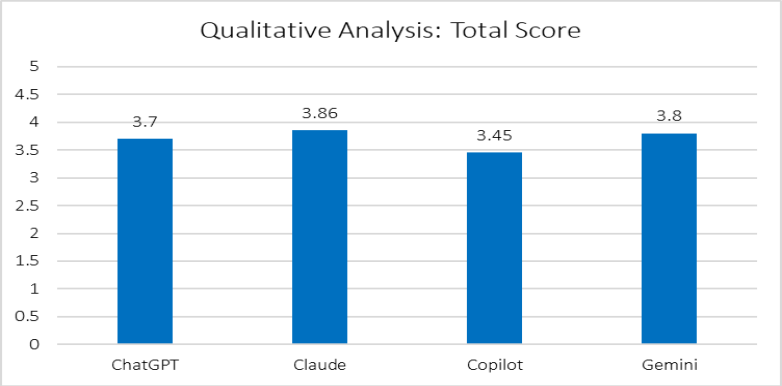

Figure 3: Total scores of the qualitative analysis

Regarding the quantitative analysis, four different ROIUGE indexes were calculated and then aggregated into an overall score, together with the BERTscore. The resulting scores are shown in Table 4.

Table 4: Quantitative scores

	ROUGE (overall) score	BERTscore
ChatGPT	0.157	0.699
Gemini	0.148	0.689
Copilot	0.168	0,709
Claude	0.171	0.702

Lastly, we compared the summaries in relation to the used language. We did not find major differences between the Italian and English summaries, especially concerning the quantitative analysis.

5. Discussion and Conclusion

The present study has thoroughly examined the topic of GenAI. Specifically, the objective was to evaluate the performance of the main GenAI systems currently used for summarising documents. First, we conducted a general overview of how these systems function. Then, the study's methodology was defined, which involved analysing six articles of different types (scientific and non-scientific), topics, and languages (English and Italian). Finally, the summaries produced by the various systems were analysed using both qualitative and quantitative criteria.

The most relevant results of the analysis are summarised as follows:

- **ChatGPT**: In the qualitative analysis, its summaries achieved the highest scores in linguistic competence. The quantitative analysis revealed that the tool rephrases summaries more than Claude and Copilot.
- **Gemini**: The quantitative analysis indicated that this tool rephrases summaries the most. For this reason, the qualitative analysis's summaries received the best scores for coherence and cohesion.
- **Claude**: The qualitative analysis identified this tool as the best in understanding the original document and capturing its key points. It also achieved the highest scores regarding alignment between the request and the response. The quantitative analysis highlighted that the tool retains the original document's style more closely. Indeed, its summaries contained more details compared to those of other tools.
- **Copilot**: In the qualitative analysis, together with Claude, it achieved the highest scores for alignment between the request and the response. The quantitative analysis results show that the tool tends to maintain the same style and use the words of the original document. However, the composition of the summaries is similar to that of ChatGPT.

Overall, no definitive conclusions can be drawn regarding which tool is the best; however, the results indicate that Copilot's performance is slightly inferior to the others.

Finally, we compared the performance of the tools in producing summaries in Italian and English. The analysis highlighted that linguistic competence is better for summaries in English, whereas logical and reasoning abilities are better for summaries in Italian. However, when analysing the summaries as a whole, no significant differences were found between the two languages.

Compared to already existing literature that mostly compares two different GenAI tools for different tasks, our paper focussed on a typical knowledge management task - document summarization, considering even four main GenAI tools for the evaluation. Furthermore, while most studies use quantitative measures for their comparison, our study combined qualitative and quantitative measures to gain a more in-depth understanding of the GenAI tool performances.

5.1 Managerial Implications

This study focused on analysing the performance of GenAI tools in summarising documents. However, it is important to remember that the overall value of a technology is determined by the sum of technological utility, the availability of complementary goods, and the customer base (Schilling, 2022).

Applying this concept to our case, we have that:

- "Technological utility" refers to the performance of generative AI tools in summarising documents, which was the subject of the analyses conducted in this study.
- "Availability of complementary goods" considers the presence of additional components and the possible integration of the tool within an ecosystem.
- "Customer base" refers to the users of GenAI tools or those using an ecosystem in which a GenAI tool can be integrated.

If an organisation wishes to integrate a GenAI tool within its corporate environment, it should consider all three aspects mentioned above. Therefore, choosing one tool over another should also depend on the business context, not just the intrinsic performance. Below are some illustrative examples.

5.1.1 Microsoft ecosystem

If an organisation decides to fully adopt the Microsoft ecosystem for document storage and management (e.g., using Microsoft Teams to share documents within a workgroup), Copilot would be the preferred choice despite its slightly inferior performance. By purchasing Copilot for Microsoft 365 (the enterprise version), the tool can be integrated into all Microsoft applications such as Word, Outlook, Teams, PowerPoint, etc.

Examples of possible activities include summarising texts directly in Word, generating summary reports from PowerPoint presentations, or summarising conversations in Outlook. Furthermore, access to the Microsoft 365 portal, where all organisational documents are stored, allows AI to perform various operations. The integration of this tool facilitates smoother workflows by reducing the time spent switching between applications compared to using other tools.

Thus, both the ecosystem and the large customer base (many companies use Microsoft applications) play a significant role in the selection.

5.1.2 Google ecosystem

Google Gemini would be the logical choice if an organisation opts to utilise the Google ecosystem for document storage and management fully. The reasons and advantages are similar to those mentioned earlier, as Google Gemini can also be integrated within the Google ecosystem, providing easy and quick access to all Google Workspace applications for summarising documents, presentations, and emails.

5.1.3 Hybrid ecosystems

If an organisation chooses to use a proprietary system (not Google or Microsoft) or a mix of multiple ecosystems for document management, the advantages of a unified and integrated ecosystem would be lost. In this case, ChatGPT or Claude would be the choice. Since OpenAI and Anthropic do not have their own ecosystems for integrating these tools, both are valid options for generating summaries, achieving similar performance.

However, it is interesting that ChatGPT has introduced a feature allowing document uploads directly from Google Drive or OneDrive. The strategies of AI tool developers differ: Microsoft and Google aim to induce users to adopt their ecosystems by integrating AI tools, whereas OpenAI seeks to expand its user base by opening up to the most widely used ecosystems.

Finally, during the study, several topics emerged that were not explored in depth as they were not essential for evaluating the tools' performance. For example, the impact of these tools on organisational processes was not analysed. The research methods themselves offer opportunities for reflection, such as conducting further quantitative analysis using a set of high-quality reference summaries instead of comparing with the original document. Another possibility is to conduct a sample study for qualitative analysis to make the study results statistically significant.

5.2 Future Research Directions

The findings provide valuable insights for future developments. Potential areas for further research include:

- Conducting the study using documents that are not STEM-related.
- Increasing the number of documents analysed to evaluate Claude's performance in understanding scientific journal articles, given that its results differed from those of other documents.
- Conducting a focused analysis of ChatGPT's linguistic competence in Italian, as it achieved unexpectedly high scores compared to other tools.

5.3 Limitations

The study presents some limitations. In particular, as a pilot study only a limited number of tools were compared based on a likewise limited number of documents, given the length of the comparison process, mainly due to the use of qualitative criteria. Another limitation relates to the subjectivity of the qualitative evaluation. Lastly, tools are continuously evolving, and hence, their performance is improving.

Ethics declaration: We did not need ethical clearance for the research referred to in this paper.

AI declaration: We used ChatGPT, Copilot, Claude and Gemini to produce the summaries that were analysed.

References

Alavi, M., Leidner, D. E., and Mousavi, R. (2024) "A Knowledge Management Perspective of Generative Artificial Intelligence", *Journal of the Association for Information Systems*, Vol. 25, No. 1, pp 1-12.

Anthropic PBC (2024) "Claude 3.5 Sonnet", [online], https://www.anthropic.com/news/claude-3-5-sonnet.

Balagopal, A. (2024, June 20). *Unlocking precision: Abstractive summarization and the power of retrieval-augmented generation (RAG)*. [online], https://www.radai.com/blogs/unlocking-precision-abstractive-summarization-and-the-power-of-retrieval-augmented-generation-rag.

Betzalel, E., Penso, C., and Fetaya, E. (2024) "Evaluation Metrics for Generative Models: An Empirical Study", *Machine Learning and Knowledge Extraction*, Vo. 6, No. 3, pp 1531-1544.

Brynjolfsson, E., Li, D., and Raymond, L. (2025) "Generative AI at work", *The Quarterly Journal of Economics*, ahead-of-print.

El-Kassas, W.S., Salama, C.R., Rafea, A.A., and Mohamed, H.K. (2021) "Automatic text summarization: A comprehensive survey", *Expert systems with applications*, Vol. 165, 113679.

Feuerriegel, S., Hartmann, J., Janiesch, C., and Zschech, P. (2024) "Generative AI", *Business and Information Systems Engineering*, Vol. 66, No. 1, pp 111-126.

Gemini Team (2024) "Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context" *arXiv preprint arXiv:2403.05530*.

Google DeepMind (2024) "Gemini Nano", [online], https://deepmind.google/technologies/gemini/nano/. Dharrao, D., Mishra, M., Kazi, A., Pangavhane, M., Pise, P., and Bongale, A.M. (2024) "Summarizing Business News: Evaluating BART, T5, and PEGASUS for Effective Information Extraction", *Revue d'Intelligence Artificielle*, Vol. 38, No. 3, pp 847-855.

Jarrahi, M.H., Askay, D., Eshraghi, A., and Smith, P. (2023) "Artificial intelligence and knowledge management: A partnership between human and AI", *Business Horizons*, Vol. 66, No. 1, pp 87-99.

Lin, C.-Y. (2004) "ROUGE: A Package for Automatic Evaluation of Summaries", *Proceedings of the 4th ACL Workshop on Text Summarization Branches Out (WAS 2004),* pp 74-81.

Lo, L.S. (2023) "The CLEAR path: A framework for enhancing information literacy through prompt engineering", *The Journal of Academic Librarianship*, Vol. 49, No. 4, 102720.

Microsoft (2024) "Microsoft Copilot", [online], https://www.microsoft.com/it-it/microsoft-copilot?market=it.

OpenAI (2024) "Models", [online], https://platform.openai.com/docs/models.

Özbolat, H. (2023) "BERTScore and ROUGE: Two metrics for evaluating text summarization systems", [online], https://haticeozbolat17.medium.com/bertscore-and-rouge-two-metrics-for-evaluating-text-summarization-systems-6337b1d98917.

Rane, N., Choudhary, S., and Rane, J. (2024) "Gemini or ChatGPT? Efficiency, performance, and adaptability of cutting-edge generative artificial intelligence (AI) in finance and accounting", [online], https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4731283.

Schilling, M. (2022) *Strategic Management of Technological Innovation*, McGraw-Hill, New York, 7th Ed.

Shukla, M., Goyal, I., Gupta, B., and Sharma, J. (2024) "A comparative study of ChatGPT, Gemini, and Perplexity", *International Journal of Innovative Research in Computer Science and Technology*, Vol. 12, No. 4, pp 10-15.

Sobo, A., Mubarak, A., Baimagambetov, A., and Polatidis, N. (2025) "Evaluating LLMs for code generation in HRI: A comparative study of ChatGPT, Gemini, and Claude", *Applied Artificial Intelligence*, Vol. 39, No. 1, 2439610.

Tewari, N., Joshi, M., Budhani, S. K., Kandpal, V., Rai, A. K., and Jeena, B. (2024) "Comparative Analysis of Generative Artificial Intelligence Tools", in *2024 1st International Conference on Advances in Computing, Communication and Networking (ICAC2N)*, pp 825-829.

White, J., Fu, Q., Hays, S., Sandborn, M., Olea, C., Gilbert, H., Elnashar, A., Spencer-Smith, I., Schmidt, D.C. (2023) "A prompt pattern catalog to enhance prompt engineering with ChatGPT", *arXiv preprint arXiv:2302.11382*.

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., and Artzi, Y. (2020) "BERTtscore: Evaluating text generation with BERT", *International Conference on Learning Representations*, arXiv preprint arXiv:1904.09675.

Zipitria, I., Larrañaga, P., Armañanzas, R., Arruarte, A., and Elorriaga, J. A. (2008) "What is behind a summary-evaluation decision?", *Behavior Research Methods*, Vol. 40, No. 2, pp 597-612.