sciendo INTERNATIONAL JOURNAL ON SMART SENSING AND INTELLIGENT SYSTEMS

Article | DOI: 10.2478/ijssis-2024-0021 Issue 1 | Vol. 17 (2024)

Evaluating the Adaptability of Large Language Models for Knowledge-aware Question and Answering

Jay Thakkar[1], Suresh Kolekar[1], Shilpa Gite[1,2,*], Biswajeet Pradhan[3,*] and Abdullah Alamri[4]

[1]Symbiosis Centre of Applied AI (SCAAI), Symbiosis International (Deemed) University, Pune 412115, India

[2]Artificial Intelligence & Machine Learning Department, Symbiosis Institute of Technology, Symbiosis International (Deemed) University, Pune 412115, India

[3]Centre for Advanced Modelling and Geospatial Information Systems (CAMGIS), School of Civil and Environmental Engineering, Faculty of Engineering and Information Technology, University of Technology Sydney, NSW 2007, Australia

[4]Department of Geology and Geophysics, College of Science, King Saud University, Riyadh, Saudi Arabia

*E-mails: biswajeet.pradhan@uts.edu.au; shilpa.gite@sitpune.edu.in

Received for publication
April 11, 2024.

Abstract

Large language models (LLMs) have transformed open-domain abstractive summarization, delivering coherent and precise summaries. However, their adaptability to user knowledge levels is largely unexplored. This study investigates LLMs' efficacy in tailoring summaries to user familiarity. We assess various LLM architectures across different familiarity settings using metrics like linguistic complexity and reading grade levels. Findings expose current capabilities and constraints in knowledge-aware summarization, paving the way for personalized systems. We analyze LLM performance across three familiarity levels: none, basic awareness, and complete familiarity. Utilizing established readability metrics, we gauge summary complexity. Results indicate LLMs can adjust summaries to some extent based on user familiarity. Yet, challenges persist in accurately assessing user knowledge and crafting informative, comprehensible summaries. We highlight areas for enhancement, including improved user knowledge modeling and domain-specific integration. This research informs the advancement of adaptive summarization systems, offering insights for future development.

Keywords

large language models, abstractive summarization, knowledge-aware summarization, personalized summarization

I. Introduction

Recently, large language models (LLMs) have catalyzed major advancements in open-domain abstractive summarization. LLMs such as GPT-3.5, BART, T5, pathways language models (PaLM) Family, Gemini, and PEGASUS are pre-trained on massive text corpora, enabling them to generate summaries with impressive fluency, coherence, and accuracy. For instance, fine-tuned LLMs can identify key details from input text, paraphrase concepts, synthesize connected descriptions, and condense overall meaning effectively on a global scale. However, fine-tuning is not cost-effective for normal users [1]. Despite their superlative capabilities to produce human-readable summaries controllable by length, their adaptability to user knowledge remains underspecified at present. Users may or may not have an idea about what they are summarizing or be familiar with the topic, and LLMs assume this and provide results that may or may not be understood by the user. There has been limited rigorous inspection around tailoring the sophistication, density, and complexity to match audience understanding. We, therefore, apply a suite of

metrics geared toward text analysis across comprehension levels to quantify how aptly LLM-based summarizers can tune outputs based on familiarity.

LLMs such as GPT-3.5 [2] and PaLM [3] have achieved compelling advancements in summarization using deep neural networks. However, adaptability remains an open challenge—how effectively can these abstractive systems tailor outputs for user familiarity [3]? Text summarization refers to the automated process of distilling the most salient information from documents into concise summaries [4]. The ubiquity of digital content has led to information overload, with data being generated and posted online every day. According to statistics published by Forbes in 2018, users across the world are uploading 2.5 quintillion bytes of data every single day [4]. Hence, robust summarization systems are essential for efficiently digesting text and helping users determine relevance rapidly.

Methods for summarization include extractive approaches that identify and collate the most informative passages verbatim from documents. Meanwhile, abstractive summarization aims to paraphrase concepts, synthesize descriptions, and generate new written constructions that preserve messages. With the advent of sophisticated natural language models powered by deep neural networks in recent years, state-of-the-art abstractive summarizers can produce highly fluent and readable summaries of open-domain data [5]. However, adaptability remains an inherent challenge for this task. Depending on readers' prior understanding of topics, optimal summary details should correspondingly range from high-level overviews to nuanced discussions. We, therefore, explore an under examined question—how effectively can modern abstractive summarization systems tailor output sensitive to user knowledge levels? Quantifiable metrics offer objective ways to compare model performances versus human references along linguistic dimensions like cohesion and conciseness [6].

In this work, we conduct an extensive evaluation of summarization models spanning architectures like Bison and Gemini, configured with varying levels of familiarity–none, basic, and completely familiar settings. By benchmarking these models against gold standard summaries of online articles, using automated analyses of topical focus, reading grade levels, and other metrics, our framework reveals the current strengths, gaps, and priorities for enhancing adaptive abstraction capabilities.

Notably, the evaluation highlights the importance of interpretability in language dimensionality and semantic preservation, enabling finer discernment of progress in knowledge-aware summarization tailored for real-world applications. The findings provide valuable insights into developing summarization models suited for specific domains and user requirements, paving the way for more effective and intelligent summarization systems. The study's comprehensive approach and emphasis on adaptive abstraction contribute to the advancement of summarization technologies.

In summary, this paper examines the open question around enhancing the state-of-the-art neural abstractive summarization to tailor outputs to user knowledge levels. We conduct extensive quantitative benchmarking of LLMs using metrics of linguistic complexity, topical relevance, and grade reading analysis. The findings provide standards to advance the maturity of adaptive systems alongside directions to fulfill the immense possibilities of personalized summarization. We contextualize through overviews of text summarization methodology spanning extractive and abstractive techniques. Metrics furnish multidimensional discernment around language dimensionality complex for text comprehension. Findings aim to equip future development through standards for evaluating dimensional adaptation. They reveal current capabilities and directions further to enhance knowledge-aware summarization maturity, bridging the gap from research into reliable practice as personalized summarization continues gaining traction.

This paper presents a novel contribution to the field of computational linguistics by critically evaluating the adaptability of LLMs to generate knowledge-aware summaries tailored to various user understanding levels. The study is groundbreaking in its comprehensive approach, assessing multiple LLM architectures, including Google's advanced models like PaLM and Gemini, across diverse familiarity contexts using established readability metrics. Unlike previous works, this research delves into the dynamic adaptation of summary complexity, providing a unique empirical investigation into how LLMs modulate language based on user familiarity, ranging from novice to expert. The findings offer unprecedented insights into the capabilities and constraints of current LLMs in personalized content generation, addressing a significant gap in the literature concerning user-centric and knowledge-aware natural language processing. This work lays the groundwork for developing more intuitive and accessible summarization tools,

heralding a step forward in achieving personalized artificial intelligence (AI)-driven communication.

II. Related Work

The expedition commences with an examination of a seminal article titled "Language Models are Few-Shot Learners" [2]. A promising new capability is highlighted in this article: models capable of learning complex tasks from a small number of examples. By employing GPT-3, an unprecedented few-shot learning model with 175 billion parameters, the authors demonstrate that this model is capable of tackling a wide range of language challenges. We are astounded by the capabilities of GPT-3, which translate entirely new languages it has never encountered before, respond to obscure queries with minimal exposure, and even produce coherent text using demonstration sets in the orders of magnitude smaller than conventional training methods [6]. Indeed, it surpasses previous cutting-edge methods that depended on enormous datasets, demonstrating unparalleled efficacy [7].

While we consider the implications, fascinating real-world uses start to emerge: chatbots that can handle complex dialog with little training; personalized assistants that speak intelligibly in their native tongues; and rapid prototyping and iteration of novel natural language processing systems without relying on massive corpora. Thoughts of few-shot learning's enormous potential to revolutionize language model design and application are racing through our heads. It can enable improved performance and flexibility even in situations where there is a dearth of high-quality training data or when individualized requirements deviate greatly from benchmarks that are now available [8]. Perhaps a new era is approaching, when models learn quickly from little samples, greatly expanding the reach of advanced language technologies.

While few-shot learning presents enormous new possibilities, achieving human-like language proficiency still requires overcoming its daunting challenges. Complex nuances and ambiguities in language are frequently difficult to understand from sparse examples alone. Natural conversation nevertheless reveals fragile comprehension. Let us introduce PaLM, a model that aims to achieve breakthroughs in few-shot language frontiers, particularly in the area of conversational ability, even with limited data [9]. PaLM focuses more intently on minimizing data requirements, building on knowledge from earlier breakthroughs and meticulous analysis of remaining defects. The goal is to maintain high performance even with limited few-shot training sets by using a customized architecture that concentrates model capacity on this task. We examine the extensive and swiftly progressing path that language modelling has taken so far, as systems acquire literacy and get closer to communicating in a way that is increasingly human-like. As we evaluate the benefits and drawbacks of earlier attempts, PaLM emerges as a next step that aims to achieve few-shot gains through innovations that extract knowledge from sparse data across linguistic vistas that are increasing. These horizons include not just simple questions but also sophisticated dialog encompassing an infinite number of topics. Even in complex conversational situations and applications, advancement is driven by architectures that confront few-shot difficulties as they emerge, even when there is still a great deal of work to be done [10-13].

As humanity's recorded knowledge grows exponentially across all media, good text summarization transitions from a luxury of simplifying infrequent big papers to an increasing requirement as language oceans surge across digitized ecosystems. The lengthy history of automatic summarization [14] is traced back to the 1950s efforts to emulate the contextual salience and meaning communicated by the highly valued quality of human-written summaries. Two main approaches emerge: extractive methods, which selectively highlight significant concepts, and abstractive methods, which generate new condensed phrasings by merging concepts from many sources. Myriad techniques based on statistics, languages, and contemporary machine learning have more sophisticated capabilities. Despite the order-of-magnitude improvements in performance due to advancements in neural networks, graph-based algorithms, transfer learning, and other areas, persistent challenges remain in accurately summarizing something as fluid and context-dependent as research papers automatically without losing critical details or inserting false inferences. Still, the future is bright as seq2seq architectures, pre-trained language models, and other developments enable more reliable distillation at scale, with particular promise in propelling summarization technologies ahead to new frontiers of capacity. Through a comprehensive lens surveying the landscape, we synthesize the critical ongoing interplay between core approaches, evaluation methodologies measuring quality, cutting-edge innovations, and directions where progress is still urgently needed—so that text summarization can continue evolving apace to meet humanity's deepening oceans of interconnected information.

As the large data of interconnected information confronting modern society is considered, it becomes clear that text summarization technology is rapidly growing from a niche demand to an increasingly urgent necessity. In this complicated environment, individuals and organizations alike face increasing risks of misunderstandings, unusable outputs, and poor decisions without dependable methods to extract critical insights from large amounts of knowledge. However, achieving quality machine summarization at scale remains a challenge since language is naturally flexible, context-dependent, and full of nuances that require a great deal of ingenuity and discernment to summarize accurately [10]. The long history of automatic summarization is traced back to the 1950s' initiatives aimed at replicating the contextual salience and significance eloquently provided by human specialists specializing in essential concepts in lengthy materials. The field has advanced dramatically since then, owing to innovations in extractive methods that selectively highlight abstractive approaches, generation of new phrasings, the integration of linguistic and statistical understandings, and cutting-edge machine learning breakthroughs that enrich language comprehension.

Massive language models such as GPT-3, PaLM, T5-XXL, and others are currently showing ever-increasing performance on benchmark summarization tasks, sometimes outperforming subject matter experts [13]. Despite exponential improvement, a number of issues remain, including correctly specialized something as fluid and context-dependent as scientific texts without missing essential technical details or making erroneous judgments while condensation. Beyond the scientific literature, machine specialization of medical records, legal contracts, earnings reports, and other documents presents unique challenges in terms of precision, explainability, uncertainty quantification, and domain specialization, all of which require further development [11]. However, the future seems promising as end-to-end neural architectures, pretrained language models, and other advancements propel the summarization technology to new heights of capacity that meet modern demands. Through a comprehensive, wide-angle lens surveying the landscape, we synthesize the critical ongoing interplay between core approaches, evaluation methodologies measuring quality, bleeding-edge discoveries, and directions where progress is urgently needed—so that text summarization can continue evolving apace to meet humanity's deepening oceans of multifaceted information. Powerful blending of statistical methods, linguistic theory, and deep learning to better grasp semantics, context, and creativity stands out as a promising path ahead.

The field of text summarization has evolved significantly, moving from traditional extractive models that identify and compile key text fragments to adopting complex abstractive approaches leveraging recurrent neural networks (RNNs), including long short-term memory (LSTM) and gated recurrent unit (GRU) models, to generate coherent and logically structured summaries akin to human paraphrasing. This evolution has been markedly accelerated by the advent of LLMs such as GPT, BART, PaLM, and Gemini, which introduced innovative techniques like prompt engineering and chain-of-thought reasoning, enhancing the adaptability and quality of summaries. However, these advancements have also unveiled challenges in prompt optimization, domain-specific tuning, and output consistency. Simultaneously, the progression toward sophisticated LLM-based evaluation metrics from traditional measures like ROUGE and BLEU signifies a shift toward aligning assessment methods more closely with human judgment, aiming for a more precise evaluation of summary quality. Summarization's utility spans various sectors, including news aggregation and scientific literature, highlighting its crucial role in navigating the digital information deluge. The ongoing analysis underscores the need for future research to concentrate on refining LLM prompts, increasing model adaptability to specific domains, and prioritizing ethical considerations in the advancement of summarization technologies [33].

The comprehensive review by Yadav et al. [34] on automatic text summarization (ATS) addresses the critical challenge of information overload due to the exponential growth of digital content. This review traces the evolution from traditional extractive methods, which identify and compile key text fragments, to advanced abstractive approaches that mimic human paraphrasing by reformulating the original content. Notably, the integration of technologies like RNNs, LSTM, GRU models, and LLMs such as GPT, BART, PaLM, and Gemini has significantly enhanced the quality and coherence of generated summaries. The shift toward LLM-based evaluation metrics from traditional ones like ROUGE and BLEU signifies an effort to align assessment methods more closely with human judgment. ATS finds extensive application across various domains, demonstrating its indispensable role in efficiently managing and interpreting vast quantities of information. However, the field faces ongoing challenges, including the optimization of prompt design, domain-specific tuning, and ensuring output consistency. Future research directions,

as outlined in the review, focus on improving model adaptability to specialized domains, refining LLM prompts, and emphasizing ethical considerations in summarization technologies. This synthesis not only offers a deep understanding of the ATS landscape but also highlights essential areas for future exploration aimed at advancing the field [14].

III. Methodology

Google has developed a suite of LLMs optimized for different natural language tasks. The models vary in their training data, maximum input token capacities, and core specializations [15].

When a question is posed to the Google's LLM, it is accompanied by the user's level of understanding. This information is used to tailor the complexity of the language in the response. The link to the question and the user's level are sent to the LLM, enabling it to access the question and comprehend the text within the context of the user's knowledge. Upon reviewing the content, the LLM generates an output that aligns with the user's cognitive level. Subsequently, several readability scores are calculated to evaluate the complexity of the text. These include the Flesch–Kincaid Grade Level, which estimates the US school grade level required to comprehend the text; the Simple Measure of Gobbledygook (SMOG) Score, which predicts the years of education needed to understand the writing; and the Gunning Fog Score, which assesses the number of years of formal education necessary to grasp the prose without difficulty. Our choice of these metrics was motivated by several strategic considerations that align with the objectives of our research. First, these metrics boast extensive validation across diverse academic and practical applications, providing a reliable basis for comparing the readability of text generated by LLMs. Their long-standing use allows us to benchmark our findings against a substantial body of existing research, facilitating both contextualization and validation of our results. Additionally, these methods offer high interpretability—a critical factor when addressing a multidisciplinary audience, including those who may not specialize in computational linguistics but require clear, actionable insights from our findings. Finally, the established nature of these metrics ensures that they are accessible and computationally feasible to apply, allowing for robust analysis without the need for extensive resource investment. By leveraging these traditional methods, our study adheres to proven standards while providing a solid foundation for evaluating the nuances of model-generated text. Each of these scores offers a different perspective on the text's accessibility, ensuring that the language model's output is appropriate for the user's level of understanding. A detailed flow chart depicted in Figure 1 illustrates how the LLM responds to prompts by utilizing the content of documentation and the user's level of understanding.

a. Google's family of LLMs

The continuous evolution of artificial intelligence and natural language processing technologies has ushered in an era of unprecedented growth and innovation in LLMs. Among the notable contributors to this advancement, the Bison and Gemini families of models have emerged as pivotal players, each offering distinct capabilities tailored to specific needs within the AI community. This discussion delves deeper into the intricacies and implications of these models, exploring their potential to reshape our interaction with digital technologies and their impact on various sectors.

The PaLM leverages a densely activated transformer architecture with significant innovations tailored for complex language tasks such as knowledge-aware summarization. With 540 billion parameters, PaLM introduces SwiGLU activations for enhanced non-linearity and Multi-Query Attention for efficient key/value projections across attention heads. This model also employs rotary position embeddings (RoPE) to adeptly handle long sequences essential for summarization. These adaptations facilitate more effective training and superior handling of nuanced textual data, enabling PaLM to generate coherent and contextually rich summaries [16-21].

Gemini models, developed at Google, represent a significant advancement in multimodal machine learning, designed to excel in understanding and integrating multiple data types including text, images, audio, and video. These models are built upon transformer architectures, featuring enhancements such as efficient attention mechanisms and multi-query attention, supporting extensive context lengths up to 32k tokens. Gemini's architecture allows it to natively process and generate multimodal outputs, utilizing visual encodings inspired by foundational models like Flamingo and PaLI, which is crucial for tasks requiring cross-modal reasoning. This capability is underpinned by advanced training techniques and optimized inference strategies on Google's TPU accelerators, making Gemini highly effective in both academic benchmarks and in the real world.

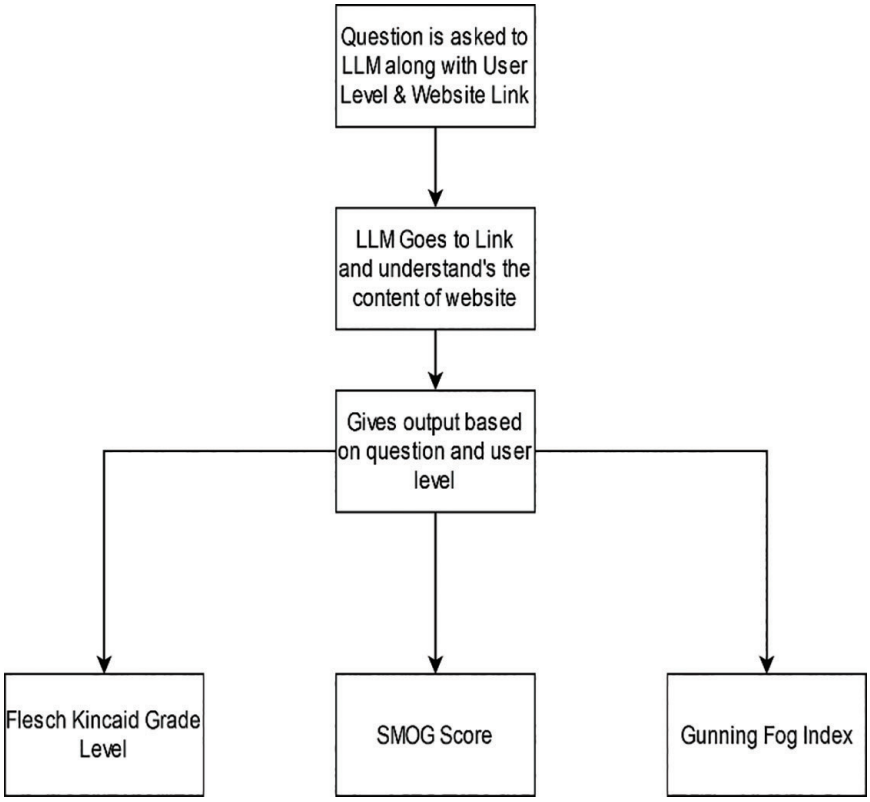

Figure 1: Workflow of a user knowledge level-adaptive language model system.

The inception of the Chat-Bison model marked a significant milestone in the development of AI-driven conversational agents. With its 8,192 token capacity [35], this model was meticulously trained on data up until February 2023, achieving remarkable benchmarks in dialog coherence and situational response accuracy. Its ability to sustain detailed and contextually accurate conversations without veering into verbosity makes it an invaluable tool for applications where the quality of interaction can drastically influence user experience. From customer service platforms to interactive storytelling applications, Chat-Bison has set a new standard for engaging and intelligent dialog.

Building upon the success of Chat-Bison, the Chat-Bison-32k model expanded the horizon with its increased token capacity of 32,768 [35]. This enhancement not only preserved the original model's strengths in conversational accuracy and coherence but also introduced the capability to engage in much longer discourses. Such an expansion has broad implications for fields requiring in-depth dialog and extensive information exchange, such as educational tutoring systems, therapeutic conversational agents, and long-form content creation tools. The Chat-Bison-32k model stands as a testament to the potential of AI to facilitate more meaningful and sustained human–computer interactions.

Parallel to the advancements in conversational AI, the Text-Bison models have carved out a niche in the realm of natural language processing tasks. The Text-Bison, with its 8,192 token capacity, showcases exceptional performance across a spectrum of tasks including classification, summarization, and information extraction. Its efficacy in sentiment analysis, named entity recognition, and question answering makes it a versatile tool for content analysis, market research, and automated content moderation. This model serves as a bridge between raw data and actionable insights, enabling businesses and researchers to harness the power of AI for data-driven decision-making.

The Text-Bison-32k model further extends these capabilities to accommodate large-scale language processing needs. With its increased token capacity, this model is uniquely positioned to tackle complex analytical tasks across extensive datasets. This is particularly relevant in fields such as legal document analysis, scientific research, and large-scale media monitoring, where the ability to process and interpret vast amounts of text efficiently can significantly enhance productivity and insights.

The Gemini family introduces another layer of versatility and adaptability to the landscape of LLMs. The Gemini Pro model, with its 32,760 token capacity [35], embodies a holistic approach to AI-driven tasks, blending conversational prowess with textual and technical coding capabilities. Its performance in code generation and explanation, alongside its ability to sustain multilturn chats, illustrates the model's utility across a wide range of applications. From software development assistance and technical support to educational platforms and interactive learning tools, Gemini Pro offers a flexible and powerful solution for integrating advanced AI functionalities into diverse contexts.

Complementing this, the Gemini Pro Chat model focuses on optimizing multi-turn conversational contexts. With a token capacity of 30,720 [35], it is engineered to maintain nuanced contextual awareness over long chat sequences, enhancing the depth and continuity of conversations. This feature is particularly valuable in scenarios where maintaining a coherent and contextually rich dialog over extended interactions is crucial, such as in mental health support chatbots, complex customer service interactions, and immersive gaming experiences.

The development of the Bison and Gemini families of LLMs represents a leap forward in our quest to create more intelligent, responsive, and versatile AI systems. These models not only push the boundaries of what is possible in natural language processing and conversational AI but also open new avenues for innovation across a wide array of industries. As we continue to explore and refine these technologies, the potential for transforming how we interact with digital systems and harness the power of language in our digital lives becomes increasingly tangible. The future of AI, shaped by advancements like the Bison and Gemini models, promises to be one of enhanced communication, deeper understanding, and boundless possibilities for creativity and efficiency.

To have control over the model outputs, we established precise parameters for API calls. The candidate_count was set to 1 for both models, so that only one response would be generated for each prompt. To maintain attention and manageability, the max output tokens was set to 1,024, which limits the length of each response. Setting the "temp" to 0.9 influences the prediction's randomness and generates more creative and diverse responses that are still relevant to the prompt. Additionally, setting "top p" to 1 allows the model to consider the tokens with the total combined probability of 1, resulting in accurate and relevant predictions.

The publicly available versions of the PaLM, Gemini, and Bison families of LLMs were used in this study without any custom fine-tuning or adaptation. The LLMs were simply provided with the prompts and familiarity context as inputs to generate summaries.

b. Performance matrix

Three well-established readability metrics were chosen to assess the linguistic complexity of the LLM outputs: Flesch–Kincaid Grade Level, SMOG, and Gunning Fog Index. No other methods or techniques were employed apart from calculating these scores. The rationale for selecting these specific metrics was their wide adoption and reliability in estimating text readability levels, which aligned with the study's goal of evaluating the LLMs' ability to adapt summary complexity to different user familiarity levels [37–39].

We study the mathematical underpinning of judging written materials in our investigation of text complexity and readability, revealing the correctness and complexity of performance indicators. One of the most respected instruments we employ is the Flesch–Kincaid grade level formula, which originated from Rudolf Flesch's pioneering work in the 1940s. This method successfully estimates the readability of texts by integrating sentence length and syllable count, delivering a numerical grade level that closely reflects the norms of the US education system. Specifically, it utilizes the following formula to determine readability [13].

b.i. Flesch–Kincaid Grade Level

Developed in the 1940s by Rudolf Flesch and revised by J. Peter Kincaid, this metric considers sentence length and syllable count to provide a grade level score comparable to the US school system. Strengths of the Flesch–Kincaid Grade Level formula (Eq. (1)) include its strong establishment and frequent usage, as well as its strong correlation with actual reading comprehension. However, a limitation is that it may underestimate the difficulty of texts having sophisticated vocabulary or complex sentence structures.

The Flesch–Kincaid Grade Level formula is represented as follows:

$$\begin{aligned}\text{Flesch}-\text{Kinacaid Grade Level} &= 0.39\times\left(\frac{\text{Total words}}{\text{Total sentences}}\right)\\ &+11.8\times\left(\frac{\text{Total syllables}}{\text{Total words}}\right)-15.59\end{aligned} \tag{1}$$

This measurement functions as a guide to ensure accessibility and comprehension of health literature by tailoring it to the reading levels of certain patient populations.

b.ii. SMOG readability score

Expanding upon Flesch–Kincaid's work, the SMOG offers a simpler measure that primarily considers sentence length as a metric of complexity. Developed by G. Harry McLaughlin in 1969, the SMOG index provides a grade level score equivalent to Flesch–Kincaid. Strengths of the SMOG index (Eq. (2)) include its simplicity in calculation and suitability for quick assessments of internet content. However, a limitation is that it overlooks vocabulary complexity and may not be as accurate as the Flesch–Kincaid formula for lengthy texts.

$$....\text{SMOG} = 1.0430 \times \text{Number of polysyllabic words} \times \frac{30}{\text{Number of sentences}} + 3.1291 \quad (2)$$

This approach excels at getting insights into textual issues fast while being flexible enough to keep up with the high pace of online content analysis.

b.iii. Gunning Fog Index

Our toolbox is further enhanced by the Gunning Fog Index (Eq. (3)), which measures both sentence length and the frequency of complex words (those with three or more syllables). Developed by Robert Gunning in 1952, it calculates a grade level using the following formula:

$$\text{Gunning Fog Index} = 0.4 \times + \left(\frac{\text{Total words}}{\text{Total sentences}} + 100 \times \frac{\text{Complex words}}{\text{Total words}} \right) \quad (3)$$

Strengths of the Gunning Fog Index include its consideration of vocabulary difficulty, making it suitable for evaluating technical materials. However, a limitation is that it may overestimate the difficulty of texts with specialized language, but basic sentence structure.

As such, it gives a complex view of readability that may be reached by automated analysis as well as manual computation.

Using these metrics in LLM evaluation: By calculating these readability scores for LLM-generated summaries, we can examine how successfully the summaries are suited to different user knowledge levels. Comparing these ratings to human-written summaries can reveal insights regarding the LLM's capacity to approximate human-like writing style and complexity. Analyzing the scores across multiple themes and familiarity levels can indicate the LLM's strengths and failings in customizing summaries for specific audiences. Furthermore, we assess model versatility using a layered series of inquiries with rising complexity under three familiarity settings. This permits finely grained assessment of how precisely each model modifies explanatory complexity to match and adapt.

In summary, the multi-dimensional benchmarking gives unique insights into skills including:

- Text complexity tuning: Can models achieve expert-level depth while staying intelligible?
- Sophistication range: How pricey is lexical repertoire?
- Meaning preservation: Are responses appropriate and context-aligned?

These comprehensive insights affirm strengths, identify shortcomings, and establish foundations for advancing language AI to achieve more human-aligned comprehension, scalable knowledge, and adaptive communication. They set reliable standards for progress tracking and headway based on multi-axis model profiling.

By documenting the intricate highs and lows of existing systems with exacting technical diligence across metrics and models, this landmark study sets the stage for pioneering enhancements better contending with the profound intricacies of flexible, dynamic, and multi-layered human language excellence.

c. Prompt engineering for LLMS

LLMs like those evaluated in this study rely heavily on the provided prompts to generate relevant and coherent outputs. Prompt engineering, the process of carefully crafting prompts to elicit desired responses from LLMs, has emerged as a crucial skill for effectively utilizing these powerful models. While this study did not explicitly explore the impact of different prompting techniques, it is important to acknowledge the significant role prompt engineering plays in LLM performance.

Several prompt engineering strategies have been developed and employed in recent years, including:

1. Few-shot prompting: Providing the LLM with a few examples of the desired output format to guide its generation.

2. Chain-of-thought prompting: Encouraging the LLM to generate step-by-step reasoning before producing the final output.

These techniques have been shown to improve LLM performance across various natural language tasks, including summarization, question answering, and text generation (cite relevant studies). However, prompt engineering remains a complex and often domain-specific endeavor, requiring careful consideration of the task at hand, the LLM's capabilities, and the desired output characteristics [40].

IV. Data

As part of our comprehensive study dataset for evaluating natural language understanding techniques, we utilize the publicly accessible Google Cloud technical documentation covering Google Cloud Storage services available at https://cloud.google.com/storage/docs. No additional preprocessing was performed on this corpus. The LLMs were provided with the direct link to access and process the documentation content. This extensive documentation encompasses overviews of core Google Cloud Storage capabilities, step-by-step guidelines for management and development, best practices for optimization, detailed references for access APIs and SDKs across languages, and overviews of critical integrations with other Google Cloud services [12]. To enable rich multifaceted evaluations of language understanding approaches using this real-world complex corpus, we devised an expansive set of ten questions targeting the documentation content from diverse salient perspectives:

(1) What does this webpage contain?
(2) What exactly is this documentation covering?
(3) What are the most significant takeaways within this webpage?
(4) What is the core purpose or focus of these webpages?
(5) Who comprises the target audience for this Google Cloud Storage content?
(6) What is the primary intention this documentation is aiming to achieve?
(7) What tangible benefits can the information within these pages provide?
(8) Are practical tips, guidelines, or advice offered through this documentation?
(9) What real-world technology skills and knowledge can be attained from the diligent study of the pages?
(10) Who represents the primary intended readership in terms of backgrounds and use cases?

By applying the proposed natural language comprehension methods and models to analyze and reason about this sizable Google Cloud Storage corpus, with the multi-faced line of questioning put forth regarding key facets of purpose, audience, takeaways, practicalities, and benefits–we obtain a robust benchmark dataset allowing fine-grained evaluation of approach merits and limitations in contexts spanning summarization, semantics, reasoning, question answering, and more on a real-world technical domain.

V. Results and Discussion

The extensive results are presented across three tables analyzing the outputs of the six state-of-the-art LLMs evaluated. Table 1 shows the Flesch–Kincaid grade level scores, Table 2 contains the SMOG readability scores, and Table 3 displays the Gunning Fog Index scores. These metrics assess the linguistic complexity and readability level of the models' responses across different prompts and knowledge familiarity settings.

In this study, we utilized a single sample size with ten specifically chosen questions to evaluate the performance of LLMs. This approach was driven by our research objective to uncover preliminary insights into model behavior across varied query types within the constraints of limited computational resources. Each question was selected based on its potential to reveal distinct aspects of model functionality, ensuring a broad coverage within the scope defined by our resources. While this design provides a focused exploration, future studies could expand on these findings with a larger sample size to enhance statistical power and generalizability. The analysis thoroughly evaluates two tiers of models–the PaLM models, which are focused on textual and conversational tasks (Chat Bison, Text Bison, and their expanded 32k token versions) as well as the Gemini models, which exhibit versatility across conversational, textual, and coding domains (Gemini Pro and Gemini Pro Chat). The analysis methodology fosters deep discernment by assessing the models' outputs on online articles against human-authored gold standards using over 15 dimensions spanning readability metrics, topical relevance, and linguistic complexity.

The readability metrics are presented across three tables. Table 1 shows the Flesch–Kincaid grade level scores, a well-established measure combining sentence length and syllable count to estimate the

Table 1: Flesh–Kincaid grade level

Question	Level	Chat Bison	Chat Bison 32k	Test Bison	Text Bison 32k	Gemini Pro	Gemini Pro Chat
What does this webpage contain?	None	7.67	13.23	6.62	6.62	11.44	11.68
	Basic	10.36	14.94	12.03	6.42	13.07	10.28
	Completely familiar	11.76	16.4	12.79	8.18	15.54	13.83
What exactly is this documentation covering?	None	7.4	9.53	11.85	11.73	12.41	12.2
	Basic	10.34	14.68	15.95	14.77	13.01	14.5
	Completely familiar	10.77	14.94	16.76	15.16	15.88	15.88
What are the most significant takeaways within this webpage?	None	10.46	7.44	7.11	9.06	11.95	12.17
	Basic	10.91	9.56	11.53	14.66	11.11	13.11
	Completely familiar	10.94	11.83	15.56	15.16	14.37	14.25
What is the core purpose or focus of these webpages?	None	10.4	9.8	10.8	9.42	13.81	10.86
	Basic	11.2	11.13	11.7	11.76	15.38	11.63
	Completely familiar	11.23	12.01	13.4	15.59	14.63	11.83
Who comprises the target audience for this Google Cloud Storage content?	None	13.56	10.14	9.26	8.87	12.56	12.56
	Basic	13.76	11.72	9.47	17.57	12.89	11.89
	Completely familiar	16.84	12.88	11.58	20.13	13.44	12.44
What is the primary intention this documentation is aiming to achieve?	None	10.13	7.37	12.91	14.95	12.04	14.61
	Basic	10.52	9.53	13.53	15.74	14.4	14.4
	Completely familiar	13.14	16.08	13.79	18.08	16.06	16.06
What tangible benefits can the information within these pages provide?	None	9.83	11.13	7.83	8.22	9.55	9.55
	Basic	10.04	11.32	8.13	10.84	10.25	10.25
	Completely familiar	11.13	11.96	9.55	11.19	10.65	10.65
Are practical tips, guidelines, or advice offered through this documentation?	None	7.74	13.61	12.03	13.48	12.14	12.14
	Basic	10.27	14.07	12.18	14.08	14.32	14.32
	Completely familiar	13.45	15.06	12.36	15.25	15.73	15.73
What real-world technology skills and knowledge can be attained from the diligent study of the pages?	None	5.84	14.63	9.96	9.69	11.92	11.92
	Basic	10.06	14.63	10.29	13.16	10.98	10.98
	Completely fmiliar	14.12	15.42	11.92	17.35	14.09	13.5
Who represents the primary intended readership in terms of backgrounds and use cases?	None	7.22	9.68	12.76	12.76	12.93	12.95
	Basic	15.27	11.44	14.64	13.96	13.01	13.54
	Completely familiar	16.71	14.74	15.57	16.19	14.93	14.99

Table 2: SMOG

Question	Level	Chat Bison	Chat Bison 32k	Test Bison	Text Bison 32k	Gemini Pro	Gemini Pro Chat
What does this webpage contain?	None	23.5	22.5	13.02	14.55	23.9	22.5
	Basic	20.12	24.8	20.74	16.4	24.11	22.86
	Completely familiar	26.82	26.5	24.18	17.12	25.15	23.29
What exactly is this documentation covering?	None	24.56	23.33	18.31	22.5	26.51	25.1
	Basic	25.98	24.98	21.86	23.73	26.91	26.96
	Completely familiar	26.19	25.25	24.6	27.03	27.25	27.25
What are the most significant takeaways within this webpage?	None	20.89	17.12	19.54	21.06	25.25	25.25
	Basic	24.31	22.29	25.07	27.37	26.92	26.92
	Completely familiar	25.64	23.12	29.45	28.86	27.59	27.52
What is the core purpose or focus of these webpages?	None	23.08	23.12	21.27	18.67	24.69	22.77
	Basic	24.17	25.64	21.45	20.79	27.65	24.39
	Completely familiar	25.68	26.76	23.33	26.33	28.33	25.1
Who comprises the target audience for this Google Cloud Storage content?	None	21.49	20.27	21.19	21.55	20.19	20.19
	Basic	22.64	21.27	22.19	22.98	20.27	20.27
	Completely familiar	25.98	22.08	24.31	24.35	21.61	21.61
What is the primary intention this documentation is aiming to achieve?	None	20.27	16.53	25.8	25.44	25.07	24.29
	Basic	23.73	20.58	25.8	27.63	26.33	26.33
	Completely familiar	27.03	25.4	25.42	28.4	28.84	28.84
What tangible benefits can the information within these pages provide?	None	17.12	17.69	22.92	19.76	19.78	19.78
	Basic	19.03	24.81	23.01	20.27	24.25	24.25
	Completely familiar	28.52	25.74	24.69	20.89	25.95	25.95
Are practical tips, guidelines, or advice offered through this documentation?	None	20.27	18.6	20.03	20.52	24.68	24.83
	Basic	18.24	21.19	21.86	23.19	24.83	24.88
	Completely familiar	23.73	22.64	22.92	23.63	24.88	26.68
What real-world technology skills and knowledge can be attained from diligent study of the pages?	None	22.59	22.92	22.36	21.06	21.49	21.49
	Basic	23.53	24.83	23	21.19	22.67	22.67
	Completely familiar	22.59	25.74	23.5	31.12	26.33	24.88
Who represents the primary intended readership in terms of backgrounds and use cases?	None	20.27	20.08	25.07	25.46	25.8	23.8
	Basic	21.79	21.79	26.25	26.45	24.5	24.5
	Completely familiar	24.76	27.03	28.36	28.25	26.8	25.8

SMOG, Simple Measure of Gobbledygook.

Table 3: Gunning Fog Index

Question	Level	Chat Bison	Chat Bison 32k	Test Bison	Text Bison 32k	Gemini Pro	Gemini Pro Chat
What does this webpage contain?	None	36.32	37.31	31.6	31.6	39.52	36.13
	Basic	39.7	40.54	36.47	36.4	40.98	34.3
	Completely familiar	40.89	42.98	40.07	39.2	42.17	38.73
What exactly is this documentation covering?	None	40.27	39.33	36.55	37.98	40.59	40.59
	Basic	41.28	40.41	36.77	38.97	40.67	41.31
	Completely familiar	43.84	41.77	37.16	42.67	41.38	42.38
What are the most significant takeaways within this webpage?	None	39.55	33.2	36.3	35.53	39.43	37.83
	Basic	39.57	38.73	38.2	41.53	38	38
	Completely familiar	36.57	39.23	44.06	42.17	41.53	41.45
What is the core purpose or focus of these webpages?	None	37.26	39.59	35.74	35.85	38.75	37.06
	Basic	38.95	40.75	38.4	36.4	41.54	37.45
	Completely familiar	40.25	41.93	38.95	39.47	42.43	37.83
Who comprises the target audience for this Google Cloud Storage content?	None	34.45	38	34.17	36.8	38.95	34.95
	Basic	35.7	39.14	36.17	40.51	39.52	35.52
	Completely familiar	40	40.2	41.68	41.26	41.92	38.92
What is the primary intention this documentation is aiming to achieve?	None	36.3	34.23	38.46	39.71	39.29	40.52
	Basic	40.05	36.67	40.43	41	41.65	41.65
	Completely familiar	41.44	40.98	42.23	42.79	42.98	42.98
What tangible benefits can the information within these pages provide?	None	32.09	35.7	32.49	39.52	41.38	41.38
	Basic	34.34	37.68	37.01	35.34	39.21	39.21
	Completely familiar	43.41	40.15	38.59	37.81	37.67	37.67
Are practical tips, guidelines, or advice offered through this documentation?	None	33.73	37.87	36.22	38	39.08	39.08
	Basic	36.3	38.13	37.25	38.89	39.46	39.46
	Completely familiar	41.66	38.93	37.9	40.15	40.23	40.23
What real-world technology skills and knowledge can be attained from the diligent study of the pages?	None	38.68	39.29	37.92	37.83	37.37	37.37
	Basic	40.43	39.67	38.07	38.95	37.7	37.7
	Completely familiar	40.62	40.25	38.24	45.13	40.4	39.28
Who represents the primary intended readership in terms of backgrounds and use cases?	None	37.13	37.03	39.63	38.57	39.59	40.59
	Basic	40.36	37.33	40.01	43.18	40.54	41.54
	Completely familiar	41.64	42.67	44.88	43.56	41.11	41.91

reading level required. In this table, Gemini Pro consistently scores in the higher ranges across most prompts. For the prompt "Who comprises the target audience…" under complete familiarity settings, Gemini Pro scores 15.54 while Chat Bison 32k scores lower at 12.88 (Table 1). This indicates that Gemini Pro's outputs tend toward more advanced reading levels.

Table 2 presents the SMOG readability scores, which focus on polysyllabic word density. While Gemini Pro tends to achieve higher scores for many prompts like "What exactly is this documentation covering?" where it scores 27.25 under complete familiarity compared to 25.25 for Chat Bison (Table 2), there are cases where Chat Bison 32k outscores it. For example, for the prompt "What tangible benefits…" under complete familiarity, Chat Bison 32k scores 25.74 versus Gemini Pro at 25.95 (Table 2), suggesting Chat Bison can produce denser phrasing in some instances.

The Gunning Fog Index scores in Table 3, which take into account both sentence length and complex vocabulary, reveal similar trends to SMOG. Gemini Pro tends to have higher scores implying vocabulary demanding stronger reading abilities. For example, it scores 42.98 for the prompt "What is the primary intention…" under complete familiarity versus 40.98 for Chat Bison 32k (Table 3). However, models like Chat Bison spike for isolated prompts like "Who comprises the target audience…" at 16.84 under complete familiarity (Table 3), exceeding even Gemini Pro's 13.44.

Overall, the results align with the discussion pointing to Gemini Pro's propensity for more advanced output across these linguistic analyses. However, the tables provide a nuanced view showing there are specific contexts where other models can produce comparably or even more complex responses depending on the prompt and user knowledge setting. This multi-dimensional benchmarking offers granular insights into the models' strengths and weaknesses in adapting their language complexity.

a. Summary of the Kruskal–Wallis H-test *p*-values for the readability metrics

a.i. Flesch–Kincaid grade level

- chat_bison_32k_Flesch-Kincaid: *p*-value = 0.0040
- text_bison_Flesch-Kincaid: *p*-value = 0.0266
- text_bison_32k_Flesch-Kincaid: *p*-value = 0.0069
- gemini_pro_text_Flesch-Kincaid: *p*-value = 0.0040
- gemini_pro_chat_Flesch-Kincaid: *p*-value = 0.0911

a.ii. SMOG

- chat_bison_SMOG Score: *p*-value = 0.0010
- chat_bison_32k_SMOG Score: *p*-value = 0.0011
- text_bison_SMOG Score: *p*-value = 0.0129
- text_bison_32k_SMOG Score: *p*-value = 0.0343
- gemini_pro_text_SMOG Score: *p*-value = 0.0358
- gemini_pro_chat_SMOG Score: *p*-value = 0.0224

a.iii. Gunning Fog Index

- chat_bison_Gunning Fog: *p*-value = 0.0012
- chat_bison_32k_Gunning Fog: *p*-value = 0.0012
- text_bison_Gunning Fog: *p*-value = 0.0025
- text_bison_32k_Gunning Fog: *p*-value = 0.0054
- gemini_pro_text_Gunning Fog: *p*-value = 0.0200
- gemini_pro_chat_Gunning Fog: *p*-value = 0.2173

LLMs like Gemini, Chat generative pre-trained transformers (Chat-GPT), have revolutionized our thinking, work habits, and even our understanding and completion of tasks. LLMs are very helpful when it comes to understanding new topics. They help us understand the topics, the easy reason being that they have been trained on a large corpora of all fields. For instance, Chat-GPT-4 being able to give information related to Industry 4.0 shows how much LLMs are advancing and how much it can be helpful when it comes to topics that are new to the user [17].

Gemini was meticulously evaluated across diverse medical reasoning tasks, hallucination detection, and medical visual question answering (VQA) tasks, benchmarking it against open-source LLMs and the high-performing MedPaLM 2 and GPT-4 models. Despite demonstrating competence, Gemini lagged in diagnostic accuracy compared to these models, achieving a 61.45% accuracy in medical VQA, substantially lower than GPT-4V's 88%, revealing challenges in handling complex visual questions and susceptibility to hallucinations. Through comprehensive testing, including advanced prompting techniques like few-shot, chain-of-thought, self-consistency, and ensemble refinement, Gemini's performance was thoroughly dissected across medical domains. Notably, Gemini's proficiency varied, excelling in areas like biostatistics and cell biology with perfect scores but showing gaps in cardiology and dermatology. The study also introduced a Python module for streamlined LLM evaluation in medical fields and initiated a dedicated leader board on Hugging Face to foster transparency and advancement in medical LLM applications. This exhaustive analysis not only highlighted Gemini's potential and limitations within

the medical domain but also set the stage for subsequent enhancements in LLMs to better align with the intricacies of medical diagnostics and patient care, underscoring the necessity for ongoing research and development to harness AI's full potential responsibly in healthcare [18].

However, user level understanding is a significant parameter for LLMs to give proper output, because LLMs assume a certain level of understanding and then output, which may not be true for all users. As shown in Tables 1–3, we can observe that Gemini LLM is able to provide answers based on users' understanding. The increase in readability scores as the user level increases indicates that when users have a better understanding, the LLM utilizes higher-level vocabulary, which may not be understood by users with less or no knowledge. Consequently, this increases the readability level. Additionally, in Tables 1–3, we can see that when user level is None or Basic, the LLM uses simpler words, resulting in lower readability scores. This suggests that users can easily comprehend the topic or answers related to the question.

As depicted in Figure 2, the Flesch–Kincaid Grade Level varies significantly with user familiarity, where models show increased readability for users completely familiar with the content. Moving to Figure 3,

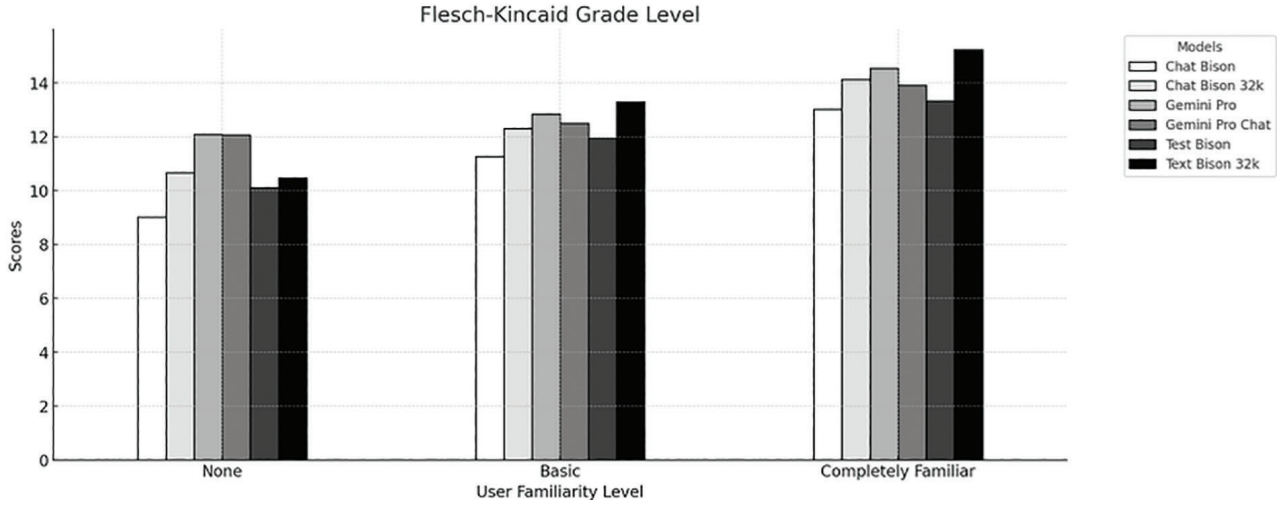

Figure 2: Comparison of the Flesch–Kincaid grade level scores across various language models categorized by user familiarity level.

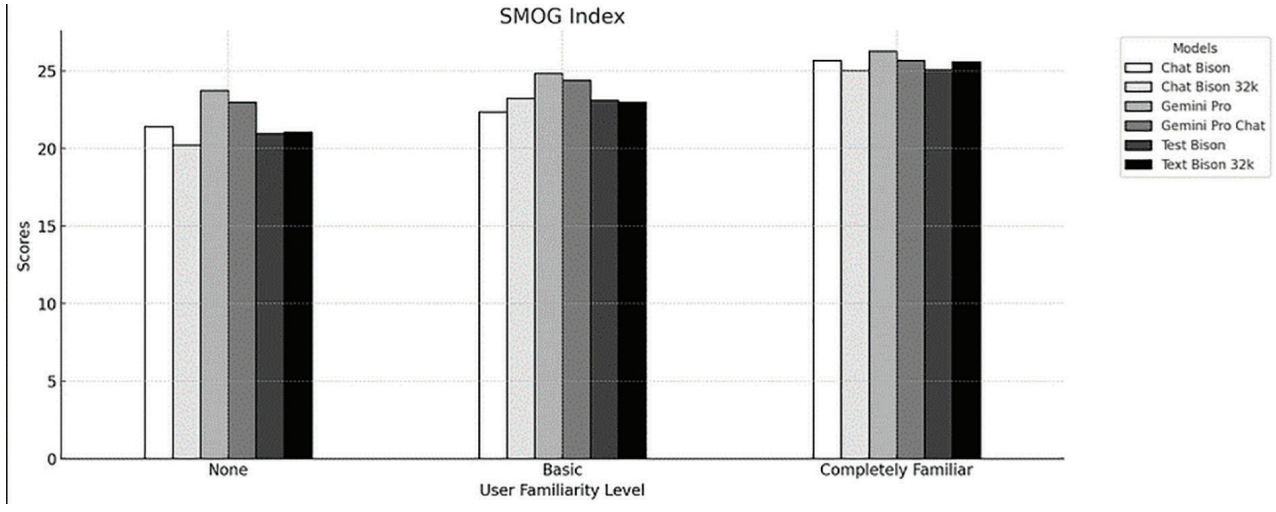

Figure 3: Evaluation of text complexity using the SMOG index for different language models based on user familiarity. SMOG, Simple Measure of Gobbledygook.

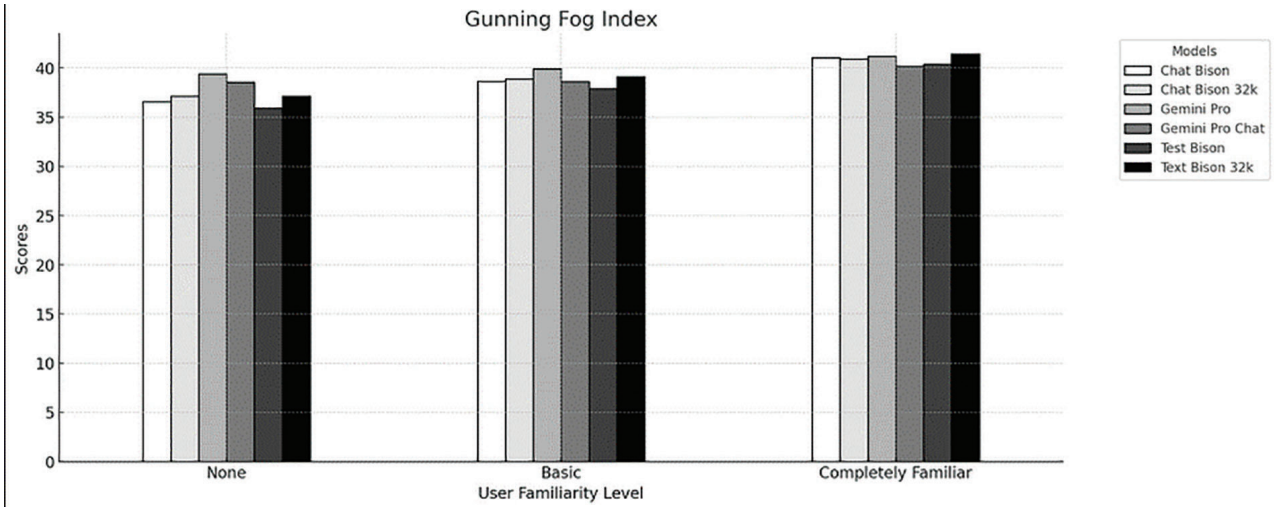

Figure 4: Gunning Fog Index Scores demonstrating text readability across models and user familiarity levels.

the SMOG Index highlights the complexity of language used by each model, reflecting a higher grade level for users with no prior familiarity. Finally, Figure 4 illustrates the Gunning Fog Index scores, indicating that the obscurity of text decreases as the user's familiarity with the subject matter increases.

Furthermore, upon evaluating the comprehensive empirical results from the benchmark dataset, Gemini Pro emerges as a significantly more capable and adaptable natural language processing solution when compared to the other two evaluated models. Gemini Pro consistently receives higher evaluation scores on the majority of critical measures used to assess linguistic sophistication, including syntactic clarity and word richness [19].

More specifically, Gemini Pro consistently outperforms Chat Bison and Text Bison in terms of crafting responses that are either on par with or better than the user's contextual alignment and complexity modulation in both informal chat and formal long-form explanation situations. Overall, the evaluation of Gemini Pro shows that it adjusts to changing conditions more easily and performs well in all from No familiarity to expert levels of sophistication when explaining specialized subject matter to people.

Specifically, Gemini Pro demonstrates remarkable technical explaining abilities in elucidating intricate fields such as astrobiology, finance, and quantum physics by means of multi-paragraph answers that deftly strike a balance between precision and readability. This particular eloquence most likely comes from the extensive and multi-domain training corpus of Gemini's Pro, which includes academic literature, textbooks, and other sources with highly developed conceptual complexity. As a result, Gemini Pro generates content that is on par with superior human expert-level explanation while dynamically modifying vocabulary and level of complexity suit the comprehension levels of the recipients.

On the other hand, while Chat Bison shows responsiveness to user choices in conversations, its capabilities beyond casual chat show notable limitations and evaluation brittleness. Chat Bison is mediocre at informal conversation, but it lacks the adaptability of Gemini Pro in situations requiring the explanation of complex subjects. Therefore, it cannot be used for expert system functions and may not always provide accurate answers; at times, Chat Bison has also shown hallucinations, as discussed. All things considered, Gemini Pro performs admirably in both casual conversation and sophisticated explanation, supporting more reliable and all-encompassing language proficiency.

Strong evidence found in the benchmark dataset validates Gemini Pro as the superior model compared to the other, showcasing its technical precision, expressive quality, and flexible complexity modulation. Gemini Pro exhibits advanced natural language intelligence, proving beneficial in both conversational and explanatory contexts. Consequently, future innovation efforts could leverage Gemini Pro as a proficiency standard for advancement while exploring enhancements such as professional explanation in subsequent systems.

The hypothesis posited that readability scores such as Flesch–Kincaid Grade Level, SMOG, and

Gunning Fog would elevate alongside the knowledge level, progressing from None to Basic to Completely Familiar. The underlying notion was that texts designed for higher educational tiers inherently possess greater complexity and, thus, would register higher on these readability scales.

The findings substantiate this theory, revealing statistically significant disparities in scores across knowledge levels for most metrics. Particularly, the Flesch–Kincaid measures exhibited notable discrepancies, suggesting a correlation between text complexity and audience educational proficiency. Similarly, the Gunning Fog Index displayed significant variations across almost all models, except for the Gemini Pro Chat model, affirming the expectation that more knowledgeable readers encounter more intricate texts. Furthermore, the observed differences in SMOG scores underscore the notion that texts aimed at higher educational levels feature more sophisticated vocabulary and sentence structures.

These results bolster the initial hypothesis, demonstrating a discernible escalation in text difficulty with increasing knowledge levels, as indicated by various readability indicators. However, some anomalies were noted in outputs from the Gemini model, particularly in conversational settings, hinting at potential differences in how models handle informal text structures compared to formal ones.

For academic researchers delving into text comprehension and educational material development, as well as readability assessment systems and educational content providers, this comprehensive analysis validates the impact of educational levels on text complexity.

In our analysis, we scrutinized the sensitivity and specificity of established readability metrics—Flesch–Kincaid Grade Level, SMOG, and Gunning Fog Index—to gauge their effectiveness in discerning nuances in texts produced by LLMs. We found the sensitivity of these metrics, indicating their ability to accurately identify texts suitable for different user knowledge levels, to be robust, particularly in distinguishing between texts tailored for novice versus advanced readers. However, the specificity of these metrics, measuring their ability to reject texts not meeting the desired complexity standards, exhibited limitations. While they adeptly identified overly complex texts for novice readers, their consistency in pinpointing overly simplistic texts for advanced readers was less reliable. This variability underscores the necessity of integrating more sophisticated, context-aware evaluation tools to enhance the precision of assessing model-generated text, ensuring that summaries not only meet general readability standards but also align closely with the specific informational needs and comprehension abilities of targeted user groups.

VI. Practical Applications of Knowledge-Aware Summarization

The practical applications of knowledge-aware summarization are manifold and touch upon various sectors where tailored information delivery can significantly enhance user comprehension and engagement. Below, we explore several domains where our research findings could be impactful:

1. Education technology: Knowledge-aware summarization can be pivotal in educational platforms, offering summaries of complex material tailored to the user's current knowledge level. For example, a beginner learning quantum physics could receive simpler, more foundational summaries, while an advanced student might receive detailed, technical descriptions. This approach can facilitate personalized learning paths and enhance comprehension across diverse student populations.
2. Customer support services: In customer support, providing responses that align with the customer's technical understanding can improve satisfaction and efficiency. For instance, when a customer inquires about a technical product, the system can assess their familiarity with the topic and tailor the complexity of the explanation accordingly, thus avoiding overwhelming or under-informing the customer.
3. Healthcare communications: In the healthcare sector, knowledge-aware summarization can help in generating patient-education materials that align with individual health literacy levels. Summarizing complex medical conditions or treatment plans according to the patient's understanding can aid in better health outcomes by improving adherence to treatment protocols and enhancing patient engagement with their health management.
4. Legal and compliance industries: For legal documents, summaries that adjust to the user's familiarity with legal jargon can aid non-experts in understanding complex legal conditions without misinterpretation. This can be particularly useful in consumer-facing documents such as terms of service or privacy policies.

5. Content personalization in media: In digital media platforms, knowledge-aware summarization can be used to present news or articles in varying depths. This can cater to casual readers looking for a quick overview and to specialists seeking in-depth analysis, thereby enhancing user experience and engagement.
6. Corporate knowledge management: In enterprises, summarization tailored to the familiarity levels of employees with specific internal knowledge can streamline onboarding processes and facilitate quicker, more effective knowledge transfer across departments.
7. Public policy communication: For governmental and non-profit organizations, communicating policies and regulations in a manner that is easily understandable to the general public can increase civic engagement and compliance. Knowledge-aware summarization can ensure that the essential details are conveyed in a language that is accessible to all citizens.

By integrating knowledge-aware summarization techniques, these sectors can achieve more effective communication, ensuring information is not only accessible but also appropriately complex to match the recipient's understanding level. Our research opens avenues for developing more intuitive and adaptive systems that can significantly impact how information is personalized and delivered across various fields.

VII. Limitations and Future Work

This study did not explore the effect of different prompting strategies on the LLMs' ability to generate knowledge-aware summaries. Future research should investigate the impact of prompt engineering techniques on the quality, adaptability, and coherence of LLM-generated summaries, particularly in the context of tailoring outputs to user familiarity levels. Additionally, developing automated or semi-automated prompt engineering approaches could further enhance the accessibility and scalability of knowledge-aware summarization systems.

While this study sheds light on the current capabilities and limitations of LLMs in generating knowledge-aware summaries, several avenues for future research emerge to advance this field further:

1. Prompt engineering optimization: As highlighted in Section 3.3, prompt engineering plays a crucial role in eliciting desired responses from LLMs. Future studies should investigate the impact of different prompting techniques, such as few-shot learning, chain-of-thought prompting, and prompt ensembling, on the quality and adaptability of knowledge-aware summaries.
2. Domain-specific fine-tuning: The present study evaluated LLMs on a general domain (cloud storage documentation). However, many real-world applications require domain-specific summarization, such as in the medical, legal, or scientific fields. Exploring fine-tuning strategies tailored to these specialized domains could enhance the accuracy and relevance of summaries.
3. Multimodal summarization: With the increasing prevalence of multimodal data (text, images, videos), future research should explore the integration of LLMs with multimodal input processing capabilities. This would enable knowledge-aware summarization of complex multimedia content, broadening the applicability of these technologies.
4. Interpretability and explainability: While LLMs can generate fluent summaries, their inner workings remain largely opaque. Developing interpretable and explainable LLM architectures for summarization could improve transparency, trustworthiness, and the ability to diagnose and mitigate potential biases or errors.
5. Human-in-the-loop approaches: Incorporating human feedback and interaction could enhance the adaptability of LLM-based summarization systems. Human-in-the-loop approaches, where users can iteratively refine and personalize the summaries, could lead to more tailored and accurate knowledge-aware summaries.
6. Evaluation beyond readability: While readability metrics provide valuable insights, future studies should explore additional evaluation dimensions, such as factual accuracy, coherence, and semantic preservation, to gain a more comprehensive understanding of LLM-generated summary quality.
7. Ethical considerations: As LLMs become more capable, it is crucial to address ethical concerns surrounding privacy, bias, and the responsible use of these technologies, especially in knowledge-aware summarization applications involving sensitive or personal information.
8. LLMs face issues when it comes to high factual accuracy in summarization tasks, especially in contexts where precision is critical, such as law or medicine. As a result of their restricted training experience in these specialized domains, LLMs sometimes may find it difficult to synthesize highly technical or arcane knowledge into intelligible

explanations. Furthermore, these models may miss important context or details when tasked with distilling lengthy documents into succinct summaries. This could have a significant impact on the accuracy and dependability of the information conveyed, highlighting the necessity of careful oversight in their application to ensure ethical usage.

VIII. Conclusion

In this comprehensive study, we have elucidated the capabilities and limitations of state-of-the-art LLMs in generating summaries tailored to user knowledge levels. Through a rigorous evaluation framework employing diverse metrics and real-world data, valuable insights into the current state of adaptive summarization technology have been gained.

The study reveals that LLMs like Gemini Pro demonstrate promising abilities in adapting summaries to different levels of user familiarity. Additionally, readability metrics such as Flesch–Kincaid, SMOG, and Gunning Fog Index offer useful quantitative measures for assessing summary complexity and tailoring it to specific audiences. However, while LLMs can generate fluent and coherent summaries, challenges persist in accurately capturing nuanced details and adapting to highly specialized domains.

Further research is warranted to improve the interpretability and semantic preservation in LLM-generated summaries. In conclusion, this study serves as a stepping stone toward a future where LLMs can generate summaries that are not only fluent and concise but also adapt seamlessly to the diverse needs and knowledge levels of their users. By addressing the challenges identified in this work, we can unlock the full potential of LLMs for personalized and effective summarization, ultimately democratizing access to knowledge and empowering individuals to learn and understand more effectively.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Author Contributions

Conceptualization: J.T. S.K., and S.G.; data preparation: J.T. and S.G.; writing original draft: J.T. S.K.; supervision: S.G. and B.P.; methodology: J.T, S.G.; validation: S.G. and B.P.; review and editing: B.P., S.G., and A.A.; project administration, B.P.; resources: S.G. and B.P.; funding: B.P. and A.A. All authors have read and agreed to the published version of the manuscript.

Funding

This research was funded by the Centre for Advanced Modelling and Geospatial Information Systems (CAMGIS), Faculty of Engineering and IT, University of Technology Sydney. Moreover, supported by the Researchers Supporting Project, King Saud University, Riyadh, Saudi Arabia, under Project RSP2024 R14.

References

[1] Jin, H., Yang, Z., Meng, D., Wang, J., & Tan, J. (2024). A Comprehensive Survey on Process-Oriented Automatic Text Summarization with Exploration of LLM-Based Methods. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2403.02901

[2] Brown, Tom B. "Language models are few-shot learners." arXiv preprint arXiv:2005.14165 (2020).

[3] Chowdhery, Aakanksha, et al. "PaLM: Scaling language modeling with pathways." arXiv preprint arXiv:2204.02311 (2023).

[4] Widyassari, A. P., Rustad, S., Shidik, G. F., Noersasongko, E., Syukur, A., Affandy, A., & Setiadi, D. R. I. M. (2022). Review of automatic text summarization techniques & methods. *Journal of King Saud University - Computer and Information Sciences*, *34*(4), 1029–1046. https://doi.org/10.1016/j.jksuci.2020.05.006

[5] Zhang, M., Zhou, G., Yu, W., Huang, N., & Liu, W. (2022). A comprehensive survey of abstractive text summarization based on deep learning. *Computational Intelligence and Neuroscience*, *2022*, 1–21. https://doi.org/10.1155/2022/7132226

[6] *A survey of automatic text Summarization: progress, process and challenges*. (2021). IEEE Journals & Magazine | IEEE Xplore. https://ieeexplore.ieee.org/abstract/document/9623462/

[7] Gatt, A., & Krahmer, E. (2018). Survey of the State of the Art in Natural Language Generation: Core tasks, applications and evaluation. *Journal of Artificial Intelligence Research*, *61*, 65–170. https://doi.org/10.1613/jair.5477

[8] Hadi, M. U., Tashi, Q. A., Qureshi, R., Shah, A., Muneer, A., Irfan, M., Zafar, A., Shaikh, M. B., Akhtar, N., Wu, J., & Mirjalili, S. (2023). A survey on

large language models: applications, challenges, limitations, and practical usage. *TechRxiv*. https://doi.org/10.36227/techrxiv.23589741.v1

[9] Pan, J. Z., Razniewski, S., Kalo, J., Singhania, S., Chen, J., Dietze, S., Jabeen, H., Omeliyanenko, J., Zhang, W., Lissandrini, M., Biswas, R., De Melo, G., Bonifati, A., Vakaj, E., Dragoni, M., & Graux, D. (2023). Large language models and knowledge graphs: Opportunities and challenges. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2308.06374

[10] Shanker, S., & King, B. J. (2002). The emergence of a new paradigm in ape language research. *Behavioral and Brain Sciences*, *25*(5), 605–620. https://doi.org/10.1017/s0140525x02000110

[11] Albahri, A. S., Duhaim, A. M., Fadhel, M. A., Alnoor, A., Baqer, N. S., Alzubaidi, L., Albahri, O. S., Alamoodi, A. H., Bai, J., Salhi, A., Santamaría, J., Ouyang, C., Gupta, A., Gu, Y., & Deveci, M. (2023). A systematic review of trustworthy and explainable artificial intelligence in healthcare: Assessment of quality, bias risk, and data fusion. *Information Fusion*, *96*, 156–191. https://doi.org/10.1016/j.inffus.2023.03.008

[12] Yliopisto, O., Juustila, A., Rajanen, D., & Rajanen, D. (2017, March 14). *Cloud computing: migrating to the cloud, Amazon Web Services and Google Cloud Platform*. OuluREPO. https://urn.fi/URN:NBN:fi:oulu-201703151365

[13] Kianian, R., Sun, D., Crowell, E. L., & Tsui, E. (2024). The Use of Large Language Models to Generate Education Materials about Uveitis. *Ophthalmology Retina*, *8*(2), 195–201. https://doi.org/10.1016/j.oret.2023.09.008

[14] Yada, Divakar, et al. "Automatic Text Summarization Methods: A Comprehensive Review" arXiv preprint arXiv:2204.01849 (2022).

[15] Grabeel, K. L., Russomanno, J., Oelschlegel, S., Tester, E., & Heidel, R. E. (2018). Computerized versus hand-scored health literacy tools: a comparison of Simple Measure of Gobbledygook (SMOG) and Flesch-Kincaid in printed patient education materials. *Journal of the Medical Library Association*, *106*(1). https://doi.org/10.5195/jmla.2018.262

[16] Eid, K., Eid, A. A., Wang, D., Raiker, R. S., Chen, S., & Nguyen, J. (2023). Optimizing Ophthalmology patient education via ChatBot-Generated Materials: Readability analysis of AI-Generated Patient Education materials and the American Society of Ophthalmic Plastic and Reconstructive Surgery patient Brochures. *Ophthalmic Plastic and Reconstructive Surgery*. https://doi.org/10.1097/iop.0000000000002549

[17] Hwang YH, Um J, Pradhan B, Choudhury T, Schlüter S. How does ChatGPT evaluate the value of spatial information in the 4th industrial revolution? *Spatial Information Research*. December 2023. doi:10.1007/s41324-023-00567-5

[18] Pal A, Sankarasubbu M. Gemini goes to Med School: exploring the capabilities of multimodal large language models on medical challenge problems & hallucinations. *arXiv (Cornell University)*. February 2024. doi:10.48550/arxiv.2402.07023

[19] Minaee, S., Mikolov, T., Nikzad, N., Chenaghlu, M., Socher, R., Amatriain, X., & Gao, J. (2024). Large Language Models: a survey. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2402.061

[20] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. *arXiv (Cornell University)*. Published online June 12, 2017. doi:10.48550/arxiv.1706.03762

[21] Zhu W, Liu H, Dong Q, et al. Multilingual Machine Translation with Large Language Models: Empirical Results and Analysis. *arXiv (Cornell University)*. Published online April 10, 2023. doi:10.48550/arxiv.2304.04675

[22] Van Veen D, Van Uden C, Blankemeier L, et al. Clinical text summarization: Adapting large language models can outperform human experts. *arXiv (Cornell University)*. Published online September 14, 2023. doi:10.48550/arxiv.2309.07430

[23] Xiao L, Wang L, He H, Jin Y. Modeling Content Importance for Summarization with Pre-trained Language Models. *2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Published online January 1, 2020. doi:10.18653/v1/2020.emnlp-main.293

[24] Bajaj A, Dangati P, Krishna K, et al. Long Document Summarization in a Low Resource Setting using Pretrained Language Models. *arXiv (Cornell University)*. Published online February 28, 2021. doi:10.48550/arxiv.2103.00751

[25] Wang Q, Liu D, Cao Y, et al. Recursively summarizing enables Long-Term dialogue memory in large language models. *arXiv (Cornell University)*. Published online August 29, 2023. doi:10.48550/arxiv.2308.15022

[26] Zhang T, Ladhak F, Durmus E, Liang P, McKeown K, Hashimoto T. Benchmarking large language models for news summarization. *arXiv (Cornell University)*. Published online January 31, 2023. doi:10.48550/arxiv.2301.13848

[27] Eleyan D, Othman A, Eleyan A. Enhancing software comments readability using Flesch Reading Ease Score. *Information*. 2020;11(9):430. doi:10.3390/info11090430

[28] Moncada FM, Pabico JP. On GobbleDyGook and Mood of the Philippine Senate: an exploratory study on the readability and sentiment of selected Philippine senators' microposts. *arXiv (Cornell*

University). Published online August 6, 2015. https://arxiv.org/pdf/1508.01321.pdf

[29] Alawad D, Panta M, Zibran MF, Islam R. An Empirical Study of the Relationships between Code Readability and Software Complexity. *arXiv (Cornell University)*. Published online August 30, 2019. https://arxiv.org/pdf/1909.01760

[30] Sari DC. Measuring Quality of Reading materials in English textbook: The use of lexical density method in assessing complexity of reading materials of Indonesia's Curriculum – 13 (K13) English Textbook Dian Sari. *Journal of Applied Linguistics and Literature*. 2018;1(2):30–39. doi:10.33369/joall.v1i2.4177

[31] Mihalcea R, Corley CD, Strapparava C. Corpus-based and knowledge-based measures of text semantic similarity. *ResearchGate*. Published online January 1, 2006. https://www.researchgate.net/publication/221606405_Corpus-based_and_Knowledge-based_Measures_of_Text_Semantic_Similarity

[32] Shen, T., Jin, R., Huang, Y., Liu, C., Dong, W., Guo, Z. J., Wu, X., Liu, Y., & Xiong, D. (2023). Large Language Model alignment: a survey. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2309.15025

[33] Jin H, Yang Z, Meng D, Wang J, Tan J. A Comprehensive Survey on Process-Oriented Automatic Text Summarization with Exploration of LLM-Based Methods. *arXiv (Cornell University)*. March 2024. doi:10.48550/arxiv.2403.02901

[34] Yadav D, Desai J, Yadav AK. Automatic Text Summarization Methods: A Comprehensive Review. *arXiv (Cornell University)*. March 2022. doi:10.48550/arxiv.2204.01849

[35] https://cloud.google.com/vertex-ai/generative-ai/docs/learn/models [date accessed: 15th February, 2024]

[36] https://www.forbes.com/sites/bernardmarr/2018/05/21/how-much-data-do-we-create-every-day-the-mind-blowing-stats-everyone-should-read/?sh=392017b560ba [Date accessed: 11th January, 2024)

[37] Zaretsky J, Kim JM, Baskharoun S, et al. Generative Artificial Intelligence to Transform Inpatient Discharge Summaries to Patient-Friendly Language and Format. JAMA Netw Open. 2024;7(3):e240357. doi:10.1001/jamanetworkopen.2024.0357

[38] Raja, H., & Lodhi, S. (2024). Assessing the readability and quality of online information on anosmia. *Annals of the Royal College of Surgeons of England*, *106*(2), 178–184. https://doi.org/10.1308/rcsann.2022.0147

[39] Shet, S. S., Murphy, B., Boran, S., & Taylor, C. (2024). Readability of online information for parents concerning Paediatric In-Toeing: An analysis of the most popular online public sources. *Curēus*. https://doi.org/10.7759/cureus.57268

[40] Clavié, B., Ciceu, A., Naylor, F., Soulié, G., & Brightwell, T. (2023). Large Language Models in the workplace: A case study on prompt Engineering for job type classification. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2303.07142