Text Summarization of Patent Documents: Model Selection, Structural Analysis, and Prompt Engineering with ChatGPT

Lavanya Pobbathi[1], Padala Shreya[1], Rahul Varma Muppalla[1], Nguyen Huy[1], Sindhuja Rajidi[1] and Ramisetty Sai Pavan[1]
[1]Dept. of Information Science, Denton, TX, USA
GROUP 04

Abstract

In the era of information overload, the need for efficient summarization of complex texts like patent documents is paramount. Recent advancements in neural network models, particularly in large language models like GPT-3.5 and GPT-4, have shown promise in automated text summarization. However, their effectiveness in summarizing lengthy, intricate texts such as patents remains an area of active research. This study evaluates the performance of various state-of-the-art neural network models, including different configurations of ChatGPT, in summarizing 1630 U.S. patent documents. Our analysis encompasses a range of input configurations, aiming to identify the most effective approaches for summarization tasks. We propose a novel methodology combining extractive and abstractive summarization techniques, guided by meticulous prompt engineering with ChatGPT. The results reveal insights into the structural analysis of patent documents and the optimal model selection for this specific domain. This research contributes to the field by offering a refined strategy for summarizing long and complex texts, demonstrating its efficacy with patent documents as a case study.

Keywords - text summarization, large language model, quality, patent document, prompt engineering

This project was completed by below dedicated team members and their contributions into it and also please find our Github URL for code. [1].

[1]

- Lavanya Pobbathi - Experiments & Research questions
- Rahul Varma, Muppalla - Research Purpose & Methodology
- Sindhuja, Rajidi - Related word & Data Preparation
- Nguyen, Huy - Data Cleaning & Evaluation Metrics
- Padala, Shreya - Prompt Engineering & Selecting Models
- Ramisetty, Mani Venkata Sai Pavan - Data Visualization & Introduction

1. Introduction

In the contemporary landscape of intellectual property, the automated summarization of patent documents represents a frontier in the intersection of legal informatics and artificial intelligence. The sheer volume of patent literature - detailed, technical, and often voluminous - poses a significant challenge for traditional text summarization methodologies. In 2022 alone, the United States Patent and Trademark Office (USPTO) granted over 382,000 patents [1], each a reservoir of innovation necessitating efficient and accurate comprehension. The complexity of these documents, comprising lengthy descriptions, intricate claims, and diverse technical jargon, escalates the need for advanced summarization techniques that can distill essential information without sacrificing context or accuracy.

This project embarks on a rigorous exploration of state-of-the-art text summarization models, including 'h2oGPT-70b', 'h2oGPT-13b', 'zephyr-7b-beta', 'gpt-3.5-turbo', and 'gpt-4', applied to a carefully curated dataset of 1630 U.S. patent documents. The dataset, meticulously cleaned and structured, encompasses patent numbers, abstracts, claims, and descriptions. The project adopts a dual-pronged approach: firstly, an empirical evaluation of these models using ROUGE and BLEU metrics; secondly, an innovative methodology combining extractive and abstractive summarization [2, 3] techniques after rigorous data preprocessing, including tokenization, lemmatization, and the elimination of superfluous elements

Furthermore, the project delves into the realm of prompt engineering, a nuanced aspect of model

- https://github.com/LavanyaPobbathi/Lavanya_INFO5731_Fall2023/tree/main/5731_Final_Project

interaction that has shown potential in refining output quality. This involves generating multiple types of prompts and evaluating the resultant summaries through a customized 'G-Eval' [4] framework and intrinsic evaluation metrics [5]. The aim is to not only assess the current capabilities of these models in handling the intricacies of patent literature but also to advance the field by identifying and addressing the gaps in existing methods.

In 2022, the United States Patent and Trademark Office (USPTO) issued a total of 382,559 patents, comprising 325,445 utility patents, 34,370 design patents, and 1,138 plant patents [6]. The swift comprehension of these patents is critical for activities such as intellectual property management, product innovation, and infringement litigation. A standard U.S. patent document encompasses sections such as application details, an abstract, prior art, and a summary of the invention. This summary, which is the crux of the document, typically spans about 10,000 words and often includes supplemental figures and tables, adding a layer of complexity to the task of automated summarization. Crucially, the claims section, which distills the essence of the invention and generally extends to about 1,000 words, remains within the processing capabilities of many state-of-the-art automated text summarization models. Despite the intricate nature of patent documents, advancements in automated text summarization, particularly through the integration of extractive and abstractive methods, have shown promising results in enhancing the efficiency and efficacy of summarizing such complex legal texts.

Through this comprehensive analysis, the project seeks to answer pivotal research questions that resonate at the core of automated text summarization.

- RQ1: What are the criteria and how to evaluate the quality of patent summarization?
- RQ2: What synergies can be achieved by combining extractive and abstractive summarization methods?
- RQ3: How does prompt engineering affect the quality of analysis? What characterizes the most and least effective prompts in this domain?

These questions guide the project's trajectory, aiming to elevate the standard of patent document summarization and contribute significantly to both the legal and AI communities. By addressing these questions, the project aspires to not only benchmark the current state of the art but also to forge pathways for future advancements in the automated summarization of complex, long-form documents.

2. Related Work

In the realm of automated text summarization, a task pivotal to natural language understanding, advancements in machine learning models have catalyzed significant progress [2, 3]. Our project intersects with these advancements by focusing on the summarization of patent documents—a domain replete with intricate terminologies and complex structures. The project leverages and builds upon various established and emergent methodologies to refine the summarization process.

- **Modeling Techniques and Challenges:** We have integrated Transformer-based models like BART and PEGASUS and addressed their limitations in processing lengthy texts by adopting models capable of managing extensive content. The 'h2oGPT-70b', 'h2oGPT-13b'[7], and 'zephyr-7b-beta' models represent an evolution from traditional RNNs, equipped with mechanisms to process the substantial length and complexity characteristic of patent documents. These models' ability to handle longer input sequences aligns with our project's goal to effectively summarize detailed patent texts.

- **Extractive [8] and Abstractive Summarization [9]:** Our project employs both extractive and abstractive summarization techniques to enhance the representation of patent documents. Extractive summarization, which selects critical information directly from the text, is complemented by abstractive summarization's ability to paraphrase and condense content creatively. This dual approach mirrors the efforts of Hsu et al. (2018) [7], who harmonized these techniques through multi-level attention mechanisms, an approach that resonates with our method of creating concise yet comprehensive patent summaries.

- **Hierarchical Summarization:** Recognizing the complex structure of patents, our project draws inspiration from hierarchical summarization strategies. This approach, exemplified by the BART-Facet model's success in decomposing

research papers, is adeptly suited to patents' multi-faceted nature. We analyze the structural components of patents, endeavoring to create summaries that reflect their multi-dimensional content, thereby providing a more holistic understanding of the invention's scope and application.

- **Prompt Engineering and Model Evaluation:** Prompt engineering[10] emerges as a critical component in our project, exploring the potential of language models such as 'gpt-3.5-turbo' and 'gpt-4' to produce high-quality summaries. Through the development and assessment of various prompts, we determine the most effective strategies for eliciting detailed and relevant content from these models. Additionally, our use of nuanced metrics, including a customized 'G-Eval' and intrinsic evaluation scores, caters to the specific evaluation needs of complex patent documents, ensuring a thorough and rigorous assessment of the summarization models' performance.

By integrating these various strands of research, our project not only builds upon the foundation laid by previous studies but also expands the scope of innovation in patent document summarization. The comprehensive evaluation of leading-edge models, along with a deep dive into prompt engineering, underscores our unique contribution to the field. This multifaceted approach ensures that the resulting summaries are not merely aggregations of information but are intelligently crafted syntheses that meet the nuanced demands of legal professionals, researchers, and patent creators.

3. Methodology

We evaluate summarization models including 'h2oGPT-70b', 'h2oGPT-13b', 'zephyr-7b-beta', 'gpt-3.5-turbo', and 'gpt-4', using a corpus of 1630 patent documents. We used both semantic and linguistic criteria such as 'G-Eval' framework and intrinsic evaluation metrics Score to assess the generated summaries. Statistical analysis is used to compare model performance across metrics to determine their effectiveness in summarizing patents. Figure 1 shows the procedure of the experiment study of our project.

The research process depicted in the image outlines a comprehensive framework for evaluating patent summarization quality, starting with the preparation and preprocessing of data. A corpus of 1630 patent documents is converted into a structured JSON format, further broken down into key components such as abstracts and claims for input data. The preprocessing phase includes the removal of punctuation and stopwords, and the division of text into manageable chunks, leading to a repository of clean data ready for processing.

The next phase employs several state-of-the-art models like GPT-4, GPT-3.5, h2oGPT-70b, h2oGPT-13b, and zephyr-7b-beta, to generate summaries of the patent documents. These models utilize both extractive and abstractive summarization methods, which could synergistically improve the efficiency and quality of the summaries by combining the precise extraction of key phrases with the generation of cohesive and novel sentences that encapsulate the patents' essence. Furthermore, the process incorporates prompt engineering, a technique that involves crafting and refining prompts to guide the AI models in generating more accurate summaries. This step raises questions about the effectiveness of different prompt strategies and their impact on the quality of the summarization, aiming to identify the attributes that distinguish the most effective prompts in the domain of patent summarization. The research seeks to optimize these prompts to improve the relevance and informativeness of the generated summaries, thereby enhancing the practical utility of the summarization models for various stakeholders in the patent ecosystem.

3.1. Data Preparation

The dataset for this study consists of a corpus of 1630 patent documents collected through web scraping of Google Patents at `https://patents.google.com`. We developed a Python program to extract the full text of each patent and store it into a JSON file. The dataset for our project comprises 1630 U.S. patent documents, meticulously sourced through web scraping techniques. The scraping process was focused on extracting comprehensive details from each document, encompassing patent numbers, abstracts, claims, and descriptions. This data was systematically compiled into an Excel spreadsheet, ensuring an organized and accessible format for further analysis.

Initial examination of the raw text from these patents revealed the presence of various special charac-

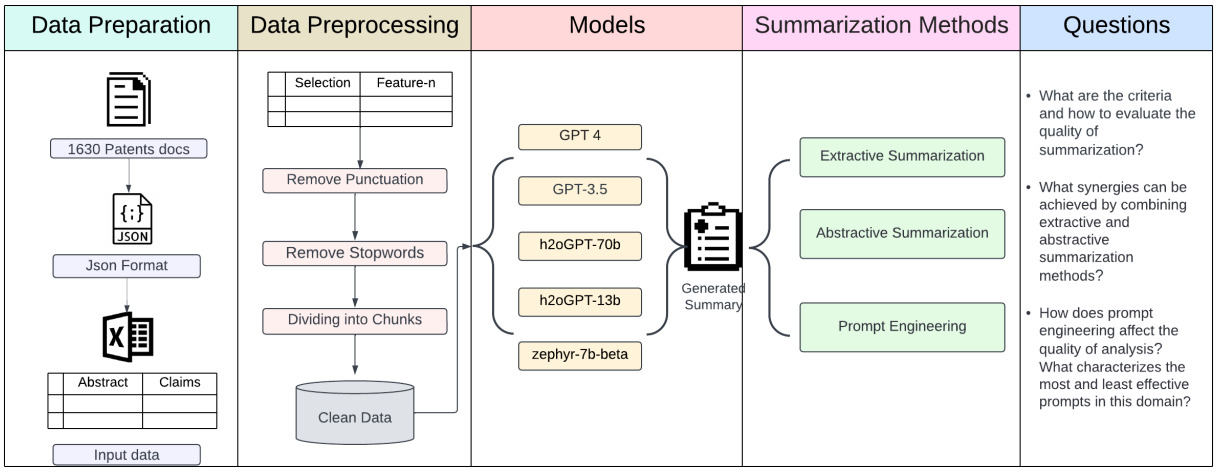

Figure 1: Overview of our project of text summarization of Patent Documents

ters, such as hashtags, dollar signs, and others, which could potentially interfere with the accuracy of the summarization models. To address this, a rigorous cleaning process was employed. The first step involved the removal of these special characters to prevent any confusion during the model training and summarization processes. This was followed by the elimination of extraneous whitespaces and newline characters, which are often overlooked but can significantly impact the consistency of data.

Further refinement of the dataset was achieved through sophisticated text processing techniques implemented in Python. Each patent document, comprising abstracts, claims, and descriptions, was subjected to a series of cleaning steps to enhance the quality of the data for text summarization. These steps included the removal of special characters and punctuations, stripping away numbers, converting all text to lowercase for uniformity, and tokenizing the text. Tokenization was followed by the removal of stopwords, using NLTK's extensive stopwords list, to eliminate common words that add little semantic value to the summaries. Additionally, each word underwent stemming and lemmatization using NLTK's PorterStemmer and WordNetLemmatizer, respectively. These processes reduce words to their root forms, aiding in the consistency and effectiveness of the summarization models.

This comprehensive data preparation strategy, leveraging both web scraping for data collection and advanced natural language processing techniques for data cleaning, sets a solid foundation for the accurate and efficient analysis of patent documents using text summarization models.

3.2. Summarization Models

In this section, we describe selected SOTA neural network models that have shown significant promise in text summarization. The models studied in this research include 'h2oGPT-70b', 'h2oGPT-13b', 'zephyr-7b-beta', 'gpt-3.5-turbo' and 'gpt-4'. These models have been utilized across various applications including text summarization. These models are different in their architectures, training data, number of parameters, and application domains.

- **H2ogpt-4096-llama2-70b-chat]:** The h2oGPT 70b [7] model represents a significant leap in the field of natural language processing. With 70 billion parameters, this model is designed for exceptional performance in understanding and generating human-like text. The larger parameter count allows for deeper and more nuanced understanding of complex language structures, making it ideal for tasks requiring a high level of language comprehension and generation. Its architecture is optimized to handle extensive context lengths, enabling it to maintain coherence over longer conversations or documents. The integration of LLaMA (Large Language Model – Advanced) technology enhances its capabilities in delivering sophisticated, context-aware responses.
- **H2ogpt-4096-llama2-13b-chat:** The h2oGPT 13b model [7], although smaller in scale compared to its 70b counterpart, is still a powerful tool in the

realm of AI-driven language models. With 13 billion parameters, it offers a balance between computational efficiency and language processing capability. This model is designed to provide reliable performance for a wide range of text generation and comprehension tasks, making it a versatile choice for applications where a balance between performance and resource usage is critical. It inherits the advanced features of the LLaMA architecture, ensuring high-quality language processing.

- **HuggingFaceH4/zephyr-7b-beta]:** The zephyr-7b-beta model [11] from HuggingFaceH4 is an innovative addition to the h2oGPT family, boasting 7 billion parameters. This model stands out for its balance between performance and efficiency, making it suitable for tasks that require high-quality language understanding without the computational load of larger models. The beta tag indicates ongoing development and optimization, with a focus on enhancing its language generation and understanding capabilities. Its moderate size makes it an ideal candidate for environments where computational resources are a consideration, without significantly compromising on the quality of output.

- **Gpt-3.5-turbo:** GPT-3.5-turbo [12], a variant of the well-known GPT-3 model by OpenAI, represents an optimized version of its predecessor. The "turbo" in its name signifies enhancements in processing speed and efficiency, allowing for faster response times without sacrificing the quality of language generation. This model is adept at understanding context and generating coherent, contextually relevant text. It excels in a variety of language tasks, including conversation, content creation, and complex problem-solving, making it a highly versatile tool in the AI language model arsenal.

- **GPT-4:** GPT-4 [13] is the latest iteration in the GPT series from OpenAI, building upon the successes of its predecessors. This model features even more sophisticated algorithms and a larger parameter count, pushing the boundaries of language model capabilities. It is designed to provide deeper understanding, more nuanced language generation, and improved context handling. GPT-4's advanced features make it particularly effective in tasks requiring complex language understanding, creative content generation, and detailed problem-solving. Its ability to generate highly coherent and contextually accurate text across a wide range of subjects sets a new standard in natural language processing.

Each of these models represents a unique combination of scale, efficiency, and capability, tailored to different needs in the domain of natural language processing and generation. Their applications range from simple conversational interfaces to complex analytical tasks, demonstrating the versatility and power of modern AI language models.

3.3. Evaluation Metrics

Many metrics has been proposed for text summarization evaluation. However, we don't have any human reference summaries. So, we selected and briefly introduce several recently used evaluation metrics which are appropriate for this study, including Intrinsic Evaluation [5]and G-Eval [4]customized evaluation metrics .

3.3.1. Intrinsic Evaluation Scores:.

- **Fluency:** Fluency pertains to the linguistic quality of the summary, including grammar, spelling, punctuation, word choice, and sentence structure. A fluent summary reads naturally and smoothly, free from language errors that could hinder comprehension. This metric evaluates how well the summarization model generates text that aligns with standard language conventions..

- **Coherence:** Coherence measures how logically and seamlessly the ideas in the summary are connected. A coherent summary presents information in a structured and logical sequence, making it easy for the reader to follow the flow of ideas. This metric assesses the summarization model's ability to maintain a logical progression of thoughts and concepts.

- **Informativeness:** Informativeness evaluates the extent to which the summary captures the essential information from the source text. An informative summary should encapsulate all critical points and details, providing a comprehensive overview of the original document's main ideas.

- **Abstraction:** Abstraction in summarization refers to the model's ability to generate paraphrased or rephrased content, rather than just extracting sentences verbatim from the source. This metric assesses the creativity and rephrasing capability of the model, crucial for abstractive summarization.

- **Consistency:** Consistency is about the factual alignment between the summary and the source

document. A consistent summary accurately reflects the information presented in the original text, without introducing errors or 'hallucinated' facts.

3.3.2. G-Eval Customized Evaluation Metrics.

- **Relevance (G-Eval):** Relevance in G-Eval measures how well the summary captures the most important content from the source document. It involves assessing the summary for the inclusion of key information and penalizing for any redundancies or superfluous details. The goal is to ensure that the summary is focused and devoid of irrelevant content.

- **Coherence (G-Eval):** Coherence in the G-Eval framework aligns with ensuring that the summary is well-structured and organized, presenting information cohesively. This metric assesses whether the summary builds upon its sentences to form a coherent body of information about the topic, mirroring the logical order and main points of the Google patent.

- **Consistency (G-Eval):** Consistency in G-Eval focuses on the factual accuracy and alignment between the summary and the Google patent. It involves checking for any factual discrepancies or errors in the summary, ensuring that all statements in the summary are supported by the original document.

- **Fluency (G-Eval):** Fluency in the G-Eval framework assesses the grammatical and linguistic quality of the summary. This metric evaluates the summary for grammatical correctness, appropriate word choice, and overall readability, ensuring that the summary is not only accurate but also well-written and clear.

Each of these metrics plays a vital role in the comprehensive evaluation of text summarization models. They collectively provide a multi-dimensional assessment of the summarization quality, covering aspects from factual accuracy to linguistic excellence. This robust evaluation framework is essential for accurately gauging the performance of sophisticated summarization models like the ones used in this project.

4. Evaluation Results

4.1. Criteria and Evaluation of Quality Improvement

Based on the Table 3 and by including the study [14] are evaluating the quality of patent summarization, the criteria must encompass a set of rigorous metrics that reflect the complexity and detailed nature of patent documents. The evaluation report provides a comprehensive overview of these metrics, assessing models such as 'h2oGPT 70b', 'h2oGPT 13b', 'zephyr-7b-beta', 'gpt-3.5-turbo', and 'gpt-4' across various dimensions.

Fluency is a cornerstone criterion, addressing the grammatical and linguistic smoothness of the generated summary. A score of 10 across all models indicates that each model produces text with excellent grammar and syntax, contributing to an overall natural and smooth reading experience. Coherence measures the logical structuring and transition of ideas within the summary. Models scoring 10 demonstrate an exceptional ability to present ideas in a clear, logical sequence, which is paramount in summarizing the complex information presented in patents. Scores of 9.5 suggest room for minor improvements in ensuring the logical flow matches the high standards required for legal documents.Informativeness assesses how effectively the summary captures the crucial information from the patent. The high scores in this category reflect the models' capabilities to identify and retain the essential points of the patent documents, providing comprehensive coverage of the content. Abstraction gauges the degree to which the summary can distill the source material into a concise and original form. The scores indicate a variation in performance, with 'h2oGPT 70b' and 'h2oGPT 13b' achieving a slightly higher degree of paraphrasing and synthesis compared to the other models. Consistency is crucial in maintaining the factual integrity of the summary. The perfect scores across all models suggest they are highly reliable in producing summaries that are factually aligned with the original patent texts, a non-negotiable aspect of legal document summarization.

4.2. Extractive and Abstractive Summarization

To address the second research question regarding the synergies achievable by combining extractive and abstractive summarization methods [15], let's focus on three pivotal visualizations: the distribution of semantic

Table 1: Evaluation scores of all models on quality of patent summarization

	H2oGPT_70b	H2oGPT_13b	zephyr7bbeta	GPT-3.5-turbo	GPT-4
Fluency	10	10	10	10	10
Coherence	10	9.5	9.5	9.5	9.5
Informativeness	10	9.5	9.5	9.5	9.5
Abstraction	7.5	7.5	5.5	5.5	5.5
Consistency	10	10	10	10	10

similarity scores, keyword coverage comparison, and the summary length versus readability score.

- **Distribution of Semantic Similarity Scores2:** The histogram showing semantic similarity scores indicates the performance of the combined summarization methods against the source text's semantic content. High similarity scores suggest that the synthesized summaries manage to retain the core semantic meaning of the original patents. This is indicative of an effective synergy, where the abstractive method's ability to paraphrase and the extractive method's direct use of source text ensure that the essence of the original content is preserved.

- **Keyword Coverage Comparison3:** The boxplot comparison of keyword coverage demonstrates another dimension of synergy. Extractive summaries are typically more reliable in covering key terms verbatim. Abstractive summaries, while sometimes missing exact terms, often capture their meaning. The combined approach increases the likelihood that all relevant concepts are included in the summaries, improving the comprehensiveness. This suggests that the combined approach mitigates the limitations inherent in each individual method, leveraging the precision of extractive summarization with the broader contextual understanding of abstractive summarization.

- **Summary Length vs. Readability Score4:** The scatter plot relating summary length to readability scores presents a nuanced view of the effectiveness of the combined methods. It shows that a synergy between extractive and abstractive methods can achieve a balance between brevity and clarity. The ideal summary is not only shorter than the extensive original patent text but also scores high in readability, making the information more accessible and digestible for the end-user..

4.3. Prompt Engineering

Analyzing the Figure 5 and 6, the impact of prompt engineering [16] on the quality of analysis in text summarization, particularly in the context of patent documents using GPT-3.5, reveals several insightful findings. The effectiveness of a prompt is significantly influenced by its clarity, specificity, and alignment with the intended task. This is evident from the experimental results where different prompts elicited varying levels of coherence, consistency, fluency, and relevance in the summaries generated by GPT-3.5.

From the provided prompts, it's clear that prompts explicitly crafted to mimic a professional context (like those of a patent attorney) and those that clearly define the scope of the summary (combining details from the description, claims, and abstract) tend to produce more coherent and relevant outputs. For instance, Prompt 1 and Prompt 3, which explicitly mention acting as a patent lawyer and ask for a synthesis of key information, led to higher coherence and relevance scores in most cases. This suggests that prompts with a clear, professional tone and specific instructions align well with the summarization of complex, technical documents like patents.

On the other hand, less effective prompts are typically those that are more general or lack specific guidance on the expected output. For example, Prompt 4, which provides a more general request for a comprehensive overview, received slightly lower scores in coherence and relevance in some experiments. This indicates that a lack of specificity can lead to summaries that are less focused and potentially overlook crucial details of the patent text.

The study [17] emphasizes the importance of providing clear information such as code intention, task, and dataset characteristics in the prompts. It demonstrates that enhanced prompts, which incorporate specific details like the nature of the problem, the dataset used, and the repair objective, significantly improve the model's performance in tasks like fault

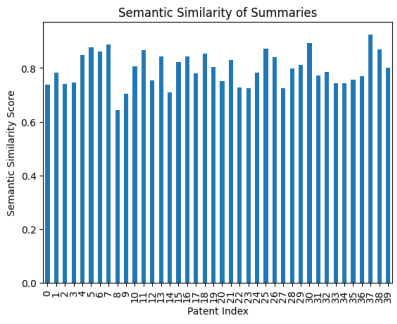

Figure 2: Semantic Similarity of Summaries and Scores

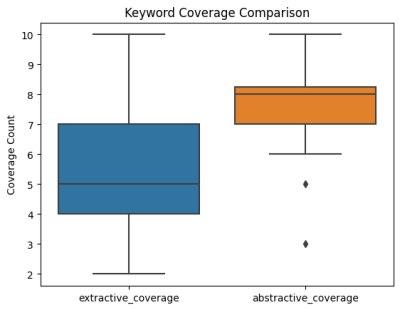

Figure 3: Keyword Coverage comparison between extractive and abstractive summarization

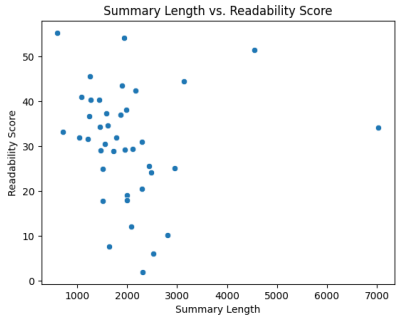

Figure 4: summary length vs readability score between extractive and abstractive summarization

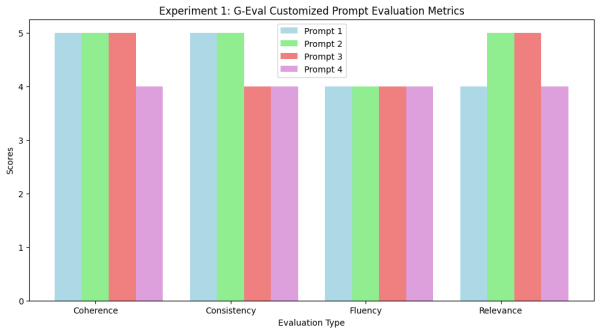

Figure 5: G-Eval Customized prompt evaluation Metrics on Experiment 1

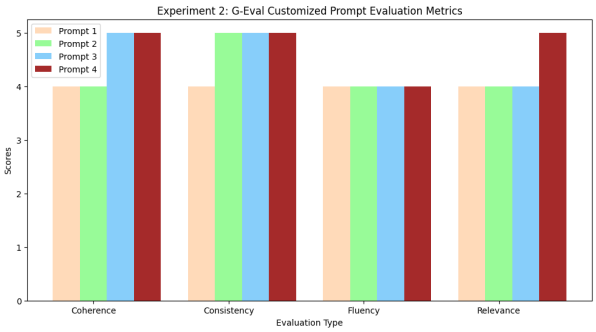

Figure 6: G-Eval Customized prompt evaluation Metrics on Experiment 2

Table 2: G-Eval Customized prompt scores of First and Second Summaries by using GPT3.5

Evaluation Type	GPT3.5_S1_P1	GPT3.5_S1_P2	GPT3.5_S1_P3	GPT3.5_S1_P4
Coherence	5	5	5	4
Consistency	5	5	4	4
Fluency	4	4	4	4
Relevance	4	5	5	4
	GPT3.5_S2_P1	GPT3.5_S2_P2	GPT3.5_S2_P3	GPT3.5_S2_P4
Coherence	4	4	5	5
Consistency	4	5	5	5
Fluency	4	4	4	4
Relevance	4	4	4	5

*S1_P1 = First Summary of Prompt1, S1_P2 = First Summary of Prompt2, S1_P3 = First Summary of Prompt3 and S1_P4 = First Summary of Prompt4. *S2_P1 = Second Summary of Prompt1, S2_P2 = Second Summary of Prompt2, S2_P3 = Second Summary of Prompt3 and S2_P4 = Second Summary of Prompt4.

detection, localization, and repair. This aligns with the observed trend in the experiment, where prompts with clearer, more directed instructions led to better-performing summaries.

In conclusion, the evaluation results highlight the exceptional capabilities of the advanced language models in summarizing patents. They provide insightful benchmarks for the critical attributes required for high-quality patent summaries—fluency, coherence, informativeness, abstraction, and consistency. These metrics collectively form a robust framework for evaluating patent summarization quality, ensuring the summaries serve the practical needs of stakeholders in the patent

Table 3: Intrinsic Evaluation Scores for First and Second Summaries

Evaluation Type	GPT3.5_S1_P1	GPT3.5_S1_P2	GPT3.5_S1_P3	GPT3.5_S1_P4
Fluency	8.5	8.5	9.5	8.5
Coherence	8.5	8.5	8.5	8.5
Informativeness	9.5	9.5	9.5	8
Abstraction	8.5	8.5	8.5	8.5
Consistency	8.5	8.5	8.5	8.5
	GPT3.5_S2_P1	GPT3.5_S2_P2	GPT3.5_S2_P3	GPT3.5_S2_P4
Fluency	10	10	10	10
Coherence	10	10	10	10
Informativeness	10	10	10	10
Abstraction	5.5	5.5	5.5	7.5
Consistency	10	10	10	10

*S1_P1 = First Summary of Prompt1, S1_P2 = First Summary of Prompt2, S1_P3 = First Summary of Prompt3 and S1_P4 = First Summary of Prompt4. *S2_P1 = Second Summary of Prompt1, S2_P2 = Second Summary of Prompt2, S2_P3 = Second Summary of Prompt3 and S2_P4 = Second Summary of Prompt4.

ecosystem effectively. On Other hand, the visualization analyses of confirm that combining extractive and abstractive summarization methods leverages their respective strengths, leading to patent summaries that are semantically aligned with the original text, comprehensive in covering critical content, and optimized for readability. This combined approach produces a synergy that neither method could achieve alone, effectively balancing detail and conciseness while ensuring the summaries remain informative and user-friendly. The synergy is particularly evident in the ability to produce summaries that are content-rich yet succinct enough to serve the practical needs of those utilizing patent information, such as researchers, legal professionals, and innovators. Finally, effective prompts in the domain of patent document summarization are characterized by their specificity, professional tone, and clear instructions that align with the task's complexity. These types of prompts enable GPT-3.5 to generate summaries that are more coherent, consistent, fluent, and relevant. Conversely, prompts that lack these characteristics tend to produce less effective outcomes. Thus, prompt engineering plays a crucial role in guiding the model to generate high-quality, task-aligned summaries.

4.4. Selection of Models:

- **Diversity in Model Architecture and Scale:** Through selection spans a range of model architectures and sizes, from 'h2oGPT-70b', a larger-scale model, to 'zephyr-7b-beta', a more moderately sized model. This diversity allows for a comparative analysis of how different scales and architectures influence the quality of text summarization. Larger models, like 'h2oGPT-70b', are typically expected to perform better in complex tasks due to their extensive training data and sophisticated architectures. However, smaller models like 'zephyr-7b-beta' might offer efficiency advantages or unique perspectives.

- **Inclusion of Cutting-Edge Models:** By including models like 'gpt-3.5-turbo' and 'gpt-4', your study incorporates some of the most advanced AI technologies available. These models are at the forefront of natural language processing and are known for their ability to generate coherent, contextually relevant text, which is crucial for effective summarization.

- **Exploring Variability:**The choice of different models helps in understanding the variability in summarization tasks. It provides insights into how different models handle the complexities of summarizing patent documents, which often contain highly technical and specialized language.

4.5. Selection of Evaluation Metrics:

- **G-Eval Customized Evaluation Prompt:** In the absence of human references, the G-Eval customized prompt provides a systematic way to assess key qualities of summarization such as coherence, consistency, fluency, and relevance. These criteria are crucial for evaluating the effectiveness of summaries in representing the original text accurately and understandably.

- **Intrinsic Evaluation:** This approach evaluates the summaries based on inherent qualities like fluency, coherence, informativeness, abstraction, and con-

sistency. It's a valuable method when external benchmarks or human evaluations are not available. The intrinsic evaluation focuses on the text's internal characteristics, offering a direct measure of the summarization quality based on the text itself rather than external references.

4.6. Justification for the Absence of Human References:

- **Resource and Time Constraints:** Human evaluation, while valuable, is often resource-intensive and time-consuming. It requires recruiting evaluators with the right expertise, especially for technical texts like patents, and can be challenging to scale for large datasets.
- **Objective and Reproducible Measures:** TAutomated metrics provide a more objective and reproducible way of evaluation compared to human judgment, which can be subjective and vary significantly between individuals.

5. Summary and Future Work

In our comprehensive study, we have scrutinized the capabilities of state-of-the-art summarization models such as 'h2oGPT-70b', 'h2oGPT-13b', 'zephyr-7b-beta', 'gpt-3.5-turbo', and 'gpt-4', focusing on their efficacy in distilling lengthy and complex patent documents. The evaluation, grounded on a robust framework comprising fluency, coherence, informativeness, abstraction, and consistency, has revealed that these models exhibit exceptional performance, particularly in fluency and consistency. This signifies their adeptness at producing summaries that are both linguistically polished and factually faithful to the source documents. However, the results also illuminate a need for improvement in abstraction, indicating a potential area for future enhancement.

Extractive and abstractive summarization methods, when applied in tandem, have shown to form a synergistic alliance that effectively captures the semantic core and critical content of patents, while maintaining high readability for diverse user groups. Prompt engineering has emerged as a pivotal factor in enhancing the quality of summaries, with specifically tailored prompts yielding more coherent and pertinent outputs. The project's findings underscore the significance of precise prompts that resonate with the sophisticated nature of patent texts.

Moving forward, the pursuit of automating the prompting process stands as a promising avenue for research. The aim will be to refine the interaction with large language models like GPT-3.5 to streamline the summarization process further. Developing a system that can intelligently craft prompts suited to the specific demands of patent documents could vastly improve the efficiency and reliability of automated patent summarization. This direction not only holds potential for academic and research applications but also promises substantial benefits for legal practitioners and innovators navigating the patent landscape.

A. Appendix A: Summary of GPT 4

This patent describes a method for encoding a video stream that contains a transparency information channel. The method involves applying a first encoding strategy, and if the encoded bitrate falls within a predetermined target bitrate, this first encoding is chosen. If not, a second encoding is applied, and if this falls within the target bitrate, it is used. If neither the first nor the second encoding strategies result in the target bitrate, a third encoding strategy is used. There are 11 claims in the patent. The primary claim involves a method for encoding a video stream with a transparency information channel. The method involves applying a first encoding strategy, and selecting it if the encoded bitrate falls within a target bitrate. If not, a second encoding strategy is applied and selected if it falls within the target bitrate. If neither first nor second encoding strategies meet the criteria, a third encoding strategy is used. Additional claims detail the specific encoding strategies used, including lossless vector graphics encoding, mathematical representation encoding, and MPEG-based encoding. The patent includes three figures. The first figure is a flowchart detailing the method of encoding based on the different encoding strategies. The second figure presents a schematic illustration of an encoding apparatus that can perform the method. The third figure presents a schematic illustration of a decoding apparatus that can decode the video stream encoded by the method.The patent details the need for high-efficiency coding for the alpha (transparency) channel of a moving picture. It suggests that current methods of encoding overlook the impact of the alpha channel, leading to inefficient coding. The patent describes a method of hybrid encoding for the alpha channel, using different encoding strategies to improve the compression of the alpha component. The patent also details the hardware and software needed to perform the encoding and decoding methods. The Claims section is the most critical as it defines the scope of the patent's protection. The claims detail the encoding method step-by-step, and outline the criteria for selecting between the different encoding strategies. The specificity of these claims is important for protecting the patent from infringement.

B. Appendix B: Summary of h2ogpt-4096-llama2-70b-chat

The present invention relates to a method and apparatus for hybrid encoding of an alpha channel for video. The in-

vention describes a system that uses a combination of vector graphics, mathematical representation, and MPEG-based encoding to efficiently compress and transmit alpha channels for video. The system is particularly useful for multiview video coding and can be applied to both the alpha channel and depth information. The claims section describes the various embodiments of the invention, including the use of a hybrid encoder for encoding the alpha channel, the use of different encoding schemes for different parts of the alpha channel, and the use of a detector to detect the encoding applied to specific portions of the alpha channel. The claims also cover the decoding apparatus and method, which includes a plurality of decoders for decoding the different portions of the alpha channel. The figures section includes a diagram of the encoding apparatus and a diagram of the decoding apparatus. The encoding apparatus includes a hybrid encoder, a detector, and a plurality of encoders for encoding the alpha channel using different schemes. The decoding apparatus includes a plurality of decoders for decoding the different portions of the alpha channel. The description section provides a detailed explanation of the invention, including the problem it solves, the advantages of the hybrid encoding scheme, and the various embodiments of the invention. The section also describes the different encoding schemes used in the invention, including vector graphics, mathematical representation, and MPEG-based encoding. Additionally, the section covers the decoding method and apparatus, including the use of a detector to detect the encoding applied to specific portions of the alpha channel. The most critical section of the patent is the Claims section, as it defines the scope of the invention and the protection sought by the inventor. The claims clearly outline the novel aspects of the invention, including the use of a hybrid encoder, the use of different encoding schemes for different parts of the alpha channel, and the use of a detector to detect the encoding applied to specific portions of the alpha channel. In summary, the patent describes a novel method and apparatus for hybrid encoding of an alpha channel for video, which combines the advantages of different encoding schemes to efficiently compress and transmit alpha channels for video. The invention is particularly useful for multiview video coding and can be applied to both the alpha channel and depth information. The claims section defines the scope of the invention and the protection sought by the inventor, making it the most critical section of the patent.

References

[1] D. Curcic, "Number of academic papers published per year," access on July 25, 2023. [Online]. Available: https://wordsrated.com/number-of-academic-papers-published-per-year/

[2] D. Bahdanau, K. H. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *3rd International Conference on Learning Representations, ICLR 2015*, 2015.

[3] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[4] C.-H. Chiang and H. yi Lee, "A closer look into automatic evaluation using large language models," 2023.

[5] F. Giner, "An intrinsic framework of information retrieval evaluation measures," 2023.

[6] "2022 u.s. patent filings statistics, by mckee voorhees & sease plc," access on July 25, 2023. [Online]. Available: https://www.lexology.com/library/detail.aspx?g=1170d66d-63b8-4901-b819-e88c67916a2f

[7] A. Candel, J. McKinney, P. Singer, P. Pfeiffer, M. Jeblick, P. Prabhu, J. Gambera, M. Landry, S. Bansal, R. Chesler, C. M. Lee, M. V. Conde, P. Stetsenko, O. Grellier, and S. Ambati, "h2ogpt: Democratizing large language models," 2023.

[8] N. Mishra, G. Sahu, I. Calixto, A. Abu-Hanna, and I. H. Laradji, "Llm aided semi-supervision for extractive dialog summarization," 2023.

[9] Z. Boukhers, O. Colas, and J.-G. Ganascia, "Abstractive text summarization techniques for patent title generation," in *Canadian Conference on Artificial Intelligence*. Springer, 2020, pp. 419–431.

[10] J. Shin, C. Tang, T. Mohati, M. Nayebi, S. Wang, and H. Hemmati, "Prompt engineering or fine tuning: An empirical assessment of large language models in automated software engineering tasks," 2023.

[11] L. Tunstall, E. Beeching, N. Lambert, N. Rajani, K. Rasul, Y. Belkada, S. Huang, L. von Werra, C. Fourrier, N. Habib, N. Sarrazin, O. Sanseviero, A. M. Rush, and T. Wolf, "Zephyr: Direct distillation of lm alignment," 2023.

[12] J. Ye, X. Chen, N. Xu, C. Zu, Z. Shao, S. Liu, Y. Cui, Z. Zhou, C. Gong, Y. Shen, J. Zhou, S. Chen, T. Gui, Q. Zhang, and X. Huang, "A comprehensive capability analysis of gpt-3 and gpt-3.5 series models," 2023.

[13] OpenAI, "Gpt-4 technical report," 2023.

[14] S. Casola and A. Lavelli, "Summarization, simplification, and generation: The case of patents," *Expert Systems with Applications*, vol. 205, p. 117627, Nov. 2022. [Online]. Available: http://dx.doi.org/10.1016/j.eswa.2022.117627

[15] V. Tretyak and D. Stepanov, "Combination of abstractive and extractive approaches for summarization of long scientific texts," 2020.

[16] D. van Zandvoort, L. Wiersema, T. Huibers, S. van Dulmen, and S. Brinkkemper, "Enhancing summarization performance through transformer-based prompt engineering in automated medical reporting," 2023.

[17] J. Cao, M. Li, M. Wen, and S. chi Cheung, "A study on prompt design, advantages and limitations of chatgpt for deep learning program repair," 2023.