Universiteit
Leiden
The Netherlands

Bachelor Computer Science & Datascience and Artificial Intelligence

Can AI make financial statements
more accessible to the general public?

Marinus Frank Hoogstrate

First supervisor: Suzan Verberne
Second supervisor: Natalia Amat Lefort

RESEARCH PROPOSAL BACHELOR THESIS

Leiden Institute of Advanced Computer Science (LIACS)
www.liacs.leidenuniv.nl

04/03/2025

Abstract

To make financial documents more accessible to people without a financial background, a summary that does not contain difficult language or financial jargon can be created. This paper evaluates how a large language model performs at creating this summary. This was tested using four prompts: a basic prompt, a question-answer based prompt and a dual prompt that separated the summary creation and text simplification. It found that a dual prompt, which first creating the summary and secondly simplified the summary, gave the best results. These summaries however contained several issues like questions were not answered, the given template of questions was not followed and answers that contained placeholders instead of information about the document itself. The language used in the summaries was also more difficult and contained more financial terms than desired.

Contents

1 Introduction

In order to have a better understanding of a company's policies and impact, it is essential to understand the financial data that the company releases like the annual report. But analysing financial data is a tedious task requiring significant effort, time and financial expertise. There have been many attempts to automatically create summaries of financial data making analysis much quicker. One example is the work by El-Haj et al. [EHAL+20] that compared the performance of different algorithms at summarizing annual reports. These algorithms tried to rank the importance of sentences in the financial reports in various ways and created the summary based on the most important sentences.

With the advancement of the technology of Artificial Intelligence (AI) and Large Language Models (LLMs), a different way to generate summaries is gaining prominence. This technology does not use the extractive method of collecting the sentences from the financial report and combining them to form a summary, but instead uses abstractive summarization by rephrasing the original text. This method can produce more fluent and coherent summaries, closer to those written by humans [DNC24]. These summaries are also shown to be preferred by humans to summaries written by fine-tuned models that use abstractive summarization [GLD23]. Summaries created by an LLM could also avoid the usual financial jargon and instead present the information in a way that is able to be understood by someone without prior experience in finance.

There are various methods of creating a summary using an LLM like multi-LLM [FLK+25] or using different prompting strategies. One of these prompting strategies is Question-answering based. This strategy asks multiple questions about the document and combines the results to create the summary [DNC24]. The benefits of this method over simple prompting (asking the LLM to create a summary) are greater control over what is put into the summary and the formatting of the summary [DNC24]. Drawbacks of this method are that the questions need to fit the document it is trying to gather information from and asking questions that are not answered in the document results in poor answers from the LLM [DNC24]. Therefore the same questions can only be used for documents that are structured in a similar way to each other and contain similar information.

In this thesis, we evaluate the question-answering based prompting strategy for the task of creating summaries of financial reports for people without a financial background. We argue that this is a suitable method because financial reports are structured similarly and the greater control over the output is important for someone who can not distinguish themselves which parts of the data are more important than other. However the performance still has to be evaluated leading to the research question:

How does a large language model perform at creating a summary of financial data that can be interpreted by people without a financial background using a question-answering based prompting strategy?

2 Background

LLMs have been shown to successfully simplify the language of complex documents. For example, Hanhart [Han25] compared multiple LLMs in simplifying Dutch legal texts without sacrificing legal accuracy. These were judged using automatic metrics (like BERTScore) and an LLM-as-a-judge

framework, and were validated by legal and linguistic experts. This showed a clear improvement in the readability of the texts, but sometimes at the cost of preserving the meaning. This capability was also demonstrated by Qiang [QHZ+25], who compared the performance of various LLMs and traditional non-LLMs in text simplification. Their work showed that LLMs perform better at lexical, syntactic, sentence, and document simplification.
Many metrics have been used to judge the quality of a summary; for instance, SummEval [FKM+21] compiled and benchmarked these metrics.
The Flesch-Kincaid test [KFJRC75] can be used to judge the difficulty of the language used, whereas LLM-as-a-judge frameworks have been shown to evaluate both difficulty and jargon. For example, a paper by Buhnila [BSA+26] evaluated how well LLMs perform at medical term simplification, judging the outputs using an LLM on the basis of correctness, completeness, and readability. These results were compared to a human evaluation, and the LLM was found to have comparable performance, aside from being more lenient in the correctness metric.

3 Study design

The data for the research comes from a dataset [ZKG+23] which has full length financial reports and corresponding summaries. First we create a set of general questions by looking at what information is mentioned in the summaries. We then prompt these questions about the financial data to an LLM and compile the output into a document. For each annual report we generate multiple summaries and judge the performance based on the following three criteria.

- Accuracy: The summary will be compared to the existing human-written summaries for accuracy using traditional summarization metrics and the LLM-as-a-Judge technique [LIX+23].
- Readability: The outputted summary will be judged on its readability using the Flesch-Kincaid readability test [KFJRC75].
- Consistency: Do the generated summaries of an annual report have similar scores in accuracy and readability.

3.1 Data

The data comes from the financial narrative summarization 2023 task [The23]. This data contains 1520 annual reports of various companies of UK firms listed on The London Stock Exchange and 3 summaries of this document. The summaries are extracted from the annual reports, they consist primarily of the statements from the chairman or chief executive, shortened to 1000 words if they were longer than 1000 words. The summaries are sorted by length. These summaries were used as a gold standard to compare other summaries to. For the task several groups created algorithms to create summaries of these annual reports and these summaries were compared to the gold standard using a Java ROUGE packet. The results were published in a paper [ZKG+23].

Create a summary containing the most important information about the company and its activities within the last year by using the information contained within the following report: document
CRITICAL RULES:
- The total number of words in the generated document must be around 900 words

Figure 1: Basic prompt

4 Experiments

4.1 Model information

For each annual report we create 3 summaries using the LLM llama 3.2:3b [AI24], and was deployed using the ollama platform. [Oll24] This model takes the prompt and multiple parameters to generate a response. We use the default setting for most parameters, but adapt the following:

- num_predict: Maximum number of tokens to predict when generating text.
- num_ctx: Context length is the maximum number of tokens that the model has access to in memory.
- temperature: Increasing the temperature will make the model answer more creatively.

The first prompt used is a basic prompt asking the LLM to create a summary of the annual report of around 900 words, this word count was chosen because the summaries created for the FNS 2023 task did not exceed 1000 words. The prompt went as follows:

The prompts using guided questions were tasked with creating a summary by answering 7 key questions about the company. We created these questions on the basis of the SWOT principle, a widely used analysis techniques that is performed by looking at the "strengths", "weaknesses", "opportunities" and "threats" a business is facing [GT17]. These questions were supplemented by questions about decisions and performance about the specific year the annual report was written.

- What were the company's key financial performance metrics for the period?
- What strategic decisions or major developments did the company announce or implement during the period, and what are their expected impacts on the business?
- What are the company's main strengths?
- What risks or challenges does the company face, and how is it addressing them?
- What are the company's future growth prospects and expansion plans, including any new markets, products, or services being pursued?
- How will the company's financial performance be affected by changes in market trends, customer behavior, or industry dynamics, and what strategies does it have in place to adapt?

- How stable is the company in terms of its financial health?

To assure that these questions were answered properly, template prompting and reflection pattern prompting techniques [WFH+23] were used to create a prompt. This prompt is shown in Figure 2.

Document to summarize: document
Answer these 7 questions with MAXIMUM detail and depth by first thinking what information would be relevant to answer the question, then answering the question based on that information:
QUESTION 1: What were the company's key financial performance metrics for the period? ANSWER: [Provide a detailed answer to this question]
QUESTION 2: What strategic decisions or major developments did the company announce or implement during the period, and what are their expected impacts on the business? ANSWER: [Provide a detailed answer to this question]
QUESTION 3: What are the company's main strengths? ANSWER: [Provide a detailed answer to this question]
QUESTION 4: What risks or challenges does the company face, and how is it addressing them? ANSWER: [Provide a detailed answer to this question]
QUESTION 5: What are the company's future growth prospects and expansion plans, including any new markets, products, or services being pursued? ANSWER: [Provide a detailed answer to this question]
QUESTION 6: How will the company's financial performance be affected by changes in market trends, customer behavior, or industry dynamics, and what strategies does it have in place to adapt? ANSWER: [Provide a detailed answer to this question]
QUESTION 7: How stable is the company in terms of its financial health? ANSWER: [Provide a detailed answer to this question] CRITICAL RULES: - If information isn't explicit, INFER from available data and context. - Be comprehensive - depth matters more than brevity. - Start response immediately with "QUESTION 1:" - The total number of words in the generated document must be around 900 words.

Figure 2: Professional prompt

The simplified summary was created using a prompt using the same guided questions except also instructing the LLM to use simple, clear language and give answers in a way so that they can be understood by people without a financial background or higher education. This prompt is shown in Figure 3.

CRITICAL: Create a simple summary of this report in a way that could be understood by a someone without a financial background or higher education. This can be done by using simple language avoiding financial jargon and if it is necessary to use financial terms, explain them. Document to summarize: document Answer these 7 questions by first thinking what information would be relevant to answer the question, then answering the question based on that information. QUESTION 1: What were the company's key financial performance metrics for the period? ANSWER: [Provide a answer to this question in simple language]
QUESTION 2: What strategic decisions or major developments did the company announce or implement during the period, and what are their expected impacts on the business? ANSWER: [Provide a answer to this question in simple language]
QUESTION 3: What are the company's main strengths? ANSWER: [Provide a answer to this question in simple language]
QUESTION 4: What risks or challenges does the company face, and how is it addressing them? ANSWER: [Provide a answer to this question in simple language]
QUESTION 5: What are the company's future growth prospects and expansion plans, including any new markets, products, or services being pursued? ANSWER: [Provide a answer to this question in simple language]
QUESTION 6: How will the company's financial performance be affected by changes in market trends, customer behavior, or industry dynamics, and what strategies does it have in place to adapt? ANSWER: [Provide a answer to this question in simple language]
QUESTION 7: How stable is the company in terms of its financial health? ANSWER: [Provide a answer to this question in simple language]
CRITICAL RULES: - If information isn't explicit, INFER from available data and context. - Be comprehensive. - Start response immediately with "QUESTION 1:" - The total number of words in the generated document must be around 900 words.

Figure 3: Simplified Prompt

4.2 Problems identified

After preliminary testing and by reading the generated summaries the following problems with the summaries were identified: 1. The LLM stating that the answer to the question could not be found in the document. An example is shown in Figure 4.

QUESTION 3: What are the company's main strengths?
Unfortunately, I couldn't find any information on the company's main strengths in the provided text.
[...]
QUESTION 5: What are the company's future growth prospects and expansion plans, including any new markets, products, or services being pursued?
The provided text does not mention specific future growth prospects or expansion plans. However, it mentions that the company will review its price agreements periodically and reassess them mutually with GAIL.
QUESTION 6: How will the company's financial performance be affected by changes in market trends, customer behavior, or industry dynamics, and what strategies does it have in place to adapt?
The provided text does not provide information on how the company's financial performance will be affected by changes in market trends, customer behavior, or industry dynamics.

Figure 4: Summary Excerpt showing an unanswered question output

The LLM should be able to find all the answers to the questions in the document, but sometimes refuses to answer a question stating that the text does not provide information, like question 3 in this example. Testing showed that the questions the LLM does not answer, are answered in other runs of the same document or when using a different prompt. To find if there are certain question the LLM refuses to answer frequently or if certain prompts results in higher rates of refusing to answer questions the summaries were analysed.
2. The summary contains a generic template for answers but does not answer questions or summarize for this specific document. An example is shown in Figure 5.

QUESTION 1: What were the company's key financial performance metrics for the period?
The company's main financial performance metric was its loss for the year, which was £[amount]. This means that the company made a profit of [amount] less than its total revenue. The company also had a significant amount of debt, which is money it owes to others.
QUESTION 2: What strategic decisions or major developments did the company announce or implement during the period, and what are their expected impacts on the business?
The company announced that it would be reducing its workforce by [number] employees. This was done to reduce costs and improve efficiency. The company also invested in new technology to improve its operations.
QUESTION 3: What are the company's main strengths?
The company's main strength is its expertise in [specific area]. The company has a long history of providing high-quality services in this area, and it has built strong relationships with its customers.

Figure 5: A summary output showing a template answer

The preliminary testing showed that this issue was found in all prompt types, but did not occur often, the frequency of this issue was also tracked to find a correlation between prompts and this issue.

3. The simplified prompt does not answer the questions using simple language. An example is shown in Figure 6.

QUESTION 6: How will the company's financial performance be affected by changes in market trends, customer behavior, or industry dynamics, and what strategies does it have in place to adapt?
The company's financial performance is likely to be impacted by changes in market trends, customer behavior, or industry dynamics, including:
* Changes in global demand for natural gas, which could impact revenue. * Shifts in government policies or regulations that affect oil and gas exploration and production.
To adapt to these changes, the company is:
* Monitoring market trends and customer behavior closely, and adjusting its strategy accordingly. * Engaging with governments and regulatory bodies to mitigate the impact of policy changes. * Diversifying its funding sources, including exploring alternative debt financing options.
QUESTION 7: How stable is the company in terms of its financial health?
The company's financial stability is strong, with:
* A solid balance sheet, including a significant cash reserve. * A proven track record of profitability and growth. * A diversified revenue stream, with a focus on natural gas production.
These factors position the company well for long-term financial stability and success.

Figure 6: A simplified prompt summary output showing financial jargon

The language used is clearly full of financial terms, and also scores high on the Flesch-Kincaid test. Adjusting the prompt and changing the parameters did not resolve this issue. To minimize this issue the values to assigned to the model parameters were adjusted using the Optuna Tree-structured Parzen Estimator [ASY+19], a Bayesian optimization algorithm that uses a probabilistic model to efficiently tune hyper parameters over the following grid:

- num_predict: 500-8000
- num_ctx: 2000-15000
- temperature: 0.1-0.9

The python module implementation of this algorithm [1] was used on the simplified prompt, with the goal of minimizing the Flesch-Kincaid score and gave the following values:

- num_predict: 4190, default -1 (infinite generation).
- num_ctx: 12035, default 2048.
- temperature: 0.810223634031622, default 0.8.

To make the LLM focus on just simplifying the language a fourth summary was generated using the professional summary as a basis and having the LLM simplify the answer in that summary. This prompt is shown in Figure 7.

[1] https://optuna.readthedocs.io/en/stable/reference/samplers/generated/optuna.samplers.TPESampler.html

CRITICAL: Reformulate the answers in simple language and in a way so that they could be understood by a someone without a financial background or higher education. This can be done by using simple language avoiding financial jargon and if it is necessary to use financial terms, explain them. Summary to simplify: document Keep the formatting of the questions and answers the same as in the original summary, but simplify the language used in the answers. The formatting is as follows: QUESTION 1: What were the company's key financial performance metrics for the period? ANSWER: [Provide a answer to this question in simple language]
QUESTION 2: What strategic decisions or major developments did the company announce or implement during the period, and what are their expected impacts on the business? ANSWER: [Provide a answer to this question in simple language]
QUESTION 3: What are the company's main strengths? ANSWER: [Provide a answer to this question in simple language]
QUESTION 4: What risks or challenges does the company face, and how is it addressing them? ANSWER: [Provide a answer to this question in simple language]
QUESTION 5: What are the company's future growth prospects and expansion plans, including any new markets, products, or services being pursued? ANSWER: [Provide a answer to this question in simple language]
QUESTION 6: How will the company's financial performance be affected by changes in market trends, customer behavior, or industry dynamics, and what strategies does it have in place to adapt? ANSWER: [Provide a answer to this question in simple language]
QUESTION 7: How stable is the company in terms of its financial health? ANSWER: [Provide a answer to this question in simple language]
CRITICAL RULES: - Simplify the language used. - Avoid financial jargon, and if it is necessary to use financial terms, explain them. - Be comprehensive. - Start response immediately with "QUESTION 1:" - The total number of words in the generated document must be around 900 words.

Figure 7: The prompt to simplify the professional summary output

An unintended consequence of this summary is that the generated summary sometimes differed from the given template, and started adding its own questions. An example of the resulting output is shown in Figure 8.

QUESTION 11: How will the company measure its success over time?
ANSWER: The company plans to measure its success by tracking key performance indicators such as revenue growth, net income, cash flow generation, production volumes, and market share. They also aim to expand their product offerings, improve customer satisfaction, and maintain a strong reputation in the industry.
QUESTION 12: What role will technology play in the company's future growth and expansion plans?
ANSWER: Technology will play an increasingly important role in the company's growth and expansion plans. The company plans to leverage new technologies to improve efficiency, reduce costs, and expand into new markets.
QUESTION 13: How will the company address its environmental and social responsibilities as part of its growth and expansion plans?
ANSWER: The company is committed to addressing its environmental and social responsibilities as part of its growth and expansion plans. They plan to implement sustainable practices, reduce their carbon footprint, and engage with local communities in a responsible manner.
QUESTION 14: What are the key partnerships and collaborations that the company will pursue in the future?
ANSWER: The company plans to pursue several strategic partnerships and collaborations in the future, including partnerships with technology companies, joint ventures with other oil and gas companies, and collaborations with suppliers and distributors.

Figure 8: Excerpt from an output with extra unwanted questions

4.3 Metrics

Considering the identified issues the following metrics were used to evaluate the summaries.

4.3.1 ROUGE

These generated summaries were graded using the same java ROUGE script as the generated summaries in the fns 2023 task. ROUGE determines the quality of a summary by checking for overlapping words and word pairs with a reference summary [Lin04].

4.3.2 BERTScore

BERTScore evaluates summaries by using the pre-trained contextual embeddings from BERT to compute the similarity between the candidate and reference sentences [ZKW+20].

4.3.3 Flesch Reading Ease

This test aims to identify the difficulty of language used by looking at the amount of syllables per word and words per sentence. With a lower score meaning it is easier to read [KFJRC75].

4.3.4 Non-answers

For the prompts with guided questions, the amount of questions that were not answered. This was achieved by searching for specific phrases with regex like: "Unfortunately", "no information" and "not provided".

4.3.5 Template answers

When the answer given does not relate to the document but is just a blank answer that can be filled in. These were found in square brackets and could be identified that way.

4.3.6 Added questions

This tracks if the summary contained more than the 7 provided questions.

4.3.7 Summary length

The amount of words used in the summary, it should be around 900 as specified in the prompts.

4.3.8 LLM-as-a-judge

The summary was evaluated using the original document and the reference summaries by llama 3.2:3b [AI24] on 4 criteria, similar to the criteria used in [BSA+26] each on a scale of 1-5, to help the model evaluate an explicit reasoning prompting strategy was used [CDM26]:

- ACCURACY (1-5): How accurate is the information in the generated summary compared to the original document and reference summaries? Consider factual correctness and absence of hallucinations. With 1 meaning having significant inaccuracies and hallucinations, and 5 meaning fully accurate with no hallucinations.
- COVERAGE (1-5): How well does the generated summary cover the key topics and important information from the original document and references? Does it capture the main points? With 1 meaning it misses most key information and 5 meaning it captures all key information effectively.
- LANGUAGE_DIFFICULTY (1-5): Evaluate how appropriate the language difficulty is for someone without a financial background or higher education. With 1 being completely incomprehensible and 5 being complete understanding.
- OVERALL_QUALITY (1-5): Overall assessment of the summary's quality as a standalone document. Consider clarity, coherence, usefulness, and informativeness. With 1 being useless to the reader and 5 meaning it is a high-quality summary that effectively and efficiently communicates the essential information.

Table 1: 500 summaries evaluation

Results by Prompt Type				
Prompt types	Basic	Professional	Simplified	Professional Simplified
BERTScore Precision	0.81	**0.82**	0.81	**0.82**
BERTScore Recall	0.80	**0.81**	0.80	0.80
BERTScore F1	0.80	**0.81**	0.80	**0.81**
Flesch Reading Ease	**34.72**	24.27	28.70	34.60
ROUGE-1 Precision	0.33	0.26	**0.35**	0.31
ROUGE-1 Recall	0.33	**0.43**	0.29	0.33
ROUGE-1 F-Score	**0.31**	**0.31**	**0.31**	0.30
ROUGE-2 Precision	**0.10**	0.06	0.08	0.06
ROUGE-2 Recall	0.12	**0.13**	0.08	0.08
ROUGE-2 F-Score	**0.10**	0.08	0.07	0.06
ROUGE-L Precision	0.25	0.24	**0.29**	0.26
ROUGE-L Recall	0.23	**0.29**	0.22	0.24
ROUGE-L F-Score	0.23	**0.26**	0.24	0.24
Word Count	699	1345	585	**824**
LLM Accuracy	3.99	**4.02**	3.98	4.01
LLM Coverage	3.07	**3.20**	3.07	3.03
LLM Language Difficulty	3.62	3.57	3.93	**3.95**
LLM Avg Overall Quality	**3.94**	**3.94**	3.84	3.93
Documents with added questions	n/a	**0**	12	35
Documents with placeholders	60	12	4	**3**
Total placeholders	304	41	6	**3**
Documents with unanswered questions	n/a	**7**	105	**7**
Total Unanswered Questions	n/a	**9**	219	20

5 Results

For 500 documents the summaries were generated and evaluated showing the results in Table 1. The BERTScore is very similar for each prompt type, only showing a slight deviation of 0.01 in each category. The Flesch Reading Ease was the highest for the basic prompt being closely followed by the professional simplified prompt, the lowest score was the professional score and simplified also had a low score. Rouge-1 shows identical F-scores for 3 prompt except for professional simplified only deviating slightly. Rouge-2 shows the highest score for the basic prompt, but the other prompts give similar results. Rouge-L shows that the professional prompt has the highest score, but the other 3 prompts do not differ from it much. Word count shows that professional simplified is the closest to the 900 word goal, with the basic and simplified prompt delivering significantly shorter summaries. The professional summary gave a significantly larger output, with 50% over the goal. The professional prompt has the highest score for accuracy as judged by the LLM, but the scores of the other prompts are very similar. The LLM scored the professional score the highest with a slight margin over the other 3 prompts who scored very similarly. The highest score for the coverage was the Professional prompt, scoring a bit higher than the other prompts, which makes sense considering it also had the longest summaries. with Professional simplified being the lowest. Professional Simplified scored the best for language difficulty as judged by the LLM, having a very close score to the simplified prompt. The Basic and Professional prompt lag significantly behind in this score. The LLM scored all prompts except simplified very similarly high in quality, however simplified did not lag far behind. Only the simplified and professional simplified suffered from the issue of added questions, with professional simplified having the most. Although all prompts suffered from the placeholder issue, the basic prompt showed the largest issue with this with 12% of generated documents containing an average of 5 placeholders. The professional prompt also suffered from this issue and the professional simplified summary the least. The simplified summary had the largest issue unanswered questions, with 21% of documents having unanswered questions. Professional had the same amount of documents with unanswered questions as professional simplified, but professional simplified has a higher number of total unanswered questions.

BERTScore The BERTScore F-score distribution in 9 shows that the scores are very concentrated, with basic having the largest distribution and professional simplified having the smallest.

Flesch Reading Ease The Flesch Reading Ease distribution in 12 shows that all prompts scoring low, with professional having the smallest distribution and simplified having the largest distribution.

word count The word count distribution in 11 show that simplified has the smallest distribution and basic the biggest distribution with the highest outliers.

LLM metrics The LLM metrics show that accuracy was almost always judged at a 4 for all prompt types, coverage almost always at 3 for all prompt types and language difficulty being spread more and overall quality was almost always 4 across all prompt types sometimes being 3.

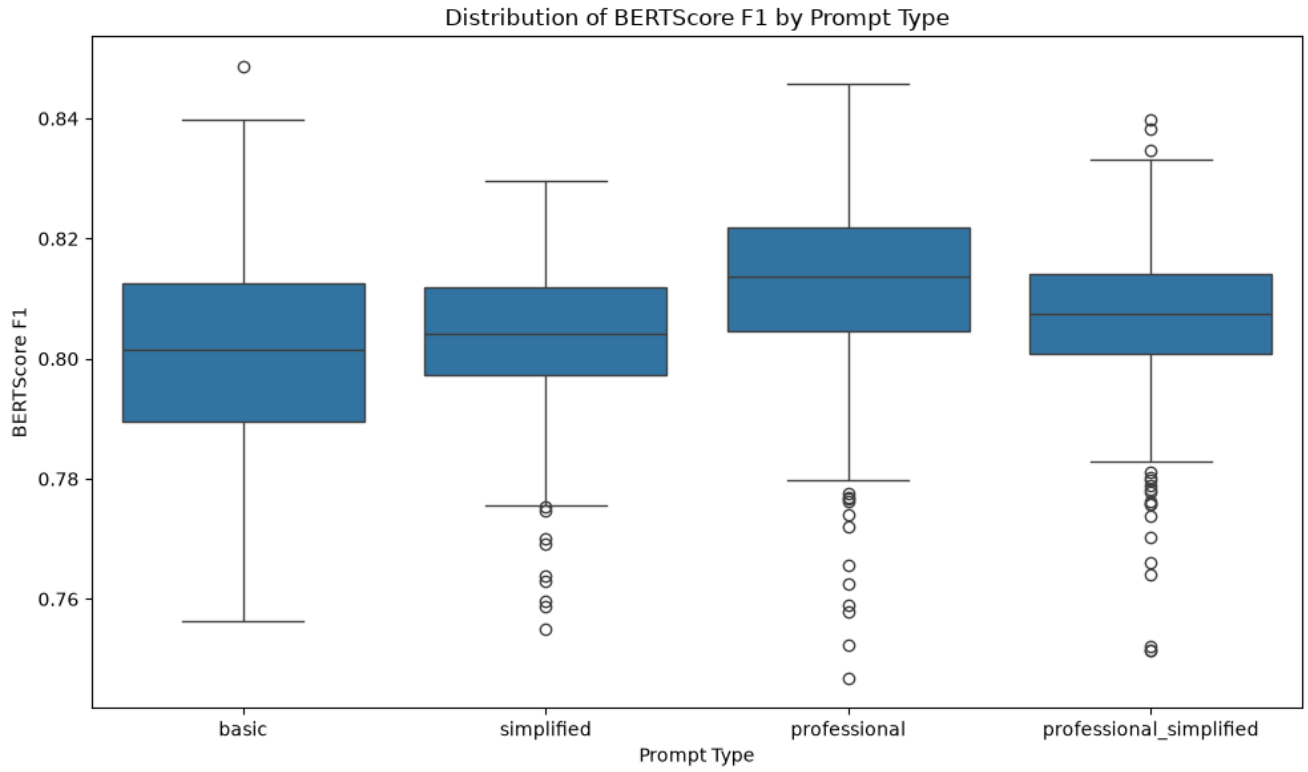

Figure 9: Distribution of BERTScore F-score per prompt type

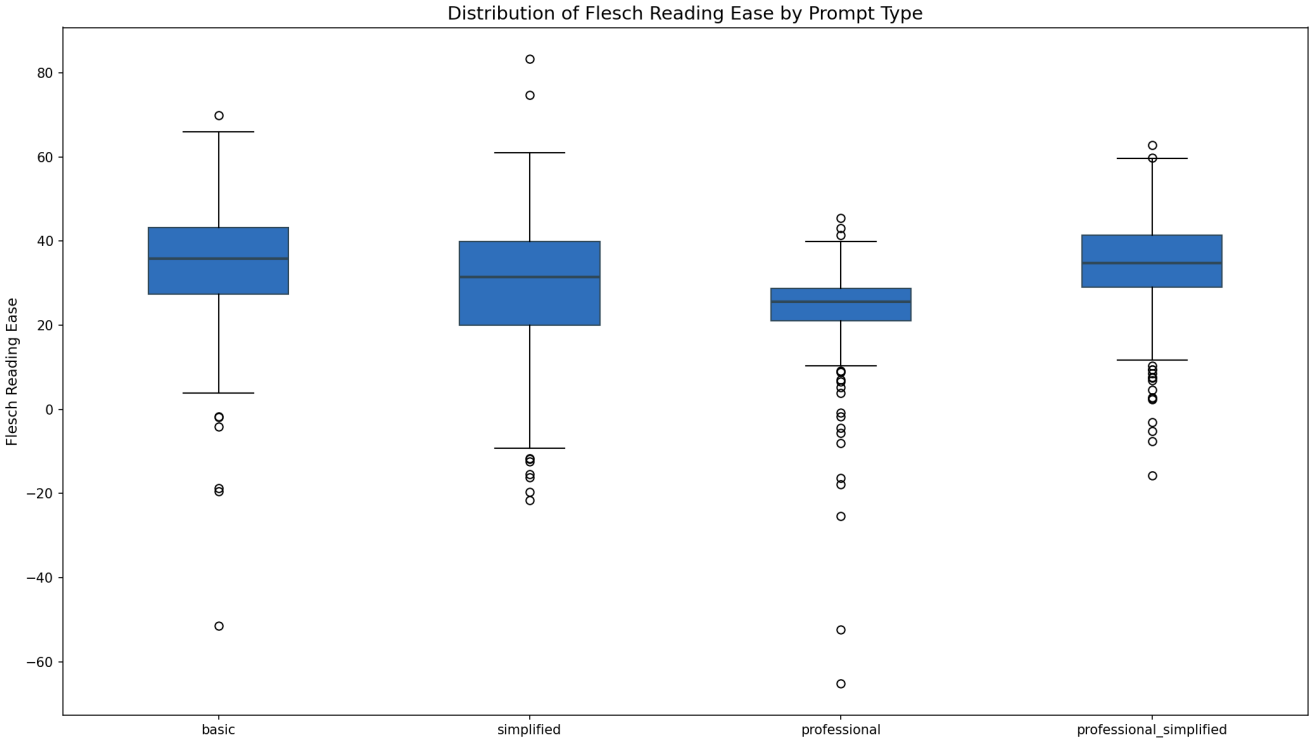

Figure 10: Distribution of flesch reading ease score per prompt type

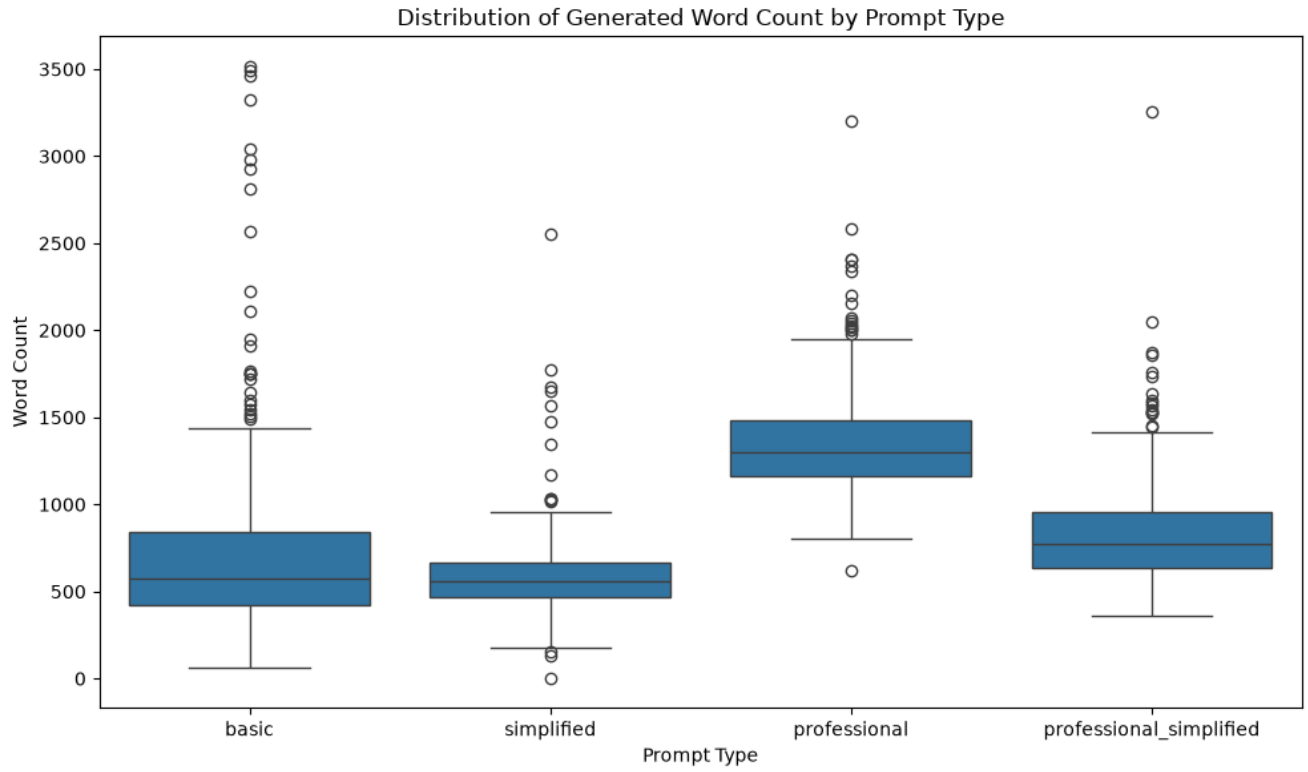

Figure 11: Distribution of word count per prompt type

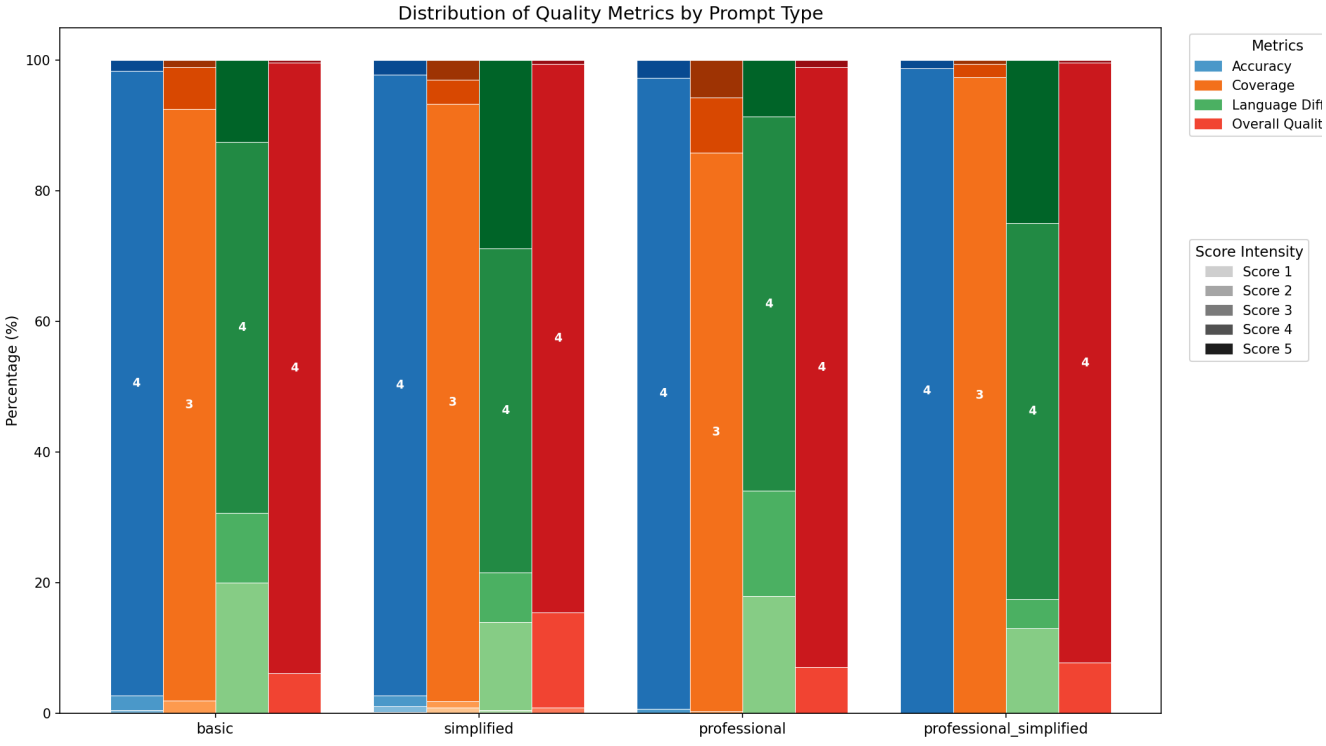

Figure 12: Distribution of LLM judged metrics per prompt type

Results by Prompt Type for Test Set					
Prompt types	MBertExtractive baseline	Basic	Professional	Simplified	Professional Simplified
ROUGE-1 Precision	0.31	0.33	0.27	**0.35**	0.31
ROUGE-1 Recall	0.27	0.31	**0.43**	0.30	0.33
ROUGE-1 F-Score	0.28	0.29	**0.32**	**0.32**	0.31
ROUGE-2 Precision	**0.11**	0.09	0.06	0.08	0.06
ROUGE-2 Recall	**0.14**	0.10	0.12	0.08	0.08
ROUGE-2 F-Score	**0.11**	0.08	0.08	0.07	0.06
ROUGE-L Precision	0.23	0.25	0.25	**0.29**	0.26
ROUGE-L Recall	0.24	0.22	**0.29**	0.23	0.24
ROUGE-L F-Score	0.23	0.22	**0.26**	0.25	0.24

Figure 13: Results by Prompt Type for Test Set and baseline

The test set in 13 show that for ROUGE-1 Professional and Simplified score the highest F-score, with the baseline having the lowest score. The scores for ROUGE-2 show the baseline having the highest score. The Rouge-L scores are improved slightly from the baseline with Professional having the highest score and only Basic falling behind, but overall very similar results.

6 Discussion

The worst summaries were the summaries generated with the basic prompt, mainly because of the large amount of placeholders, but even though it scores higher in certain metrics it is not because it is a better summary. This is because the metrics naturally favour this prompt output. It does not need to answer questions of an economic nature and use difficult words so the Flesch Reading Ease is higher, it does not need to hold to a template so it can fully utilize its own style giving it an edge in all llm-as-a-judge metrics because what it decided itself what would be important to mention would also be considered important by the LLM. After this the worst summary is the summary generated by the simplified prompt, it tries to balance out giving detailed answers to the questions and giving answers that could be understood by someone without a financial background. This results in the LLM not being able to think of a fitting answer and it deciding to not answer the question. This large amount of unanswered questions is also the reason for the lower word count, lower llm-as-a-judge scores for accuracy, coverage and overall quality. The best performing summary was generated by combining the professional prompt and professional simplified prompt. The professional summary best follows the template with low amounts of added questions, placeholders and unanswered questions. It gives detailed answers for most questions leading to higher BERTScore, Rouge scores and llm-as-a-judge scores, however it does create summaries that far exceed the desired summary length. However this summary still contains very difficult language, as shown by the Flesch Reading Ease and LLM Language difficulty. This is where Professional Simplified prompt improves that, it simplifies the language and adjusts the length of the summary to be around the desired length, this adjusting did lead to adding extra questions when the summary was too short. Or not answering the questions when the summary was too long

to save on space. Although this combination still holds to the template better compared to the other prompts it does show the recurring issue with the LLMs of not following the instructions and rules, the length is wrong, the template is wrong or economic terms are used when told to avoid usage. This could mean that adjusting the prompt would not lead to improvements in the quality of the generated summaries and that this dual task of both summarizing and simplification is too much for llama 3.2:3b to handle.

6.1 Limitations

The experiment had several limitations:

- self-bias: The LLM that judged the output was the same model that generated the summaries, research has shown that a model rates its own output higher because they favor texts that adhere to their inherent styles [XZZ+24]
- Model size: larger LLM could perform better at simplifying the language.
- Document: The experiment was only run on annual reports, other financial documents could lead to other results.

7 Conclusions and Further Research

In conclusion a llama 3.2:3b does perform well in creating a summary of a financial document using a question-answering based prompt, but struggles to write the summary in a way that could be understood by people without a financial background having issues with the language difficulty, answering the questions and holding to the desired template. Future research could include researching the performance of other (larger) LLMs, other prompt styles and performance for other financial document types. It could also be beneficial to create a new metric of measuring how much experience in reading financial documents is required to properly understand a given text, as using a llm-as-a-judge to judge a LLM generated text is prone to bias.

References

[AI24] Meta AI. Llama 3.2: A collection of multilingual large language models. https://ai.meta.com/blog/llama-3-2/, 2024.

[ASY+19] Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 2623–2631, 2019.

[BSA+26] Ioana Buhnila, Aman Sinha, Rohit Agarwal, Dilip K. Prasad, and Mathieu Constant. Evaluating LLM-as-a-judge for medical term simplification. In Dina Demner-Fushman, Sophia Ananiadou, Kirk Roberts, and Junichi Tsujii, editors, *BioNLP 2026*, pages 687–694, San Diego, California, July 2026. Association for Computational Linguistics.

[CDM26] Alessia Cantini and Andrea De Mauro. Better prompts, better usefulness: A systematic review and experimental evaluation of structured prompting techniques in large language models. *Big Data and Cognitive Computing*, 10(7), 2026.

[DNC24] Junhua Ding, Huyen Nguyen, and Haihua Chen. Evaluation of question-answering based text summarization using llm invited paper. pages 142–149, 07 2024.

[EHAL+20] Mahmoud El-Haj, Ahmed AbuRa'ed, Marina Litvak, Nikiforos Pittaras, and George Giannakopoulos. The financial narrative summarisation shared task (FNS 2020). In Dr Mahmoud El-Haj, Dr Vasiliki Athanasakou, Dr Sira Ferradans, Dr Catherine Salzedo, Dr Ans Elhag, Dr Houda Bouamor, Dr Marina Litvak, Dr Paul Rayson, Dr George Giannakopoulos, and Nikiforos Pittaras, editors, *Proceedings of the 1st Joint Workshop on Financial Narrative Processing and MultiLing Financial Summarisation*, pages 1–12, Barcelona, Spain (Online), December 2020. COLING.

[FKM+21] Alexander R. Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. Summeval: Re-evaluating summarization evaluation, 2021.

[FLK+25] Jiangnan Fang, Cheng-Tse Liu, Jieun Kim, Yash Bhedaru, Ethan Liu, Nikhil Singh, Nedim Lipka, Puneet Mathur, Nesreen K. Ahmed, Franck Dernoncourt, Ryan Rossi, and Hanieh Deilamsalehy. Multi-LLM text summarization. In Galia Angelova, Maria Kunilovskaya, Marie Escribe, and Ruslan Mitkov, editors, *Proceedings of the 15th International Conference on Recent Advances in Natural Language Processing - Natural Language Processing in the Generative AI Era*, pages 352–362, Varna, Bulgaria, September 2025. INCOMA Ltd., Shoumen, Bulgaria.

[GLD23] Tanya Goyal, Junyi Jessy Li, and Greg Durrett. News summarization and evaluation in the era of gpt-3, 2023.

[GT17] Emet Gürel and Merba Tat. Swot analysis: A theoretical review. *Journal of International Social Research*, 10(51), 2017.

[Han25] Maurits Hanhart. Translating legal texts to b1 dutch language level. Master's thesis, 2025.

[KFJRC75] J Peter Kincaid, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Technical report, 1975.

[Lin04] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004.

[LIX+23] Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-eval: NLG evaluation using gpt-4 with better human alignment. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore, December 2023. Association for Computational Linguistics.

[Oll24] Ollama. Ollama: Large language model runner. https://ollama.com, 2024. Accessed: 2026-06-26.

[QHZ+25] Jipeng Qiang, Minjiang Huang, Yi Zhu, Yunhao Yuan, Chaowei Zhang, and Kui Yu. Redefining simplicity: Benchmarking large language models from lexical to document simplification, 2025.

[The23] The Fin AI. en-fns-2023-long dataset. https://huggingface.co/datasets/TheFinAI/en-fns-2023-long, 2023. Accessed: 2026-06-26.

[WFH+23] Jules White, Quchen Fu, Sam Hays, Michael Sandborn, Carlos Olea, Henry Gilbert, Ashraf Elnashar, Jesse Spencer-Smith, and Douglas C. Schmidt. A prompt pattern catalog to enhance prompt engineering with chatgpt, 2023.

[XZZ+24] Wenda Xu, Guanglei Zhu, Xuandong Zhao, Liangming Pan, Lei Li, and William Wang. Pride and prejudice: LLM amplifies self-bias in self-refinement. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15474–15492, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[ZKG+23] Elias Zavitsanos, Aris Kosmopoulos, George Giannakopoulos, Marina Litvak, Blanca Carbajo-Coronado, Antonio Moreno-Sandoval, and Mo El-Haj. The financial narrative summarisation shared task (fns 2023). In *2023 IEEE International Conference on Big Data (BigData)*, pages 2890–2896, 2023.

[ZKW+20] Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*, 2020.