Prototype-as-Query RAG for Financial Report Summarisation

Anonymous ACL submission

Abstract

The increasing volume and complexity of financial documents pose significant challenges for automated summarisation systems. Large language models (LLMs), while capable of handling long inputs, often struggle to maintain accuracy and coherence when summarising such lengthy and specialised documents. To address these limitations, we introduce PRAGSum, a cost-efficient, language-agnostic retrieval-augmented generation (RAG) system that leverages prototype-as-query retrieval to generate concise and coherent summaries of extended financial reports. In experiments on the Financial Narrative Summarisation (FNS) 2023 dataset, PRAGSum achieves state-of-the-art ROUGE-2 F-score of 0.28. Additionally, we present SummQQ, a novel LLM-based evaluation framework that assesses summaries across five linguistic dimensions without the need for reference summaries. On the DUC 2007 dataset, SummQQ demonstrates a considerable improvement in correlation with human judgements over existing readability and fluency metrics, attaining an average Spearman's ρ of 0.543.[1]

1 Introduction

Automatic summarisation is a critical tool in today's information-rich environment, enabling users to quickly grasp the essence of lengthy texts without reading them in full (Dang, 2006). This is particularly important in areas like finance, where decisions rely on timely insights from complex documents such as financial reports, earnings statements, and regulatory filings. These documents, especially financial reports, often contain dense numerical data and industry-specific terminology, making them significantly more complex than typical long-form documents.

For example, the average document in the Financial Narrative Summarisation (FNS) 2023 dataset contains around 54k words (Zavitsanos et al., 2023), compared to around 9,400 words in the GovReport dataset (Huang et al., 2021). Despite the success of models such as BERT (Devlin et al., 2019), these documents present a significant challenge due to their length and specialised content. Although large language models (LLMs) like GPT-4 and Gemini can process longer inputs (Achiam et al., 2023; Reid et al., 2024), often surpassing 100k tokens, they still struggle to maintain accuracy and coherence throughout the summarisation process (Shukla et al., 2023; Azizov et al., 2023). These LLMs are capable of producing fluent text but are prone to generating irrelevant or hallucinated information.

To tackle these limitations, we propose a new approach to retrieval-augmented generation (RAG)-based summarisation, namely *Prototype-as-Query*. Instead of requiring a targeted user query as in traditional RAG, our method creates a prototype vector from a set of reference summaries which represents an idealised summary for documents in the particular domain. The prototype then serves as a query to retrieve document sections that are semantically aligned with the idealised summary's focus. This reduces the likelihood of irrelevant or missing content in the final output, addressing the trade-offs between content accuracy, summarisation, and readability.

In addition to ensuring content accuracy, the readability and coherence of a summary are vital, especially in complex domains like finance. While traditional readability metrics (e.g., Flesch-Kincaid (Kincaid et al., 1975)) assess text complexity, they do not fully capture important linguistic qualities like fluency and consistency. Furthermore, existing metrics for summarisation evaluation, such as ROUGE and BERTScore (Lin, 2004; Zhang et al., 2019), rely on the presence of reference summaries, limiting their utility.

To address this gap, we introduce SummQQ, an LLM-based evaluation framework for assessing the linguistic quality of machine-generated summaries.

[1] Source code available here.

Inspired by human-centric evaluation frameworks like the Pyramid method (Nenkova et al., 2007), which manually evaluates content coverage by focusing on important information units, SummQQ provides a scalable, automated alternative. Using LLMs (Liu et al., 2023; Fu et al., 2024), SummQQ evaluates summaries based on five key linguistic dimensions: grammaticality, non-redundancy, referential clarity, focus, and coherence. This enables SummQQ to offer a detailed, more nuanced evaluation of linguistic quality, without the need for reference summaries or source texts.

The contributions of this work can be summarised as follows:

- We present a new language-agnostic Prototype-as-Query RAG system, **PRAGSum**, for abstractive and extractive financial report summarisation.
- We conduct thorough experimental assessment of PRAGSum on the FNS 2023 dataset (Zavitsanos et al., 2023) demonstrating state-of-the-art performance, surpassing the previous best ROUGE-2 score by six points.
- We introduce **SummQQ** (**Summ**ary **Q**uality **Q**uestions), a novel LLM-as-a-judge evaluation framework that assesses five aspects of linguistic quality using a form-filling paradigm and probabilistic scoring function.
- We conduct a meta-evaluation of SummQQ, examining various prompting strategies, and comparing a range of readability and summary evaluation metrics on the DUC 2007 dataset (Witte et al., 2007). SummQQ presents a substantial improvement over comparable metrics in correlation with human scores.

2 Related Work

Automatic Text Summarisation Automatic text summarisation generates concise, informative summaries from input texts (Nenkova, 2011; Gupta and Gupta, 2019; Zmandar et al., 2021).

Financial text summarisation has gained research interest due to the increasing availability of complex financial narratives, such as annual reports and earnings releases (de Oliveira et al., 2002; Filippova et al., 2009). Recent efforts are highlighted by the Financial Narrative Summarisation (FNS) task, which challenges researchers to devise multilingual annual report summarisation systems (El-Haj, 2019; El-Haj et al., 2020). Among the best performing models of the task, *DiMSum* (Shukla et al., 2022) identifies and weighs key sections of financial reports, generating summaries based on these weights, while *Positional Language Model (PLM)* (Vanetik et al., 2023) uses positional encoding to create hierarchical summaries, aligning them with essential topics in the reports.

Retrieval-Augmented Generation Retrieval-augmented generation (RAG) (Lewis et al., 2020) integrates information retrieval with generative AI models. The retriever uses query and document encoders to compare input queries with indexed documents, returning the most relevant documents (Gao et al., 2023b,a). RAG allows models to access external, dynamic knowledge sources, grounding generated text in up-to-date information. Recent methods like Self-RAG (Asai et al., 2023) and Self-mem (Cheng et al., 2023) enhance RAG by enabling self-retrieval and memory-based learning, allowing models to autonomously select and store relevant information. M-RAG (Wang et al., 2024) introduces a multi-partition paradigm and multi-agent reinforcement learning framework aimed at enhancing memory quality. In the summarisation domain, RAG has been applied to code summarisation (Parvez et al., 2021; Choi et al., 2023) and query-based summarisation of health records (Saba et al., 2024). Extended financial document summarisation using RAG remains an under-explored research area (Yepes et al., 2024).

Summary Evaluation Strategies Summary evaluation metrics assess generated summaries against gold-standard summaries or source texts. They can be categorised into lexical similarity (e.g., ROUGE (Lin, 2004)), embedding similarity (e.g., BERTScore (Zhang et al., 2019)), factual consistency (e.g., SummaC (Laban et al., 2022)), and comprehensive metrics (e.g., BLANC (Vasilyev et al., 2020)). Readability metrics estimate the ease of understanding a text by considering factors like word complexity and sentence length, typically outputting a score corresponding to years of education required for comprehension. Common metrics include Flesch-Kincaid grade level (Kincaid et al., 1975), Automated Readability Index (Smith and Senter, 1967), Coleman-Liau Index (Coleman and Liau, 1975), Gunning Fog Index (Gunning, 1952), and SMOG (G. Harry McLaughlin, 1969).

A range of reference-free approaches have been proposed for automatic summary coherence evaluation, including learning from human judgements (Xenouleas et al., 2019; Mesgar et al., 2021), exploiting the *shuffle* task (Jwalapuram et al., 2022), and unsupervised metrics that utilise heuristics (Zhu and Bhat, 2020) or LLMs (Yuan et al., 2021).

LLMs are increasingly used as evaluators for summarisation and machine translation. A number of evaluation schemes have been suggested, including Reason-then-Score, MCQ Scoring, and Head-to-Head comparisons (Shen et al., 2023). GPTScore (Fu et al., 2024) uses the probability of generating the candidate text to produce a score, while G-Eval (Liu et al., 2023) utilises the probability of different output score tokens. Despite their promise, LLM evaluators face challenges such as inconsistent ratings and bias (Liu et al., 2023; Panickssery et al., 2024).

3 PRAGSum

PRAGSum combines RAG and prompt engineering techniques to create concise and relevant summaries of long financial reports. The system features three main components: a custom embedding model, a vector database with performant index, and a prototype vector query. We divide our methodology into two stages: (i) *Preparation*, in which we fine-tune the embedding model, populate the vector database, and compute the prototype, and (ii) *Execution*, where we use the prototype to retrieve pertinent extracts from the vector database, and generate summaries with an LLM.

3.1 Preparation

Fine-tune Embedding Model We train an embedding model to obtain an embedding space where relevant and non-relevant chunks are well-separated, and the intra-cluster distance among relevant chunks is reduced. We thus create a dataset of document chunks with binary labels indicating their relevance.

To annotate the document chunks, we employ longest common substring (LCS) (Crochemore et al., 2015) ratio thresholding with threshold t, which we observed to be the most reliable text similarity measure in initial experiments on the FNS dataset. Given the set of document chunks C and set of reference summaries R, the label for a chunk $c \in C$ is defined as:

$$\text{label}_c = \begin{cases} 1 & \text{if } \exists\, r \in R : \frac{LCS(c,r)}{len(c)} \geq t \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

In training, we apply a triplet loss function, which takes a batch of (chunk, label) pairs, a distance function d, and a margin m, and generates all possible (anchor a, positive p, negative n) triplets before computing the loss for each of them. The loss for a triplet is defined as:

$$L = \max\left(0, d(a,p) - d(a,n) + m\right) \quad (2)$$

Populate Vector Database We iterate through each report, conducting the following:

1. Extract metadata (company name and fiscal year) from the report using an LLM, forming a metadata dictionary including the document's unique ID. This is used later on to provide context in the summarisation prompt.
2. Split the report into chunks using a text splitter
3. Use our fine-tuned embedding model to embed the chunks.
4. Save the (chunk text, chunk embedding, metadata) triples in vector database.

Compute Prototype Next, we borrow the concept of the prototype from Snell et al. (2017) to create a global vector representation of the ideal reference summary that can be used to query our vector database of report text chunks. The prototype $\mathbf{p}$ is the mean vector of all the training set reference summary chunk embeddings $\mathcal{R}$:

$$\mathbf{p} = \frac{1}{|\mathcal{R}|} \sum_{r_i \in \mathcal{R}} r_i \quad (3)$$

3.2 Execution

The execution stage consists of retrieving the K most pertinent chunks from the report by similarity to our prototype vector and producing a summary either abstractively or extractively.

Chunk Retrieval Given source document $\mathbf{s}$ and distance function d, the top k chunks can be expressed as:

$$\mathcal{C}_k(\mathbf{s}) = \arg\min_{c_1,c_2,\ldots,c_k \in \mathcal{C}} \sum_{i=1}^{k} d(\mathbf{p}, c_i) \quad (4)$$

where $\mathcal{C}$ represents the set of all chunk embeddings for document $\mathbf{s}$.

Summary Synthesis In extractive mode (EXT), we simply concatenate the top K chunk texts and truncate to length $l = 1000$ words.

In abstractive mode (ABS), given the retrieved chunks $C_k(\mathbf{s})$, metadata m, and prompt template T, the summary $\mathbf{h}$ is generated by a language model L as follows:

$$\mathbf{h} = L(T(C_k(\mathbf{s}), m)) \quad (5)$$

The prompt template (Prompt 1) instructs the model to take on a role of an expert financial analyst with the aim of helping it to access key parametric knowledge and focus on the most relevant parts of the input (White et al., 2023; Dong et al., 2024).

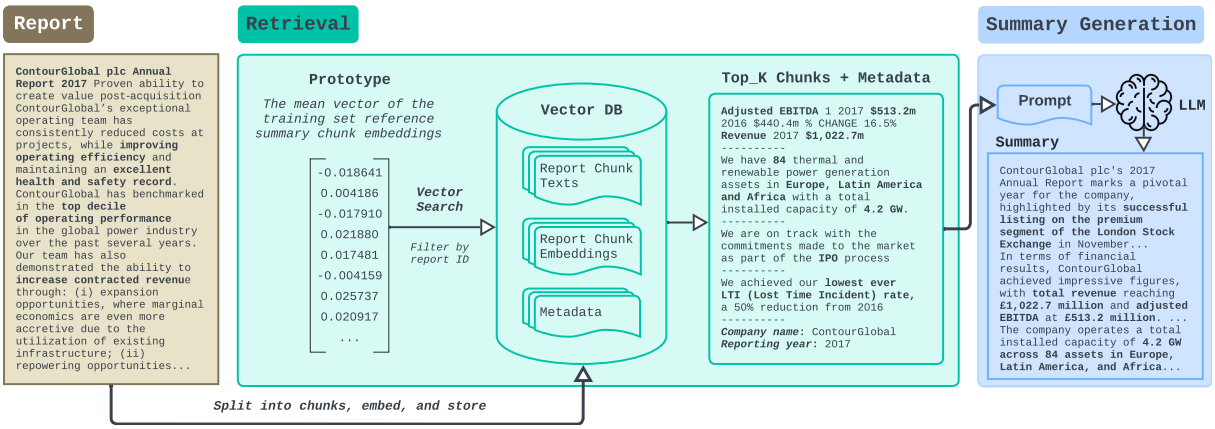

Figure 1: The proposed PRAGSum financial report summarisation system.

> **Prompt 1:** You are an expert financial analyst with extensive experience writing summaries of UK company annual reports. Please write a clear and engaging summary of {company}'s annual report using the following narrative extracts, ...

We assert the summary length requirement in the prompt, ensuring conciseness, and truncate the output. Figure 1 presents a high-level diagram of the system. Appendix C shows the prompts used for each of the datasets in full.

4 SummQQ

SummQQ is an LLM-as-a-judge summary quality evaluation framework that builds on G-Eval (Liu et al., 2023) and the DUC linguistic quality questions (Dang, 2006).

While evaluating PRAGSum, we found that traditional metrics did not adequately capture differences in readability between extractive and abstractive summaries. Moreover, these metrics could not assess various readability aspects without requiring the original source or reference summaries. DUC linguistic questions offer a comprehensive view of readability and fluency, but recent frameworks like SummEval (Fabbri et al., 2021) have not replicated this nuanced assessment. To fill this gap, we propose SummQQ, which uses GPT along with prompt engineering to automatically assess summary quality in a human-like manner.

4.1 Methodology

The linguistic quality questions (grammaticality (*GRA*), non-redundancy (*NRE*), referential clarity (*REF*), focus (*FOC*), and structure and coherence (*COH*)) are each scored on a scale from 1-5, with 5 indicating that the summary is very good with respect to the property in question, and 1 indicating that the summary is very poor with respect to the stated property. All properties and their original definitions are reported in Appendix D.

Our basic prompt template (Prompt 2), inspired by Liu et al. (2023), commences with a short description of the task (to "rate the summary on one metric"), before detailing the quality aspect definition and the rating scale. We utilise a simple form to procure a numerical score from the LLM. We compose and test a set of four prompts: zero-shot, zero-shot chain-of-thought (CoT), few-shot, and few-shot CoT (Kojima et al., 2022; Wei et al., 2022).

> **Prompt 2:** You will be given one summary written for a {source_description}. Your task is to rate the summary on one aspect of linguistic quality.
> Evaluation Criteria:
> {DUC_property} (1-5) - "{DUC_definition}"
> - 1: Very Poor
> - 2: Poor
> - 3: Barely Acceptable
> - 4: Good
> - 5: Very Good
>
> Summary:{summary}
> Evaluation Form (scores ONLY):
> - {DUC_property}:

For the CoT prompts, we adjust the evaluation form to incorporate a *Rationale* field, guiding the LLM to generate a reasoning chain with supporting examples from text before giving a final score (see Appendix C for complete prompt examples).

Then, rather than taking the raw output as our final score, which can result in low variance and low correlation with human scores (Liu et al., 2023), we fetch the top N most likely tokens at the final position of the output, filter them for numerical or worded scores, and then take the probability-weighted sum of these scores as the final result. Note that if a score appears more than once in the most probable tokens (e.g. "2" and later "II" or "two"), we sum the probabilities. In the rare

Language	Split	File Count	
		Reports	*Summaries*
English	Train	3050	10 007
	Validation	413	1383
	Test	550	1804
Spanish	Train	162	324
	Validation	50	100
	Test	50	100
Greek	Train	212	424
	Validation	50	100
	Test	50	100

Table 1: FNS 2023 English, Spanish and Greek dataset file counts.

case that none of the most likely tokens denote a score, we increment the decoding temperature until a score token appears in the top N tokens.

Formally, given the set of unique scores $S = \{s_1, \ldots, s_k\}$, with $k \leq 5$, and the total probability of each score $p(s_i)$, the final score is:

$$score = \sum_{i=1}^{k} p(s_i) \times s_i \tag{6}$$

5 Experimental Assessment

Dataset. The Financial Narrative Summarisation (FNS) 2023 shared task (Zavitsanos et al., 2023) introduced a multilingual financial report summarisation dataset consisting of English, Greek and Spanish reports and reference summaries. The reports, extracted from the PDFs using optical character recognition, are provided as plain text. The reference summaries consist of narrative sections extracted from the reports, such as CEOs' statements and Chairman's letters. We list the file counts for each of the languages in Table 1.

Baselines. We include a variety of baselines: first, we evaluate PRAGSum against two simple baselines to isolate the effects of the retriever and generator components of our system. Then, we include results for a Naive RAG (Gao et al., 2023a) strategy, and a state-of-the-art RAG framework. Lastly, we compare PRAGSum with the top solutions from teams that participated in the FNS 2023 shared task:

No-Context. No retrieved report extracts are provided in the prompt.

Long-Prompt. The entire report (truncated to 60k words) is included in the prompt instead of specific retrieved chunks.

Naive RAG. The user query (summarisation prompt) is embedded and used to retrieve relevant extracts for each report. Base/default embedding model is used[2].

Self-RAG (Asai et al., 2023). Self-RAG adaptively retrieves extracts and reflects on retrieved passages and its own generations with reflection tokens[3].

SSC AI DiMSum (*DiMSum*) (Shukla et al., 2022). DiMSum identifies and weighs key sections of financial reports, then distributes and combines partial summaries based on these weights.

SCE Positional Language Model (*PLM*) (Vanetik et al., 2023). This model uses positional encoding to capture sentence importance and generate hierarchical summaries aligned with key topics in financial reports.

MBZUAI Rocky T5 (*Rocky*) (Azizov et al., 2023). Rocky leverages multilingual T5 models fine-tuned for financial document summarisation.

For abstractive summarisation, we include four LLMs: OpenAI GPT-4o mini (*GPT*) (Achiam et al., 2023), Google Gemini 1.5 Flash (*Gemini*) (Reid et al., 2024), Claude 3 Haiku (*Claude*) and Mistral NeMo Instruct 2407 (*Mistral*), using default parameters.

Implementation. We fine-tuned a multilingual embedding model `gte-multilingual-base` on a text chunk binary classification dataset derived from the English training set annual reports. In labelling the text chunks, we used *difflib*'s `SequenceMatcher` for LCS and threshold $t = 0.5$. We split the dataset with validation size 0.1 and did not shuffle or stratify in order to retain grouping of chunks from the same report. The counts of relevant and non-relevant chunks in our training and validation data are shown in Table 2. We utilised the *Sentence-Transformers* library and applied the `BatchAllTripletLoss` loss function with distance function $d = cosine\ similarity$ and margin $m = 0.25$. We computed validation loss at the end of each epoch and used the checkpoint with the lowest validation loss at the end of training. We utilise Milvus Lite (Wang et al., 2021) for our vector database and HNSW (Malkov and Yashunin, 2020) as our vector index with parameters $M = 18$, $efConstruction = 240$ and $metric = IP$. Training lasted 6.5 hours on a single NVIDIA A10 GPU. Appendix E details base model information and training parameters.

[2]We use `gte-multilingual-base` (Zhang et al., 2024) as the embedding model and GPT-4o mini as generator with basic zero-shot prompts.

[3]We use fine-tuned `selfrag_llama2_7b` with `facebook/mcontriever-msmarco` (Izacard et al., 2021) and Faiss (Douze et al., 2024).

Split	Relevant	Non-relevant	Total
Train	69 385	836 015	905 400
Validation	7215	93 386	100 601

Table 2: Report chunk classification dataset, derived from the FNS 2023 English training set.

Evaluation Metrics. We evaluate PRAGSum and the baselines using the ROUGE family of metrics as used in the FNS 2023 task evaluation[4].

5.1 Results and Discussions

We present and discuss our results on each FNS 2023 dataset along with a summary of our retrieval optimisation study, the details of which are contained in Appendix B. We use † to denote results sourced directly from the FNS 2023 paper (Zavitsanos et al., 2023). * signifies that the result is statistically significantly better than the best FNS baseline on that dataset[5].

FNS 2023 English. Table 3 presents the performance of PRAGSum on the FNS 2023 English test set. Since the reference summaries for this dataset consist of sections of text extracted verbatim from the reports and maximising syntactic similarity (ROUGE) is the goal, it follows that we focus purely on extractive summarisation and thus exclude PRAGSum (ABS) and the no-context/long-prompt baselines. In extractive mode, PRAGSum achieves state-of-the-art results in ROUGE-2 and ROUGE-SU4,with scores of 0.33 and 0.38, respectively, outperforming the baselines. It is also highly competitive in ROUGE-1 and ROUGE-L, with scores of 0.46 for both, closely matching the top-performing baseline DiMSum. The Naive RAG baseline roughly matches the performance of Rocky T5, the worst-performing FNS baseline, while Self-RAG performs relatively poorly on all metrics.

We offer further evaluation in Appendix A and display example output summaries in Appendix F. We present a financial cost analysis of PRAGSum vs. long-prompting in Appendix G.

FNS 2023 Spanish. The FNS 2023 Spanish dataset differs from its English counterpart in that the reference summaries are based on Chairman's statements, which have been removed from the source reports. To address this challenge, we adjust our prompt to tell the model to generate summaries in the style of a Chairman's letter. Note that we use the same embedding model, fine-tuned on the English dataset, for the Spanish and Greek datasets. The prototype used is the mean vector of the reference summary chunk embeddings from both the English and Spanish datasets. We do not present results for one-shot and few-shot prompting for PRAGSum on the English or Spanish datasets as preliminary tests showed inferior performance using these strategies.

System	R-1	R-2	R-L	R-SU4
Self-RAG	0.18	0.07	0.18	0.10
Naive RAG	0.28	0.11	0.25	0.15
Rocky†	0.24	0.12	0.26	0.14
PLM†	0.43	0.27	0.41	0.33
DiMSum†	**0.48**	0.32	**0.47**	0.31
PRAGSum (EXT)	0.46	**0.33***	0.46	**0.38***

Table 3: PRAGSum ROUGE F1 scores on the FNS 2023 English test set compared with the baseline systems. EXT denotes that the system is operating in extractive mode. Rows marked with † are sourced directly from the official FNS 2023 results.

Table 4 shows the ROUGE scores for the baseline systems and PRAGSum on the Spanish test set. We see that despite the reference summaries being abstractive, PRAGSum (EXT) still demonstrates the best performance. The no-context baseline performs poorly, suggesting that the LLMs lack detailed and timely parametric knowledge about the companies. Besides, the long-prompt strategy is competitive, but only with Gemini and GPT. PRAGSum (ABS) is competitive, especially with Gemini as the LLM, and matches PRAGSum (EXT) on ROUGE-1. As before, the naive RAG baseline shows similar overall performance to Rocky T5. Self-RAG displays weak scores, likely owing to the lack of non-English training data used in pretraining and Self-RAG fine-tuning.

FNS 2023 Greek. Each report in the Greek dataset has two reference summaries: (i) a declaration of responsibility from the board of directors, confirming that the financial statements have been prepared accurately and in accordance with standards, and (ii) an extended management discussion of the report. Given the considerable length of the management discussions, achieving high ROUGE recall is challenging, as pointed out in Zavitsanos et al. (2023). Thus we focus on the declarations, instructing the LLM to write a declaration of responsibility. After observing acutely poor perfor-

[4]We utilise the ROUGE 2.0 Java library with $beta = 1$ and stopword removal, configuration aligned with the FNS 2023 shared task evaluators.

[5]All results shown in bold in Tables 3, 4 and 5 have been tested for statistical significance, with * indicating that the result is significantly better than the highest-scoring FNS baseline on that dataset using approximate randomisation with $R = 10000$ and $\alpha = 0.05$ (Graham et al., 2014).

System	Model	R-1	R-2	R-L	R-SU4
Self-RAG	-	0.06	0.02	0.03	0.02
Naive RAG	-	0.35	0.10	0.19	0.15
Rocky†	-	0.39	0.08	0.15	0.14
DiMSum†	-	0.40	0.11	0.16	0.17
PLM†	-	0.41	0.14	0.25	0.20
PRAGSum (EXT)	-	**0.47***	**0.17**	**0.27**	**0.23**
No-Context	GPT	0.41	0.09	0.20	0.16
	Gemini	0.35	0.07	0.19	0.13
	Claude	0.39	0.09	0.21	0.16
	Mistral	0.36	0.08	0.19	0.15
Long-Prompt	GPT	0.45	0.14	0.24	0.20
	Gemini	0.46	0.15	0.26	0.21
	Claude	0.37	0.11	0.23	0.16
	Mistral	0.23	0.06	0.13	0.10
PRAGSum (ABS)	GPT	0.46	0.13	0.24	0.20
	Gemini	**0.47***	0.15	0.26	0.22
	Claude	0.40	0.12	0.24	0.17
	Mistral	0.45	0.15	0.25	0.21

Table 4: ROUGE F1 scores for the baselines and PRAGSum (ABS) on the FNS 2023 Spanish test set with different LLMs. We used top K = 10 for PRAGSum and zero-shot prompting for all systems.

System	Model	R-1	R-2	R-L	R-SU4
Self-RAG	-	0.04	0.01	0.03	0.02
Naive RAG	-	0.26	0.07	0.20	0.12
DiMSum†	-	0.29	0.12	0.20	0.16
Rocky†	-	0.31	0.13	0.25	0.16
PLM†	-	0.32	0.13	0.26	0.17
PRAGSum (EXT)	-	0.32	0.11	0.25	0.16
No-Context (One-Shot)	GPT	0.42	0.26	0.38	0.29
	Gemini	0.31	0.15	0.29	0.17
	Claude	0.42	0.26	**0.40***	0.28
	Mistral	0.42	0.26	0.39	0.28
No-Context (Few-Shot)	GPT	0.42	0.27	0.39	0.30
	Gemini	0.34	0.16	0.31	0.19
	Claude	0.42	0.26	0.39	0.28
	Mistral	0.36	0.25	0.34	0.27
PRAGSum (ABS) (One-Shot)	GPT	**0.43***	0.27	0.39	0.29
	Gemini	0.42	0.27	0.39	0.28
	Claude	**0.43***	0.26	**0.40***	0.29
	Mistral	0.39	0.24	0.35	0.26
PRAGSum (ABS) (Few-Shot)	GPT	**0.43***	**0.29***	**0.40***	**0.31***
	Gemini	0.41	0.27	0.38	0.28
	Claude	0.42	0.28	0.39	0.29
	Mistral	0.38	0.24	0.35	0.26

Table 5: ROUGE F1 scores for the baselines and PRAGSum (ABS) on the FNS 2023 Greek test set with different LLMs. We set top K = 10 and $n = 3$ examples in the few-shot prompt.

mance with zero-shot prompting, we investigate one-shot and few-shot prompting on this dataset. To select examples ("shots") for the prompts, we compute embeddings of the training set declarations, perform K-means clustering, and select the example(s) closest to the cluster centre(s). We compute our prototype over the reference summary chunk embeddings for both the English and Greek datasets. Table 5 highlights a sizeable improvement in ROUGE scores with PRAGSum (ABS) over the FNS baselines, and the best performance achieved using GPT and few-shot prompting. Naive RAG underperforms the FNS baselines, and Self-RAG gives near-zero scores. Conversely, the strong performance of the no-context baseline, particularly with Claude and GPT, suggests that retrieval plays a limited role in this dataset. This may be attributed to the boilerplate nature of the summaries; the LLM only has to substitute in the correct named entities to generate accurate synopses, which are provided by broader parametric knowledge.

Note that we exclude the long-prompt baseline for this dataset as the formulaic nature of the generated summaries means that providing entire reports in the prompt has diminutive benefit.

FNS 2023 Overall. We summarise the ROUGE-2 results of PRAGSum and the FNS and RAG baselines in Table 6. PRAGSum betters the prior top submission by six points overall.

Retrieval Optimisation Study. To optimise the performance of PRAGSum's retriever component, we examined the effect of different retrieval configurations on key quality metrics, including mean precision@k, mean reciprocal rank, and mean normalised discounted cumulative gain. Full details are provided in Appendix B.

5.2 Meta-Evaluation of SummQQ

Here we compare existing readability and summarisation evaluation metric types and SummQQ with human judgements of several aspects of linguistic quality using the DUC 2007 quality questions dataset. Our aim is to provide a fresh evaluation and comparison of common summary quality and readability metrics, whilst appraising SummQQ against comparable (denoted with ✓) metrics.

System	EN	EL	ES	W-AVG
Self-RAG	0.07	0.01	0.02	0.04
Naive RAG	0.11	0.07	0.10	0.10
Rocky†	0.12	0.13	0.08	0.11
PLM†	0.27	0.13	0.14	0.20
DiMSum†	0.32	0.12	0.11	0.22
PRAGSum	**0.33***	**0.29***	**0.17**	**0.28**

Table 6: Baseline system and PRAGSum ROUGE-2 F1 scores across all FNS 2023 datasets. The English (EN) scores are given a weight of 0.5 in computing the average scores as in Zavitsanos et al. (2023). The PRAGSum results given are the scores of the optimal configuration on each dataset.

Metric Type	Metric	GRA		NRE		REF		FOC		COH		**AVG**	
		ρ	τ	ρ	τ	ρ	τ	ρ	τ	ρ	τ	ρ	τ
Lexical Similarity	ROUGE-1	0.316	0.242	0.031	0.019	0.284	0.217	**0.335**	**0.258**	0.409	0.316	0.275	0.210
	ROUGE-2	0.306	0.235	0.008	0.001	0.302	0.229	0.304	0.232	0.396	0.304	0.263	0.200
	ROUGE-L	**0.329**	**0.252**	**0.074**	**0.054**	**0.306**	**0.233**	0.333	0.255	**0.424**	**0.328**	**0.293**	**0.224**
Embedding Similarity	BERTScore	**0.390**	**0.300**	**0.108**	**0.079**	**0.322**	**0.246**	0.350	0.269	0.411	0.316	**0.316**	**0.242**
	MoverScore	0.343	0.262	0.083	0.061	**0.322**	0.245	**0.383**	**0.293**	**0.435**	**0.334**	0.313	0.239
Readability ✓	FK	0.126	0.093	0.035	0.026	0.181	0.134	0.191	0.141	0.186	0.140	0.143	0.107
	GFI	**0.155**	**0.116**	0.042	0.032	0.195	0.144	0.211	0.157	**0.215**	**0.162**	0.164	0.122
	ARI	0.127	0.094	0.057	0.042	0.179	0.132	0.191	0.141	0.168	0.126	0.144	0.107
	CLI	0.116	0.086	**0.062**	**0.046**	**0.212**	**0.155**	**0.245**	**0.182**	0.200	0.147	**0.167**	**0.123**
Comprehensive ✓	BARTScore	0.046	0.034	0.047	0.035	0.021	0.016	**0.058**	**0.043**	0.017	0.011	0.038	0.028
	BLANC	**0.049**	**0.036**	**0.076**	**0.056**	**0.060**	**0.046**	0.030	0.024	**0.080**	**0.062**	**0.059**	**0.045**
Ours ✓	*SummQQ-mini*												
	- Zero-Shot	0.654	0.517	0.393	0.303	0.468	0.355	**0.482**	**0.370**	0.596	0.464	0.519	0.402
	- Zero-Shot-CoT	0.623	0.490	0.455	0.350	0.448	0.338	0.446	0.339	0.560	0.433	0.506	0.390
	- Few-Shot	**0.658**	**0.521**	0.419	0.322	0.474	0.361	0.473	0.359	0.559	0.433	0.517	0.399
	- Few-Shot-CoT	0.629	0.496	0.483	0.374	0.462	0.351	0.408	0.305	0.558	0.431	0.508	0.392
	SummQQ												
	- Zero-Shot	0.604	0.474	**0.501**	**0.392**	**0.533**	**0.410**	0.470	0.361	**0.605**	**0.474**	**0.543**	**0.422**

Table 7: Summary-level absolute Spearman (ρ) and Kendall-Tau (τ) correlations of existing metrics and SummQQ variants with human judgements of five aspects of summary readability and fluency on the DUC 2007 dataset.

Dataset. The DUC 2007 dataset consists of human-written and machine-generated summaries for topic clusters of news articles[6]. Every summary is annotated with human ratings of the five DUC linguistic quality aspects.

Baselines. We include a variety of widely used summary evaluation and readability metrics in our investigation:

Readability: Flesch-Kinkaid (FK), Gunning Fog Index (GFI), Automated Readability Index (ARI), and Coleman-Liau Index (CLI), using the `py-readability-metrics` package.

Lexical similarity: ROUGE-1, ROUGE-2, and ROUGE-L, implemented using the `rouge-score` package with default settings.

Embedding similarity: BERTScore (`bert-score` package with `roberta-large-mnli`) and MoverScore v2 (Zhao et al., 2019) (`distilbert-base-uncased`) with batch size 32.

Comprehensive: BARTScore (Yuan et al., 2021) (using `facebook/bart-large-cnn` with batch size 4) and BLANC (batch size 32).

For SummQQ, we use OpenAI's GPT-4o and GPT-4o mini, fixing the temperature at 0 and setting $top\_logprobs = 20$.

Results and Discussions. Table 7 presents Spearman and Kendall-Tau correlations of SummQQ and the baseline metrics against human judgements across five linguistic quality aspects.

The readability metrics exhibit consistently weak correlations with human ratings, particularly for non-redundancy (NRE). Among them, the Coleman-Liau Index performs best but still shows limited alignment with human evaluations, especially when considering more nuanced linguistic elements. The comprehensive metrics, intended to assess multiple quality attributes including fluency and coherence (Yuan et al., 2021; Vasilyev et al., 2020), perform remarkably poorly overall, highlighting the limitations of multi-faceted reference-less metrics in measuring isolated attributes.

SummQQ, particularly in its zero-shot GPT-4o configuration, outperforms all other comparable (*Readability* and *Comprehensive*) metrics in alignment with human judgements. Introducing CoT results in improved correlation on NRE but slightly worse correlation on the other four aspects, while appending shots to the prompt offers improved performance on the finer-grained properties (GRA, NRE and REF), but is detrimental on higher-level aspects (FOC and COH).

6 Conclusion

In this paper, we introduced two novel contributions to automatic summarisation: PRAGSum, a prototype-as-query RAG system for extractive and abstractive summarisation of financial annual reports, and SummQQ, an LLM-as-a-judge framework for intrinsic summary quality evaluation. PRAGSum achieved state-of-the-art ROUGE scores on the FNS 2023 shared task and outperformed other modern RAG approaches. SummQQ demonstrated a substantial improvement over comparable automatic metrics in correlation with human scores on the DUC 2007 dataset.

[6]The dataset features 45 clusters of 25 related news articles. Each topic has 32 machine-generated summaries and 4 human reference summaries, resulting in a total of 1620 summaries.

Limitations

Our novel summarisation system relies on the existence of a training set of documents and gold standard summaries in the relevant domain that can be used to calculate the prototype vector. Testing of PRAGSum on a wider range of long document summarisation datasets and in heterogeneous languages could be conducted, though we heed the critical lack of these datasets in many common languages. There are several prompting techniques, such as Auto CoT, that could present benefits to both our proposed systems. Owing to financial constraints, evaluation of few-shot and CoT prompting strategies for SummQQ with GPT-4o have been left for future work.

Ethical Considerations

As with any automatic summarisation system, if erroneous or false data form part of the input to PRAGSum, it is possible that our system will include this data in its output. In the case of financial report summarisation, a summary containing fraudulent information about a company could be used to purposely mislead investors and manipulate stock prices. LLMs are prone to hallucinations, which could result in PRAGSum-generated summaries containing incorrect information (Zhang et al., 2023). We acknowledge the tendency of the LLM evaluator in SummQQ to assign higher ratings to its own outputs (Panickssery et al., 2024). LLMs are vulnerable to malicious attacks including prompt injection, jailbreak prompting, and token manipulation, which can be exploited for pharming, malware distribution, and other cyber threats (Zou et al., 2023).

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2023. Self-rag: Learning to retrieve, generate, and critique through self-reflection. *arXiv preprint arXiv:2310.11511*.

Dilshod Azizov, Jiyong Li, Hilal AlQuabeh, and Shangsong Liang. 2023. Advanced nlp techniques for summarizing multilingual financial narratives from global annual reports. In *2023 IEEE International Conference on Big Data (BigData)*, pages 2802–2804. IEEE.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. O'Reilly Media, Inc.

Harrison Chase. 2022. Langchain.

Xin Cheng, Di Luo, Xiuying Chen, Lemao Liu, Dongyan Zhao, and Rui Yan. 2023. Lift yourself up: Retrieval-augmented text generation with self-memory. In *Advances in Neural Information Processing Systems*, volume 36, pages 43780–43799. Curran Associates, Inc.

Yunseok Choi, Cheolwon Na, Hyojun Kim, and Jee-Hyong Lee. 2023. Readsum: Retrieval-augmented adaptive transformer for source code summarization. *IEEE Access*, 11:51155–51165.

Meri Coleman and T. L. Liau. 1975. A computer readability formula designed for machine scoring. *Journal of Applied Psychology*, 60(2):283–284.

Maxime Crochemore, Costas S. Iliopoulos, Alessio Langiu, and Filippo Mignosi. 2015. The longest common substring problem. *Mathematical Structures in Computer Science*, 27(2):277–295.

Hoa Trang Dang. 2006. Duc 2005: evaluation of question-focused summarization systems. In *Proceedings of the Workshop on Task-Focused Summarization and Question Answering - SumQA '06*, SumQA '06, pages 48–55. Association for Computational Linguistics.

Paulo Cesar Fernandes de Oliveira, Khurshid Ahmad, and Lee Gillam. 2002. A financial news summarization system based on lexical cohesion. In *Proceedings of the International Conference on Terminology and Knowledge Engineering, Nancy, France*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics*.

Yihong Dong, Xue Jiang, Zhi Jin, and Ge Li. 2024. Self-collaboration code generation via chatgpt. *ACM Transactions on Software Engineering and Methodology*.

Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. 2024. The faiss library. *arXiv preprint arXiv:2401.08281*.

Mahmoud El-Haj. 2019. Multiling 2019: Financial narrative summarisation. In *Proceedings of the Workshop MultiLing 2019: Summarization Across Languages, Genres and Sources*, pages 6–10.

Mahmoud El-Haj, Ahmed Ghassan Tawfiq AbuRa'ed, Marina Litvak, Nikiforos Pittaras, and George Giannakopoulos. 2020. The financial narrative summarisation shared task (fns 2020). In *Proceedings*

of the 1st Joint Workshop on Financial Narrative Processing and MultiLing Financial Summarisation.

Alexander R Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. Summeval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409.

Katja Filippova, Mihai Surdeanu, Massimiliano Ciaramita, and Hugo Zaragoza. 2009. Company-oriented extractive summarization of financial news. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pages 246–254.

Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2024. Gptscore: Evaluate as you desire. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*. Association for Computational Linguistics.

G. Harry McLaughlin. 1969. Smog grading - a new readability formula. *The Journal of Reading*.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Qianyu Guo, Meng Wang, and Haofen Wang. 2023a. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, and Haofen Wang. 2023b. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*.

Yvette Graham, Nitika Mathur, and Timothy Baldwin. 2014. Randomized significance tests in machine translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, page 266–274. Association for Computational Linguistics.

Robert Gunning. 1952. *The Technique of Clear Writing*. McGraw-Hill.

Som Gupta and S. K Gupta. 2019. Abstractive summarization: An overview of the state of the art. *Expert Systems with Applications*, 121:49–65.

Luyang Huang, Shuyang Cao, Nikolaus Parulian, Heng Ji, and Lu Wang. 2021. Efficient attentions for long document summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2021. Unsupervised dense information retrieval with contrastive learning. *arXiv preprint arXiv:2112.09118*.

Prathyusha Jwalapuram, Shafiq Joty, and Xiang Lin. 2022. Rethinking self-supervision objectives for generalizable coherence modeling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6044–6059, Dublin, Ireland. Association for Computational Linguistics.

J. P. Kincaid, Jr. Fishburne, Rogers Robert P., Chissom Richard L., and Brad S. 1975. *Derivation of New Readability Formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel*.

Takeshi Kojima, Shixiang (Shane) Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems*, volume 35, pages 22199–22213. Curran Associates, Inc.

Philippe Laban, Tobias Schnabel, Paul N. Bennett, and Marti A. Hearst. 2022. Summac: Re-visiting nli-based models for inconsistency detection in summarization. *Transactions of the Association for Computational Linguistics*, 10:163–177.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: Nlg evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522. Association for Computational Linguistics.

Yu A. Malkov and D. A. Yashunin. 2020. Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs. *IEEE Trans. Pattern Anal. Mach. Intell.*, 42(4):824–836.

Mohsen Mesgar, Leonardo F. R. Ribeiro, and Iryna Gurevych. 2021. A neural graph-based local coherence model. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2316–2321, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. 2023. Mteb: Massive text embedding benchmark. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics.

Ani Nenkova. 2011. Automatic summarization. *Foundations and Trends® in Information Retrieval*, 5(2):103–233.

Ani Nenkova, Rebecca Passonneau, and Kathleen McKeown. 2007. The pyramid method: Incorporating human content selection variation in summarization evaluation. *ACM Transactions on Speech and Language Processing*, 4(2):4–es.

Arjun Panickssery, Samuel R Bowman, and Shi Feng. 2024. Llm evaluators recognize and favor their own generations. *arXiv preprint arXiv:2404.13076*.

Md Rizwan Parvez, Wasi Ahmad, Saikat Chakraborty, Baishakhi Ray, and Kai-Wei Chang. 2021. Retrieval augmented code generation and summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2021*. Association for Computational Linguistics.

Machel Reid, Nikolay Savinov, Denis Teplyashin, et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *Google AI Whitepaper*.

Walid Saba, Suzanne Wendelken, and James Shanahan. 2024. Question-answering based summarization of electronic health records using retrieval augmented generation. *arXiv preprint arXiv:2401.01469*.

Chenhui Shen, Liying Cheng, Xuan-Phi Nguyen, Yang You, and Lidong Bing. 2023. Large language models are not yet human-level evaluators for abstractive summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4215–4233, Singapore. Association for Computational Linguistics.

Neelesh Shukla, Amit Vaid, Raghu Katikeri, Sangeeth Keeriyadath, and Msp Raja. 2022. DiMSum: Distributed and multilingual summarization of financial narratives. In *Proceedings of the 4th Financial Narrative Processing Workshop @LREC2022*, pages 65–72, Marseille, France. European Language Resources Association.

Neelesh K Shukla, Raghu Katikeri, Msp Raja, Gowtham Sivam, Shlok Yadav, Amit Vaid, and Shreenivas Prabhakararao. 2023. Generative ai approach to distributed summarization of financial narratives. In *2023 IEEE International Conference on Big Data (BigData)*, page 2872–2876. IEEE.

E A Smith and R. Senter. 1967. Automated readability index. *AMRL-TR. Aerospace Medical Research Laboratories*, pages 1–14.

Jake Snell, Kevin Swersky, and Richard Zemel. 2017. Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Natalia Vanetik, Elizaveta Podkaminer, and Marina Litvak. 2023. Summarizing financial reports with positional language model. In *2023 IEEE International Conference on Big Data (BigData)*, volume 33, page 2877–2883. IEEE.

Oleg Vasilyev, Vedant Dharnidharka, and John Bohannon. 2020. Fill in the blanc: Human-free quality estimation of document summaries. In *Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems*. Association for Computational Linguistics.

Jianguo Wang, Xiaomeng Yi, Rentong Guo, Hai Jin, Peng Xu, Shengjun Li, Xiangyu Wang, Xiangzhou Guo, Chengming Li, Xiaohai Xu, Kun Yu, Yuxing Yuan, Yinghao Zou, Jiquan Long, Yudong Cai, Zhenxiang Li, Zhifeng Zhang, Yihua Mo, Jun Gu, Ruiyi Jiang, Yi Wei, and Charles Xie. 2021. Milvus: A purpose-built vector data management system. In *Proceedings of the 2021 International Conference on Management of Data*, SIGMOD '21, page 2614–2627, New York, NY, USA. Association for Computing Machinery.

Zheng Wang, Shu Teo, Jieer Ouyang, Yongjun Xu, and Wei Shi. 2024. M-RAG: Reinforcing large language model performance through retrieval-augmented generation with multiple partitions. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1966–1978, Bangkok, Thailand. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc.

Jules White, Quchen Fu, Sam Hays, Michael Sandborn, Carlos Olea, Henry Gilbert, Ashraf Elnashar, Jesse Spencer-Smith, and Douglas C. Schmidt. 2023. A prompt pattern catalog to enhance prompt engineering with chatgpt. *Preprint*, arXiv:2302.11382.

René Witte, Ralf Krestel, and Sabine Bergler. 2007. Generating update summaries for duc 2007. In *Proceedings of the Document Understanding Conference*, pages 1–5.

Stratos Xenouleas, Prodromos Malakasiotis, Marianna Apidianaki, and Ion Androutsopoulos. 2019. SUM-QE: a BERT-based summary quality estimation model. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6005–6011, Hong Kong, China. Association for Computational Linguistics.

Antonio Jimeno Yepes, Yao You, Jan Milczek, Sebastian Laverde, and Renyu Li. 2024. Financial report chunking for effective retrieval augmented generation. *arXiv preprint arXiv:2402.05131*.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. Bartscore: Evaluating generated text as text generation. In *Advances in Neural Information Processing Systems*, volume 34, pages 27263–27277. Curran Associates, Inc.

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

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Xin Zhang, Yanzhao Zhang, Dingkun Long, Wen Xie, Ziqi Dai, Jialong Tang, Huan Lin, Baosong Yang, Pengjun Xie, Fei Huang, Meishan Zhang, Wenjie Li, and Min Zhang. 2024. mgte: Generalized long-context text representation and reranking models for multilingual text retrieval. *Preprint*, arXiv:2407.19669.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, et al. 2023. Siren's song in the ai ocean: a survey on hallucination in large language models. *arXiv preprint arXiv:2309.01219*.

Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M. Meyer, and Steffen Eger. 2019. Moverscore: Text generation evaluating with contextualized embeddings and earth mover distance. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics.

Wanzheng Zhu and Suma Bhat. 2020. GRUEN for evaluating linguistic quality of generated text. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 94–108, Online. Association for Computational Linguistics.

Nadhem Zmandar, Abhishek Singh, Mahmoud El-Haj, and Paul Rayson. 2021. Joint abstractive and extractive method for long financial document summarization. In *Proceedings of the 3rd Financial Narrative Processing Workshop*, pages 99–105, Lancaster, United Kingdom. Association for Computational Linguistics.

Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J Zico Kolter, and Matt Fredrikson. 2023. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*.

A Further PRAGSum Summary Quality Evaluation

System	Model	BERTScore	AlignScore	GRA	NRE	REF	FOC	COH
Ref. Summaries	-	-	-	3.13	3.38	3.74	3.72	3.47
PRAGSum (EXT)	-	**0.80**	**0.64**	2.67	2.66	3.48	2.87	2.87
No-Context	GPT	0.73	0.33	**4.99**	**4.03**	**4.93**	**4.85**	**4.96**
	Gemini	0.72	0.37	4.77	3.93	4.26	4.17	4.41
	Claude	0.73	0.34	4.98	4.00	4.84	4.57	4.71
	Mistral	0.74	0.31	4.96	4.00	4.77	4.56	4.56
Long-Prompt	GPT	**0.76**	0.40	**4.98**	4.00	4.38	4.29	4.10
	Gemini	**0.76**	0.47	4.88	3.99	4.27	4.10	4.03
	Claude	0.74	0.42	4.37	3.94	4.12	4.28	3.93
	Mistral	0.67	0.35	2.79	2.22	2.77	2.42	2.30
PRAGSum (ABS)	GPT	**0.76**	0.49	**4.98**	4.00	4.26	4.24	4.10
	Gemini	0.75	**0.54**	4.91	3.99	4.20	4.09	4.02
	Claude	**0.76**	0.53	4.79	3.99	4.12	4.15	4.00
	Mistral	0.75	0.53	4.82	4.00	4.08	4.05	4.01

Table 8: PRAGSum and long-prompt baseline summary evaluation results (BERTScore, AlignScore and SummQQ) with different LLMs on the FNS 2023 English test set. We utilised top K $= 10$ and zero-shot prompting for PRAGSum and GPT-4o mini and zero-shot prompt with SummQQ.

We evaluated PRAGSum on semantic similarity and factual alignment with reference summaries, along with linguistic quality, using the following metrics:

- BERTScore (BS). Computes semantic similarity by aligning token-level contextual embeddings. We use the Python *bert-score* package, the `roberta-large-mnli` model and batch size of 32.
- AlignScore (AS). Assesses factual consistency via a unified information alignment function. We employ the `roberta-base` model at the AlignScore-base checkpoint, with a batch size of 64 and the default `nli_sp` 3-way evaluation mode.
- SummQQ. Our proposed evaluation metric that uses an LLM-as-a-judge framework to assess linguistic quality (§4). For this, we apply the zero-shot prompt with GPT-4o mini as the evaluator.

Table 8 compares the results for PRAGSum in both extractive (EXT) and abstractive (ABS) modes with the no-context and long-prompt baselines across the four LLMs. In extractive mode (EXT), PRAGSum achieves the highest scores for reference-based metrics like BERTScore (0.80) and AlignScore (0.64), confirming its strong accuracy and factual consistency when compared to the reference summaries. These metrics reflect PRAGSum's ability to maintain semantic similarity with the reference texts while ensuring the factual correctness of the generated content.

The SummQQ ratings, designed to assess multiple aspects of readability and fluency, reveal that abstractive summaries produced by PRAGSum (ABS) and the two baselines are significantly better-written than the extractive summaries. Also, the no-context baseline attains substantially higher scores on referential clarity, focus and coherence, which may imply that LLMs produce better-structured output when drawing exclusively on parametric knowledge as opposed to information contained in the prompt Specifically, the no-context baseline with GPT-4o scored 4.85 on focus and 4.96 on coherence, compared to 4.24 and 4.10 for PRAGSum (ABS). Intriguingly, PRAGSum (ABS) and both baselines exceed the reference summaries' scores across all five linguistic quality aspects. This suggests that LLM-generated summaries can compete with and even surpass the reference summaries in terms of structure and readability. Conversely, Mistral NeMo struggled with handling very long contexts, which significantly impacted its performance on the long-context baseline. Lastly, the consistent higher scoring of GPT is suggestive of the evaluator favouring its own generations (Panickssery et al., 2024).

B Retrieval Optimisation Study

This study was designed to investigate the effect of different configurations on the performance of PRAGSum's retriever component. We utilise the FNS 2022 validation set with these metrics: mean precision@k (MP@k), mean reciprocal rank (MRR) and mean normalised discounted cumulative gain

(MNDCG). LCS ratio thresholding with $t = 0.35$ is used to decide whether a chunk is relevant. Table 9 delineates the results of our ablation tests. We employ OpenAI's `text-embedding-3-small` as the embedding model, and chunk size 4096 and top K of 5 for the initial two tests.

First, we test if splitting the reference summaries into chunks before computing the prototype improves retrieval performance, We then examine the impact of performing basic cleaning with RegEx. Specifically, we (a) remove non-ASCII characters, (b) remove unnecessary punctuation, and (c) normalise whitespace. Next, we evaluate different text chunking strategies, as follows:

- Fixed-size: splits into chunks of a fixed size regardless of sentence/semantic boundaries.
- Recursive: recursively splits on progressively smaller delimiters until size limit is satisfied.
- Sentence: splits at sentence boundaries.
- Semantic: splits based on semantic meaning.

We utilise NLTK's `sent_tokenize` (Bird et al., 2009) for sentence chunking, and LangChain (Chase, 2022) text splitters for the other strategies. In all cases, we set a minimum chunk size of 100 characters. Following this, we investigate different chunk sizes with the best strategy (recursive), modulating K so that the volume of text retrieved between chunk sizes is constant.

Finally, we assess the benefit of our custom embedding model `gte-multilingual-base-fns`. We include two pre-trained embedding models, `stella_en_1.5B_v5`, which ranks #1 on the MTEB (Muennighoff et al., 2023) benchmark at the time of writing, and OpenAI's `text-embedding-3-small`.

Setting	Value	MP@k	MRR	MNDCG
Split Reference Summaries	Disabled	0.33	0.55	0.34
	Enabled	**0.38**	**0.62**	**0.40**
RegEx Cleaning	**Disabled**	**0.38**	**0.62**	**0.40**
	Enabled	0.29	0.49	0.30
Chunking Strategy	Semantic	0.31	0.57	0.32
	Fixed-size	0.48	0.71	0.50
	Sentence	0.48	0.74	0.49
	Recursive	**0.51**	**0.75**	**0.54**
Chunk Size	512	0.37	**0.78**	0.42
	1024	**0.39**	0.76	**0.43**
	2048	0.38	0.68	0.41
	4096	0.38	0.62	0.40
Embedding Model	`stella_en_1.5B_v5`	0.40	0.59	0.40
	`text-embedding-3-small`	0.51	0.75	0.54
	`gte-multilingual-base-fns`	**0.91**	**0.96**	**0.92**

Table 9: Retrieval system optimisation study results on the FNS 2022 English validation set.

C Example Prompts

C.1 PRAGSum

Listing 1: Zero-shot prompt used for abstractive summarisation of the FNS 2023 English reports.

```
You are an expert financial analyst with extensive experience crafting summaries of UK company
annual reports.

Please write a clear and engaging summary of {company}'s annual report using the following narrative
extracts, focusing on key financial information, strategic highlights, and significant developments.

Further instructions:

    -   Your summary should be close to, but no more than, 1000 words in length.
    -   Present any monetary amounts in £ unless another currency symbol is stated in the extract.
    -   Include as much relevant information as possible from the extracts provided.
    -   Format your response as plain text in paragraphs without any headings or subheadings.

Narrative extracts:

{context}
```

Listing 2: Zero-shot prompt used for summarisation of the FNS 2023 Spanish reports.

```
You are an expert Spanish financial analyst with extensive experience crafting summaries of annual
reports for Spanish companies.

Please write a clear and engaging summary of {company}'s annual report in Spanish, using the
provided narrative extracts. Your summary should take the form of a Chairman's letter and focus on
key financial information, strategic highlights, and significant developments.

Further instructions:

    -   The summary should be comprehensive and well-structured, with a maximum length of 1000 words
    (aim for at least 800 words).
    -   Present any monetary amounts in euros € unless another currency is specified.
    -   Include all relevant details from the extracts.
    -   Format your response as plain text, avoiding headings and/or subheadings.

Narrative extracts:

{context}
```

Listing 3: Few-shot prompt used for summarisation of the FNS 2023 Greek reports.

```
You are an expert Greek financial analyst with extensive experience in preparing annual reports and
financial statements.

Please write a declaration of responsibility for a Greek company annual report using the provided
narrative extracts below.

Use the following examples to guide you, adjusting details like names and dates as necessary:

{examples}


Narrative extracts:

{context}
```

C.2 SummQQ

Listing 4: Zero-shot CoT prompt for assessing referential clarity of summaries written for public company annual reports.

```
You will be given one summary written for a public company annual report.
```

```
Your task is to rate the summary on one aspect of linguistic quality, providing a step-by-step
rationale for your rating.

Evaluation Criteria:

Referential clarity (1-5) - "It should be easy to identify who or what the pronouns and noun phrases
in the summary are referring to. If a person or other entity is mentioned, it should be clear what
their role in the story is. So, a reference would be unclear if an entity is referenced but its
identity or relation to the story remains unclear."

    - 1: Very Poor
    - 2: Poor
    - 3: Barely Acceptable
    - 4: Good
    - 5: Very Good


Summary:

{summary}


Evaluation Form:

- Rationale (with examples):
- Final Score (score ONLY):
```

Listing 5: Few-shot prompt for evaluating grammaticality of summaries written for series of news articles.

```
You will be given one summary written for a series of related news articles.

Your task is to rate the summary on one aspect of linguistic quality.

Evaluation Criteria:

Grammaticality (1-5) - "The summary should have no datelines, system-internal formatting,
capitalisation errors or obviously ungrammatical sentences (e.g., fragments, missing components)
that make the text difficult to read."

    - 1: Very Poor
    - 2: Poor
    - 3: Barely Acceptable
    - 4: Good
    - 5: Very Good


Examples:

Example Summary 1:

Greenhouse gas emissions - including carbon dioxide created by the burning of coal, gas and oil, are
believed by most atmospheric scientists to cause the warming of the Earth's surface and a change in
the global climate. "The thing that screams out here is that from the inter-tidal zone to the open
ocean, from the Antarctic to the Arctic, we are seeing worrisome things everywhere that seem to be
most closely correlated with climate change". The answer, many experts believe, may depend on how
much fresh water flows into the North Atlantic Ocean as a result of melting Arctic ice and the
runoff from an increase in Northern Hemisphere precipitation that some scientists say is already
resulting from global warming. "The speed of change caused by the change in climate is greater than
most ecosystems are going to be able to adapt to". "Significant loss of species must be considered
as one of the most important impacts of climate change," the study said. He quoted the Second
Assessment Report compiled by the Inter-Governmental Panel on Climate Change as saying that "as long
as emissions (of greenhouse gases into the atmosphere) are allowed to continue in a business as
usual manner, global warming will occur and occasion changes in the climate system". Dudley said
that the climate change or global warming up threatens the wellbeing of all mankind, and that the
climate change impacts are more evident every day in all the world. The report, Climate Change
Impacts on the United

Final Score: 1

Example Summary 2:
```

```
Among the spending Clinton will seek: $300 million to build at least six new reservation schools...

...

----------------


Summary:

{summary}


Evaluation Form (scores ONLY):

- Grammaticality:
```

D DUC Quality Questions

1. **Grammaticality** (*GRA*): "The summary should have no datelines, system-internal formatting, capitalisation errors or obviously ungrammatical sentences (e.g., fragments, missing components) that make the text difficult to read."
2. **Non-redundancy** (*NRE*): "There should be no unnecessary repetition in the summary. Unnecessary repetition might take the form of whole sentences that are repeated, or repeated facts, or the repeated use of a noun or noun phrase (e.g., "Bill Clinton") when a pronoun ("he") would suffice."
3. **Referential clarity** (*REF*): "It should be easy to identify who or what the pronouns and noun phrases in the summary are referring to. If a person or other entity is mentioned, it should be clear what their role in the story is. So, a reference would be unclear if an entity is referenced but its identity or relation to the story remains unclear."
4. **Focus** (*FOC*): "The summary should have a focus; sentences should only contain information that is related to the rest of the summary."
5. **Structure and Coherence** (*COH*): "The summary should be well-structured and well-organised. The summary should not just be a heap of related information, but should build from sentence to sentence to a coherent body of information about a topic."

E Model Specification and Training Arguments

Parameter	Value
Name	`gte-multilingual-base`
Size (parameters)	305M
Embedding dimensions	768
Max. input tokens	8192

Table 10: Base model information.

Parameter	Value
Base model	`gte-multilingual-base`
Epochs	3
Batch size	32
Grad. accum. steps	16
Learning rate	2×10^{-5}
Warm-up ratio	0.1
Weight decay	0.01
LR scheduler	Cosine
Mixed precision	Enabled

Table 11: Training parameters used for `gte-multilingual-base-fns`

F Example Summaries

The summaries below are of the 2022 annual report of Barratt Developments PLC, included in the FNS 2023 English test set. We display the summary generated by each baseline system (excluding the FNS

baselines, as we do not have access to their original generations) and by PRAGSum in EXT and ABS modes. The LLM employed for the abstractive summaries is GPT-4o mini.

F.1 PRAGSum Extractive

of strong housing demand. I would, once capacity to deliver more than 1,000 home by nutrient neutrality. We are currently again, like to thank our employees, sub- completions per year. engaging with the consultation around contractors and supply chain partners for future planning reform. We would urge the To support our site-based construction their hard work and commitment, which Government to ensure any changes deliver activity, address the longer-term challenge enabled us to successfully grow our site- a planning system that is responsive to of labour availability in the industry based construction activity, notwithstanding housing need, predictable and timely, and and build the most energy-efficient and the significant supply chain challenges, well-resourced at local authority level, to sustainable homes for the future, Oregon, and deliver high-quality homes and great ensure a flow of consented land, which will our in-house timber frame manufacturing the Boards agenda. Full details around conditions. engagement during the year can be found On behalf of the Board, I would like to in pages 41 to 51. Following the excellent performance of the thank you for the confidence you have business throughout FY22 and our strong shown in the Group during the past year and resilient balance sheet, the Board has and for your continued support. approved a return of surplus capital of 200m in FY23 through the implementation of a share buyback programme which will John Allan start shortly with an initial tranche of 50m Chairman to be completed by the end of the calendar 6 September 2022 year and the total programme completed no later than 30 June 2023. www.barrattdevelopments.co.uk 09 England moving annual planning consents and net new build home additions ('000s) STRATEGIC REPORT Marketplace were 194,060 in the last reported 12-month has not been maintained and in the 12 UK economy 6 our build quality and customer service. The the first time we have won this award and Board continues to seek ways of further We delivered 17,908 high quality, energy reaffirms both our progress to date and developing and advancing the positive efficient new homes (including JVs) across our commitment to be the leading national culture of our business and recognises Britain in FY22. This performance is 3.9% sustainable housebuilder. that the Groups culture is driven by its ahead of last year and also ahead of the More information on our sustainability leadership. For further information, see 17,856 homes we completed pre-pandemic strategy is included in the Chief Executives page 80. in FY19. We achieved adjusted profit statement on pages 16 to 33. before tax of 1,054.8m, a new record for Building sustainably the Group. Building safety Our Building Sustainably framework is the I would like to express my thanks to all our We have always been clear that we do not Sheffield and Anglia, in our Northern The land market remains attractive with David Thomas and East regions respectively to support a steady supply of opportunities. Despite Chief Executive our future growth. Both divisions are dual some planning delays during the year, branded, offering both Barratt and David planning consents have remained ahead of Wilson homes and, following a period home building activity at a national level. of land bank assembly, offer attractive Planning delays are however becoming opportunities for additional growth over more commonplace, reflecting constrained the coming years. Once operating at scale, planning resources, the delayed impacts Introduction over the next five to seven years, we believe of the pandemic and emerging land use We have made excellent progress in a year these two divisions combined will have the issues, notably the challenges created of strong housing demand. I would, once capacity to deliver more than 1,000 home House Laboratories, University of Salford www.barrattdevelopments.co.uk 15 STRATEGIC REPORT Chief Executives statement We continue to lead the industry on At the end of January 2022, we acquired sustainability, with a particular focus on Gladman Developments Limited. Gladman reducing our environmental impact, and is the countrys largest land promoter, wehave clear targets and plans for the which brought into the Group an industry- years ahead. leading team of experts in land sourcing, promotion and planning. Gladman, at the Housing market fundamentals time of its acquisition, held a portfolio of Despite the continued macroeconomic 406 land promotion sites encompassing uncertainties, the housing market more than 98,000 plots, which will provide fundamentals remain attractive. Strong an additional route to both grow the demand for high-quality, energy-efficient Groups strategic land bank and accelerate homes has been evident across the UK the strategic land bank conversion. 160 bps increase in adjusted Ongoing build optimisation Land acquisition at a Gross margin gross margin to 24.8% (FY21: and focus on build cost minimum 23% gross 23.2%). inflation control. margin and ongoing build optimisation and 390 bps decrease in gross Delivering continued performance. margin to 17.1% (FY21: operational improvements 21.0%). across our business. 220 bps increase in ROCE to Disciplined and controlled Minimum of 25% delivered ROCE 30.0% (FY21: restated 27.8%). land and work in progress through continued operating investment to support growth. framework discipline. www.barrattdevelopments.co.uk 17 STRATEGIC REPORT Chief Executives statement CONTINUED We are working with the HBF and the Keeping people safe The Barratt Foundation Department for Levelling Up, Housing Now in its second year of operation, the Our fundamental priority is always to and Communities (DLUHC) to agree Barratt Foundation was particularly active provide a safe environment for all our builds and places it creates. Its record in practice have each taught me something winning so many Pride in the Job awards different that I can bring to the role at shows the commitment of site teams Barratt, including my understanding of & to delivering the highest standards of the housebuilding sector. Ill be using my Q A construction, safety on site and customer experience to ensure our finance function satisfaction. Barratt is also a clear thought Mike joined the Company continues to bring actionable insight to the leader on sustainability in housebuilding business whilst maintaining our strong in December 2021 from and is helping to

F.2 PRAGSum Abstractive

Barratt Developments PLC has reported a robust performance for the financial year 2022 (FY22), demonstrating resilience amid ongoing macroeconomic uncertainties. The company completed 17,908 high-quality, energy-efficient homes, marking a 3.9% increase from the previous year and surpassing pre-pandemic levels. This achievement is underpinned by a strong demand for housing, which remains a critical focus for the UK government, aiming to build 300,000 new homes annually.

Financially, Barratt achieved an adjusted profit before tax of £1,054.8 million, setting a new record for the Group. The gross margin improved to 24.8%, up from 23.2% in FY21, reflecting effective cost management and operational efficiencies. The return on capital employed (ROCE) also saw a significant increase, rising to 30.0% from a restated 27.8% in the previous year. The company's strong balance sheet and cash generation capabilities have enabled the Board to approve a return of surplus capital amounting to £200 million through a share buyback program, with an initial tranche of £50 million expected to be completed by the end of the calendar year.

Strategically, Barratt is committed to sustainability and has made significant strides in reducing its environmental impact. The company's Building Sustainably framework focuses on energy efficiency and waste reduction, aligning with broader industry goals to combat climate change. Barratt has set clear targets for the coming years, aiming to enhance its sustainability practices while maintaining high standards of customer service and build quality. The company was recognized as the Sustainable Housebuilder of the Year at the Housebuilder Awards 2021, highlighting its leadership in the sector.

The acquisition of Gladman Developments Limited, the UK's largest land promoter, has further strengthened Barratt's position in the market. This acquisition brought in a portfolio of 406 land promotion sites, providing access to over 98,000 plots, which will facilitate the growth of the Group's strategic land bank. The integration of Gladman is expected to enhance Barratt's land sourcing and planning capabilities, allowing for incremental completions of 500 homes per annum from FY25.

Despite challenges such as planning delays and supply chain issues, Barratt remains optimistic about future growth. The company has opened two new divisions in Sheffield and Anglia, which are expected to contribute significantly to its operational capacity. The land market continues to present attractive opportunities, with planning consents remaining ahead of homebuilding activity at a national level.

Barratt's commitment to its employees is evident in its focus on fostering a positive workplace culture and investing in talent development. The company aims to attract and retain top talent by promoting diversity and inclusion, alongside providing career development opportunities. The Barratt Foundation has also been active in supporting students facing financial hardship, demonstrating the company's commitment to social responsibility.

Looking ahead, Barratt has set ambitious targets for FY23, expecting to grow total home completions to between 18,400 and 18,800 homes, including approximately 750 from joint ventures. The company is focused on maintaining its operational improvements and ensuring a disciplined approach to land acquisition and investment.

In summary, Barratt Developments PLC has demonstrated strong financial performance and strategic growth in FY22, underpinned by a commitment to sustainability, operational excellence, and employee engagement. The company is well-positioned to navigate the challenges of the housing market while contributing to the UK's housing needs through the delivery of high-quality, energy-efficient homes.

F.3 No-Context

Barratt Developments PLC has delivered a robust performance in its latest annual report, showcasing resilience and strategic growth in a challenging economic environment. The company reported a revenue of £4.8 billion, reflecting a slight increase from the previous year, driven by strong demand for new homes and a focus on quality and customer satisfaction. The operating profit stood at £1.1 billion, resulting in an operating margin of 23%, which underscores the efficiency of Barratt's operations and its ability to manage costs effectively.

The company's pre-tax profit reached £1 billion, demonstrating a solid return on investment and a commitment to delivering value to shareholders. Barratt declared a total dividend of 32.4 pence per share, an increase of 5% compared to the prior year, highlighting its strong cash generation capabilities and commitment to returning capital to shareholders. The company's net cash position improved to £1.2 billion, providing a strong financial foundation for future growth and investment.

Strategically, Barratt has continued to focus on its commitment to sustainability and environmental responsibility. The company has made significant strides in reducing its carbon footprint, with a target to achieve net-zero carbon emissions by 2040. This commitment is reflected in its investment in energy-efficient building practices and the use of sustainable materials in construction. Barratt has also enhanced its community engagement initiatives, ensuring that new developments meet the needs of local residents and contribute positively to the communities in which they operate.

In terms of operational performance, Barratt completed the construction of over 17,000 homes during the year, maintaining its position as one of the largest housebuilders in the UK. The company has successfully navigated supply chain challenges and labor shortages, demonstrating agility and resilience in its operations. Barratt's focus on quality has been recognized through various industry awards, reinforcing its reputation as a trusted builder.

The company has also made significant progress in its land acquisition strategy, securing over 10,000 plots for future development. This proactive approach positions Barratt well to capitalize on the ongoing demand for housing, particularly in key growth areas across the UK. The company's strategic land bank now stands at approximately 80,000 plots, providing a solid pipeline for future growth.

Barratt has embraced digital transformation as part of its strategic initiatives, investing in technology to enhance customer experience and streamline operations. The introduction of virtual tours and online sales platforms has improved accessibility for potential buyers, particularly in the wake of the COVID-19 pandemic. This digital focus not only enhances customer engagement but also drives operational efficiencies.

The report also highlights Barratt's commitment to diversity and inclusion within its workforce. The company has implemented various initiatives aimed at promoting a diverse workplace, recognizing that a varied team contributes to better decision-making and innovation. Barratt's efforts in this area have been acknowledged through various awards, reflecting its dedication to fostering an inclusive culture.

Looking ahead, Barratt Developments remains optimistic about the housing market, despite potential economic headwinds. The company anticipates continued demand for new homes, driven by demographic trends and government initiatives aimed at increasing housing supply. Barratt is well-positioned to navigate any challenges, thanks to its strong financial position, strategic land bank, and commitment to quality and sustainability.

In summary, Barratt Developments PLC has demonstrated strong financial performance and strategic growth in its latest annual report. With a focus on sustainability, operational efficiency, and customer satisfaction, the company is well-equipped to meet the evolving needs of the housing market. Its commitment to returning value to shareholders, coupled with a proactive approach to land acquisition and digital transformation, positions Barratt for continued success in the years to come.

F.4 Long-Prompt

Barratt Developments PLC's Annual Report and Accounts for 2022 outlines the company's robust performance and strategic initiatives aimed at leading the future of housebuilding in the UK. The report emphasizes the company's commitment to sustainability, customer satisfaction, and operational excellence, while also addressing the challenges posed by the current economic environment.

In the financial year ending June 30, 2022, Barratt achieved total home completions of 17,908, a 3.9% increase from the previous year (FY21: 17,243). This includes 13,327 private completions and 3,835 affordable homes, reflecting a strong demand for high-quality, energy-efficient housing. The company reported an adjusted profit before tax of £1,054.8 million, up from £919.7 million in FY21, marking a record performance for the Group. The adjusted gross margin improved to 24.8%, driven by house price inflation outpacing build cost inflation, which was reported at 17.1% for the year.

Barratt's strategic priorities focus on putting customers first, creating great places, leading construction, and investing in people. The company has been recognized for its commitment to customer service, achieving a 5-star rating in the Home Builders Federation (HBF) customer satisfaction survey for the 13th consecutive year. The report highlights the importance of customer feedback in shaping the company's offerings and improving the home-buying experience.

Sustainability is a core component of Barratt's strategy, with the company aiming to achieve net-zero carbon emissions across its operations by 2040. The report details the Group's Building Sustainably framework, which integrates environmental, social, and governance (ESG) considerations into its business model. Barratt has committed to reducing its greenhouse gas emissions by 29% by 2025 from 2018 levels and has made significant progress in this area, achieving a carbon intensity of 1.53 tCO2e per 100m² of legally completed build area.

The acquisition of Gladman Developments in January 2022 has further strengthened Barratt's position in the market by enhancing its land promotion capabilities. This strategic move is expected to contribute to the Group's growth by providing access to a larger portfolio of land and development opportunities.

The report also addresses the challenges posed by the current economic climate, including rising inflation and interest rates, which could impact housing demand and affordability. Barratt is actively monitoring these trends and has implemented measures to mitigate potential risks, such as adjusting its land acquisition strategy and focusing on operational efficiencies.

In terms of governance, Barratt has established a Sustainability Committee to oversee its climate-related initiatives and ensure alignment with the Taskforce for Climate-related Financial Disclosures (TCFD) recommendations. The Group is committed to transparency in its reporting and has undertaken a thorough review of its climate-related risks and opportunities.

The report concludes with a positive outlook for the future, emphasizing Barratt's strong financial position, substantial net cash reserves of £1,138.6 million, and a well-diversified land bank. The company is poised to continue its growth trajectory, with a target of 21,500 total home completions in the medium term, while maintaining its commitment to sustainability and customer satisfaction.

Overall, Barratt Developments PLC's Annual Report for 2022 reflects a company that is not only focused on delivering financial performance but is also dedicated to creating sustainable communities and enhancing the customer experience in the UK housing market.

F.5 Naive RAG

Barratt Developments PLC Annual Report Summary 2022

Overview:
Barratt Developments PLC, the UK's largest housebuilder, has demonstrated resilience and strategic growth in its 2022 annual report. The company has focused on delivering high-quality, sustainable homes while navigating economic challenges, including rising energy costs and inflation.

Key Financial Highlights:
- **Home Completions:** Total home completions increased by 3.9% to 17,908, surpassing the pre-pandemic level of 17,856 homes achieved in FY19. Wholly owned completions also rose by 3.9% to 17,162 homes.
- **Financial Position:** The company reported a year-end net cash position of £1,138.6 million, down from £1,317.4 million in FY21. However, net indebtedness improved to a surplus of £405.0 million, compared to £659.1 million in the previous year.
- **Gross Margin:** Barratt maintained a gross margin target of at least 23%, reflecting its commitment to operational efficiency and quality.
- **Return on Capital Employed (ROCE):** ROCE increased by 220 basis points to 30.0%, indicating effective

capital utilization.
- **Total Comprehensive Income:** The total comprehensive income for the year was £515.2 million, a decrease from £658.3 million in FY21, primarily due to macroeconomic pressures.

Strategic Highlights:
- **Sustainability Commitment:** Barratt is committed to sustainable development, aligning with the UK Government's Streamline Energy and Carbon Reporting (SECR) requirements. The company aims to enhance biodiversity and local environments through its developments.
- **Growth Targets:** The company has set a new medium-term target of 21,500 home completions, reflecting its ambition to address the housing shortage in the UK.
- **Employee Engagement:** Barratt continues to invest in its workforce, with a focus on attracting and retaining talent. The Barratt Foundation provided £111,000 in scholarships and bursaries to support students facing financial hardship.

Significant Developments:
- **Regulatory Changes:** The introduction of the Residential Property Developer Tax (RPDT) from April 2022 resulted in a charge of £8.8 million, applicable to profits exceeding £25 million annually for a ten-year period.
- **Operational Resilience:** The company has adapted to post-pandemic challenges and macroeconomic impacts, including those stemming from the war in Ukraine, while maintaining a focus on build quality and customer service.

Conclusion:
Barratt Developments PLC's annual report for 2022 reflects a strong financial position and a commitment to sustainable growth. The company is well-positioned to meet its ambitious targets while navigating the complexities of the current economic landscape. With a focus on quality, efficiency, and stakeholder engagement, Barratt continues to play a pivotal role in the UK housing market.

F.6 Self-RAG

Barratt Developments PLC is a leading UK housebuilder that has been delivering high-quality homes for over 60 years. The company's annual report for the year ended 31 December 2022 provides a comprehensive overview of its financial performance, strategic highlights, and significant developments. Key Financial Information:

Barratt Developments' financial performance for the year ended 31 December 2022 was strong, with revenue increasing by 11% to £3. 8 billion and profit before tax rising by 21% to £1. 2 billion. The company's net cash generation increased by 22% to £1. 1 billion, and its net debt was reduced by £1. 2 billion to £1. 1 billion. The company's dividend per share increased by 10% to 22. 5 pence. Strategic Highlights:

Barratt Developments' strategic priorities for 2022-2025 include:

*Delivering high-quality homes and creating thriving communities
*Continuing to invest in innovation and technology to improve efficiency and customer experience
*Ensuring a sustainable approach to development and construction
*Continuing to invest in talent and diversity

Significant Developments:

During the year, Barratt Developments made significant progress in several areas. The company launched its new brand, "Barratt Homes," which reflects its commitment to building high-quality homes and creating thriving communities. The company also launched its "Building for Tomorrow" initiative, which aims to reduce the carbon footprint of its new homes by 50% by 2030. In addition, Barratt Developments acquired the land and assets of the former Bovis Homes Group, which expanded the company's presence in the South of England and provided opportunities for growth. Conclusion:

Barratt Developments' annual report for the year ended 31 December 2022 provides a comprehensive overview of the company's financial performance, strategic highlights, and significant developments. The company's financial performance was strong, with revenue, profit before tax, and net cash generation all increasing. The company's strategic priorities for 2022-2025 focus on delivering high-quality homes, innovation and technology, sustainability, and talent and diversity. The acquisition of the land and assets of the former Bovis Homes Group provides opportunities for growth and expansion.

G Cost Analysis of PRAGSum vs. Long-Prompt Baseline

We investigate the extent to which PRAGSum (in abstractive mode) diminishes the cost of generating summaries of FNS 2023 reports compared to our long-prompt baseline. Amounts are based on cost per input token and do not include output token costs which we presuppose as being equivalent. We truncated reports to 60k words before counting the tokens in the prompt. For PRAGSum, we retrieved the top ten chunks for each report and concatenated them before inserting into the prompt. The tokeniser for Claude 3 Haiku is not publicly available, and thus we omit Claude from this analysis. Our results are displayed in Table 12.

Model	Language	Cost		Reduction
		Long-Prompt	*PRAGSum*	
GPT	EN	$ 5.92	$ 0.20	96.6%
	ES	$ 0.28	$ 0.01	96.4%
Gemini	EN	$ 3.35	$ 0.11	96.7%
	ES	$ 0.15	$ 0.01	93.3%
Mistral	EN	$13.88	$ 0.45	**96.8%**
	ES	$ 0.63	$ 0.03	95.2%

Table 12: Costs of generating summaries of FNS 2023 English and Spanish test set reports with PRAGSum compared to long-prompting. The stated cost is over the entire set.

We find that PRAGSum consistently presents more than a 90% reduction in the total cost of input tokens compared to long-prompting whilst constructing summaries that match or exceed the quality of the long-prompt summaries. Whilst long-prompting results in strong summarisation performance with three of the four LLMs we tested, our system offers significant cost efficiencies without sacrificing quality, rendering it a more scalable and economically viable solution for large-scale summarisation tasks.