

SUSGEN-GPT: A Data-Centric LLM for Financial NLP and Sustainability Report Generation

Anonymous ACL submission

Abstract

The rapid growth of the financial sector and the increasing emphasis on Environmental, Social, and Governance (ESG) considerations have highlighted the need for advanced natural language processing (NLP) tools. Despite significant advancements, there remains a lack of open-source Large Language Models (LLMs) proficient across both general finance and ESG domains, such as generating ESG reports. To address this gap, we propose SUSGEN-30K, a high-quality, category-balanced dataset that comprises seven financial NLP tasks and ESG report generation. Additionally, we propose TCFD-BENCH, a benchmark designed to enhance the evaluation of sustainability report generation. Employing a data-centric methodology, we developed a suite of models, referred to as SUSGEN-GPT. When trained on our curated dataset, these suites of models achieved state-of-the-art performance, surpassing the benchmarks set by models of significantly larger size. By doing so, we introduce a data-centric approach to effectively address the aforementioned existing challenges, aiming to fostering continual development in the financial and ESG research community.

1 Introduction

As automation in the financial sector gains more attention and climate change becomes an increasingly urgent issue, there is a growing need for more advanced tools capable of solving these problems. The ability to efficiently process and generate financial reports, analyze ESG metrics, and produce comprehensive TCFD-format¹ reports is crucial for maintaining transparency and accountability. Recently, LLMs (Brown et al., 2020; Ouyang et al., 2022; Touvron et al., 2023a; OpenAI, 2023a; Tou-

¹<https://www.fsb-tcfd.org/>. We investigated various sustainability reporting guidelines, including GRI, SASB, EU CSRD, etc., and ultimately chose TCFD because it is more standardized and universally applicable.

vron et al., 2023b) have gained widespread popularity due to their ability to solve complex tasks such as commonsense reasoning and machine translation. However, there remains a substantial gap in LLMs tailored towards specialized domains such as finance and ESG (Liu et al., 2023; Wang et al., 2023a). This is likely due to the training dataset being exposed to mostly general text documents and lack data of specialized information.

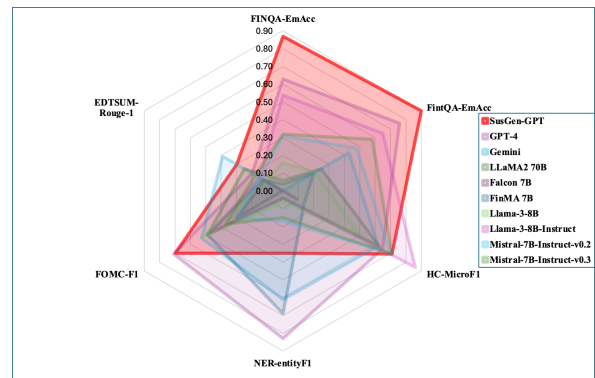


Figure 1: Performance comparison with other baseline models on various financial NLP tasks. The radar graph shows that the best model in SUSGEN-GPT achieves state-of-the-art in most benchmarks.

In light of these challenges, we propose SUSGEN-30K, a meticulously curated dataset that is designed to balance the size of each downstream task across both financial and ESG domains. These includes tasks such as Sentiment Analysis (SA), Named Entity Recognition (NER) for finance and Task Force on Climate-related Financial Disclosures (TCFD (Financial Stability Board, 2017)) report generations. More importantly, we introduce a suite of LLMs, trained on our proposed dataset, which we refer to as SUSGEN-GPT. Due to the balanced nature behind the training data, SUSGEN-GPT is capable of achieving superior performance across multiple downstream tasks simultaneously, when compared against models that are orders of

magnitude larger.

Additionally, we propose a new benchmark, TCFD-BENCH, tailored specially to assess Annual reports on their ability to generate concise and accurate ESG reports. We likewise conduct experiments on the proposed benchmark using SUSGEN-GPT. In total, our contributions include the following:

1. SUSGEN-30K, a large-scale high quality dataset in both financial & ESG domain.
2. A suite of fine-tuned LLMs, SUSGEN-GPT, shown to achieve state-of-the-art performance across both general financial and ESG NLP benchmarks as shown in Figure 1.
3. We propose and release a well-curated benchmark, tailored towards ESG report generation, TCFD-BENCH.

2 Related Work and Background

NLP for Finance & ESG Natural Language Processing (NLP) has found extensive applications in various financial tasks, demonstrating its versatility and depth in addressing diverse financial issues (Masson and Paroubek, 2024; Aguda et al., 2024). The key tasks in the financial domain include Question Answering (QA), Headline Classification (HC), and report generation. More notably, there exists a gap in achieving an acceptable level of proficiency in automating the generation of ESG reports. One such effort, ChatReport (Ni et al., 2023) is developed to perform summarization and analysis on ESG reports. However, these tools face challenges such as generating reports that are overly simplified and lacking important details. Other attempts (Bronzini et al., 2024; Zou et al., 2023; Luccioni et al., 2020) mainly rely on existing data extraction techniques and face difficulty in processing unstructured data. Our dataset aims to bridge these gaps by providing a data-centric approach to training LLMs in a multi-task manner.

General Large Language Models Given the increased accessibility to large amounts of publicly available data, there has been a constant upward trend in releasing instruct-tuned LLMs. These models include Alpaca (Taori et al., 2023), an LLM trained on a dataset augmented with GPT-3. The latest model to join the open-source community is Mistral-v0.3 (Jiang et al., 2023) and Llama3 (AI@Meta, 2024), which have demonstrated impressive human-like capabilities across various domains. However, these models are not tailored

towards any specific domain and often underperform on specialized domains such as finance and ESG.

Financial Large Language Models Financial Large Language Models (FinLLMs) are specifically developed to handle financial text data, offering more precise financial analysis and predictions. One of the earlier efforts, BloombergGPT (Wu et al., 2023), is a 50B model trained on a massive dataset comprising a mixture of financial and general text data. However, it is not publicly accessible and hence there is a call for more open and inclusive alternatives. Other open-source alternatives includes FinGPT (Yang et al., 2023; Liu et al., 2023) and CFGPT (Lei et al., 2024), which introduce tools focused on data acquisition, cleaning, and preprocessing. Their goal is to democratize financial data and the development of FinLLMs, offering a wide range of potential applications. Nonetheless, these efforts have not addressed issues such as the imbalance in training data and lack capabilities in the ESG domain. CFGPT faces limitations in language such as only being limited to the Chinese language.

Financial Benchmarks As FinLLMs rapidly advance, the importance of financial evaluation benchmarks has grown significantly. These benchmarks include FinGPT Benchmarks and FLUE, focused on assessing NLP tools on a wide array of tasks such as NER and SA. PIXIU (Xie et al., 2023, 2024) is a large-scale multitask dataset containing 136K data samples as well as offering benchmarks covering five downstream tasks. However, these evaluation frameworks lack specialized ESG content. We aimed to bridge this gap by introducing TCFD-BENCH.

3 SUSGEN-GPT

3.1 Framework

The SUSGEN-GPT framework is designed to handle both document processing and direct question-answering in the financial and ESG domains. These models are trained to process raw unstructured documents or answer financial questions directly.

To process raw documents, our pipeline utilizes Retrieval-Augmented Generation (RAG) technology for summarization, followed by the Report Summarization (RS) module to generate the ESG report. The RAG component retrieves relevant information from a vector database and synthesizes

161 it to create a comprehensive summary of the com- 206
162 pany’s Annual report. This summary is then pro- 207
163 vided as context to the RS Module in the input 208
164 prompt to generate an ESG report. The provided 209
165 summary ensures the generated report complies 210
166 with TCFD standards. Additionally, the model is 211
167 capable of answering ESG-related queries concern-
168 ing the company report.

169 To provide an overview of SusGen-GPT’s capa- 212
170 bilities, it is equipped to handle the following finan- 213
171 cial tasks: sentiment analysis Relation Extraction 214
172 (RE), named entity recognition, headline classifi- 215
173 cation, Financial Question Answering (FIN-QA), 216
174 Financial Table Question Answering (FIN-TQA), 217
175 and Text Summarization (SUM). These function- 218
176 alities are illustrated in Figure 2, which depicts 219
177 the overall pipeline of the model, from user input 220
178 and internal processing to report generation and 221
179 answering questions. 222

180 3.2 Task Definition 218

181 SUSGEN-GPT is designed to perform a wide 219
182 range of tasks within the financial and ESG do- 220
183 mains. These tasks include: 221

- 184 • **Headline Classification:** Categorizing news 222
185 headlines or report titles into predefined 223
186 classes. 224
- 187 • **Named Entity Recognition (NER):** Identifying 225
188 and classifying proper nouns such as 226
189 names of individuals, organizations, locations, 227
190 and financial entities. 228
- 191 • **Relation Extraction:** Detecting relationships 229
192 between entities in text, such as acquisitions, 230
193 partnerships, or regulatory actions. 231
- 194 • **Sentiment Analysis:** Assessing sentiment in 232
195 text to determine if it is positive, negative, or 233
196 neutral. 234
- 197 • **Financial Question Answering:** Providing 235
198 accurate answers to queries related to financial 236
199 data and reports. 237
- 200 • **Financial Table Question Answering:** Ex- 238
201 tending question answering to structured data 239
202 by extracting and interpreting information 240
203 from financial tables. 241
- 204 • **Text Summarization:** Generating concise 242
205 summaries of longer documents. 243

- **Sustainability Report Generation:** Creating 206
comprehensive and coherent reports on a com- 207
pany’s environmental, social, and governance 208
(ESG) performance by integrating data from 209
various sources and adhering to guidelines 210
such as the TCFD recommendations. 211

As Appendix B, figure 8 provides short exam- 212
ples of financial tasks handled by our model. These 213
tasks are essential in the financial domain for ex- 214
tracting, analyzing, and summarizing critical infor- 215
mation, which aids in decision-making and report- 216
ing processes. 217

218 3.3 Data Construction 218

219 **SUSGEN-30K** The data construction process for 219
220 SUSGEN-30K involves a comprehensive pipeline 220
221 that starts with data collection from various sources 221
222 such as company reports (including annual and 222
223 ESG reports), publicly available financial datasets, 223
224 and automated content crawlers that scrape finan- 224
225 cial data from the web, shown as the Figure 5. 225

226 This raw data undergoes thorough preprocessing 226
227 steps, including manual annotation to extract use- 227
228 ful content, machine-translated data to augment the 228
229 dataset with multilingual data, and other augmen- 229
230 tation techniques to generate novel data samples. 230
231 We also include anonymization to remove sensitive 231
232 information and comply with privacy regulations. 232
233 Finally, the preprocessed data is reformatted into a 233
234 format compatible with the SFT dataset, ensuring 234
235 it is well-balanced and ready for training models 235
236 in financial NLP and Sustainability Report Gener- 236
237 ation. This structured approach ensures that the 237
238 dataset is robust, diverse, and high-quality, suitable 238
239 for advancing the field of sustainable finance. The 239
240 collected data are then split into the seven afore- 240
241 mentioned financial tasks in Sec. 3.2. We introduce 241
242 an additional task in the ESG domain, Sustainabil- 242
243 ity Report Generation (SRG). 243

244 Inspired by the scaling law (Kaplan et al., 2020) 244
245 and Common-7B (Li et al., 2024), we perform 245
246 scaling on the dataset to address the imbalance 246
247 in sample size between the different tasks in the 247
248 dataset. For large-scale category data, we down- 248
249 sample them based on data quality to create a well- 249
250 balanced dataset. Finally, we concatenate all the 250
251 samples to form the SUSGEN-30K instruction 251
252 dataset, which can be used for the financial NLP 252
253 domain. For detailed information regarding the 253
254 data sources and composition, please refer to Fig- 254
255 ure 3 in Appendix A. 255

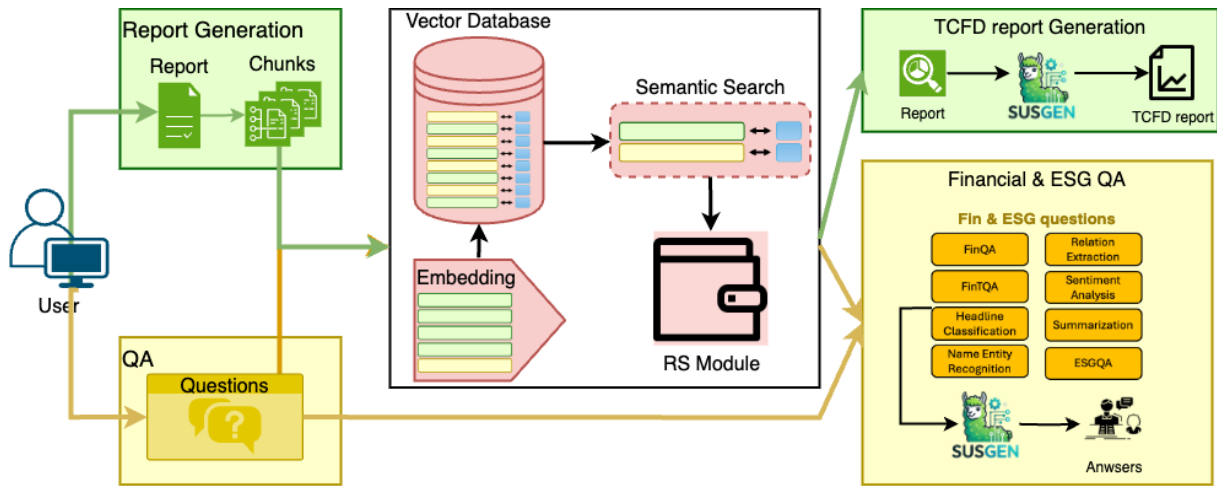


Figure 2: Overview of the SUSGEN-GPT Pipeline.

File: “Wolfspeed_2022.pdf”

Context: Wolfspeed, Inc. is a leading global company specializing in the development and production of silicon carbide (SiC) technology..... Company’s corporate governance structure is designed to ensure transparency, integrity, and accountability in all its operations.....
(All 205 words)

Input: As an expert in TCFD, with the context information, you are required to answer the following input question on the topic of governance. Describe the organization 2019’s governance around climate-related risks and opportunities.

Output: The company’s board of directors is responsible for all sustainability matters, including climate change.....
(All 292 words)

Figure 3: A sample in TCFD-BENCH. Models are expected to generate TCFD format sustainability report when given questions and context.

TCFD-BENCH This benchmark offers significant potential to streamline and enhance the quality of ESG reports in the TCFD format. A sample is shown in Figure 3, illustrating that each report includes three main parts: context, input, and output. The context section provides a detailed introduction to the company, outlining its specific structure regarding the given topic (governance). The input consists of the instruction and question, while the output presents the answer from the report. All text

is extracted using GPT-4o and expert analysis, as depicted in Figure 4. The context information is generated from annual reports using GPT-4o, while the TCFD reports are manually extracted to obtain the questions and answers. An additional instruction component is included to guide the model’s performance. This dual approach leverages both automated large language models and human expertise to build a comprehensive dataset for ESG reporting.

266
267
268
269
270
271
272
273
274
275

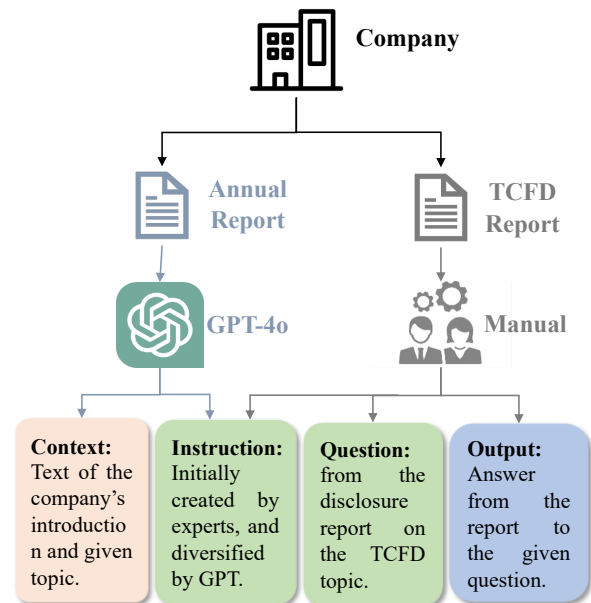


Figure 4: The pipeline of TCFD-BENCH construction for one company. Extraction from GPT-4o and experts are two mainly methods we used for the benchmark.

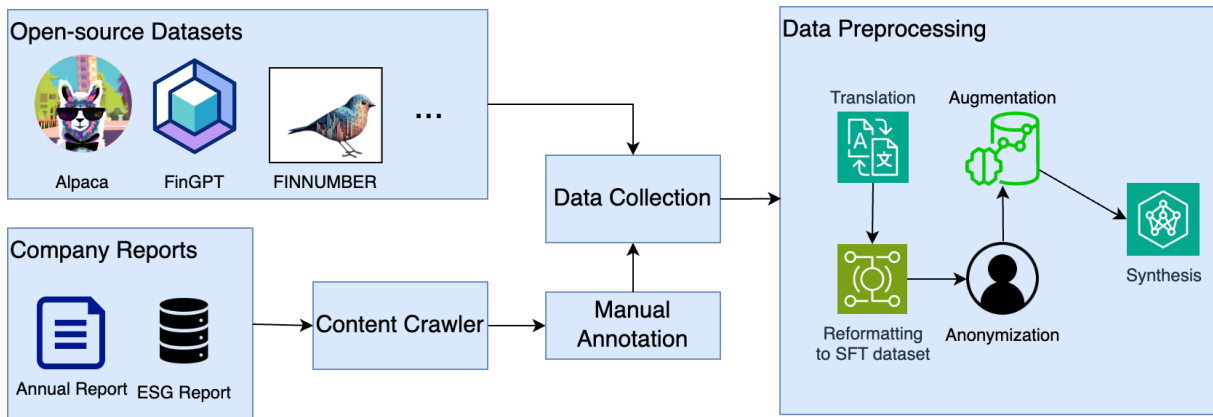


Figure 5: SUSGEN-30k Data Construction. The process involves collecting data from open-source datasets and company reports, followed by manual annotation and various preprocessing steps including translation, augmentation, anonymization, and reformatting to create the final SFT dataset.

3.4 Statistics

In this section, we present the statistical information about our training dataset, SUSGEN-30k. As illustrated in Figure 6, the dataset is well-balanced across various task categories, ensuring comprehensive coverage of financial and ESG domains.

The distribution of categories is as follows: Financial Question Answering (12%), Financial Table Question Answering (12%), Sentiment Analysis (12%), Relation Extraction (12%), Named Entity Recognition (11%), Headline Classification (11%), Sustainability Report Generation (10%), General (10%), and Math (10%). This balanced distribution allows the model to learn effectively from each task without bias, contributing to the robustness and versatility of SUSGEN-GPT in handling diverse financial and ESG-related tasks.

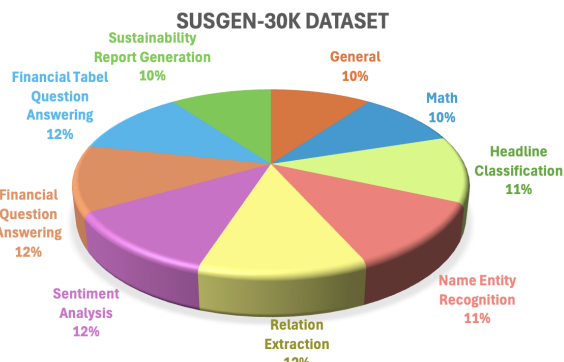


Figure 6: SUSGEN-30k Category Distribution. Highlight the proportion of data dedicated to each specific task area in financial NLP.

3.5 Evaluations

In this section, we introduce the metrics used to evaluate the performance of SUSGEN-GPT on various financial and ESG tasks. For Financial Question Answering and Financial Table Question Answering, we employed Exact Match Accuracy and F1 score to measure the precision of the answers. For Headline Classification, we used the Micro F1-score to balance precision and recall across all classes. Named Entity Recognition was assessed using the Entity F1-score to evaluate the accuracy of recognizing and classifying named entities. Sentiment Analysis used the F1-score and Accuracy to measure the balance between precision and recall for predicted sentiments. For Text Summarization, we utilized the ROUGE (Lin, 2004) and BERTScor (Zhang et al., 2019) to evaluate the quality of the summaries by comparing the overlap of unigrams between the generated and reference summaries. For sustainable report generation, BERTScore, ROUGE, METEOR (Banerjee and Lavie, 2005), and BLEU-N (Papineni et al., 2002) Score were used to assess model performance.

These metrics were used to measure the similarity of the machine-generated text to the expert reference content, ensuring that the evaluations are robust and reliable. The chosen metrics are relevant and effective in capturing the performance nuances of each task, offering a detailed view of how well the model performs in each area. Evaluation scores were calculated for each sub-task individually to provide specific insights into each task’s performance. Additionally, we provide the average score for each category to offer a holistic view of SUSGEN-GPT’s performance across dif-

ferent tasks, highlighting the model’s strengths and areas for improvement.

4 Experiments & Analysis

4.1 Experimental Setup

Our experiments employed the SUSGEN-GPT models, leveraging four baseline models: Mistral-v0.3-7B, Mistral-Instruct-v0.3-7B², LLaMA-3-8B, and LLaMA-3-8B-Instruct³, using the QLoRA (Dettmers et al., 2024) method for supervised fine-tuning due to computational efficiency. The experiments were conducted on two NVIDIA RTX 24GB 3090 Ti GPUs. We use different scale datasets of our curated SUSGEN-30K as the training data. During training, we employ 32-bit Paged AdamW optimizer with a cosine learning rate schedule for total 3 epochs of training. The learning rate is set to 2e-5, 10% warmup steps, 8 batchsize per device with 8 gradient accumulation steps. The maximum token length is set 2048 tokens with alpaca prompt template. And we use 4-bit quantization with double quantization enabled and bfloat16 as the compute data type, set lora rank to 16 and alpha to 32 with a dropout rate of 0.1. Out of twelve models we trained, the most resource-intensive one, involving 30K data records and 8B model, takes around 10 hours on our device.

During evaluation, We use same prompt as used in Training and combing vllm inference optimization techniques. We use LangChain⁴ to manage vector-database retriever. We use all-mpnet-base-v2⁵ for text chunk embedding, split reports into chunks of 1024 tokens and retrieve the top 10 related chunks. We set the temperature to 0.2, top_p to 0.9, top_k to 40 and repetition_penalty to 1.2.

4.2 Benchmarks & Baseline Models

In this section, we introduce the benchmarks and baseline models used to evaluate SUSGEN-GPT’s performance. Benchmarks consist of 14 datasets across 8 tasks to evaluate the performance of SusGen-GPT and baseline models.

Benchmarks For financial Q&A task, the FinQA (Chen et al., 2021) dataset focuses on multi-step numerical reasoning through financial reports. In

²The model is released by Mistral AI under the Apache 2.0 license for both commercial and non-commercial usage.

³LLaMA3 models are licensed under a bespoke commercial license by Meta AI.

⁴<https://python.langchain.com/>

⁵<https://huggingface.co/sentence-transformers/all-mpnet-base-v2>

financial table Q&A, the TATQA (Zhu et al., 2021) dataset addresses multi-step numerical reasoning through financial tables, while the ConvFinQA (Chen et al., 2022) dataset involves multiple rounds of Q&A based on earnings reports and tables. Sentiment analysis examines linguistic and economic meanings in financial texts, using FinQASA (Maia et al., 2018) for sentiment extraction and FOMC (Shah et al., 2023a) to categorize sentences as "hawkish" or "dovish." As for news headlines classification, the MultiFin (Jørgensen et al., 2023) dataset classifies financial texts like analyst reports, news stories, and investor commentary. The MLESG (Chen et al., 2023) dataset detects ESG issues. Named entity recognition extracts entities from financial agreements and SEC documents using NER (Alvarado et al., 2015) and FINER-ORD (Shah et al., 2023b) datasets. Relation extraction uses the FINRED (Sharma et al., 2022) dataset to identify relationships in financial news and earnings records, such as "products produced" and "manufacturers." The SC (Mariko et al., 2020) dataset discerns causal relationships in news and SEC filings. For text summarisation, EDTSUM (Zhou et al., 2021) dataset abstracts financial news articles into concise summaries. And in sustainability report generation, the AnnualReport dataset extracts TCFD-compliant ESG content from annual reports of listed companies, assessing the model’s ability to generate ESG quizzes and reports.

Baseline Models We compare against the most capable model from OpenAI, GPT-4 (OpenAI, 2023b) which excels across multiple NLP tasks. We use GPT-4 as a benchmark to compare SUSGEN-GPT’s performance. We also compare against Mistral 7B-Inst-V0.2/V0.3 (Jiang et al., 2023), which are open-source LLMs known for their general NLP task performance. LLaMA3 (AI@Meta, 2024) is a state-of-the-art open-source LLM. Developed by Meta, they enhance accuracy and efficiency in text generation and comprehension. Launched by Google, Gemini (Team et al., 2023) is a multimodal model processing text and images, improving cross-modal task performance. FinMA7B (Xie et al., 2023) is optimized for the financial sector, with 7 billion parameters, enhancing financial text analysis accuracy. Falcon7B (Almazrouei et al., 2023) is an open-source model with 7 billion parameters, designed for diverse NLP tasks, providing efficient text generation and comprehension.

4.3 Main Results and Comparison

The performance of SUSGEN-GPT on various financial tasks is summarized in Table 1. We evaluate SUSGEN-GPT on several financial NLP tasks using well-established benchmarks, comparing its performance against models like GPT-4, Gemini, and various versions of LLaMA and Mixtral.

Our model achieves competitive results across multiple benchmarks. For instance, in Sentiment Analysis, SUSGEN-GPT scores an F1 of 0.64 on the FiQASA dataset and 0.70 on the FOMC dataset, closely trailing behind GPT-4, which shows slightly higher scores. This demonstrates SUSGEN-GPT’s robust performance in capturing sentiment from financial texts.

In HC, SUSGEN-GPT outperforms several other models, achieving a Micro F1 score of 0.71 on the MultiFin dataset and 0.50 on the MLESG dataset, indicating its proficiency in accurately categorizing financial news headlines.

For NER, the model scores 0.35 and 0.09 for EntityF1 on the NER and FINER-ORD datasets, respectively. While these scores are lower compared to GPT-4, they highlight the potential for further improvement in entity recognition tasks.

In RE, SUSGEN-GPT excels with an F1 score of 0.96 on the SC dataset, significantly outperforming other models, though it shows a modest performance of 0.19 on the FinRED dataset.

For FinQA and FinTQA, SUSGEN-GPT demonstrates strong capabilities with an Exact Match Accuracy (EmAcc) of 0.87 on the FinQA dataset and 0.69 on the TATQA dataset, surpassing most competing models. This highlights the model’s effectiveness in understanding and accurately answering complex financial queries.

The performance of SUSGEN-GPT in generating sustainability reports is evaluated using the TCFD-BENCH dataset and compared against the CHATREPORT model which are shown in Table 2.

SUSGEN-GPT demonstrates superior performance across various metrics. It achieves a Rouge-L score of 0.18, a Bert-Score of 0.40, and a METEOR score of 0.22. These results indicate that SUSGEN-GPT generates more accurate and comprehensive sustainability reports compared to CHATREPORT, which trails in most metrics.

These findings underscore the advanced capabilities of SUSGEN-GPT in producing high-quality, TCFD-compliant sustainability reports, making it a valuable tool for organizations aiming to enhance

their ESG reporting.

4.4 Ablation Study

In this section, we investigate the data scaling effect on SUSGEN-GPT using datasets of 3k, 10k, and 30k scales. Comprehensive results are provided in Table 4 in the Appendix C. Figure 7 illustrates the performance trends for FinTQA and Sentiment Analysis tasks.

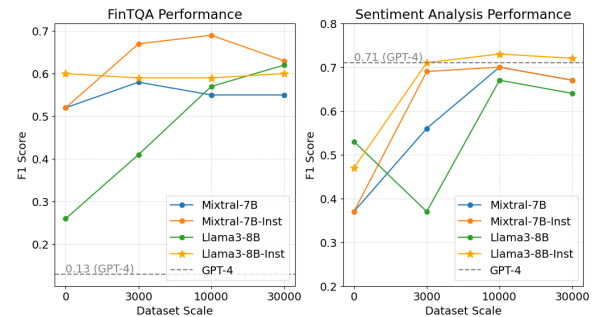


Figure 7: Performance of SUSGEN-GPT on FinTQA and sentiment analysis (SA) across different scales of supervised fine-tuning datasets.

The results indicate that increasing the dataset size significantly improves the performance of SUSGEN-GPT, especially after supervised fine-tuning. For FinTQA (left plot), the models show notable gains with larger datasets after fine-tuning: Mixtral-7B-Inst and Mixtral-7B models exhibit consistent improvement, with Mixtral-7B-Inst achieving the highest F1 score. Llama3-8B-Inst shows substantial gains from the base model, reaching an F1 score close to 0.83 at the 30k scale, suggesting strong potential. For Sentiment Analysis (right plot), Mixtral-7B-Inst leads in performance across all dataset scales. Llama3-8B-Inst also shows significant improvement, highlighting the benefits of larger datasets. These findings demonstrate that scaling the dataset enhances the model’s ability to capture complex patterns, leading to better performance across financial and ESG tasks. Larger datasets provide more diverse and comprehensive training data, crucial for achieving state-of-the-art results.

4.5 Qualitative Analysis

Generated Sustainability Report Analysis A comparison of the outputs obtained by SusGen-GPT and ChatReport based on the company’s annual report after inputting the same TCFD standard ESG-related questions is shown in Appendix D. ChatReport is a robust, publicly available text

Datasets	Tasks	Metrics	SusGen	GPT4	Gemini	LLaMA3 8B	LLaMA2 70B	LLaMA3 8B-Inst	FinMA 7B	Falcon 7B	Mixtral 7B-Inst-V0.2	Mixtral 7B-Inst-V0.3
FiQASA (Maia et al., 2018)	SA	F1	0.64	0.80	0.81	0.82	0.83	0.82	0.79	0.77	0.75	0.84
FOMC (Shah et al., 2023a)	SA	F1	0.70	0.71	0.53	0.53	0.49	0.47	0.49	0.30	0.30	0.37
MultiFin (Jørgensen et al., 2023)	HC	MicroF1	0.71	0.65	0.62	0.50	0.63	0.86	0.14	0.09	0.70	0.71
MLESG (Chen et al., 2023)	HC	MicroF1	0.50	0.35	0.34	0.23	0.31	0.48	0.00	0.06	0.57	0.59
NER (Alvarado et al., 2015)	NER	EntityF1	0.35	0.83*	0.61	0.06	0.04	0.04	0.69	0.00	0.17	0.15
FINER-ORD (Shah et al., 2023b)	NER	EntityF1	0.09	0.77	0.14	0.06	0.07	0.04	0.00	0.00	0.08	0.14
FinRED (Sharma et al., 2022)	RE	F1	0.19	0.02	0.00	0.04	0.00	0.08	0.00	0.00	0.13	0.14
SC (Mariko et al., 2020)	RE	F1	0.96	0.81	0.74	0.93	0.61	0.90	0.19	0.67	0.90	0.85
FinQA (Chen et al., 2021)	FINQA	EmAcc	0.87	0.63*	0.00	0.16	0.06	0.54	0.04	0.00	0.31	0.32
TATQA (Zhu et al., 2021)	FINTQA	EmAcc	0.69	0.13	0.18	0.26	0.01	0.60	0.00	0.00	0.48	0.52
ConvFinQA (Chen et al., 2022)	FINTQA	EmAcc	0.90	0.76*	0.43	0.21	0.25	0.65	0.20	0.00	0.48	0.58
EDTSUM (Zhou et al., 2021)	SUM	Rouge-1	0.30	0.20	0.39	0.11	0.25	0.15	0.13	0.12	0.15	0.18
		BertScore	0.55	0.67	0.72	0.41	0.68	0.47	0.38	0.51	0.48	0.49

Table 1: The zero-shot and few-shot performance of different LLMs on the general financial tasks. “*” represents the result from the previous paper.

Models	Rouge-L	Bert-Score	METEOR	BLEU-1	BLEU-2	BLEU-3	BLEU-4
ChatReport	0.14	0.32	0.12	0.41	0.10	0.03	0.02
SUSGEN-GPT	0.18	0.40	0.22	0.37	0.12	0.06	0.04

Table 2: The sustainability report generation performance on TCFD-BENCH with the comparison of CHATREPORT. All results via our evaluations are the average of three runs.

generation model for creating ESG reports in the TCFD format, leveraging the capabilities of GPT-3.5. However, SUSGEN-GPT surpasses ChatReport in several key areas. It offers a detailed breakdown, specific action points, enhanced contextual clarity, and a comprehensive approach.

SUSGEN-GPT’s structured format provides significant advantages, making it a superior tool for generating TCFD-format ESG reports. These improvements ensure that reports produced by SusGen-GPT are not only more informative but also more actionable and easier to understand. By breaking down the risks and opportunities into specific time horizons, SusGen-GPT provides a clear roadmap for addressing climate-related challenges and leveraging opportunities. Each identified risk and opportunity is linked to specific actions and goals, ensuring that the report is practical and aligned with the company’s strategic objectives. The enhanced contextual clarity provided by SusGen-GPT allows users to better understand the implications of each risk and opportunity, aiding stakeholders in making informed decisions. The detailed action points help in formulating concrete sustainability strategies, improving operational efficiency, enhancing reputation, and ensuring regulatory compliance.

In contrast, while ChatReport provides a broad overview of climate-related risks and opportunities, it lacks the depth and specificity found in SUSGEN-GPT’s output. ChatReport’s responses

are less structured and do not consistently associate risks with specific time horizons, which can make it harder for users to prioritize actions and understand the timeline for implementation. In summary, SUSGEN-GPT’s comprehensive and structured approach to generating ESG reports in the TCFD format makes it a more effective tool for companies aiming to address climate-related risks and opportunities in a clear, actionable, and strategically aligned manner.

5 Conclusion and Future Work

In conclusion, our work specifically addresses the significant gap in specialized language models for the financial and ESG sectors by introducing SUSGEN-30K and SUSGEN-GPT. The balanced nature of SUSGEN-30K allows SUSGEN-GPT to excel in multiple downstream financial tasks including sustainability report generation, achieving superior performance even than other larger scale LLMs. Additionally, our proposed TCFD-BENCH benchmark provides a rigorous evaluation framework for ESG report generation. Future work will focus on expanding the dataset to cover more specialized tasks within the ESG domain, enhancing the model’s capabilities, and refining the benchmark to include more diverse evaluation metrics. This work lays a robust foundation for developing advanced tools that meet the growing demands of the financial sector and contribute to climate-related financial disclosures.

570 **Limitations**

571 **Limited Model Performance** One of the limita- 620
572 tions of our work is the performance of our models, 621
573 which is influenced by resource constraints. The 622
574 large models we utilize are typically in the range 623
575 of 7B/8B in terms of parameters. Due to limited 624
576 resources, we have not had the opportunity to ex-
577 plore the potential benefits of 70B parameter or
578 larger models. As a result, our models may not
579 achieve their full potential performance, and this
580 is an aspect we are mindful of as we continue our
581 work. We aim to address these limitations in the
582 future and strive to improve the performance of our
583 models as resources allow.

584 **Limited Evaluation from Expert** While experts 625
585 have conducted quality analysis for certain cases, 626
586 the evaluation of large-scale data relies on auto- 627
587 mated scores such as BLEU and ROUGE metrics. 628
588 These automated evaluations, while valuable, may 629
589 potentially introduce biases and lack the nuanced 630
590 understanding that human expertise provides, par- 631
591 ticularly in the ESG (Environmental, Social, and 632
592 Governance) domain. Furthermore, there is a short- 633
593 age of expert resources in the ESG domain, which 634
594 limits the comprehensive evaluation of the data 635
595 with respect to ESG factors. As a result, the evalua- 636
596 tion may not fully capture the depth and complexity 637
597 of the ESG-related aspects of the data. 638

598 **Unsuitable for Vary ESG Subtasks** The model’s 639
599 performance may exhibit significant variability 640
600 across different subfields, highlighting the neces- 641
601 sity for targeted research and optimization for spe- 642
602 cific ESG-related subtasks. For example, certain 643
603 subfields, such as renewable energy investment or 644
604 sustainable supply chain management, might re- 645
605 quire more customized approaches to ensure that 646
606 the model’s performance meets the expected stan- 647
607 dards. These variations necessitate a more granular 648
608 understanding of each subfield’s unique character- 649
609 istics and requirements, demanding further data 650
610 gathering and model adjustments. 651

611 **Insufficient Diverse Report Template** Despite the 652
612 significant progress achieved by our model, there 653
613 are still notable limitations concerning the compre- 654
614 hensiveness of the dataset. Firstly, our model was 655
615 predominantly trained on a limited number of re- 656
616 ports in the TCFD (Task Force on Climate-related 657
617 Financial Disclosures) format, with sparse repre- 658
618 sentation from other key standards such as the GRI 659
619 (Global Reporting Initiative), SASB (Sustainability 660

Accounting Standards Board), and CDP (Carbon 620
Disclosure Project). This narrow data source lim- 621
its the model’s generalization capabilities and may 622
hinder its performance when dealing with reports 623
adhering to different standards and formats. 624

625 **Ethical Considerations**

626 **False Information from LLMs** One of the press- 626
627 ing issues in this work is the phenomenon of model 627
628 hallucination, where the model generates informa- 628
629 tion that is not present in the input data. This prob- 629
630 lem is particularly significant when applying ad- 630
631 vanced techniques like LLMs to financial data, as 631
632 the generation of false information can have serious 632
633 implications. For instance, inaccurate or mislead- 633
634 ing financial reports can lead to incorrect business 634
635 decisions, regulatory non-compliance, and loss of 635
636 stakeholder trust. Addressing this issue requires 636
637 ongoing research and development to improve the 637
638 reliability and accuracy of NLP models. 638

639 **Bias towards Firm Perspective** Another consider- 639
640 ation is the inherent bias towards the firm’s perspec- 640
641 tive in the extracted information from corporate 641
642 sustainability reports. This bias arises because the 642
643 data predominantly originates from the companies 643
644 themselves, potentially leading to a one-sided view 644
645 that may overlook critical aspects such as stake- 645
646 holder opinions and third-party assessments. To 646
647 mitigate this, future work should aim to incorporate 647
648 a more diverse set of data sources, including inde- 648
649 pendent audits and reports from non-governmental 649
650 organizations (NGOs), to provide a more balanced 650
651 view of corporate sustainability practices. 651

652 **License of the Tool** To ensure that our tool is ac- 652
653 cessible and adaptable by all stakeholders, we have 653
654 chosen to release it under the Apache License 2.0. 654
655 This open-source license allows for wide distribu- 655
656 tion, usage, and modification of the tool, thereby 656
657 facilitating collaborative development and continu- 657
658 ous improvement. By doing so, we aim to foster a 658
659 community of practice that can collectively address 659
660 the challenges and leverage the opportunities pre- 660
661 sented by the use of NLP in financial contexts. This 661
662 approach not only democratizes access to advanced 662
663 technologies but also encourages transparency and 663
664 accountability in their application. 664

665
666
667
668
669
670
671
672
673
674

675

676
677
678
679
680
681

682
683
684
685
686

687
688
689
690
691
692

693
694
695
696
697

698
699
700
701
702
703

704
705
706
707
708
709
710

711
712
713
714
715
716
717

718
719
720

References

Toyin D. Aguda, Suchetha Siddagangappa, Elena Kochkina, Simerjot Kaur, Dongsheng Wang, and Charese Smiley. 2024. [Large language models as financial data annotators: A study on effectiveness and efficiency](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 10124–10145, Torino, Italia. ELRA and ICCL.

AI@Meta. 2024. [Llama 3 model card](#).

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, M  rouane Debbah,   tienne Goffinet, Daniel Hessel, Julien Launay, Quentin Malartic, et al. 2023. The falcon series of open language models. *arXiv preprint arXiv:2311.16867*.

Julio Cesar Salinas Alvarado, Karin Verspoor, and Timothy Baldwin. 2015. Domain adaption of named entity recognition to support credit risk assessment. In *Proceedings of the Australasian Language Technology Association Workshop 2015*, pages 84–90.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Marco Bronzini, Carlo Nicolini, Bruno Lepri, Andrea Passerini, and Jacopo Staiano. 2024. Glitter or gold? deriving structured insights from sustainability reports via large language models. *EPJ Data Science*, 13(1):41.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Chung-Chi Chen, Yu-Min Tseng, Juyeon Kang, Ana  s Lhuissier, Min-Yuh Day, Teng-Tsai Tu, and Hsin-Hsi Chen. 2023. Multi-lingual esg issue identification. In *Proceedings of the Fifth Workshop on Financial Technology and Natural Language Processing and the Second Multimodal AI For Financial Forecasting*, pages 111–115.

Zhiyu Chen, Wenhu Chen, Charese Smiley, Sameena Shah, Iana Borova, Dylan Langdon, Reema Moussa, Matt Beane, Ting-Hao Huang, Bryan R Routledge, et al. 2021. Finqa: A dataset of numerical reasoning over financial data. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3697–3711.

Zhiyu Chen, Shiyang Li, Charese Smiley, Zhiqiang Ma, Sameena Shah, and William Yang Wang. 2022. [Con-
vfnqa: Exploring the chain of numerical reasoning in](#)

[conversational finance question answering](#). *Preprint*, arXiv:2210.03849. 721
722

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*. 723
724
725
726
727
728

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2024. Qlora: Efficient finetuning of quantized llms. *Advances in Neural Information Processing Systems*, 36. 729
730
731
732

Financial Stability Board. 2017. Task force on climate-related financial disclosures: Overview. *Task force on climate-related financial disclosures*. 733
734
735

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, L  lio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timoth  e Lacroix, and William El Sayed. 2023. [Mistral 7b](#). *Preprint*, arXiv:2310.06825. 736
737
738
739
740
741
742
743

Rasmus J  rgensen, Oliver Brandt, Mareike Hartmann, Xiang Dai, Christian Igel, and Desmond Elliott. 2023. Multifin: A dataset for multilingual financial nlp. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 864–879. 744
745
746
747
748

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. [Scaling laws for neural language models](#). *Preprint*, arXiv:2001.08361. 749
750
751
752
753

Yang Lei, Jiangtong Li, Dawei Cheng, Zhijun Ding, and Changjun Jiang. 2024. [Cfbenchmark: Chinese financial assistant benchmark for large language model](#). *Preprint*, arXiv:2311.05812. 754
755
756
757

Chen Li, Weiqi Wang, Jingcheng Hu, Yixuan Wei, Nanning Zheng, Han Hu, Zheng Zhang, and Houwen Peng. 2024. [Common 7b language models already possess strong math capabilities](#). *Preprint*, arXiv:2403.04706. 758
759
760
761
762

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81. 763
764
765

Xiao-Yang Liu, Guoxuan Wang, Hongyang Yang, and Daochen Zha. 2023. Data-centric finqpt: Democratizing internet-scale data for financial large language models. *NeurIPS Workshop on Instruction Tuning and Instruction Following*. 766
767
768
769
770

Alexandra Luccioni, Emily Baylor, and Nicolas Duchene. 2020. [Analyzing sustainability reports using natural language processing](#). *Preprint*, arXiv:2011.08073. 771
772
773
774

775	Macedo Maia, Siegfried Handschuh, André Freitas,	Soumya Sharma, Tapas Nayak, Arusarka Bose,	831
776	Brian Davis, Ross McDermott, Manel Zarrouk, and	Ajay Kumar Meena, Koustuv Dasgupta, Niloy Gan-	832
777	Alexandra Balahur. 2018. Wwv'18 open challenge:	guly, and Pawan Goyal. 2022. Finred: A dataset for	833
778	financial opinion mining and question answering. In	relation extraction in financial domain. In <i>Compan-</i>	834
779	<i>Companion proceedings of the the web conference</i>	<i>ion Proceedings of the Web Conference 2022</i> , pages	835
780	2018, pages 1941–1942.	595–597.	836
781	Dominique Mariko, Hanna Abi Akl, Estelle Labidurie,	Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann	837
782	Stephane Durfort, Hugues De Mazancourt, and Mah-	Dubois, Xuechen Li, Carlos Guestrin, Percy Liang,	838
783	mahmoud El-Haj. 2020. Financial document causality	and Tatsunori B. Hashimoto. 2023. Stanford alpaca:	839
784	detection shared task (fincasual 2020). <i>arXiv preprint</i>	An instruction-following llama model. https://	840
785	<i>arXiv:2012.02505</i> .	github.com/tatsu-lab/stanford_alpaca .	841
786	Corentin Masson and Patrick Paroubek. 2024. Evalu-	Gemini Team, Rohan Anil, Sebastian Borgeaud,	842
787	ating topic model on asymmetric and multi-domain	Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu,	843
788	financial corpus . In <i>Proceedings of the 2024 Joint</i>	Radu Soricut, Johan Schalkwyk, Andrew M Dai,	844
789	<i>International Conference on Computational Linguis-</i>	Anja Hauth, et al. 2023. Gemini: a family of	845
790	<i>tics, Language Resources and Evaluation (LREC-</i>	highly capable multimodal models. <i>arXiv preprint</i>	846
791	<i>COLING 2024</i>), pages 6515–6529, Torino, Italia.	<i>arXiv:2312.11805</i> .	847
792	ELRA and ICCL.	Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier	848
793	Jingwei Ni, Julia Bingler, Chiara Colesanti-Senni, Math-	Martinet, Marie-Anne Lachaux, Timothée Lacroix,	849
794	ias Kraus, Glen Gostlow, Tobias Schimanski, Do-	Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal	850
795	minik Stammbach, Saeid Ashraf Vaghefi, Qian Wang,	Azhar, et al. 2023a. Llama: Open and effi-	851
796	Nicolas Webersinke, et al. 2023. Chatreport: Democ-	cient foundation language models. <i>arXiv preprint</i>	852
797	ratizing sustainability disclosure analysis through	<i>arXiv:2302.13971</i> .	853
798	llm-based tools. <i>arXiv preprint arXiv:2307.15770</i> .	Hugo Touvron, Louis Martin, Kevin Stone, Peter Al-	854
799	OpenAI. 2023a. Gpt-4 technical report . Technical re-	bert, Amjad Almahairi, Yasmine Babaei, Nikolay	855
800	port, OpenAI.	Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti	856
801	OpenAI. 2023b. Gpt-4 technical report . <i>Preprint</i> ,	Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton	857
802	<i>arXiv:2303.08774</i> .	Ferrer, Moya Chen, Guillem Cucurull, David Esiobu,	858
803	Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida,	Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller,	859
804	Carroll Wainwright, Pamela Mishkin, Chong Zhang,	Cynthia Gao, Vedanuj Goswami, Naman Goyal, An-	860
805	Sandhini Agarwal, Katarina Slama, Alex Ray, et al.	thony Hartshorn, Saghar Hosseini, Rui Hou, Hakan	861
806	2022. Training language models to follow instruc-	Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa,	862
807	tions with human feedback. <i>Advances in Neural</i>	Isabel Kloumann, Artem Korenev, Punit Singh Koura,	863
808	<i>Information Processing Systems</i> , 35:27730–27744.	Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Di-	864
809	Kishore Papineni, Salim Roukos, Todd Ward, and Wei-	ana Liskovich, Yinghai Lu, Yuning Mao, Xavier Mar-	865
810	Jing Zhu. 2002. Bleu: a method for automatic evalua-	tinet, Todor Mihaylov, Pushkar Mishra, Igor Moly-	866
811	tion of machine translation. In <i>Proceedings of the</i>	bog, Yixin Nie, Andrew Poulton, Jeremy Reizen-	867
812	<i>40th annual meeting of the Association for Computa-</i>	stein, Rashi Rungta, Kalyan Saladi, Alan Schelten,	868
813	<i>tional Linguistics</i> , pages 311–318.	Ruan Silva, Eric Michael Smith, Ranjan Subrama-	869
814	Agam Shah, Suvan Paturi, and Sudheer Chava. 2023a.	nian, Xiaoqing Ellen Tan, Binh Tang, Ross Tay-	870
815	Trillion dollar words: A new financial dataset, task &	lor, Adina Williams, Jian Xiang Kuan, Puxin Xu,	871
816	market analysis . In <i>Proceedings of the 61st Annual</i>	Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan,	872
817	<i>Meeting of the Association for Computational Linguis-</i>	Melanie Kambadur, Sharan Narang, Aurelien Rod-	873
818	<i>tics (Volume 1: Long Papers)</i> , pages 6664–6679,	riguez, Robert Stojnic, Sergey Edunov, and Thomas	874
819	Toronto, Canada. Association for Computational Lin-	Scialom. 2023b. Llama 2: Open foundation and	875
820	guistics.	fine-tuned chat models . <i>Preprint</i> , <i>arXiv:2307.09288</i> .	876
821	Agam Shah, Ruchit Vithani, Abhinav Gullapalli, and	Neng Wang, Hongyang Yang, and Christina Dan	877
822	Sudheer Chava. 2023b. Finer: Financial named en-	Wang. 2023a. Fingpt: Instruction tuning benchmark	878
823	ity recognition dataset and weak-supervision model.	for open-source large language models in financial	879
824	<i>arXiv preprint arXiv:2302.11157</i> .	datasets. <i>NeurIPS Workshop on Instruction Tuning</i>	880
825	Raj Sanjay Shah, Kunal Chawla, Dheeraj Eidnani,	<i>and Instruction Following</i> .	881
826	Agam Shah, Wendi Du, Sudheer Chava, Natraj Ra-	Neng Wang, Hongyang Yang, and Christina Dan	882
827	man, Charese Smiley, Jiaao Chen, and Diyi Yang.	Wang. 2023b. Fingpt: Instruction tuning benchmark	883
828	2022. When flue meets flang: Benchmarks and	for open-source large language models in financial	884
829	large pre-trained language model for financial do-	datasets . <i>Preprint</i> , <i>arXiv:2310.04793</i> .	885
830	main . <i>Preprint</i> , <i>arXiv:2211.00083</i> .	Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski,	886
		Mark Dredze, Sebastian Gehrmann, Prabhanjan Kam-	887
		badur, David Rosenberg, and Gideon Mann. 2023.	888

889 [Bloomberggpt: A large language model for finance.](#)
890 *Preprint*, arXiv:2303.17564.

891 Qianqian Xie, Weiguang Han, Zhengyu Chen, Ruoyu
892 Xiang, Xiao Zhang, Yueru He, Mengxi Xiao, Dong
893 Li, Yongfu Dai, Duanyu Feng, Yijing Xu, Haoqiang
894 Kang, Ziyang Kuang, Chenhan Yuan, Kailai Yang,
895 Zheheng Luo, Tianlin Zhang, Zhiwei Liu, Guojun
896 Xiong, Zhiyang Deng, Yuechen Jiang, Zhiyuan Yao,
897 Haohang Li, Yangyang Yu, Gang Hu, Jiajia Huang,
898 Xiao-Yang Liu, Alejandro Lopez-Lira, Benyou Wang,
899 Yanzhao Lai, Hao Wang, Min Peng, Sophia Anani-
900 adou, and Jimin Huang. 2024. [The finben: An holistic financial benchmark for large language models.](#)
901 *Preprint*, arXiv:2402.12659.

903 Qianqian Xie, Weiguang Han, Xiao Zhang, Yanzhao
904 Lai, Min Peng, Alejandro Lopez-Lira, and Jimin
905 Huang. 2023. [Pixiu: A large language model, instruction data and evaluation benchmark for finance.](#)
906 *Preprint*, arXiv:2306.05443.

908 Hongyang Yang, Xiao-Yang Liu, and Christina Dan
909 Wang. 2023. [Fingpt: Open-source financial large language models.](#) *FinLLM Symposium at IJCAI 2023*.

911 Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q
912 Weinberger, and Yoav Artzi. 2019. [Bertscore: Evaluating text generation with bert.](#) *arXiv preprint*
913 *arXiv:1904.09675*.

915 Zhihan Zhou, Liqian Ma, and Han Liu. 2021. [Trade the event: Corporate events detection for news-based event-driven trading.](#) *Preprint*, arXiv:2105.12825.

918 Fengbin Zhu, Wenqiang Lei, Youcheng Huang, Chao
919 Wang, Shuo Zhang, Jiancheng Lv, Fuli Feng, and
920 Tat-Seng Chua. 2021. [Tat-qa: A question answering benchmark on a hybrid of tabular and textual content in finance.](#) *arXiv preprint* *arXiv:2105.07624*.

923 Yi Zou, Mengying Shi, Zhongjie Chen, Zhu Deng,
924 ZongXiong Lei, Zihan Zeng, Shiming Yang, HongX-
925 iang Tong, Lei Xiao, and Wenwen Zhou. 2023. [Esgreveal: An llm-based approach for extracting structured data from esg reports.](#) *arXiv preprint*
926 *arXiv:2312.17264*.

929
930
931
932
933
934
935
936
937
938
939
940
941
942
943
944
945
946
947
948
949
950
951
952
953
954
955
956
957
958
959
960
961
962
963
964
965
966
967
968
969
970
971
972
973
974
975
976
977

A Composition of our SUSGEN dataset

This section details the various components and sources of data included in the SUSGEN dataset, providing an overview of the dataset’s structure and content.

Each dataset listed in Table 3 is meticulously curated and split to optimize model training and evaluation, covering a diverse range of scenarios and applications in financial NLP. This approach ensures that the model performs well across these specific tasks and integrates seamlessly with real-world financial applications, providing robust and reliable insights.

B Financial Tasks Examples

This appendix provides examples of various financial tasks handled by SUSGEN-GPT. The tasks include Financial Question Answering, where the model answers specific queries based on financial reports; Financial Table Question Answering, which involves extracting data from financial tables; Headline Classification, categorizing news headlines; Named Entity Recognition, identifying proper nouns; Relation Extraction, detecting relationships between entities; Sentiment Analysis, assessing sentiment in text; and Summarization, generating concise summaries of longer documents. These examples illustrate the diverse capabilities of SUSGEN-GPT in processing and analyzing financial data.

C Our Ablation Study Results about Training Data Scale

This section showcases the results of our ablation studies, focusing on the impact of different training data scales on the model’s performance.

Table 4 presents the zero-shot and few-shot performance of various LLMs, including different configurations of SUSGEN-GPT with dataset sizes of 3k, 10k, and 30k. The metrics evaluated include F1 score, Micro F1, Entity F1, Exact Match Accuracy, Rouge, and BertScore across multiple datasets.

The results indicate that increasing the dataset size significantly improves the performance of SUSGEN-GPT, especially after supervised fine-tuning. For instance, on the FiQASA dataset, SUSGEN-GPT with Mistral-7B-Instruct v0.3 at 30k scale achieves an F1 score of 0.87, the highest among all configurations. Similarly, for the FOMC dataset, the model attains an F1 score of 0.73 at 10k scale, outperforming other versions.

In the MultiFin dataset for Headline Classification, SUSGEN-GPT with 10k data and Mistral-7B-Instruct v0.3 achieves a Micro F1 score of 0.71, demonstrating the model’s improved ability to categorize financial news headlines with more extensive training data. For the NER task on the FINER-ORD dataset, the model shows a significant improvement in Entity F1, reaching 0.35 at 10k scale.

For Relation Extraction on the SC dataset, SUSGEN-GPT with 30k data and Mistral-7B-Instruct v0.3 achieves an F1 score of 0.96, indicating robust performance. In Financial Question Answering tasks such as FinQA and TATQA, the model consistently performs well, with EmAcc reaching 0.87 and 0.69, respectively, at 10k and 30k scales.

The sustainability report generation results on the TCFD-Bench dataset also show that SUSGEN-GPT excels with larger datasets. The model achieves a Rouge-L score of 0.18 and a BertScore of 0.40, demonstrating superior performance compared to smaller dataset scales.

D Sustainability Report Generation Examples

We provide examples of sustainability reports generated by SUSGEN-GPT, demonstrating the model’s ability to create comprehensive and accurate ESG reports. SusGen-GPT’s structured format offers significant advantages, establishing it as a superior tool for generating TCFD-format ESG reports. These enhancements ensure that reports produced by SusGen-GPT are not only more informative but also more actionable and easier to comprehend. By categorizing risks and opportunities into specific time horizons (short-term, medium-term, and long-term), SusGen-GPT provides a clear roadmap for addressing climate-related challenges and leveraging opportunities. Each identified risk and opportunity is linked to specific actions and goals, making the report practical and aligned with the company’s strategic objectives.

Additionally, the improved contextual clarity provided by SusGen-GPT allows users to better understand the implications of each risk and opportunity. This is essential for stakeholders who depend on these reports to make informed decisions. The detailed action points offered by SusGen-GPT assist in developing concrete strategies for sustainability, enhancing operational efficiencies, improving reputation, and ensuring compliance with regu-

978
979
980
981
982
983
984
985
986
987
988
989
990
991
992
993
994
995
996
997
998
999
1000
1001
1002
1003
1004
1005
1006
1007
1008
1009
1010
1011
1012
1013
1014
1015
1016
1017
1018
1019
1020
1021
1022
1023
1024
1025
1026
1027

latory requirements.

In contrast, while ChatReport delivers a general overview of climate-related risks and opportunities, it lacks the depth and specificity present in SusGen-GPT's output. ChatReport's responses are less structured and do not consistently link risks to specific time horizons, making it more challenging for users to prioritize actions and comprehend the timeline for implementation.

Task	Dataset	Train	Language	Test	Final	Comment
General	Alpaca-52k (Taori et al., 2023)	52,000	EN	✗	3,000	
Arithmetic	GSM-8k (Cobbe et al., 2021)	7,473	EN	1,319	3,000	
HC	finppt-headline-cls (Wang et al., 2023b)	82,200	EN	20,500	1,500	CLS
HC	finppt-headline (Wang et al., 2023b)	82,200	EN	20,500	1,500	Instr Diff
HC	FLUE-headline (Shah et al., 2022)	80,000	EN	✗	0	✗
HC	flare-multifin-en (Xie et al., 2024)	✗	EN	546	500	CLS
HC	flare-mlesg-en (Xie et al., 2024)	✗	EN	300	300	ESG-CLS
NER	finppt-ner-cls (Wang et al., 2023b)	13,500	EN	3,500	2,700	CLS
NER	finppt-ner (Wang et al., 2023b)	511	EN	98	500	
NER	flare-ner (Xie et al., 2024)	408	EN	98	300	valid103
NER	flare-finer-ord (Xie et al., 2024)	✗	EN	1,075	1,075	
RE	finppt-finred (Wang et al., 2023b)	27,600	EN	5,112	5,112	RE+CLS
RE	finppt-finred-re (Wang et al., 2023b)	11,400	EN	2,140	1,750	RE
RE	finppt-finred-cls (Wang et al., 2023b)	48,500	EN	8,930	1,750	CLS
RE	flare-finarg-ecc-auc-test (Xie et al., 2024)	✗	EN	969	0	RE+CLS
RE	flare-causal20-sc-test (Xie et al., 2024)	✗	EN	8,628	8,628	RE+CLS
SA	esg-sentiment	611	EN	93	843	ESG
SA	enhanced-financial-phrasebank	4,850	EN	✗	1,457	
SA	FIN_NUMBER	4,680	KO	✗	xx	ESG
SA	finppt-sentiment (Wang et al., 2023b)	76,800	EN	✗	800	Duplicate?
SA	finppt-sentiment-cls (Wang et al., 2023b)	47,600	EN	✗	400	CLS
SA	FLUE-sentiment (Shah et al., 2022)	4850	EN	✗	0	✗
SA	flare-fiqasa (Xie et al., 2024)	750	EN	235	235	valid188
SA	flare-fomc (Xie et al., 2024)	✗	EN	496	496	valid188
FIN-QA	FIN_NUMBER-EQA/train	400	KO	✗	400	
FIN-QA	FIN_NUMBER-BQA/train	400	KO	✗	400	CLS
FIN-QA	FIN_NUMBER-MCQA/train	400	KO	✗	398	CLS
FIN-QA	FIN_NUMBER-NQA-ARI/train	400	KO	✗	398	
FIN-QA	FIN_NUMBER-NQA-COM/train	400	KO	✗	399	
FIN-QA	FIN_NUMBER-NQA-EXT/train	400	KO	✗	397	
FIN-QA	flare-cfa/test (Xie et al., 2024)	✗	EN	1030	0	CLS
FIN-QA	finppt-fiqqa (Wang et al., 2023b)	17,100	EN	✗	708	
FIN-QA	finppt-fineval (Wang et al., 2023b)	1,060	ZH	265	0	CLS
FIN-QA	flare-finqa (Xie et al., 2024)	6250	EN	1147	400	
FIN-QA	flare-fsrl (Xie et al., 2024)	✗	EN	97	97	
FIN-TQA	finppt-convfinqa (Wang et al., 2023b)	11,100	EN	1,490	1,000	
FIN-TQA	flare-convfinqa (Xie et al., 2024)	8890	EN	1,490	2,500	
FIN-TQA	flare-tatqa (Xie et al., 2024)	✗	EN	1,668	1,668	
SUM	flare-edtsum-test (Xie et al., 2024)	✗	EN	2000	2000	
SRG	ESG-Chat	914	EN	✗	914	
SRG	TCFD_QA	260	EN	✗	1669	
SRG	salmasally	417	FR	✗	417	

Table 3: **Composition of our SUSGEN dataset.** We report the list of datasets and associated splits used to build the dataset. We mainly focus on eight following tasks in the datasets in order to let the model cover most applications in the financial NLP domain. HC: Headline Classification. NER: Named Entity Recognition. RE: Relation Extraction. SA: Sentiment Analysis. FIN-QA: Financial Question Answering. FIN-TQA: Financial Table Question Answering. SUM: Text Summary. SRG: Sustainability Report Generation.

Task: Financial Question Answering

Context: entergy corporation and subsidiaries management 2019s financial discussion and analysis a result of the entergy Louisiana and entergy gulf states Louisiana business combination , results of operations for 2015 also include two items that occurred in October 2015 : 1) a deferred tax asset and resulting net increase in tax basis of approximately \$ 334 million and 2) a regulatory liability of \$ 107 million (\$ 66 million net-of-tax) as a result of customer credits to be realized by electric customers of entergy Louisiana , consistent with the terms of the stipulated settlement in the business combination proceeding .

...

Question: what is the net change in net revenue during 2015 for entergy corporation?"

Response: 94.0

Task: Financial Table Question Answering

Context:

...

	2007	2006	2005
Weighted average exercise price per share	\$60.94	\$37.84	\$25.14
Expected annual dividends per share	\$0.96	\$0.80	\$0.66
Expected life in years	5.0	5.1	5.5
Expected volatility	27% (27%)	28% (28%)	28% (28%)
Risk-free interest rate	4.1% (4.1%)	5.0% (5.0%)	3.8% (3.8%)
Weighted average grant date fair value of stock option awards granted	\$17.24	\$10.19	\$6.15

Question: What was the weighted average exercise price per share in 2007?

Response: The weighted average exercise price per share in 2007 was \$60.94.

Task: Headline Classification

Context: Broadband with Speeds Up to 1 GIG is Provided at No Cost to Families Through the FCC's Emergency Connectivity Fund: What's the news? Los Angeles Unified School District and AT&T* are providing high-speed broadband to students' homes at no cost to their families. As we reimagine the future of education, connectivity is the new pen and paper. And through this joint effort, more students and households in Los Angeles will have the reliable internet needed to fully participate in education and digital life.

Response: Access to Communications. This news article focuses on providing high-speed broadband to students' homes, enhancing connectivity for education.

Task: Named Entity Recognition

Context: The actual disbursement date of the last installment of the Loan to the Borrower shall be the "Effective Date".

Response: "Borrower, ORG", "Effective Date, DATE"

Task: Relation Extraction

Context: For instance, the Google Chromecast or the Teewe 2 Android streaming device or, for that matter, any Android box is a small computer that is attached to a television.

Response: Chromecast - product/material produced by - Google

Task: Sentiment Analysis

Context: \ \$BBRY Actually lost .03c per share if U incl VZ as no debt and 3.1 in Cash.

Response: Negative

Task: Summarization

Context: LONDON--(BUSINESS WIRE)--Technavio has been monitoring the all-season tire market in Europe and it is poised to grow by USD 3.42 billion during 2020-2024, progressing at a CAGR of almost 9% during the forecast period.

...

This growing client base relies on Technavio's comprehensive coverage, extensive research, and actionable market insights to identify opportunities in existing and potential markets and assess their competitive positions within changing market scenarios.

Response: Technavio's report highlights the growth of the all-season tire market in Europe, projecting an increase of USD 3.42 billion from 2020 to 2024 with a CAGR of nearly 9%. The study discusses market fragmentation, the impact of COVID-19, and key market players, emphasizing the importance of nanotechnology and regulatory changes driving market growth.

Figure 8: Financial Tasks Examples. Examples of main tasks in financial domain and the corresponding context and response.

Datasets	Metrics	SusGen	SusGen	SusGen	SusGen	SusGen	SusGen	SusGen	SusGen	SusGen	SusGen	SusGen	
		GPT-3k Llama3 8B Instruct	GPT-3k Llama3 8B	GPT-3k Mistral 7B Instruct v0.3	GPT-3k Mistral 7B v0.3	GPT-10k Llama3 8B Instruct	GPT-10k Llama3 8B	GPT-10k Mistral 7B Instruct v0.3	GPT-10k Mistral 7B v0.3	GPT-30k Llama3 8B Instruct	GPT-30k Llama3 8B	GPT-30k Mistral 7B Instruct v0.3	GPT-30k Mistral 7B v0.3
FiQASA	F1	0.73	0.60	0.73	0.87	0.66	0.56	0.64	0.76	0.64	0.46	0.60	0.63
	FOMC	0.71	0.37	0.69	0.56	0.73	0.67	0.70	0.70	0.72	0.64	0.67	0.67
MultiFin	MicroF1	0.50	0.00	0.60	0.39	0.56	0.66	0.71	0.70	0.41	0.32	0.50	0.41
	MLESG	0.30	0.03	0.51	0.06	0.25	0.08	0.50	0.40	0.22	0.04	0.32	0.04
NER	EntityF1	0.37	0.10	0.02	0.27	0.31	0.35	0.35	0.34	0.31	0.35	0.25	0.31
	FINER-ORD	EntityF1	0.28	0.00	0.14	0.11	0.06	0.09	0.09	0.07	0.11	0.04	0.08
FinRED	F1	0.04	0.02	0.06	0.05	0.19	0.21	0.19	0.17	0.16	0.23	0.11	0.09
	SC	F1	0.77	0.23	0.95	0.85	0.52	0.48	0.96	0.84	0.83	0.41	0.96
FinQA	EmAcc	0.65	0.56	0.33	0.37	0.38	0.64	0.87	0.84	0.38	0.53	0.26	0.57
TATQA	EmAcc	0.59	0.41	0.67	0.58	0.59	0.57	0.69	0.55	0.60	0.62	0.63	0.55
ConvFinQA	EmAcc	0.68	0.35	0.64	0.53	0.39	0.70	0.90	0.84	0.49	0.73	0.68	0.58
EDTSUM	Rouge-1	0.20	0.12	0.24	0.10	0.20	0.14	0.30	0.19	0.19	0.12	0.25	0.13
	Rouge-2	0.08	0.03	0.10	0.04	0.08	0.04	0.13	0.08	0.07	0.03	0.10	0.05
	Rouge-L	0.10	0.15	0.19	0.08	0.16	0.10	0.25	0.15	0.15	0.09	0.20	0.10
	BertScore	0.44	0.67	0.53	0.45	0.50	0.46	0.55	0.50	0.49	0.44	0.52	0.47
TCFD-BENCH	Rouge-L	0.14	0.10	0.17	0.13	0.13	0.11	0.18	0.16	0.13	0.10	0.18	0.13
	BertScore	0.32	0.28	0.35	0.30	0.31	0.29	0.40	0.36	0.31	0.28	0.38	0.31
	METEOR	0.25	0.15	0.24	0.25	0.20	0.19	0.22	0.21	0.22	0.19	0.27	0.23
	BLEU-1	0.22	0.21	0.27	0.18	0.22	0.19	0.37	0.30	0.20	0.17	0.29	0.19
	BLEU-2	0.05	0.03	0.09	0.05	0.04	0.03	0.12	0.09	0.04	0.02	0.10	0.04
	BLEU-3	0.02	0.01	0.05	0.02	0.01	0.01	0.06	0.04	0.01	0.00	0.05	0.02
BLEU-4	0.01	0.00	0.03	0.01	0.01	0.01	0.04	0.02	0.01	0.00	0.03	0.01	

Table 4: The zero-shot and few-shot performance of different LLMs on data scaling effects. All results via our evaluations are the average of three runs. “-” represents the result that is currently unable to yield due to model size or availability, and “*” represents the result from the previous paper.

Models	Example Prompt	Example Output
ChatReport	<p>Instruction: You are a senior equity analyst with expertise in climate science, evaluating a company's sustainability report, you will answer the question in detail based on the company's ESG Annual Report. Report:</p> <p>Question: What are the most relevant climate-related risks and opportunities that the organization has identified over the short, medium, and long term? Are risks clearly associated with a horizon?</p>	<p>Walmart has identified climate change, regeneration of natural resources, waste management, and people in supply chains as relevant climate-related risks and opportunities. The report does not clearly associate these risks with specific time horizons.</p>
SusGen	<p>Instruction: You are a senior equity analyst with expertise in climate science, evaluating a company's sustainability report, you will answer the question in detail based on the company's ESG Annual Report. Report: Walmart's ESG Report for FY2022 highlights its sustainability efforts, aligned with global standards like GRI, SASB, and TCFD. Efforts Product Supply Chain Sustainability Sourcing: Walmart sets standards for key commodities and engages suppliers through initiatives like Project Gigaton™.....</p> <p>Question: What are the most relevant climate-related risks and opportunities that the organization has identified over the short, medium, and long term? Are risks clearly associated with a horizon?</p>	<p>Answer:</p> <p>The organization, Walmart, identifies several climate-related risks and opportunities in various time horizons as follows:</p> <p>Short Term (Up to One Year):</p> <ul style="list-style-type: none"> * Continuous improvement in carbon footprint due to ongoing investments in energy efficient technologies and processes. * Enhancing supply chain resilience against physical climate risks such as extreme weather events. * Complying with regulatory requirements relating to Green House Gas (GHG) emission reductions. <p>Medium Term (One To Ten Years):</p> <ul style="list-style-type: none"> * Transitioning towards net zero emissions by implementing more aggressive decarbonization measures. * Expanding capacity for storing and utilizing Renewable Energy (RE), given increasing adoption and decreasing costs. * Adapting infrastructure and logistics networks to be better prepared for intensifying effects of climate change. <p>Long Term (More Than Ten years):</p> <ul style="list-style-type: none"> * Exploration into emerging low-carbon transportation solutions to offset reliance on fossil fuels within their value chains. * Investment in research & development partnerships focused on sustainable production methods and circular economy principles. * Long-term goal alignment with limiting warming to 1.5°C above preindustrial levels under the Paris agreement. <p>Each of these climate-related risks and opportunities is explicitly linked with specific time frames throughout the text of the provided context. The nature of each risk or opportunity also varies; some being primarily challenges to business continuity while others present potential benefits via cost savings, improved operational efficiencies, enhanced reputation, etc.</p>

Figure 9: A generated Sustainability Report Example. A comparison of the output obtained by our SUSGEN-GPT and ChatReport with the same question in the TCFD-BENCH.