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 focus on Environmental, Social, and Governance (ESG) considerations have created a pressing need for advanced natural language processing (NLP) tools. Despite recent advancements, there is still a notable absence of open-source Large Language Models (LLMs) that are proficient across both general finance and ESG domains, such as generating ESG reports. To address this gap, we introduce SusGen-30k, a high-quality, category-balanced dataset comprising seven financial NLP tasks and ESG report generation. In addition, we propose TCFD-Bench, a benchmark designed to improve the evaluation of sustainability report generation. Our data-centric approach led to the development of a suite of models, SusGen-GPT, trained on the curated dataset. These models were evaluated across six adapted tasks and two off-the-shelf tasks, showing state-of-the-art performance, surpassing all other models except GPT-4. Remarkably, SusGen-GPT achieved an average score only 0.02 below GPT-4, despite using models with only 7-8B parameters compared to GPT-4's 1,700B. This demonstrates the efficiency of our approach in delivering high performance with significantly fewer resources, addressing existing challenges and fostering further advancements in the financial and ESG research community. Code and data would release.

1 Introduction

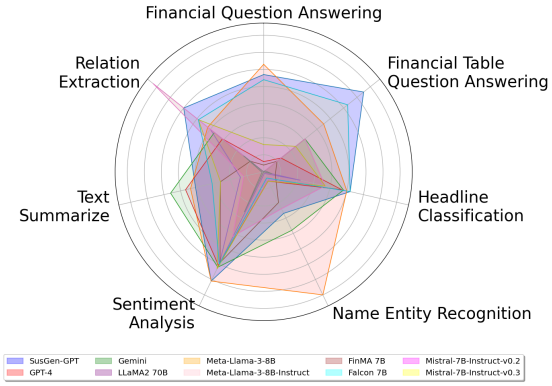
As automation in the financial sector gains more attention and climate change becomes an increasingly urgent issue, there is a growing need for more advanced tools capable of solving these problems. The ability to efficiently process and generate financial reports, analyze ESG metrics, and produce comprehensive TCFD-format[1] reports is crucial for maintaining transparency and accountability. Recently, LLMs (Brown et al., 2020; Ouyang et al., 2022; Touvron et al., 2023a; OpenAI, 2023a; Touvron et al., 2023b) have gained widespread popularity due to their ability to solve complex tasks such as commonsense reasoning, machine translation or even performing self-training (Yeo et al., 2024). 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.

In light of these challenges, we propose SusGen-30k, a meticulously curated dataset that is designed to tackle problems task across both financial and ESG domains. More importantly, we introduce a suite of LLMs, trained on our proposed dataset, which we refer to as SusGen-GPT. SusGen-GPT is capable of achieving superior performance across multiple downstream tasks simultaneously, when compared against models that are orders of magnitude larger.

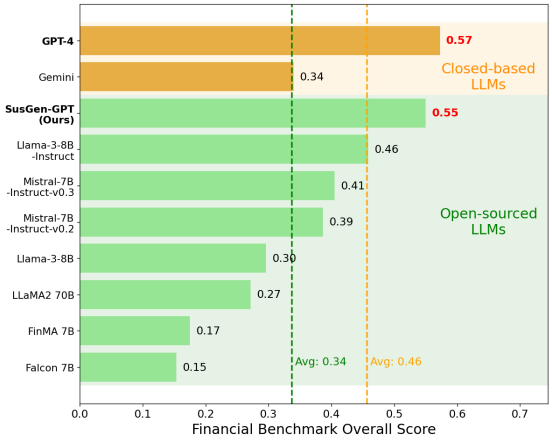
Additionally, we propose a new benchmark, TCFD-Bench, specifically designed to assess models' ability to generate concise and accurate ESG reports from annual reports. We likewise conduct experiments on the proposed benchmark using SusGen-GPT. In total, our contributions include the following:

1. SusGen-30k, a large-scale high quality dataset in both financial & ESG domain.
2. We propose and release a well-curated benchmark, tailored towards ESG report generation, TCFD-Bench.
3. A suite of fine-tuned LLMs, SusGen-GPT, shown to achieve comparable state-of-the-art performance to GPT-4 across both general fi-

[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.



(a) Performance comparison (Relative) with baseline models on each financial NLP task with average score.



(b) Overall scores of financial NLP benchmark across 6 tasks. Dashed lines for two LLM types' average.

Figure 1: An overview of model comparison with both open-source and closed-base baseline models on six financial NLP tasks. The two sub-figures show that our model SUSGEN-GPT achieves state-of-the-art performance in most benchmarks.

nancial and ESG NLP benchmarks as shown in Figure 1 when most of open-source models struggle to perform well in these domains. Remarkably, our models only have only 7-8B parameter models even with quantization comparing to the GPT4's 1,700B parameters or so, making it a computational efficient solution.

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 (Dubey et al., 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 key concerns on the imbalance in training data and lack of knowledge 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. For example, Fin-

GPT 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 system is designed to handle both document processing and direct question-answering in the financial and ESG domains when utilizing SusGen-GPT as the core as shown in Figure 2 to solve problems multi financial NLP tasks. All tasks are first prompted and then fed into SUSGEN-GPT to generate responses. For the sustainability report generation task, the SUSGEN system processes raw, unstructured documents using Retrieval-Augmented Generation (RAG) to extract relevant information from annual reports. This information is then combined into pre-defined prompts, which SUSGEN-GPT uses to generate a comprehensive TCFD report. The provided summary ensures the generated report complies with TCFD standards. Additionally, the model is capable of answering ESG-related queries concerning the company report.

3.2 Data Construction

SUSGEN-30K The dataset is originally sourced from two primary sources: open-source datasets available on Hugging Face[2] and annual reports sourced from TCFDHub[3] Database. The construction process for SUSGEN-30K involves a comprehensive automatic pipeline that starts with data collection from various sources such as company reports (including annual and ESG reports), publicly available financial datasets, and automated content crawlers that scrape financial data from the web, shown as the Figure 4.

This raw data undergoes thorough preprocessing steps, including manual annotation to extract useful content, machine-translated data to augment the dataset with multilingual data, and other augmentation techniques to generate novel data samples. We also include anonymization to remove sensitive information and comply with privacy regulations. Finally, the preprocessed data is reformatted into a format compatible with the Supervised-FineTuning(SFT) dataset, ensuring it is well-balanced and ready for training models in financial NLP and Sustainability Report Generation. This structured approach ensures that the dataset is robust, diverse, and high-quality, suitable for advancing the field of sustainable finance. The collected data are then split into the seven aforementioned financial tasks in Appendix B. We introduce an additional task in the ESG domain, Sustainability Report Generation (SRG).

Inspired by the scaling law (Kaplan et al., 2020) and Common-7B (Li et al., 2024), we perform scaling on the dataset to address the imbalance in sample size between the different tasks in the dataset. For large-scale category data, we downsample them based on data quality to create a well-balanced dataset. Finally, we concatenate all the samples to form the SUSGEN-30K instruction dataset, which can be used for the financial NLP domain. For detailed information regarding the data sources and composition, please refer to Figure 3 in Appendix C.

TCFD-BENCH The benchmark includes a balanced distribution of tasks related to ESG reporting, ensuring coverage of key areas like governance, risk management, and strategic planning, which 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 manual effort, as depicted in Figure 3. The context information is generated from annual reports using GPT-4o, while the TCFD reports are manually extracted to obtain the questions and answers. To be specific, the explicit question-answer pairs are extracted from 14 ESG PDF reports in the TCFD format. Then we anonymize the data to cover sensitive information, and use Mistral 7B to generate diversified instructions 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.

[2] https://huggingface.co/
[3] https://www.tcfdhub.org/reports

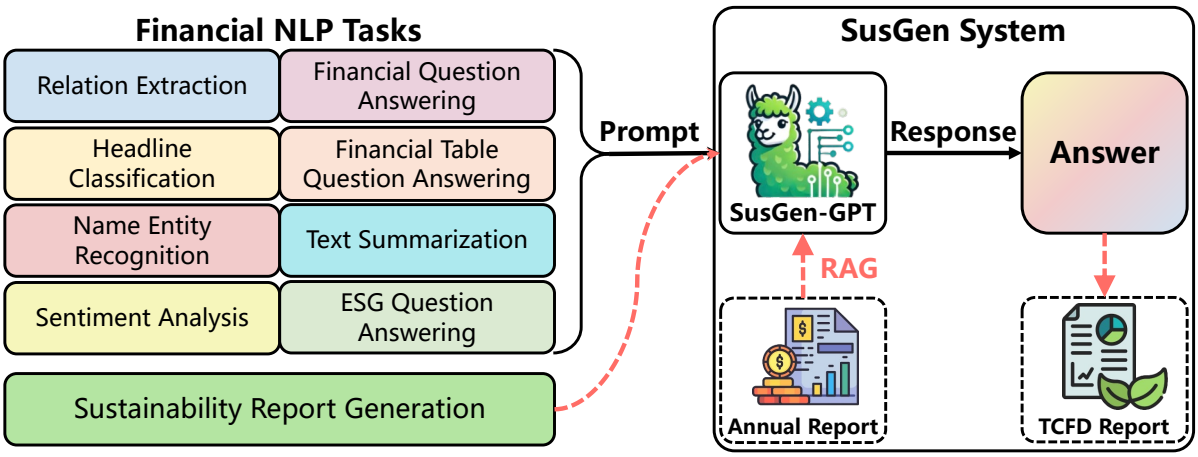


Figure 2: Overview of the SUSGEN System Pipeline.

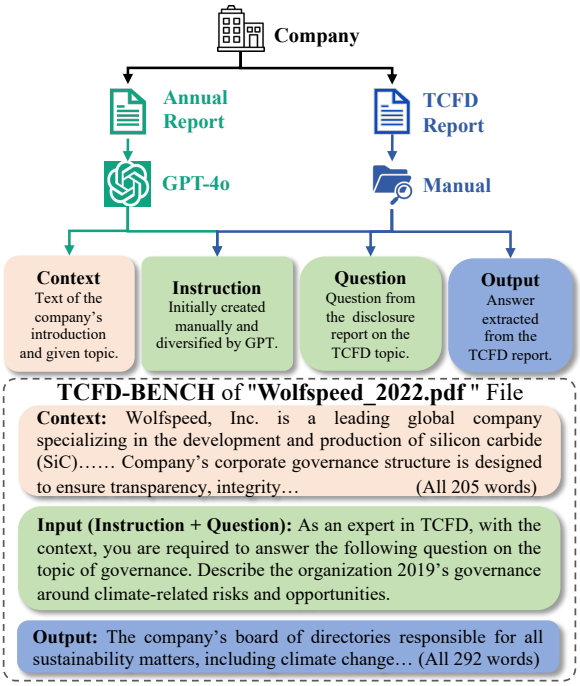


Figure 3: The data construction pipeline of TCFD-BENCH, illustrated with an example extracted and processed from the Wolfspeed_2022.pdf reports.

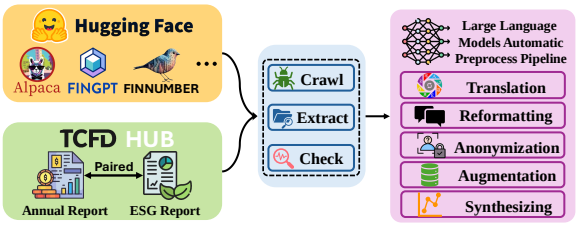


Figure 4: The pipeline of SUSGEN-30K data construction. The process involves collecting data open-source datasets from hugging-face and company reports from TCFD-Hub Database, followed by quality control and various automatic LLMs pre-processing steps to create the final instruction-following format dataset.

3.3 Statistics

In this section, we present the statistical information about our training dataset, SUSGEN-30K. As illustrated in Figure 5, the dataset is well-balanced across various task categories, ensuring comprehensive coverage of financial and ESG domains. Specific details regarding the dataset's categories can be found in the Appendix C.1, including Table 3, which provides information on the data sources and quantities for each category, and Figure 8, which shows the token length distribution for each task in the final SUSGEN-30K dataset. This balanced distribution allows the model to learn effectively from multi-tasks with low bias, contributing to the robustness and versatility of SusGen-GPT in handling diverse financial and ESG-related tasks.

3.4 Evaluations

Our evaluation includes six adapted tasks and two non-adapted (off-the-shelf) tasks, the latter consisting of Text Summarization and Sustainability Report Generation, which were not present in the training dataset. The metrics used to evaluate the performance of SusGen-GPT on various financial and ESG tasks are as follows.

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 classify-

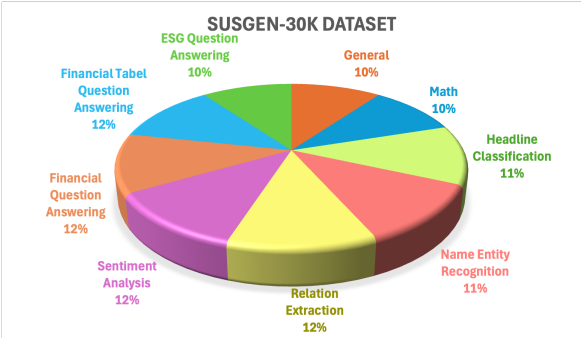


Figure 5: SUSGEN-30K Category Distribution. Highlight the proportion of data dedicated to each specific task area in financial NLP.

ing 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 BERTScore (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 different tasks, highlighting the model's strengths and areas for improvement.

4 Experiments & Analysis

4.1 Experimental Setup

Based on the understanding that LoRA can effectively retain learned knowledge without significant forgetting (Biderman et al., 2024), we chose the QLoRA method over full fine-tuning to preserve the model's general capabilities while ensuring computational efficiency. Our experiments employed the SUSGEN-GPT models, leveraging four baseline models: Mistral-v0.3-7B, Mistral-Instruct-v0.3-7B[4], LLaMA-3-8B, and LLaMA-3-8B-Instruct[5], using the QLoRA method for supervised fine-tuning due to its 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 alpaca prompt shown in Appendix A as used in Training and combing vllm inference optimization techiques. We use LangChain[6] to manage vector-database retriever. We use all-mpnet-base-v2[7] 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 which consist of 14 datasets across 8 tasks in financial and esg NLP domain and baseline models used to evaluate SusGen-GPT's performance.

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

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

[5]LLaMA3 models are licensed under a bespoke commercial license by Meta AI.

[6]https://python.langchain.com/

[7]https://huggingface.co/sentence-transformers/all-mpnet-base-v2

(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 with the most capable model from OpenAI, GPT-4 (OpenAI, 2023b) which excels across multiple NLP tasks. Similarly, Mistral 7B-Inst-v0.2/v0.3 (Jiang et al., 2023) is also used to be compared. LLaMA3 (Dubey et al., 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 7B parameters, designed for diverse NLP tasks, providing efficient text generation.

4.3 Main Results and Comparison

We evaluate SUSGEN-GPT on several financial NLP tasks using well-established benchmarks, comparing its performance against baseline models. The performance of SUSGEN-GPT on 7 financial tasks is summarized in Table 1 and Figure 1.

Our model achieves competitive results across multiple benchmarks. For example, in SA, SUSGEN-GPT scores an F1 of 0.72 on the FiQASA dataset, close to GPT-4's 0.70. However, on the FOMC dataset, GPT-4 slightly surpasses SUSGEN-GPT, with scores of 0.71 vs 0.70. This demonstrates SUSGEN-GPT's robust performance in financial sentiment analysis.

In HC, SUSGEN-GPT performs well on the MultiFin dataset, achieving a Micro F1 score of 0.52, though GPT-4 leads with 0.65. On the MLESG dataset, SUSGEN-GPT takes the lead with a score of 0.51, demonstrating its strength in financial headline classification.

For NER, SUSGEN-GPT scores 0.35 on the NER dataset but falls behind GPT-4's 0.83. On the FINER-ORD dataset, it scores 0.18, compared to GPT-4's 0.77. These results suggest that SUSGEN-GPT performs well but could benefit from improvement in entity recognition.

In RE, SUSGEN-GPT excels on the SC dataset, achieving an F1 score of 0.96, significantly outperforming others. However, it performs modestly on the FinRED dataset, with a score of 0.23.

For FinQA and FinTQA, SUSGEN-GPT shows strong capabilities, achieving an EmAcc of 0.57 on FinQA, slightly behind GPT-4's 0.63, and an impressive 0.80 on TATQA, surpassing other models. This highlights the model's proficiency in financial question-answering.

In sustainability report generation, SUSGEN-GPT is evaluated using the TCFD-BENCH dataset and compared with CHATREPORT, as shown in Table 2.

SUSGEN-GPT achieves a Rouge-L score of 0.20, a Bert-Score of 0.40, and a METEOR score of 0.27, outperforming CHATREPORT, which scores 0.14, 0.32, and 0.12. While CHATREPORT leads on BLEU-1 with 0.41, SUSGEN-GPT surpasses it in BLEU-2, BLEU-3, and BLEU-4 scores.

These results emphasize the advanced capabilities of SUSGEN-GPT in generating high-quality, TCFD-compliant sustainability reports, making it a valuable asset for organizations focused on ESG reporting.

4.4 Ablation Study

In this section, we investigate the effect of dataset scaling on SUSGEN-GPT across eight financial NLP tasks using datasets of 10k, 20k, and 30k samples. Comprehensive results can be found in Table 4 in Appendix D, and the performance trends for all tasks are illustrated in Figure 6.

Overall, the results show a clear trend: increasing the dataset size consistently improves performance across all tasks. This pattern is evident as larger datasets allow the models to better capture

Datasets	Tasks	Metrics	SusGen-GPT (Ours)	GPT4*	Gemini*	LLaMA2 70B	LLaMA3 8B	LLaMA3 8B-Inst	FinMA 7B	Falcon 7B	Mistral 7B-Inst-v0.2	Mistral 7B-Inst-v0.3
FiQASA (Maia et al., 2018)	SA	F1	0.72	0.70	0.71	0.73	0.72	0.72	0.69	0.67	0.65	**0.74**
FOMC (Shah et al., 2023a)	SA	F1	0.70	**0.71**	0.53	0.49	0.53	0.47	0.49	0.30	0.30	0.37
MultiFin (Jørgensen et al., 2023)	HC	MicroF1	0.52	**0.65**	0.62	0.63	0.50	0.56	0.14	0.09	0.50	0.51
MLESG (Chen et al., 2023)	HC	MicroF1	**0.51**	0.35	0.34	0.31	0.23	0.48	0.00	0.06	0.47	0.49
NER (Alvarado et al., 2015)	NER	EntityF1	0.35	**0.83**	0.61	0.04	0.06	0.04	0.39	0.00	0.17	0.15
FINER-ORD (Shah et al., 2023b)	NER	EntityF1	0.18	**0.77**	0.14	0.07	0.06	0.04	0.00	0.00	0.08	0.14
FinRED (Sharma et al., 2022)	RE	F1	**0.23**	0.02	0.00	0.00	0.04	0.08	0.00	0.00	0.13	0.14
SC (Mariko et al., 2020)	RE	F1	**0.96**	0.81	0.74	0.61	0.93	0.90	0.19	0.67	0.90	0.85
FinQA (Chen et al., 2021)	FINQA	EmAcc	0.57	**0.63**	0.00	0.06	0.16	0.54	0.04	0.00	0.31	0.32
TATQA (Zhu et al., 2021)	FINTQA	EmAcc	**0.80**	0.13	0.18	0.01	0.26	0.60	0.00	0.00	0.48	0.52
ConvFinQA (Chen et al., 2022)	FINTQA	EmAcc	0.69	**0.76**	0.43	0.25	0.21	0.65	0.20	0.00	0.48	0.58
EDTSUM (Zhou et al., 2021)	SUM	Rouge-1	0.27	0.20	**0.39**	0.25	0.11	0.15	0.13	0.12	0.15	0.18
		BertScore	0.54	0.67	**0.72**	0.68	0.41	0.47	0.38	0.51	0.48	0.49

Table 1: Comparison of zero-shot and few-shot performance between our model and baseline LLMs on seven general financial tasks. "*" represents the evaluation result from the previous paper FinBen (Xie et al., 2024).

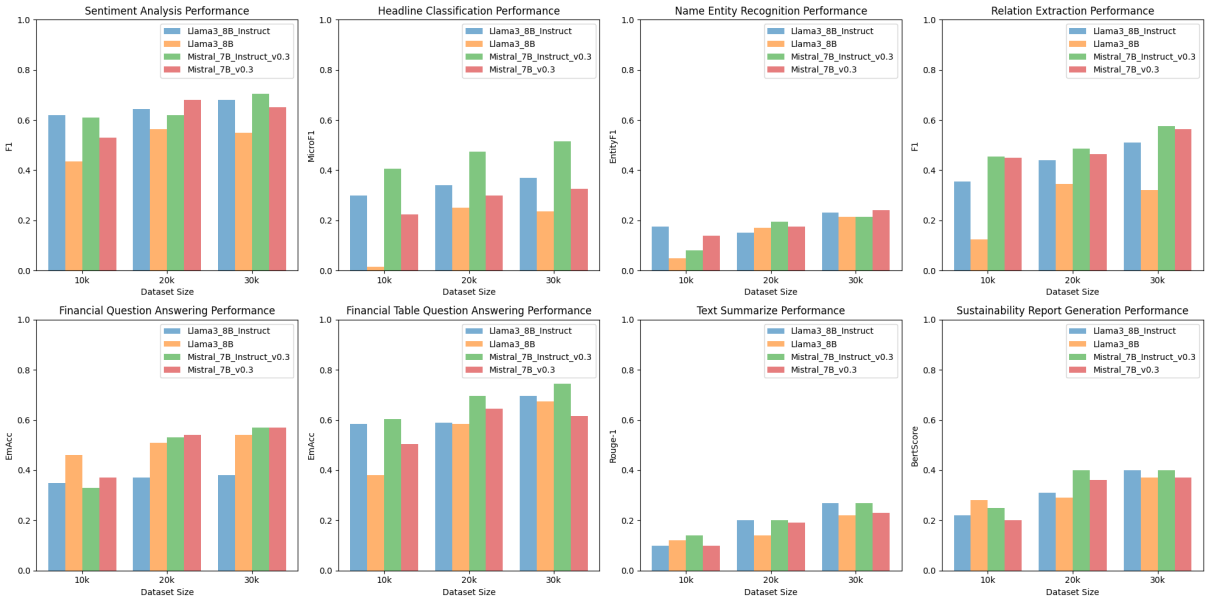


Figure 6: Ablation study results of our models SUSGEN-GPT trained on 10k, 20k, and 30k subsets of our dataset SUSGEN-30K, illustrating the data scaling effect across eight financial NLP tasks.

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 (Ours)	**0.20**	**0.40**	**0.27**	0.32	**0.12**	**0.07**	**0.05**

Table 2: Comparison of sustainability report generation performance on TCFD-BENCH between our model and CHATREPORT.

complex patterns in financial and ESG-related data, leading to higher scores in various metrics such as F1 and MicroF1.

In FiQASA, SUSGEN-GPT reaches an F1 score of 0.72 at the 30k scale, up from 0.66 at 10k. Similarly, in the FOMC dataset, SUSGEN-GPT achieves an F1 score of 0.70 at 30k, compared to 0.61 at 10k. In Relation Extraction tasks like FinRED, the model improves from an F1 score of 0.04 at 10k to 0.23 at 30k, showing substantial performance gains as the dataset grows. Likewise, in NER tasks, the EntityF1 score increases from 0.27 at 10k to 0.35 at 30k. These findings clearly demonstrate that increasing the dataset size enables SUSGEN-GPT to capture more complex patterns, leading to better overall performance in financial NLP tasks.

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 E. The detailed comparison of these models is illustrated in Figure 7. ChatReport is a robust, publicly available text generation model for creating ESG reports in the TCFD format, leveraging the capabilities of GPT-3.5. However, SUSGEN-GPT

Example Prompt	Model	Example Output
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 Sustainable Sourcing: Walmart sets standards for key commodities and engages suppliers through initiatives like Project Gigaton™ ..... ***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?	ChatReport	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.
	SusGen	The organization, Walmart, identifies several climate-related risks and opportunities in various time horizons as follows: **Short Term (Up to One Year):** * 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. **Medium Term (One To Ten Years):** * Transitioning towards net zero emissions by implementing more aggressive decolonization 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. **Long Term (More Than Ten Years):** * Exploration into emerging low-carbon transportation solutions to offset reliance on fossil fuels within their value chains. * Investment in research and 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. 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.

Figure 7: 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.

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.

Limitations

Limited Model Performance One of the limitations of our work is the performance of our models, which is influenced by resource constraints. The large models we utilize are typically in the range of 7B/8B in terms of parameters. Due to limited resources, we have not had the opportunity to explore the potential benefits of 70B parameter or larger models. As a result, our models may not achieve their full potential performance, and this is an aspect we are mindful of as we continue our work. We aim to address these limitations in the future and strive to improve the performance of our models as resources allow.

Limited Evaluation from Expert While experts have conducted quality analysis for certain cases, the evaluation of large-scale data relies on automated scores such as BLEU and ROUGE metrics. These automated evaluations, while valuable, may potentially introduce biases and lack the nuanced understanding that human expertise provides, particularly in the ESG (Environmental, Social, and Governance) domain. Furthermore, there is a shortage of expert resources in the ESG domain, which limits the comprehensive evaluation of the data with respect to ESG factors. As a result, the evaluation may not fully capture the depth and complexity of the ESG-related aspects of the data.

Unsuitable for Vary ESG Subtasks The model's performance may exhibit significant variability across different subfields, highlighting the necessity for targeted research and optimization for specific ESG-related subtasks. For example, certain subfields, such as renewable energy investment or sustainable supply chain management, might require more customized approaches to ensure that the model's performance meets the expected standards. These variations necessitate a more granular understanding of each subfield's unique characteristics and requirements, demanding further data gathering and model adjustments.

Insufficient Diverse Report Template Despite the significant progress achieved by our model, there are still notable limitations concerning the comprehensiveness of the dataset. Firstly, our model was predominantly trained on a limited number of reports in the TCFD (Task Force on Climate-related Financial Disclosures) format, with sparse representation from other key standards such as the GRI (Global Reporting Initiative), SASB (Sustainability Accounting Standards Board), and CDP (Carbon Disclosure Project). This narrow data source limits the model's generalization capabilities and may hinder its performance when dealing with reports adhering to different standards and formats.

Ethical Considerations

False Information from LLMs One of the pressing issues in this work is the phenomenon of model hallucination, where the model generates information that is not present in the input data. This problem is particularly significant when applying advanced techniques like LLMs to financial data, as the generation of false information can have serious implications. For instance, inaccurate or misleading financial reports can lead to incorrect business decisions, regulatory non-compliance, and loss of stakeholder trust. Addressing this issue requires ongoing research and development to improve the reliability and accuracy of NLP models.

Bias towards Firm Perspective Another consideration is the inherent bias towards the firm's perspective in the extracted information from corporate sustainability reports. This bias arises because the data predominantly originates from the companies themselves, potentially leading to a one-sided view that may overlook critical aspects such as stakeholder opinions and third-party assessments. To mitigate this, future work should aim to incorporate a more diverse set of data sources, including independent audits and reports from non-governmental organizations (NGOs), to provide a more balanced view of corporate sustainability practices.

License of the Tool To ensure that our tool is accessible and adaptable by all stakeholders, we have chosen to release it under the Apache License 2.0 later. This open-source license allows for wide distribution, usage, and modification of the tool, thereby facilitating collaborative development and continuous improvement. By doing so, we aim to foster a community of practice that can collectively address the challenges and leverage the opportunities presented by the use of NLP in financial contexts. This approach not only democratizes access to advanced technologies but also encourages transparency and accountability in their application.

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.

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Mérouane Debbah, Étienne Goffinet, Daniel Hesslow, 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.

Dan Biderman, Jose Gonzalez Ortiz, Jacob Portes, Mansheej Paul, Philip Greengard, Connor Jennings, Daniel King, Sam Havens, Vitaliy Chiley, Jonathan Frankle, et al. 2024. Lora learns less and forgets less. *arXiv preprint arXiv:2405.09673*.

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, Anaïs 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. Convfinqa: Exploring the chain of numerical reasoning in conversational finance question answering. *Preprint*, arXiv:2210.03849.

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*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar

Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaoqing Ellen Tan, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aaron Grattafiori, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alex Vaughan, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Franco, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, Danny Wyatt, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Firat Ozgenel, Francesco Caggioni, Francisco Guzmán, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Govind Thattai, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Karthik Prasad, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kun Huang, Kunal Chawla, Kushal Lakhotia, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Maria Tsimpoukelli, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikolay Pavlovich Laptev, Ning Dong, Ning Zhang, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Rohan Maheswari, Russ Howes, Ruty Rinott, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Kohler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vítor Albiero, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaofang Wang, Xiaojian Wu, Xiaolan Wang, Xide Xia, Xilun Wu, Xinbo Gao, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yuchen Hao, Yundi Qian, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, and Zhiwei Zhao. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Rasmus Jørgensen, 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.

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.

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.

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.

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

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

Alexandra Luccioni, Emily Baylor, and Nicolas Duchene. 2020. Analyzing sustainability reports using natural language processing. *Preprint*, arXiv:2011.08073.

Macedo Maia, Siegfried Handschuh, André Freitas, Brian Davis, Ross McDermott, Manel Zarrouk, and Alexandra Balahur. 2018. Www'18 open challenge: financial opinion mining and question answering. In *Companion proceedings of the the web conference 2018*, pages 1941–1942.

Dominique Mariko, Hanna Abi Akl, Estelle Labidurie, Stephane Durfort, Hugues De Mazancourt, and Mahmoud El-Haj. 2020. Financial document causality detection shared task (fincausal 2020). *arXiv preprint arXiv:2012.02505*.

Corentin Masson and Patrick Paroubek. 2024. Evaluating topic model on asymmetric and multi-domain financial corpus. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 6515–6529, Torino, Italia. ELRA and ICCL.

Jingwei Ni, Julia Bingler, Chiara Colesanti-Senni, Mathias Kraus, Glen Gostlow, Tobias Schimanski, Dominik Stammbach, Saeid Ashraf Vaghefi, Qian Wang, Nicolas Webersinke, Tobias Wekhof, Tingyu Yu, and Markus Leippold. 2023. CHATREPORT: Democratizing sustainability disclosure analysis through LLM-based tools. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 21–51, Singapore. Association for Computational Linguistics.

OpenAI. 2023a. Gpt-4 technical report. Technical report, OpenAI.

OpenAI. 2023b. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

OpenAI. 2023c. Instructgpt: Ai for generating instructions. https://openai.com/research/instructgpt/.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Agam Shah, Suvan Paturi, and Sudheer Chava. 2023a. Trillion dollar words: A new financial dataset, task & market analysis. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6664–6679, Toronto, Canada. Association for Computational Linguistics.

Agam Shah, Ruchit Vithani, Abhinav Gullapalli, and Sudheer Chava. 2023b. Finer: Financial named entity recognition dataset and weak-supervision model. *arXiv preprint arXiv:2302.11157*.

Raj Sanjay Shah, Kunal Chawla, Dheeraj Eidnani, Agam Shah, Wendi Du, Sudheer Chava, Natraj Raman, Charese Smiley, Jiaao Chen, and Diyi Yang. 2022. When flue meets flang: Benchmarks and large pre-trained language model for financial domain. *Preprint*, arXiv:2211.00083.

Soumya Sharma, Tapas Nayak, Arusarka Bose, Ajay Kumar Meena, Koustuv Dasgupta, Niloy Ganguly, and Pawan Goyal. 2022. Finred: A dataset for relation extraction in financial domain. In *Companion Proceedings of the Web Conference 2022*, pages 595–597.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal

Azhar, et al. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023b. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Neng Wang, Hongyang Yang, and Christina Dan Wang. 2023a. Fingpt: Instruction tuning benchmark for open-source large language models in financial datasets. *NeurIPS Workshop on Instruction Tuning and Instruction Following*.

Neng Wang, Hongyang Yang, and Christina Dan Wang. 2023b. Fingpt: Instruction tuning benchmark for open-source large language models in financial datasets. *Preprint*, arXiv:2310.04793.

Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg, and Gideon Mann. 2023. Bloomberggpt: A large language model for finance. *Preprint*, arXiv:2303.17564.

Qianqian Xie, Weiguang Han, Zhengyu Chen, Ruoyu Xiang, Xiao Zhang, Yueru He, Mengxi Xiao, Dong Li, Yongfu Dai, Duanyu Feng, Yijing Xu, Haoqiang Kang, Ziyan Kuang, Chenhan Yuan, Kailai Yang, Zheheng Luo, Tianlin Zhang, Zhiwei Liu, Guojun Xiong, Zhiyang Deng, Yuechen Jiang, Zhiyuan Yao, Haohang Li, Yangyang Yu, Gang Hu, Jiajia Huang, Xiao-Yang Liu, Alejandro Lopez-Lira, Benyou Wang, Yanzhao Lai, Hao Wang, Min Peng, Sophia Ananiadou, and Jimin Huang. 2024. The finben: An holistic financial benchmark for large language models. *Preprint*, arXiv:2402.12659.

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

Hongyang Yang, Xiao-Yang Liu, and Christina Dan Wang. 2023. Fingpt: Open-source financial large language models. *FinLLM Symposium at IJCAI 2023*.

Wei Jie Yeo, Teddy Ferdinan, Przemyslaw Kazienko, Ranjan Satapathy, and Erik Cambria. 2024. Self-training large language models through knowledge detection. *arXiv preprint arXiv:2406.11275*.

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

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

Fengbin Zhu, Wenqiang Lei, Youcheng Huang, Chao Wang, Shuo Zhang, Jiancheng Lv, Fuli Feng, and 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*.

Yi Zou, Mengying Shi, Zhongjie Chen, Zhu Deng, ZongXiong Lei, Zihan Zeng, Shiming Yang, HongXiang Tong, Lei Xiao, and Wenwen Zhou. 2023. Esgreveal: An llm-based approach for extracting structured data from esg reports. *arXiv preprint arXiv:2312.17264*.

A Prompt Template

We use alpaca (Taori et al., 2023) prompt template to train and test our models as following.

Alpaca Prompt Template

Below is an instruction that describes a task, paired with an input that provides further context. Write a response that appropriately completes the request.

<begin of special token>
Instruction:
<Instruction>

Input:
<Input>

<end of special token>
Response:
<Output>

For the Llama3 model, the special tokens are *<|begin_of_text|>* and *<|eot_id|>*, while for the Mistral model, they are *[INST]* and *[/INST]*.

B Financial Tasks Definition and Examples

This appendix outlines the definitions of various financial tasks, along with corresponding examples within the financial and ESG domains, as addressed by SUSGEN-GPT. These tasks include Headline Classification (HC), Named Entity Recognition (NER), Relation Extraction (RE), Sentiment Analysis (SA), Financial Question Answering (FIN-QA), Financial Table Question Answering (FIN-TQA), Text Summarization (SUM), and Environmental, Social, and Governance Question Answering (ESG-QA). Each task is designed to tackle specific challenges, providing detailed analysis and insights within the financial and ESG contexts.

B.1 Headline Classification (HC)

Headline Classification refers to the process of categorizing news headlines or report titles into predefined classes, aiming to systematically group them based on their content, context, or subject matter for easier identification and analysis.

Headline Classification

Instruction:
In this task, you're working with English headlines from the MULTIFIN dataset. This dataset is made up of real-world article headlines from a large accounting firm's websites. Your objective is to categorize each headline according to its primary topic. The potential categories are 'Finance', 'Technology', 'Tax & Accounting', 'Business & Management', 'Government & Controls', and 'Industry'. Your response should only include the category that best fits the headline.
Input:
PwC and Alpina Publisher have announced publication of the book The Critical Few
Output:
Industry.

B.2 Named Entity Recognition (NER)

Named Entity Recognition is the task of identifying and classifying proper nouns, such as names of individuals, organizations, locations, and financial entities, within a given text.

Named Entity Recognition

Instruction:
In the sentences extracted from financial agreements in U.S. SEC filings, identify the named entities that represent a person ('PER'), an organization ('ORG'), or a location ('LOC'). The required answer format is: 'entity name, entity type'.
Input:
SUBORDINATED LOAN AGREEMENT HERBERT SMITH LLP Page 1 of 12 7 - December 2007 TABLE OF CONTENTS Clause Headings Page 1.
Output:
HERBERT SMITH, PER

B.3 Relation Extraction (RE)

Relation Extraction involves detecting relationships between entities in text, such as acquisitions, partnerships, regulatory actions, or other interactions.

Relation Extraction

Instruction:
Given the input sentence, please extract the subject and object containing a certain relation in the sentence according to the following relation types, in the format of r̈elation1: word1, word2; relation2: word3, word4.̈ Relations include: product/material produced; manufacturer; distributed by; industry; position held; original broadcaster; owned by; founded by; distribution format; headquarters location; stock exchange; currency; parent organization; chief executive officer; director/manager; owner of; operator; member of; employer; chairperson; platform; subsidiary; legal form; publisher; developer; brand; business division; location of formation; creator.
Input:
While chief executive officer Tim Cook has succeeded in introducing an entirely new category with the Apple Watch, sales remain modest, indicating that Apple will have to keep relying on the iPhone to fuel growth.
Output:
position_held: Tim Cook, chief executive officer

B.4 Sentiment Analysis (SA)

Sentiment Analysis is the task of evaluating the emotional tone in a text to determine whether it conveys a positive, negative, or neutral sentiment.

Sentiment Analysis

Instruction:
What sentiment is expressed in this statement? Choose from Positive, Negative, or Neutral.
Input:
Amidst high expectations, Qualcomm's first-quarter projection fell short of analyst estimates, indicating a potential dip in profitability. The company's estimated profit range of 46 to 50 cents per share, excluding certain items, contrasts sharply with the analyst consensus of 61 cents per share, signaling a possible negative market response.
Output:
Neutral

B.5 Financial Question Answering (FIN-QA)

Financial Question Answering is the task of providing accurate responses to queries based on financial data, reports, and documents, helping users understand complex financial information and make informed decisions.

Financial Question Answering

Instruction:
Read the text below and answer the questions by performing the appropriate numeric operation.
Input:
The NASDAQ index climbed up to the beginning of the opening, but the government bond yields reached a level of nearly 2%.
KOSDAQ is trading at 841.36, down 11.47 points (1.34%) from the previous trading day.Individuals and institutions bought 80 billion won and 18.5 billion won, respectively.On the other hand, foreigners sell 92.8 billion won.
Question: Where did you buy more individuals and institutions in the KOSDAQ market?
Output:
In the KOSDAQ market, more individuals and institutions are bought.The individual bought 80 billion won, while the agency bought 18.5 billion won.Therefore, individuals bought more than institutions.Therefore, the answer is an individual.

B.6 Financial Table Question Answering (FIN-TQA)

Financial Table Question Answering extends traditional question answering by focusing on structured data, specifically extracting and interpreting information from financial tables.

Financial Table Question Answering

Instruction:
Please provide your answer to the financial question based on the context.
Input:
Entergy Corporation and Subsidiaries Management 2019's 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. See Note 2 to the financial statements for further discussion of the business combination and customer credits..... Net Revenue Utility: Following is an analysis of the change in net revenue comparing 2015 to 2014.

Amount (in millions)	
2014 Net Revenue	$5735
Retail Electric Price	187
Volume/Weather	95
Waterford 3 Replacement Steam Generator Provision	-32 (32)
MISO Deferral	-35 (35)
Louisiana Business Combination Customer Credits	-107 (107)
Other	-14 (14)
2015 Net Revenue	$5829

The retail electric price variance is primarily due to: 2022 Formula Rate Plan increases at Entergy Louisiana, as approved by the LPSC, effective December 2014 and January 2015; 2022 an increase in Energy Efficiency Rider revenue primarily due to increases in the Energy Efficiency Rider at Entergy Arkansas, as approved by the APSC, effective July 2015 and July 2014, and new Energy Efficiency Riders at Entergy Louisiana and Entergy Mississippi that began in the fourth quarter of 2014; and 2022 an annual net rate increase at Entergy Mississippi of $16 million, effective February 2015, as a result of the MPSC order in the June 2014 rate case. See Note 2 to the financial statements for a discussion of rate and regulatory proceedings.
Question: What is the net change in net revenue during 2015 for Entergy Corporation?

Output: 94.0

B.7 Text Summarization

Text Summarization involves generating concise and coherent summaries from longer documents, focusing on extracting the most relevant information while preserving the overall meaning.

Text Summarization

Instruction:
You are given a text consists of multiple sentences. Your task is to perform abstractive summarization on this text. Use your understanding of the content to express the main ideas and crucial details in a shorter, coherent, and natural sounding text.
Input:
BEIJING, April 9, 2020 /PRNewswire/ – Baidu, Inc. (NASDAQ: BIDU) (B̈aiduör the C̈ompany), a leading search engine, knowledge and information centered Internet platform and AI company, today announced that pursuant to directives of relevant PRC regulators, Baidu has suspended updating its content on certain newsfeeds channels within Baidu App and conduct maintenance, beginning from April 8, 2020. The Company expects that the suspension may have impact on the marketing services revenue related to the suspended channels. The Company will undertake additional measures to fully comply with the directives of the regulators, ... Chinese governmental policies relating to the Internet and Internet search providers, and general economic conditions in China and elsewhere. ... SOURCE Baidu, Inc.
Output:
Baidu Takes Measures to Comply with Government Directives

B.8 Environmental, Social, and Governance Question Answering (ESG-QA)

Environmental, Social, and Governance (ESG) Question Answering is a specialized task aimed at answering questions related to sustainability, corporate responsibility, and governance practices.

ESG Question Answering

Instruction:
As an ESG specialist at NASDAQ, you help companies improve their ESG scores and can answer any ESG-related questions.
Input:
What are the necessary steps to implement free trade policies in my company?
Output:
Implementing free trade policies in a company involves several necessary steps. Here is a general outline of the process:
1. Research and Evaluate: Start by researching the concept of free trade understanding its impact benefits and challenges. Evaluate how adopting free trade policies will align with your company's goals vision and industry dynamics.
2. Assess Current Trade Policies: Evaluate your current trade policies and identify any barriers or restrictions that could hinder free trade. Analyze your existing agreements tariff structures trade regulations and any protectionist measures.
...
11. Monitor and Review: Continuously monitor the impact of your free trade policies evaluate their effectiveness and make necessary adjustments. Stay updated on evolving trade regulations market trends and any political developments that might affect international trade.
Remember that specific steps may vary depending on your industry company size and market conditions. Consulting with trade experts and professionals in your specific field can provide the necessary guidance tailored to your unique circumstances.

These examples illustrate the diverse capabilities of SUSGEN-GPT in processing and analyzing financial data.

C The Composition of SUSGEN-30K

C.1 Data Statistics

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.

We also provide the token length distribution regarding individual task in our SUSGEN-30K dataset, as shown in Figure 8.

C.2 LLMs automatic pre-processing pipeline

This section explains how we construct the automatic pre-processing pipeline utilizing large language models (LLMs) to handle our aggregated data, which includes five steps in total.

C.2.1 Translation

The first step in our pre-processing pipeline is handling multilingual data, as the aggregated dataset contains content in several languages, including German, French, and Korean, etc. We begin by detecting these non-English portions of the data. Once identified, LLMs are used to automatically translate the non-English text into English. This translation process ensures that all data, regardless of its original language, is standardized in English, which is essential for consistent downstream processing in subsequent steps.

C.2.2 Reformatting

In the second step of our pipeline, we prepare the data for supervised fine-tuning LLMs. To achieve this, all data is reformatted into an instruction-following format, similar to the prompt templates shown in Appendix A. This involves structuring each data point with clear sections for instruction, input, and output. By converting the data into this format, we ensure that it aligns with the instruction-based learning paradigm (Taori et al., 2023; OpenAI, 2023c), optimizing it for fine-tuning large models to follow and execute tasks as instructed.

C.2.3 Anonymization

As illustrated in Figure 4, our dataset is composed of two key parts, one of which is sourced from TCFD Hub. This portion of the data was extracted from publicly available PDFs and reformatted into an instruction-following format. However, the output sections of this data often contain a significant amount of company-specific information. Although these reports are publicly accessible, we aim to minimize potential model biases and protect data privacy. To achieve this, we employ LLMs to anonymize the entity-related information, ensuring that all company names and identifying details are removed. This step helps safeguard privacy while

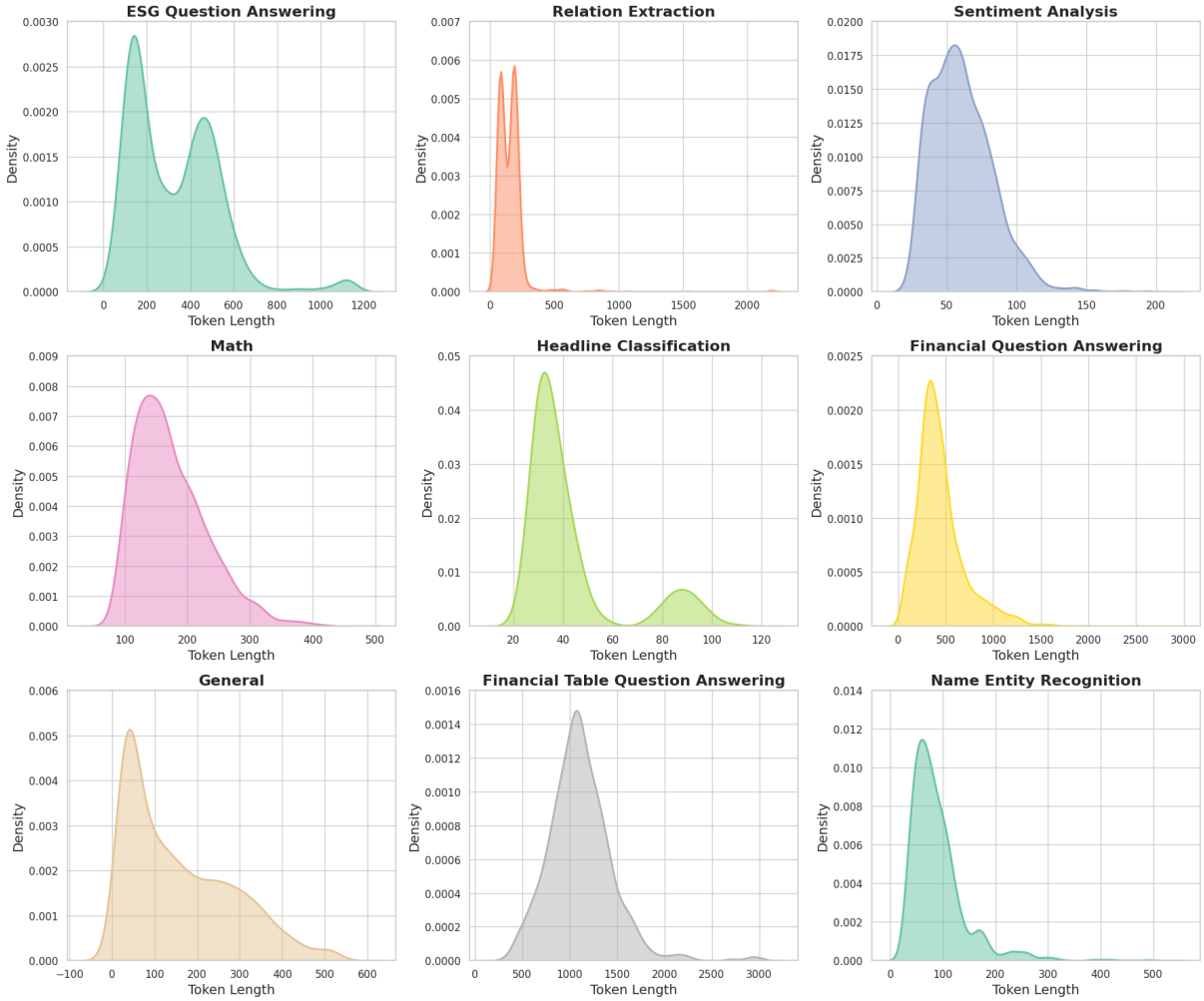


Figure 8: Token length distribution across the 9 sections of SUSGEN-30K dataset.

maintaining the integrity of the dataset. We use Mixtral-8x22B-Instruct-v0.1[8] for this step using the following prompt to process those output:

Prompt for Anonymization

[INST]
Process the text following the instructions below:
1.Replace all the specific company entity name with "we" or "our company".
2.Replace other private information with generic terms.
Text: {Text}
[/INST]

C.2.4 Augmentation

After processing the collected data, we noticed that many datasets contain fixed instruction formats. For example, in Sentiment Analysis, the instruction might always be something like: "Please analyze the sentiment in the input and answer 'negative,' 'positive,' or 'neutral'." This lack of variation in instructions can lead to over-fitting during model training, as the model may become overly accustomed to fixed prompts. To address this issue, we augment the instruction sections by introducing diverse prompts. By generating varied instructions, we inject noise and variation into the data, reducing the likelihood of the model over-fitting to a single fixed instruction format. This augmentation step ensures the model is exposed to a broader range of instructions, improving its generalization capabilities. The prompt that we used is shown below.

Prompt for Augmentation

[8] https://huggingface.co/mistralai/Mixtral-8x7B-v0.1

[INST]
Process the text following the instructions below:
1.Rephrase the whole text without change original meaning and elements.
2.Adjust the processed text to similar length as the original text.
3.Ensure the text is coherent and fluent and output the final text.
Text: {Text}
[/INST]

C.2.5 Synthesizing

For certain tasks, such as ESG-QA, the dataset contains fewer samples compared to other tasks, creating a challenge for effective model training. Drawing inspiration from Common-7B (Li et al., 2024), which demonstrated that training LLMs on synthetic data can improve performance in specific domains, we applied a similar approach to augment the data for these low-sample tasks. The data synthesis process follows a similar methodology to the third (Anonymization) and fourth (Augmentation) steps, but with some key differences. We increased the temperature in the generation process to produce more diverse outputs, and we applied augmentation to both the instruction and output sections. Finally, as with the anonymization step, we ensured that all synthesized outputs were anonymized to protect privacy. This approach resulted in an expanded dataset for the tasks with fewer data samples, enhancing model performance in those areas while maintaining data integrity.

D Our Ablation Study Results about Training Data Scale

D.1 Quantitative Ablation Study

This section showcases the quantitative results of our ablation studies, focusing on the training data scaling effect on our models' performance.

Table 4 presents the zero-shot and few-shot performance of various LLMs, including different configurations of SusGen-GPT with different training dataset sizes of 10k, 20k, and 30k respectively. The metrics evaluated include F1 score, Micro F1, Entity F1, Exact Match Accuracy, Rouge, and BertScore across 8 tasks across multiple financial nlp benchmarks. The results indicate that increasing the dataset size significantly improves the performance of SusGen-GPT, especially after supervised fine-tuning.

D.2 Qualitative Ablation Study

In this section, we present the qualitative ablation results of our models to test their ability to generalize to open-ended tasks. Since many tasks are close-ended and can be demonstrated through quantitative results, we selected an open-ended case to showcase qualitative performance.

The prompt we give to the model is "What is tcfd format in the context of climate change?". As shown in the Table 5, revealing clear trends in model performance as dataset size increases from 10k to 30k. As demonstrated in the table, the SusGen-GPT-30k-Mistral7B-Instruct-v0.3 model produces the most coherent, well-structured, and detailed response to the open-ended prompt. It exhibits a deep understanding of the TCFD framework, provides an appropriate text length, and features minimal noise, reflecting its strong ability to handle complex tasks. The SusGen-GPT-20k model also demonstrates a strong grasp of TCFD, but lacks the text length and detail seen in the 30k model. Meanwhile, the 10k model, while offering a solid response, presents a less structured and slightly less clear explanation. Lastly, the untrained or minimally trained models show noticeable gaps in their answers, with shorter, less precise responses. This progression suggests that increasing the training dataset size significantly enhances the model's capacity to deliver high-quality, open-ended outputs.

E Sustainability Report Generation Examples

We provide examples of sustainability reports generated by SusGen-GPT, as shown in Figure 7 ,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 regulatory 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	fingpt-headline-cls (Wang et al., 2023b)	82,200	EN	20,500	1,500	CLS
HC	fingpt-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	fingpt-ner-cls (Wang et al., 2023b)	13,500	EN	3,500	2,700	CLS
NER	fingpt-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	fingpt-finred (Wang et al., 2023b)	27,600	EN	5,112	5,112	RE+CLS
RE	fingpt-finred-re (Wang et al., 2023b)	11,400	EN	2,140	1,750	RE
RE	fingpt-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-SA/train (link)	4,680	KO	✗	xx	ESG
SA	fingpt-sentiment (Wang et al., 2023b)	76,800	EN	✗	800	
SA	fingpt-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	fingpt-fiqa_qa (Wang et al., 2023b)	17,100	EN	✗	708	
FIN-QA	fingpt-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	fingpt-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	
ESG-QA	ESG-Chat	914	EN	✗	914	
ESG-QA	TCFD_QA	260	EN	✗	1669	
ESG-QA	salmasally	417	FR	✗	417	

Table 3: **Composition of our SusGen-30k 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. ESG-QA: Environmental, Social, and Governance Question Answering. Additionally, we integrate portions of Alpaca and GSM8K into our dataset to mitigate the risk of model over-fitting.

Datasets	Metrics	SusGen GPT-10k Llama3 8B Instruct	SusGen GPT-10k Llama3 8B	SusGen GPT-10k Mistral 7B Instruct v0.3	SusGen GPT-10k Mistral 7B v0.3	SusGen GPT-20k Llama3 8B Instruct	SusGen GPT-20k Llama3 8B	SusGen GPT-20k Mistral 7B Instruct v0.3	SusGen GPT-20k Mistral 7B v0.3	SusGen GPT-30k Llama3 8B Instruct	SusGen GPT-30k Llama3 8B	SusGen GPT-30k Mistral 7B Instruct v0.3	SusGen GPT-30k Mistral 7B v0.3
FiQASA	F1	0.63	0.50	0.63	0.60	0.66	0.56	0.64	0.76	0.66	0.46	**0.72**	0.63
FOMC	F1	0.61	0.37	0.59	0.46	0.63	0.57	0.60	0.60	**0.70**	0.64	0.69	0.67
MultiFin	MicroF1	0.30	0.00	0.40	0.39	0.43	0.42	0.51	0.50	0.46	0.43	**0.52**	0.51
MLESG	MicroF1	0.30	0.03	0.41	0.06	0.25	0.08	0.44	0.10	0.28	0.04	**0.51**	0.14
NER	EntityF1	0.27	0.10	0.02	0.17	0.21	0.25	0.25	0.24	**0.35**	0.33	0.25	0.31
FINER-ORD	EntityF1	0.08	0.00	0.14	0.11	0.09	0.09	0.14	0.11	0.11	0.10	**0.18**	0.17
FinRED	F1	0.04	0.02	0.06	0.05	0.16	0.21	0.11	0.09	0.16	**0.23**	0.19	0.17
SC	F1	0.67	0.23	0.85	0.85	0.72	0.48	0.86	0.84	0.86	0.41	**0.96**	**0.96**
FinQA	EmAcc	0.35	0.46	0.33	0.37	0.37	0.51	0.53	0.54	0.38	0.54	**0.57**	**0.57**
TATQA	EmAcc	0.59	0.41	0.67	0.58	0.59	0.57	0.69	0.65	0.69	0.62	**0.80**	0.65
ConvFinQA	EmAcc	0.58	0.35	0.54	0.43	0.59	0.60	0.70	0.64	0.70	**0.73**	0.69	0.58
EDTSUM	Rouge-1	0.10	0.12	0.14	0.10	0.20	0.14	0.20	0.19	0.21	0.22	**0.27**	0.23
	Rouge-2	0.08	0.03	0.10	0.04	0.08	0.04	0.11	0.10	0.08	0.05	**0.13**	0.11
	Rouge-L	0.10	0.10	0.19	0.08	0.16	0.15	**0.22**	0.10	0.16	0.19	**0.22**	0.15
	BertScore	0.44	0.47	0.43	0.45	0.50	0.49	0.45	0.50	0.51	0.48	**0.54**	0.51
TCFD-BENCH	Rouge-L	0.13	0.10	0.17	0.13	0.14	0.11	0.18	0.16	0.14	0.12	**0.20**	0.15
	BertScore	0.22	0.28	0.25	0.20	0.31	0.29	**0.40**	0.36	0.31	0.29	**0.40**	0.37
	METEOR	0.15	0.15	0.14	0.15	0.20	0.19	0.22	0.21	0.24	0.21	**0.27**	0.23
	BLEU-1	0.12	0.11	0.17	0.18	0.20	0.19	0.27	0.20	0.20	0.17	**0.32**	0.19
	BLEU-2	0.05	0.03	0.05	0.05	0.05	0.03	**0.12**	0.09	0.05	0.03	**0.12**	0.10
	BLEU-3	0.02	0.01	0.03	0.02	0.02	0.01	0.06	0.04	0.02	0.01	**0.07**	0.04
	BLEU-4	0.01	0.00	0.02	0.01	0.01	0.01	0.04	0.02	0.01	0.01	**0.05**	0.03

Table 4: Ablation study results of our models on 8 different financial NLP tasks, using three different subset datasets of 10k, 20k, and 30k examples. The suffix in each model name denotes the base model used for initialization.

Models	Example Prompt	Example Output
Mistral7B-Instruct-v0.3	***Question Answering:*** TCFD stands for Task Force on Climate-related Financial Disclosures. It's an initiative by the Financial Stability Board (FSB) to develop a framework for companies and other organizations to disclose climate-related financial risks and opportunities. The goal of this framework is to help investors, lenders, and insurance underwriters assess and price these risks more accurately, thereby promoting more informed decision making and better management of climate risk. The TCFD format provides guidance on how organizations can report their exposure to climate-related risks and opportunities, including physical risks such as extreme weather events or sea level rise, transition risks related to policy changes or technological shifts, and business opportunities arising from low carbon technologies or new markets. By adopting the TCFD format, organizations can improve transparency and comparability of climate-related information, which can ultimately lead to more effective management of climate risk and more sustainable investment decisions.	Basic understanding of TCFD, but suboptimal text length and unclear structure.
SUSGEN-GPT-10k-Mistral7B-Instruct-v0.3	***Question Answering:*** TCFD stands for Task Force on Climate-related Financial Disclosures. It's an initiative by the Financial Stability Board (FSB) to develop a framework for companies and organizations to disclose information related to climate-related risks and opportunities, with the goal of improving transparency and understanding of these factors for investors, lenders, insurers, and other stakeholders. The TCFD recommendations cover four areas: Governance, Strategy, Risk Management, Metrics and Targets. Companies that adopt TCFD reporting provide detailed information about their approach to managing climate risk and opportunity, including their governance structures, strategy for addressing climate issues, and metrics used to measure progress towards sustainability goals. This information helps stakeholders assess the resilience of an organization's business model and its ability to manage climate-related risks effectively. By providing clear, consistent, and comparable information, TCFD aims to enhance market efficiency, reduce financial risks, and support informed decision-making. In summary, TCFD is a set of recommendations aimed at enhancing an organization's transparency regarding climate-related risks and opportunities, enabling better investment decisions and improved management of climate-related risks.	Good understanding of TCFD, improved text length, but unclear explanation structure.
SUSGEN-GPT-20k-Mistral7B-Instruct-v0.3	***Question Answering:*** TCFD stands for Task Force on Climate-related Financial Disclosures. It's an international organization founded by the Financial Stability Board (FSB) to develop recommendations for companies and other organizations to disclose climate-related risks and opportunities in their financial reporting. The TCFD framework includes four core areas of climate-related risk management: Governance, Strategy, Risk Management, Metrics and Targets. Companies that adopt TCFD disclosures are expected to provide transparent and consistent information about how they identify, assess, and manage climate-related risks and opportunities, enabling investors, lenders, and other stakeholders to make informed decisions based on this information. By using TCFD format, organizations can demonstrate their commitment to transparency and accountability regarding climate change and its potential impact on their business operations and financial performance.	Excellent TCFD understanding, clear explanation structure, but poor text length.
SUSGEN-GPT-30k-Mistral7B-Instruct-v0.3	***Question Answering:*** TCFD stands for Task Force on Climate-related Financial Disclosures. It is an international organization founded by the Financial Stability Board (FSB) in 2015 to develop recommendations for companies and other organizations to disclose climate-related risks and opportunities in a more transparent way. The TCFD framework focuses on four key areas of climate-related risk management: Governance, Strategy, Risk Management, Metrics and Targets. Companies that adopt TCFD reporting are expected to provide detailed information about their climate-related risks and opportunities, enabling investors and stakeholders to make informed decisions based on this data. TCFD reports are typically presented in a standardized format, allowing for easy comparison between different organizations. This promotes greater transparency and accountability when it comes to addressing climate-related issues within financial portfolios. By providing clear and consistent information, TCFD aims to improve the overall understanding of climate-related risks and opportunities, and help organizations better manage these factors to mitigate potential negative impacts and capitalize on emerging opportunities. In summary, TCFD format refers to the specific disclosure requirements set forth by the TCFD for organizations to report on climate-related risks and opportunities in a transparent and comparable manner.	Deep understanding of TCFD, appropriate text length, clear answer structure, and minimal noise.

Table 5: Ablation study results of our models on open-ended question answering, using three different subset datasets of 10k, 20k, and 30k examples. The suffix in each model name denotes the base model used for initialization.