KRX-Bench: Automating Financial Benchmark Creation via Large Language Models

Guijin Son[1], **Hyunjun Jeon**[1], **Chami Hwang**[2], **Hanearl Jung**[1]
OneLineAI[1], Korea Exchange[2],
{guijn.son, hyunjun.jeon, earl}@onelineai.com,
hcharm2ing@krx.co.kr

Abstract

Inaccuracies or outdatedness of large language models (LLMs) in the finance domain may lead to misguided decisions and substantial financial losses, highlighting the importance of appropriate tools to evaluate and identify LLMs ready for production. In this work, we introduce **KRX-Bench**, an automated pipeline for creating financial benchmarks via GPT-4. To demonstrate the effectiveness of the pipeline, we create **KRX-Bench-POC**, a benchmark assessing the knowledge of LLMs in real-world companies. This dataset comprises 1,002 questions, each focusing on companies across the U.S., Japanese, and Korean stock markets. Our findings indicate that KRX-Bench can autonomously produce accurate benchmarks, achieving a minimal "false positive" rate of 1%. Notably, we find that despite leveraging GPT-4 as the generator, our pipeline can supplement enough knowledge to create questions beyond its limitations. Finally, we explore various applications of KRX-Bench, including generating open-ended, multilingual questions and reasoning benchmarks, showcasing its versatility in creating comprehensive evaluation tools for LLMs. We make our pipeline and dataset publicly available and integrate the evaluation code into EleutherAI's Language Model Evaluation Harness.

Keywords: Large Language Model, Benchmark,Finance

1. Introduction

With the advent of highly capable large language models (LLMs), the financial industry now faces pre-industrial adoption across diverse tasks (Son et al., 2023a; Callanan et al., 2023). However, key concerns surrounding the accuracy, reasoning skills, and safety of the content generated by LLMs raise diverse concerns (Wei et al., 2023; Bang et al., 2023; Alkaissi and McFarlane, 2023). While certain fields, such as arts or music, may tolerate or even embrace a degree of imaginative deviation (or "hallucination") in the outputs of LLMs, sectors like Medicine and Finance are notably intolerant of such inaccuracies. In the financial domain, hallucinations by LLMs can propagate misinformation, potentially leading to misguided investment decisions and consequent financial losses. However, existing research has predominantly focused on assessing financial LLMs' reasoning capabilities (Chen et al., 2021, 2022) or proficiency in singular tasks (Son et al., 2023b; Malo et al., 2014; Loukas et al., 2022), leaving a critical gap in understanding their comprehension of the real-world financial landscape.

To bridge this gap, we introduce **KRX-Bench**, a pipeline for the automated creation of financial benchmarks. The automated nature of KRX-Bench is ideally suited for generating a dynamic benchmark that can self-update, making it uniquely capable of capturing the rapidly changing financial sector. To demonstrate its effectiveness, we create **KRX-Bench-POC** a benchmark comprising 1,002 instances, each about companies across the U.S., Japanese, and Korean stock markets. Our assessment confirms that KRX-Bench can autonomously produce accurate benchmarks. We apply machine-learned techniques and verify that the benchmark is free of unwanted artifacts. Furthermore, a qualitative review highlights an exceptionally low "false positive" rate of 1%, indicating that human annotators deem the vast majority of questions reliable and answerable. We observe the best performing openly available LLMs (e.g., *Qwen1.5-72B*, and *Llama-2-70B*) to score below 80% suggesting room for improvement. Surprisingly, *GPT-4-Turbo* the most capable LLM available and the generator of the benchmark scores below 90% suggesting that the pipeline is capable of creating beyond the knowledge of the generator.

Finally, we demonstrate diverse applications of KRX-Bench, including creating open-ended, multilingual, and reasoning-focused benchmarks, with only minor modifications to the prompts or input documents. Our findings suggest that the pipeline can be readily adapted to generate more challenging questions simply by updating the input documents. Our contributions are twofold:

1. We present **KRX-Bench** an automated pipeline for creating financial benchmarks.

2. We introduce **KRX-Bench-POC**, to our knowledge, the first benchmark evaluating the knowledge of LLMs across multiple stock markets.[1]

[1] https://anonymous.4open.science/r/KRX-Bench-1FCE/

Proceedings of the Joint Workshop of the 7th FinNLP, the 5th KDF, and the 4th ECONLP, pages 10–20

2. Related Works

2.1. Financial Large Language Models

The financial industry has shown interest in adopting LLMs, demonstrated by the launch of BloombergGPT (Wu et al., 2023), a 50 billion parameter model specifically trained for Finance. An array of openly-available financial LLMs has followed the model, each focusing on reading comprehension (Cheng et al., 2023), financial task solving (Wang et al., 2023), or multimodality (Bhatia et al., 2024). Furthermore, multiple research have explored the possibility of LLMs to replace human analysts by either training open-source language models on tailored datasets (Son et al., 2023a) or prompting proprietary language models to solve CFA exams (Callanan et al., 2023). However, adopting LLMs in Finance faces hurdles, primarily due to their tendency to generate inaccurate information, known as *hallucinations* (Huang et al., 2023). This issue is critical in Finance, where incorrect data can lead to poor decision-making and significant financial losses. Furthermore, the risk of spreading false information through LLMs could be considered unethical or even fraudulent, slowing their integration into financial operations.

2.2. Evaluation of Financial LLMs

LLM evaluation tools have progressed from basic question-answering tasks (Rajpurkar et al., 2016) to complex reasoning (Cobbe et al., 2021) or large-scale knowledge benchmarks (Hendrycks et al., 2020; Son et al., 2024). The assessment of financial LLMs has followed a similar path, initially focusing on evaluating specific tasks (Chen et al., 2021, 2022; Loukas et al., 2022) to employing a comprehensive set of benchmarks (Xie et al., 2024; Shah et al., 2022) for a more thorough evaluation. However, the field lacks appropriate tools to accurately assess financial LLMs' grasp of the real-world financial environment, such as knowledge of company details, business objectives, and financial regulations. Furthermore, the financial market changes quickly over time—new companies emerge, and existing ones transform, quickly rendering benchmarks focused on real-life knowledge obsolete (Son et al., 2023b).

To this end, we introduce **KRX-Bench**, a pipeline for the automated generation of financial benchmarks, designed to adapt continuously to the dynamic financial market. Additionally, we provide a set of questions generated through the pipeline, which, to the best of our knowledge, is the first to evaluate LLMs across multiple stock markets and regulatory environments.

3. KRX-Bench

In this section, we elaborate on the **KRX-Bench** pipeline (3.1)) and conduct a proof of concept study leveraging the pipleine (Section 3.2).

3.1. KRX-Bench Pipeline

The **KRX-Bench** is an automated pipeline designed for generating financial benchmarks. It leverages *GPT-4-Turbo* to craft challenging questions from existing corpora, encompassing three main steps.

Question and Answer Generation In this step, we provide a document to *GPT-4-Turbo* and prompt it to generate Q&A pairs from the text. The document may be annual reports, documentation on financial lawsuits, or anything of the user's choice. While the model's cognitive capacity bounds the question generation, it can still craft questions extending beyond its pre-trained knowledge by leveraging the supplementary materials.

Creation of Distractors To reformat the Q&A pairs generated in the prior step to multiple-choice questions, we generate distractors (wrong answer choices). Simply choosing random answers as distractors could make them too easily distinguishable, so we employ *GPT-4-Turbo* to create distractors of high quality. For each question Q^*, we use the BM25 algorithm to find 10 similarly worded questions $[Q^1...Q^{10}]$ and then instruct *GPT-4-Turbo* to adapt the corresponding answers $[A^1...A^{10}]$ into plausible incorrect options for Q^*. To ensure the distractors' quality, we filter by two heuristic rules:

1. Exclude options mentioning companies irrelevant to the question.

2. Remove any answer option whose length significantly deviates from the average length of incorrect answers to maintain a uniform answer structure.

If the filtering process yields more than four distractors, we randomly select four from the remaining options.

Quality Control A critical condition for a fully automated pipeline for benchmark creation without a human in the loop is to minimize the inclusion of "false positives" or unanswerable questions. Accordingly, in this final step, we prompt *GPT-4-Turbo* to identify and eliminate unanswerable questions. For a comprehensive list of criteria used to determine unanswerability, see Figure 1 for the prompts used throughout the pipeline.

Question and Answer Generation

{CONTEXT}

Instruction: From the given the text generate an English question and answer pair. Do not ask for quantitative questions ask about the details about the company. Make sure to include the name of the company in the question. The name of the company is *{COMPANY}*.

Generate in Q: <str> A: <str> format
Question:

Creation of Distractors

You will given a question, gold answer and one irrelevant answer. Your job is to transform the irrelvant answer to a well-designed wrong answer. Replace mentions of different companies to the company of interest in the question. Alter the detail a bit to make good wrong answer.

Question: *{QUESTION}*
Gold Answer: *{GOLD}*
Irrelevant Answer: *{CANDIDATE}*
Good Wrong Answer:

Quality Control

You are the final sensitivity reader for a benchmark that is about to be published.
Read through a question included in the benchmark and evaluate whether the question is answerable.
A question is deemed unanswerable if:

1. The question does not include the name of a company.
2. The question is based on information that is outdated or no longer relevant.
3. The question assumes knowledge that is not commonly available or requires specialized expertise beyond the scope of the benchmark audience.
4. The question's phrasing is ambiguous or can lead to multiple equally valid interpretations.
5. The information needed to answer the question correctly is not present within the context provided in the benchmark.
6. The question contains biases or assumptions that could disadvantage certain groups of people or promote stereotypes.

Please ensure that each question in the benchmark meets these criteria to be considered answerable and appropriate for publication. Explain whether the question meets each criteria and return [[Yes]] for answersable questions and [[No]] for unanswerable questions.

Question: *{QUESTION}*
Gold Answer: *{GOLD}*
Wrong Answers:
1. *{WRONG ANSWER1}*
2. *{WRONG ANSWER2}*
3. *{WRONG ANSWER3}*
4. *{WRONG ANSWER4}*

Decision:

Figure 1: Prompts used throughout the **KRX-Bench** pipeline.

3.2. Proof of Concept

To demonstrate the KRX Bench pipeline's effectiveness in practice, we introduce **KRX-Bench-POC**, a benchmark dataset of 1,002 questions from companies of three nations: the United States, Japan, and Korea.

KRX-Bench-POC Initially, we compiled a dataset from annual reports across three nations: the United States, Japan, and Korea. For the U.S. (Loh) and Japan (chakki), we collect from existing resources, while Korean reports are from DART[2], a digital repository for company filings. The selection is not based on the latest fiscal data—U.S. reports are from 2022, and Japan's from 2018. This is because this section aims to showcase the capability of the pipeline rather than currently creating up-to-date benchmarks. We plan to release updated versions of the benchmark in the future. To ensure consistency, we randomly chose 500 annual reports each from Japan and Korea. For details on

[2] https://dart.fss.or.kr/main.do

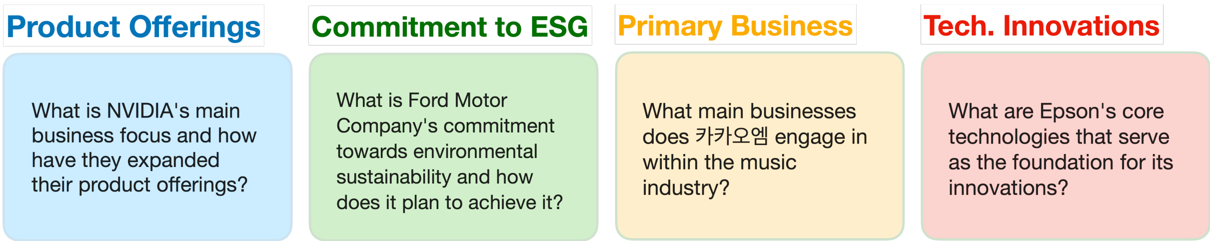

Figure 2: Selected samples of questions included in the **KRX-Bench-POC**.

the dataset composition, see Table 1.

Following this step, we execute the KRX Bench pipeline on the collected annual reports and generate multiple-choice questions. Following a quality filtering process, we retain a total of 1003 questions: 373 for the US, 319 for Korea, and 311 for Japan.

Country	# of Doc	Av. Length	Fiscal Year
United States	494	55479	2022
Korea	2896	5158	2023
Japan	3718	1339	2018

Table 1: Statistics on the collected annual reports.

Diversity We evaluate the diversity of KRX-Bench-POC, by randomly selecting 99 questions, 33 from each nation, and categorizing each by topic. This survey reveals that the pipeline yields a broad spectrum of 15 distinct categories, including Business Goals, Product Offerings, Financial Policy, and Business Strategy, with no single category predominating. Primary Business emerged as the most represented category. For a detailed breakdown of each category and sample questions, refer to Table 2 and Figure 2, respectively.

Category	Frequency
Primary Business	33
Business Goals	14
Company Mission	10
Industry	10
Business Operations	8
Product Offerings	7
ETC	17

Table 2: A survey on the category of generated questions. The ETC category includes the following: Financial Policy, Innovation, Business Strategy Commitment to ESG, Long Term Strategy, Mid Term Strategy, Global Strategy, Company History.

Quality In this section, we analyze the quality of the generated dataset. First, we test whether the benchmark includes potentially exploitable artifacts(e.g., shortcuts or patterns) that LLMs might abuse to solve the question. Specifically, we test two machine-learned features: (1) Similarity-Based Feature: We evaluate if the option most similar to the question, using BM25 and BERT[3] for similarity measurements, is likely to be correct; (2) Zero-Shot Classifier Feature: We employ a zero-shot classifier, trained on natural language inference tasks, to determine if it can accurately solve the questions without specific training (Laurer et al., 2023). Table 3 presents a performance comparison between the machine-learned features on our KRX-Bench-POC and Hellaswag (Zellers et al., 2019), a widely adopted benchmark for commonsense reasoning. Similarity-based measures on KRX-Bench-POC outperform random guessing but achieve similar or lower scores than their performance on Hellaswag. This indicates that KRX-Bench-POC maintains a comparable level of challenge and avoids introducing excessive artifacts compared to established benchmarks.

Feature	KRX-Bench-POC	Hellaswag
Random Baseline	20.0%	25.0%
Similarity (BM-25)	37.3%	**54.1%**
Similarity (BERT)	**39.8%**	32.2%
Zero-Shot Classifier	20.4%	**25.1%**

Table 3: Accuracy of machine-learned models on the **KRX-Bench-POC** and *Hellaswag*.

Furthermore, we assess the quality control step introduced in Section 3.1 through qualitative analysis, examining both answerable and unanswerable questions classified by *GPT-4-Turbo*. Two of the authors review 200 randomly sampled questions (100 deemed answerable and 100 deemed unanswerable by *GPT-4-Turbo*) without prior knowledge of *GPT-4*'s judgments. Results, shown in figure 2, reveal a remarkably low "false positive" rate of only 1%, indicating that very few unanswerable questions were incorrectly labeled as answerable. Although achieving a 0% "false positive" rate would be ideal, even human-curated datasets struggle to meet this standard. The observed 1% rate is

[3] We use *all-MiniLM-L6-v2* from the Sentence Transformers library (Reimers and Gurevych, 2019).

sufficiently low for reliable evaluation. Additionally, the "true negative" rate of 43% highlights the effectiveness of our pipeline's quality control in mirroring human judgment, ensuring the pipeline's benchmark generation abilities.

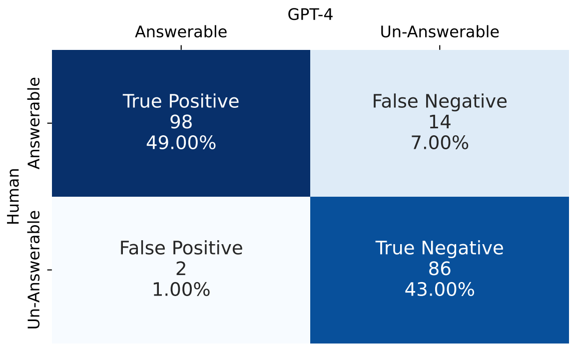

Figure 3: A confusion matrix comparing the decision of human annotators against the quality control step by *GPT-4-Turbo*.

4. Experimental Setup

In this section, we explain our experimental setup for evaluating different LLMs on the **KRX-Bench-POC**.

4.1. Models

In this work we evaluate 12 different LLMS ranging in different size for evaluation. The evaluated models include: (1) *Llama-2* (7B, 13B, 70B)) (Touvron et al., 2023) (2) *Qwen1.5* (0.5B, 1.8B, 4B, 7B, 14B, 72B) (Team) and (3) *GPT-3.5-Turbo* and *GPT-4-Turbo* (OpenAI, 2023). We also evaluate *Japanese-StableLM-Base-Beta-7B* (Lee et al.) and *Llama-2-KOEN-7B* (Junbum, 2023), which are variations of *Llama-2* each continually pre-trained on Japanese and Korean correspondingly.

4.2. Evaluation Methods

For evaluation, we prompt a model to generate the most plausible option via greedy decoding. All models are evaluated in full precision in a 3-shot setting on 8 X A100 80GB GPUs. See Figure 4 for the prompt used in our evaluation. For reproducibility, the evaluation codes used in our research are implemented via LM-Eval-Harness (Gao et al., 2023).

Direct Evaluation

Question: *{QUESTION}*
Options:
A. {OPTION A}
B. {OPTION B}
C. {OPTION C}
D. {OPTION D}
E. {OPTION E}
Answer:

Figure 4: Prompt used in our Direct Evaluation.

5. Results on KRX-Bench-POC

Model Size and Performance Table 4 presents the evaluation results for various models on the KRX-Bench-POC. Larger models consistently outperform smaller ones, indicating a linear scaling trend. This pattern holds for both *Qwen1.5* and *Llama-2* model families, demonstrating that our benchmark aligns with typical benchmark behaviors. Notably, the top models, *Qwen1.5-72B*, and *Llama-2-70B* achieve scores under 80%, indicating room for improvement. This suggests that our pipeline successfully generates challenging benchmarks for state-of-the-art open models without any human supervision.

Regional Bias In Figure 5, we notice a regional bias in model performance; despite all questions being in English, models perform better on questions about U.S. companies than those about Japanese or Korean companies. This trend is consistent across all models, with leading models like *Qwen1.5-72B* and *Llama-2-70B* scoring around 90% for U.S. companies but only about 70% for Japanese and Korean companies. This pattern is also evident in proprietary models such as *GPT-3.5-Turbo* and *GPT-4-Turbo*. Several factors could contribute to this disparity, including the scarcity of English resources on Japanese and Korean companies, which limits the models' ability to acquire knowledge about these companies during pretraining. This implies that leveraging more difficult documents as input, internal documents, for example, could easily elevate the benchmark's difficulty.

Surprisingly, models specifically trained on additional Japanese and Korean data, such as *Japanese-StableLM-Base-Beta-7B* and *Llama-2-KOEN-7B*, show decreased performance across all subsets. Despite being trained on an extra 100B tokens of Japanese and 60B tokens of Korean, these models do not improve scores for questions related to their targeted nations; instead, their overall scores drop. This unexpected outcome may be attributed to two main reasons. Firstly, the added web-crawled tokens might not provide sufficient information about the companies featured in the benchmark. Secondly, further pretraining on dedicated national data could induce catastrophic forgetting, weakening the models' English language problem-solving abilities. This observation chal-

Models	N=3			
	US	KO	JR	Total
Pre-Trained Models				
Qwen1.5-0.5B	20.38	17.87	18.06	18.77
Qwen1.5-1.8B	39.68	24.14	20.97	28.26
Qwen1.5-4B	58.45	31.35	30.65	40.15
Qwen1.5-7B	81.77	47.34	48.06	59.06
Qwen1.5-14B	87.13	57.68	60.65	68.49
Qwen1.5-72B	87.40	72.10	72.58	77.36
Llama-2-7B	42.09	20.38	23.23	28.56
Llama-2-13B	85.52	52.98	51.94	63.48
Llama-2-70B	93.30	71.16	73.23	79.23
Continual Pretrained Models				
Japanese-StableLM-Base-Beta-7B	32.98	21.00	23.87	25.95
Llama-2-KOEN-7B	17.16	19.44	18.06	18.22
Proprietary Models				
GPT-3.5-Turbo	87.13	63.32	66.13	72.19
GPT-4-Turbo	95.44	84.33	84.84	88.20

Table 4: Average accuracy(%) calculated using the Direct method in a 3-shot setting across the entire test set. We report the macro-average accuracy across subjects within each category.

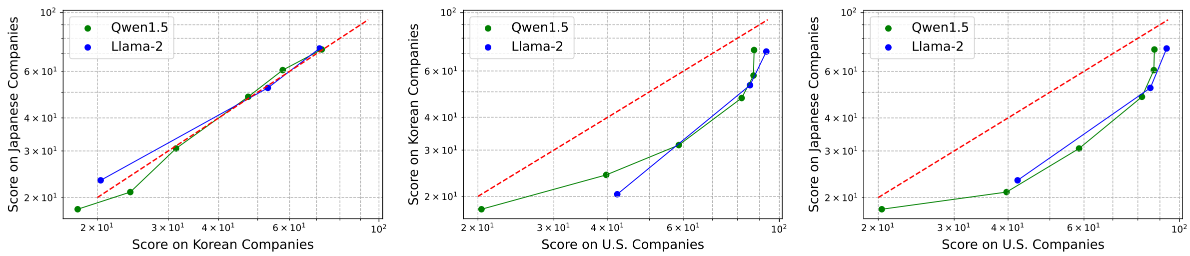

Figure 5: Comparison of Qwen1.5 and Llama-2 models based on the nationality of the companies.

lenges the assumption that language-transferred models can effectively substitute dedicated language models, revealing they may lack comprehensive knowledge.

Can GPT-4 Ace Itself? The KRX-Bench pipeline employs *GPT-4-Turbo* throughout its stages, leading to the following question: Can this pipeline effectively evaluate *GPT-4-Turbo* itself? To address this, we also assess *GPT-4-Turbo*'s performance on the benchmark. Although *GPT-4-Turbo* achieves the highest scores among tested models, it does not excel across all subsets, particularly for the Korean and Japanese questions. This outcome suggests that using supplementary materials for dataset generation enables the generator to formulate questions that extend beyond its knowledge cutoff. Consequently, the pipeline demonstrates its capacity to create benchmarks capable of challenging the model itself, including *GPT-4-Turbo*, thereby validating its utility in generating effective evaluation tools.

6. Practitioner's Guide for Implementing KRX-Bench

In this section, we extend beyond **KRX-Bench-POC** to show different use-cases of the **KRX-Bench** pipeline in generating financial Benchmarks.

6.1. Open-Ended Generation

While the KRX-Bench pipeline primarily generates multiple-choice questions, it can also assess the generative capabilities of language models by simply providing a question and prompting an LLM to generate an answer. This approach offers a more nuanced evaluation of a model's generation ability. To illustrate this, we sample 70 questions from the KRX-Bench-POC dataset to create **KRX-Bench-Gen**. The questions span various categories: Primary Business, Industry, Product Offerings, Business Strategy, Technology, Business Goals, Financial Policy, Commitment to ESG, and Risk. Primary

Question

What significant actions did HOYA Corporation take in 2017 concerning the medical devices sector?

Gold Answer

In 2017, HOYA Corporation established a joint venture company with China's Aohua for a medical flexible endoscope business and acquired ReadSpeaker, a leading company in cloud-based text-to-speech services. They also set up a new production base in Thailand for intraocular lenses used in cataract surgery, and acquired Performance Optics, LLC along with its subsidiaries VISION EASE and Daemyung Optical.

GPT-3.5-Turbo

In 2017, HOYA Corporation made a significant move in the medical devices sector by acquiring the medical unit of the US-based company, Performance Optics. This acquisition allowed HOYA to expand its presence in the ophthalmic lens market and strengthen its position in the medical devices industry. The deal included Performance Optics' manufacturing facilities …

Score:3

GPT-4-Turbo

In 2017, HOYA Corporation, a global technology company based in Tokyo, Japan, known for its innovations in optics and medical devices, took a significant step in the medical devices sector by acquiring Performance Optics, LLC, including its subsidiaries VISION EASE, a global ophthalmic lens marketer and manufacturer, and Daemyung Optical Co. Ltd., a leading optical lens manufacturer in South Korea…

Score:4

Figure 6: Example generations of *GPT-3.5-Turbo* and *GPT-4-Turbo* on open-ended questions.

Business is the largest category with 11 questions, while Risk is the smallest with four, averaging 7.8 questions per category.

Given that pre-trained models without further tuning might struggle with open-ended questions, we focuses on GPT variants. We assess *GPT-3.5-Turbo* and *GPT-4-Turbo*, employing an "LLM-as-a-Judge" approach based on Zheng et al. (2024)'s implementation. This judge model is prompted to rate answers on a scale from 1 to 5. In Table 5 we observe *GPT-4-Turbo* to score slightly higher than *GPT-3.5-Turbo*.

Models	Open-Ended Generation
GPT-3.5-Turbo	3.28
GPT-4-Turbo	3.55

Table 5: Evaluation results of GPT variants on open-ended questions.

To provide additional insight, we include an example of a question and the generated responses in Figure 6. In this example, both models accurately identify HOYA Corporation's acquisition of Performance Optics, yet *GPT-4-Turbo* provides a more detailed response by noting the inclusion of subsidiaries in the acquisition. This illustrates how our benchmark can be utilized to assess both generative capabilities and knowledge depth. The accuracy of evaluations could be further improved by employing more knowledgeable LLM judges with expertise in finance or by incorporating human evaluators.

6.2. Multilinguality

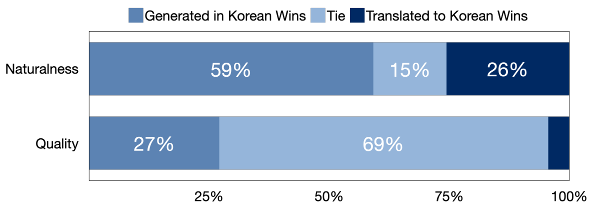

Figure 7: Win rate analysis between generating in Korean and translating a material generated in English to Korean.

OpenAI (2023) reports *GPT-4-Turbo* to have robust multilingual capabilities. Accordingly, we explore where the identical benchmark generation pipeline can be applied to generate benchmarks in languages other than English, specifically Korean. We adapt the pipeline by incorporating "Generate in Korean" into our prompts, generating 250 questions in Korean. We conduct a comparative quality analysis to assess the effectiveness of generating questions directly in Korean versus translating questions from English. We randomly select 250 questions from the KRX-Bench-POC dataset and hire two annotators for evaluation. Presented with pairs of questions—one generated in Korean and the other translated—they are tasked to identify the question that sounds more natural to native Korean speakers and is of higher quality, without knowledge of the questions' generated methodology. The annotators have the option to choose one of the options or declare a tie. Figure 7 indicates that annotators consistently find the directly generated Korean samples more natural for native speakers. We suspect that direct generation allows *GPT-4-Turbo* to leverage its in-context learning abilities to learn from the provided Korean document, thereby commanding better Korean than the translation approach. Quality-wise, annotators considered both methods to yield questions of similar quality 69% of the time, but in 29% of cases, the directly generated samples were preferred. These results demonstrate that our pipeline can be seamlessly adapted to produce high-quality multilingual benchmarks with minimal adjustments.

In Table 6, we report the evaluation results for the subset generated in Korean. Interestingly, unlike our previous experiments *Llama-2-KOEN-7B* outperforms *Llama-2-7B*. We attribute this improvement primarily to the language advantage. Unlike the assessments reported in Table 4, which involved questions about Korea in English, this experiment presented questions in the Korean language. This context likely favored *Llama-2-KOEN-7B*, benefiting from its targeted continual pretraining in Korean.

6.3. Beyond Knowledge Benchmarks

저축은행의 특수관계자가 설정한 사모펀드에 저축은행이 투자하는 것이 가능한가요?

(Can a savings bank invest in a private equity fund established by an affiliated party?)

금융회사가 암호화된 개인신용정보를 전송하기 위해 제3자의 통신회선 서비스를 이용하려고 할 때, 이는 신용정보법 제17조의 개인신용정보 처리 위탁에 해당하나요?

(Does the transmission of encrypted personal information via a third-party communication service qualify as outsourcing under Article 17 of the Credit Information Act?)

Figure 8: Examples of the generated reasoning benchmark. English translations are added for broader accessibility.

This section explores whether the KRX-Bench pipeline can be leveraged to create reasoning benchmarks. Previously, we introduced KRX-Bench-POC to showcase the pipeline's ability to generate benchmarks evaluating LLMs' knowledge of real-world companies. Alongside such knowledge benchmarks, reasoning benchmarks are crucial for a comprehensive assessment of LLMs, focusing on their capacity to apply knowledge logically to solve problems. For this purpose, we compile a set of Korean documents related to financial lawsuits and process them through the same pipeline, producing 100 questions that challenge LLMs to conduct legal reasoning on financial disputes. We choose to generate questions in Korean to preserve the intricate details crucial in legal contexts, concerned that translation might compromise these subtleties. We present an example of the generated question in Figure 8.

Table 6 presents the evaluation results for the reasoning subset, where *Llama-2-KOEN-7B* continues to outperform *Llama-2-7B*. Notably, *GPT-4-Turbo* achieves a near-perfect score on the reasoning subset. This performance could stem from various factors. Firstly, the lawsuit collection used for this subset, sourced from the internet and dating back to the 1980s, may have been part of GPT variants' pretraining data. Secondly, LLMs might struggle to generate challenging distractor options that surpass their reasoning capabilities. While supplying reference materials enables the generation of questions beyond the model's knowledge, our current pipeline might fail to guide models to create sufficiently complex distractors effectively. Future research is required to better understand these dynamics. However, despite these considerations, the benchmarks still provide a rigorous test for evaluating the capabilities of leading open LLMs.

Models	Multilingual (Kor)	Reasoning
Llama-2-KOEN-7B	38.8	58.0
Llama-2-7B	34.8	24.0
Llama-2-13B	50.4	48.0
Llama-2-70B	63.2	81.0
GPT-3.5-Turbo	58.4	92.0
GPT-4-Turbo	84.8	96.0

Table 6: Evaluation results of selected models on subsets generated in Section (6.2) and Section (6.3).

7. Conclusion

In this study, we introduce **KRX-Bench**, an automated pipeline designed for generating financial benchmarks. We validate the pipeline's effectiveness and reliability by developing **KRX-Bench-POC**, at the best of our knowledge, the first dataset aimed at evaluating LLMs' understanding of companies across diverse stock markets. Our findings confirm that the proposed pipeline can autonomously produce trustworthy benchmarks. This feature suits the fast-changing dynamics of the financial sector, enabling the generation of benchmarks that evolve in tandem with market changes. Additionally, we illustrate its broad applicability through various use cases, including creating open-ended, multilingual, and reasoning-based questions, highlighting our method's versatile utility.

8. Acknowledgements

This work was done in colaboration with OneLineAI[4] and Korea Exchange(KRX)[5].

9. Bibliographical References

Hussam Alkaissi and Samy I McFarlane. 2023. Artificial hallucinations in chatgpt: implications in scientific writing. *Cureus*, 15(2).

Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. 2023. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, et al. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023*.

Gagan Bhatia, El Moatez Billah Nagoudi, Hasan Cavusoglu, and Muhammad Abdul-Mageed. 2024. Fintral: A family of gpt-4 level multimodal financial large language models. *arXiv preprint arXiv:2402.10986*.

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.

Ethan Callanan, Amarachi Mbakwe, Antony Papadimitriou, Yulong Pei, Mathieu Sibue, Xiaodan Zhu, Zhiqiang Ma, Xiaomo Liu, and Sameena Shah. 2023. Can gpt models be financial analysts? an evaluation of chatgpt and gpt-4 on mock cfa exams. *arXiv preprint arXiv:2310.08678*.

chakki. Coarij: Corpus of annual reports in japan. https://github.com/chakki-works/CoARiJ. Accessed: 2024-02-27.

Zhiyu Chen, Wenhu Chen, Charese Smiley, Sameena Shah, Iana Borova, Dylan Langdon, Reema Moussa, Matt Beane, Ting-Hao Huang, Bryan Routledge, et al. 2021. Finqa: A dataset of numerical reasoning over financial data. *arXiv preprint arXiv:2109.00122*.

[4] https://www.onelineai.com

[5] https://global.krx.co.kr/main/main.jsp

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. *arXiv preprint arXiv:2210.03849*.

Daixuan Cheng, Shaohan Huang, and Furu Wei. 2023. Adapting large language models via reading comprehension. *arXiv preprint arXiv:2309.09530*.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. 2023. A framework for few-shot language model evaluation.

Neel Guha, Julian Nyarko, Daniel E Ho, Christopher Ré, Adam Chilton, Aditya Narayana, Alex Chohlas-Wood, Austin Peters, Brandon Waldon, Daniel N Rockmore, et al. 2023. Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models. *arXiv preprint arXiv:2308.11462*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. 2023. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *arXiv preprint arXiv:2311.05232*.

Pranab Islam, Anand Kannappan, Douwe Kiela, Rebecca Qian, Nino Scherrer, and Bertie Vidgen. 2023. Financebench: A new benchmark for financial question answering. *arXiv preprint arXiv:2311.11944*.

Hamish Ivison, Yizhong Wang, Valentina Pyatkin, Nathan Lambert, Matthew Peters, Pradeep Dasigi, Joel Jang, David Wadden, Noah A Smith, Iz Beltagy, et al. 2023. Camels in a changing climate: Enhancing lm adaptation with tulu 2. *arXiv preprint arXiv:2311.10702*.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421.

L. Junbum. 2023. llama-2-ko-7b (revision 4a9993e).

Hyunwoong Ko, Kichang Yang, Minho Ryu, Taekyoon Choi, Seungmu Yang, Sungho Park, et al. 2023. A technical report for polyglot-ko: Open-source large-scale korean language models. *arXiv preprint arXiv:2306.02254*.

Rik Koncel-Kedziorski, Michael Krumdick, Viet Lai, Varshini Reddy, Charles Lovering, and Chris Tanner. 2023. Bizbench: A quantitative reasoning benchmark for business and finance. *arXiv preprint arXiv:2311.06602*.

Moritz Laurer, Wouter van Atteveldt, Andreu Casas, and Kasper Welbers. 2023. Building efficient universal classifiers with natural language inference. *arXiv preprint arXiv:2312.17543*.

Meng Lee, Fujiki Nakamura, Makoto Shing, Paul McCann, Takuya Akiba, and Naoki Orii. Japanese stablelm base beta 7b.

Yuanzhi Li, Sébastien Bubeck, Ronen Eldan, Allie Del Giorno, Suriya Gunasekar, and Yin Tat Lee. 2023. Textbooks are all you need ii: **phi-1.5** technical report. *arXiv preprint arXiv:2309.05463*.

Jerry Loh. sp500-edgar-10k. `https://huggingface.co/datasets/jlohding/sp500-edgar-10k`. Accessed: 2024-02-27.

Lefteris Loukas, Manos Fergadiotis, Ilias Chalkidis, Eirini Spyropoulou, Prodromos Malakasiotis, Ion Androutsopoulos, and Georgios Paliouras. 2022. Finer: Financial numeric entity recognition for xbrl tagging. *arXiv preprint arXiv:2203.06482*.

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.

Pekka Malo, Ankur Sinha, Pekka Korhonen, Jyrki Wallenius, and Pyry Takala. 2014. Good debt or bad debt: Detecting semantic orientations in economic texts. *Journal of the Association for Information Science and Technology*, 65(4):782–796.

OpenAI. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Tomohiro Sawada, Daniel Paleka, Alexander Havrilla, Pranav Tadepalli, Paula Vidas, Alexander Kranias, John J Nay, Kshitij Gupta, and Aran Komatsuzaki. 2023. Arb: Advanced reasoning benchmark for large language models. *arXiv preprint arXiv:2307.13692*.

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. *arXiv preprint arXiv:2211.00083*.

Guijin Son, Hanearl Jung, Moonjeong Hahm, Keonju Na, and Sol Jin. 2023a. Beyond classification: Financial reasoning in state-of-the-art language models. *arXiv preprint arXiv:2305.01505*.

Guijin Son, Hanwool Lee, Nahyeon Kang, and Moonjeong Hahm. 2023b. Removing non-stationary knowledge from pre-trained language models for entity-level sentiment classification in finance. *arXiv preprint arXiv:2301.03136*.

Guijin Son, Hanwool Lee, Sungdong Kim, Seungone Kim, Niklas Muennighoff, Taekyoon Choi, Cheonbok Park, Kang Min Yoo, and Stella Biderman. 2024. Kmmlu: Measuring massive multitask language understanding in korean. *arXiv preprint arXiv:2402.11548*.

Guijin Son, Hanwool Lee, Suwan Kim, Jaecheol Lee, Je Won Yeom, Jihyu Jung, Jung Woo Kim, and Songseong Kim. 2023c. Hae-rae bench: Evaluation of korean knowledge in language models. *arXiv preprint arXiv:2309.02706*.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V Le, Ed H Chi, Denny Zhou, et al. 2022. Challenging big-bench tasks and whether chain-of-thought can solve them. *arXiv preprint arXiv:2210.09261*.

Qwen Team. Introducing qwen1.5. https://qwenlm.github.io/blog/qwen1.5/. Accessed: 2024-02-27.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

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

Alexander Wei, Nika Haghtalab, and Jacob Steinhardt. 2023. Jailbroken: How does llm safety training fail? *arXiv preprint arXiv:2307.02483*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

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. *arXiv preprint arXiv:2303.17564*.

Qianqian Xie, Weiguang Han, Zhengyu Chen, Ruoyu Xiang, Xiao Zhang, Yueru He, Mengxi Xiao, Dong Li, Yongfu Dai, Duanyu Feng, et al. 2024. The finben: An holistic financial benchmark for large language models. *arXiv preprint arXiv:2402.12659*.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. Hellaswag: Can a machine really finish your sentence? *arXiv preprint arXiv:1905.07830*.

Liwen Zhang, Weige Cai, Zhaowei Liu, Zhi Yang, Wei Dai, Yujie Liao, Qianru Qin, Yifei Li, Xingyu Liu, Zhiqiang Liu, et al. 2023. Fineval: A chinese financial domain knowledge evaluation benchmark for large language models. *arXiv preprint arXiv:2308.09975*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2024. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36.