ESG Classification by Implicit Rule Learning via GPT-4

Hyo Jeong Yun[1], Chanyoung Kim[2], Moonjeong Hahm[1], Kyuri Kim[3], Guijin Son[4]*
Chung-ang University[1], Konkuk University[2], Seoul Women's University[3], Yonsei University[4]
dbsgywjd@cau.ac.kr, spthsrbwls123@yonsei.ac.kr

Abstract

Environmental, social, and governance (ESG) factors are widely adopted as higher investment return indicators. Accordingly, ongoing efforts are being made to automate ESG evaluation with language models to extract signals from massive web text easily. However, recent approaches suffer from a lack of training data, as rating agencies keep their evaluation metrics confidential. This paper investigates whether state-of-the-art language models like GPT-4 can be guided to align with unknown ESG evaluation criteria through strategies such as prompting, chain-of-thought reasoning, and dynamic in-context learning. We demonstrate the efficacy of these approaches by ranking 2nd in the Shared-Task ML-ESG-3 *Impact Type* track for Korean without updating the model on the provided training data. We also explore how adjusting prompts impacts the ability of language models to address financial tasks leveraging smaller models with openly available weights. We observe longer general pre-training to correlate with enhanced performance in financial downstream tasks. Our findings showcase the potential of language models to navigate complex, subjective evaluation guidelines despite lacking explicit training examples, revealing opportunities for training-free solutions for financial downstream tasks.

Keywords: Large Language Model, Benchmark,Finance

1. Introduction

In recent years, there has been a noticeable increase in investors factoring environmental, social, and governance (ESG) considerations into their investment choices. Recent studies, through meta-analysis, have shown that improved ESG performance correlates with better corporate financial outcomes, potentially leading to higher investment returns (Cort and Esty, 2020; Friede et al., 2015). Assessing ESG performance involves nuanced analysis, and, as a result, the industry relies on rating agencies like MSCI[1], Sustainalytics[2], and Bloomberg[3] to evaluate and rank companies. Ongoing efforts to automate the ESG evaluation process exist, mainly through leveraging language models as substitutes for human analysts (Mehra et al., 2022). However, the specific methodologies used by each rating agency are not widely disclosed, leading to a lack of understanding of the detailed metrics necessary for evaluation. The closed nature of these agencies presents significant challenges when training language models to accurately replicate their evaluation criteria. This is particularly problematic for earlier language models, such as BERT (Devlin et al., 2018), which heavily rely on explicit training data on the output distribution to accurately approximate the underlying function. Without access to the specific criteria and data used by these agencies, it is difficult to teach language models to make judgments that align with past standards. Researchers have sought to enhance training datasets through synthetic data to address this issue (Glenn et al., 2023). Nonetheless, several hurdles exist. First, the lack of transparency in the evaluation methodologies used by rating agencies, which often include subjective assessments, makes it difficult for researchers to generate realistic datasets. Moreover, the creation of large-scale, high-quality labeled datasets is resource-intensive. Manually annotating extensive text collections requires considerable time and skilled professionals. Furthermore, the accurate classification of sentences poses challenges due to the subjective nature of interpretation, which can vary even among experts (Auzepy et al., 2023). Finally, the rapid evolution of ESG criteria requires regular updates on the training dataset and retraining the model to align with changing investor expectations, emerging trends, and new reporting standards.

In this paper, we investigate whether state-of-the-art language models can be guided to align with unknown values (specifically, ESG evaluation standards) without learning from explicit training data. We employ multiple strategies, such as prompting, Chain-of-Thought reasoning (Wei et al., 2022), and dynamic in-context learning (Dong et al., 2022) with *GPT-4* (OpenAI), to participate in the Shared-Task ML-ESG-3 and rank second place in the *Impact Type* track for Korean. Our findings underscore the efficacy of these strategies in approximating unknown guidelines, showcasing their potential in navigating the complexities of ESG criteria alignment. Furthermore, we extend our investigation

* Corresponding author.

[1] https://www.msci.com

[2] https://www.sustainalytics.com

[3] https://www.bloomberg.com

[…] 스카버러 유전에서 25년간 이산화탄소 3억7000만톤이 배출될 것이라고 예측한 것. 이는 호주에서 연간 발생하는 이산화탄소 배출량보다 3배나 많은 양이다. 또 재단은 호주 신임 연방 환경장관인 타냐 플리버섹에게, 이번 프로젝트가 기후 변화를 악화시켜 그레이트 배리어 리프에 악영향을 미칠지 여부를 판단할 수 있을 때까지 사업 시작을 중단시켜 줄 것을 요구했다. 호주보존재단의 켈리 오샤나시 대표는 “스카버러 가스 프로젝트는 국가 환경법의 허점을 파고들어 승인을 따낸 사업”이라며, […] , 호주에서는 노후화된 지역 석탄화력발전소에서도 정전이 발생한 바 있다. 이에 최근 선거에서 승리한 캔버라의 중도 좌파 정부는 “환경을 우선 생각해야 하지만 상업적인 부분도 무시할 수 없다”며 지속적으로 화석 연료 프로젝트를 지원할 것이라고 밝혔다.

[…] It is predicted that the Scarborough gas field will emit 370 million tons of CO2 over 25 years, three times more than Australia's annual emissions. The Australian Conservation Foundation has asked Environment Minister Tanya Plibersek to delay the project's start until its impact on climate change and the Great Barrier Reef can be assessed, criticizing the project for exploiting legal loopholes to gain approval. […] ,In Australia, power outages have occurred at aging coal-fired power plants. The centrist-left government in Canberra, after winning the recent election, stated it would prioritize the environment but cannot ignore commercial aspects, pledging continued support for fossil fuel projects.

Figure 1: An example from the ML-ESG dataset. Sentences highlighted in red indicate negative implications for ESG, while those in blue denote positive ESG implications. The gold label for the ESG type of this text is "Opportunity." English translations are added for broader accessibility.

Category	Opp.	Risk	Cannot Dist.	Total.
Sustainable Econ.	160	57	41	258
Corporate Govern.	134	31	40	205
Env. & Society	71	79	6	156
Disclosure & Eval.	87	55	11	153
ESG Life	7	3	10	20
Opinion	3	4	1	8
Total	462	229	109	800

Table 1: Statistics on the *Impact Type* of Shared-Task ML-ESG-3 for Korean.

Category	< 2 Yrs	2-5 Yrs	> 5 Yrs	Total
Sustainable Econ.	101	54	103	258
Corporate Govern.	137	36	32	205
Env. & Society	67	26	63	156
Disclosure & Eval.	119	23	11	153
ESG Life	16	1	3	20
Opinion	6	2	0	8
Total	446	212	142	800

Table 2: Statistics on the *Impact Duration* of Shared-Task ML-ESG-3 for Korean.

to include two smaller models with publicly accessible weights, examining how slight modifications in prompts influence their performance and calibration. To the best of our knowledge, this study represents the first attempt to explore how adjustments in prompts can impact the ability of language models to address financial problems.

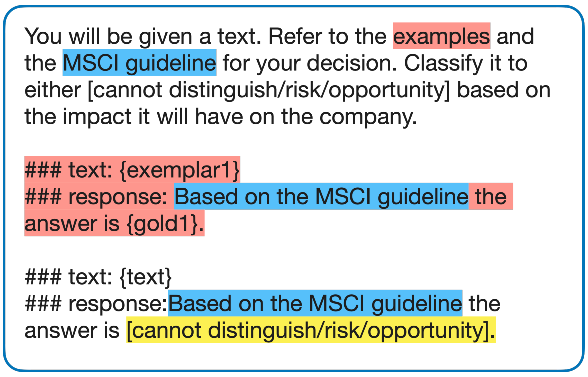
2. Shared Task ML-ESG-3

The Shared-Task ML-ESG-3 for Korean consists of two downstream tasks: *Impact Type* and *Impact Duration*. The *Impact Type* task involves classifying given ESG news articles to one of *Opportunity*, *Risk*, or *Cannot Distinguish*. The *Impact Duration* task involves classifying the impact duration of a news article as one of *Less than 2 years*, *2 to 5 years*, or *More than 5 years*. The dataset includes separate training and testing sets, with 800 Korean articles in the training set and 200 articles in the testing set.

In Table 1 we illustrate the distribution of impact types across categories in the training dataset. We observe significant data imbalance across multiple columns. For instance, while the largest category, "Sustainable Economics" feature 258 samples, the smallest category "Opinions," only include eight. Furthermore, *Opportunity* category comprises 462 entries, roughly four times the count of the *Cannot Distinguish* category, which has 109 entries. The imbalance of data could potentially be attributed to either: 1) a sampling error arising from the small dataset size, or 2) the real-world distribution of ESG-related news being skewed, as press may be more reluctant to report negative issues due to associated risks. Regardless of the underlying cause, this imbalanced training set poses a critical challenge for traditional approaches to training language models, as they will inevitably learn skewed representations from the biased data distribution. Similar patterns can be found also for the *Impact Duration* subset as shown in Table 2. The *Less than 2 years* category is the largest with 446 entries, nearly three times more than the *More than 5 years* category, which is the least represented with 142 entries.

3. Main Results

In this section, we elaborate on our methodology(Section 3.1) and report observed performances (Section 3.2).



You will be given a text. Refer to the examples and the MSCI guideline for your decision. Classify it to either [cannot distinguish/risk/opportunity] based on the impact it will have on the company.

text: {exemplar1}
response: Based on the MSCI guideline the answer is {gold1}.

text: {text}
response:Based on the MSCI guideline the answer is [cannot distinguish/risk/opportunity].

Figure 2: An example prompt with one examplar (highlighted in red) and prompts to follow the MSCI guidelines (highlghted in blue). We calculate the chance for the gold answer to follow "the answer is".

3.1. Methodology

Predicting the ESG types and their impact duration from texts is a non-trivial task that traditionally relies on human experts. However, the criteria these experts use are mostly kept confidential. This ambiguity fence researchers from developing precise rules for LLMs to learn to perform such tasks. Accordingly, this leads to a question: **Can LLMs implicitly approximate unknown rules, without a comprehensive understanding of the task?** To address this question, we employ *GPT-4*, a state-of-the-art language model. To align the model with the implicit rules we leverage the following approaches:

In-Context Learning (Dong et al., 2022): In-context learning (ICL) is an approach where LLMs are provided with exemplars demonstrating the desired behavior. Instead of updating parameters through backpropagation, the model infers patterns from the examples and generalizes during inference. In our work, we dynamically alter the provided examples using the BM-25 algorithm. For a given input sample, we retrieve five relevant articles from the training set and provide them for ICL to the model during inference.

Chain-of-Thought (Wei et al., 2022):Chain-of-thought guide models to generate a series of intermediate reasoning steps while solving a task. In an autoregressive structure, one forward pass is calculated per generated token; accordingly, allowing a model to generate intermediate reasoning allows it to leverage more forward passes as needed.

Prompt Engineering (White et al., 2023):Prompt engineering involves creating prompts or prefixed to guide LLMs during inference. A prompt engineers the LLM to follow a desired behavior and output format. In this work, we prompt the language model to follow the MSCI guidelines for classification.

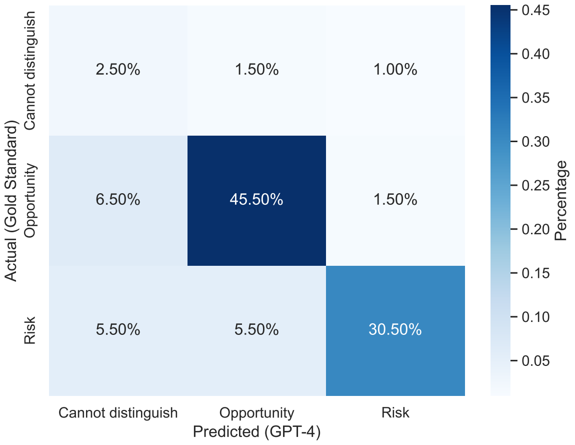


Figure 3: A confusion matrix analyzing the performance of *GPT-4* on the *Impact Type* subset.

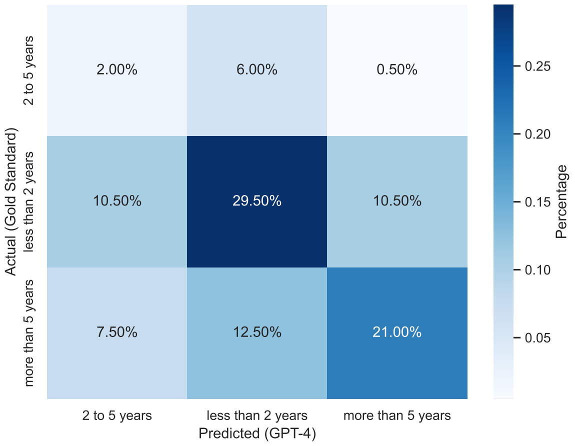


Figure 4: A confusion matrix analyzing the performance of *GPT-4* on the *Impact Duration* subset.

3.2. Evaluation Results

Table 4 showcases the performance of selected models on the Korean subset for the Shared Task ML-ESG-3. Notably, our approach, which utilizes 5-shot exemplars and prompt engineering based on MSCI guidelines, ranks second in *Impact Type* classification. However, it falls short in accurately predicting *Impact Duration*. An initial analysis of the outputs, presented in Figures 3 and 4, reveals a tendency of *GPT-4* to incorrectly classify impact durations as *less than 2 years*. Further qualitative examination shows that articles containing multiple perspectives and events often mislead the model. This observation is consistent with findings that LLMs struggle with comprehending and referencing longer text inputs (Levy et al., 2024).

Task	Model	Min	Max	Mean	△ (Max - Min)
Impact Duration	EEVE-Korean-10.8B	38.0	48.5	44.9	10.5
Impact Type	EEVE-Korean-10.8B	35.0	55.5	48.9	20.5
Impact Duration	Yi-Ko-6B	44.0	51.5	47.9	7.5
Impact Type	Yi-Ko-6B	59.0	65.5	63.2	6.5

Table 3: Performance summary of *Yi-Ko-6B* and *EEVE-Korean-10.8B* with ten different prompts. We report the accuracy (%) of each models.

Submission	Impact Type	Impact Duration
Ours	76.13	43.98
3idiots_3	**79.85**	61.54
Jetsons_1	-	**66.24**
Tredence_2	75.95	58.18

Table 4: Performance of selected models. The highest-scoring model is highlighted in **bold**, and second-highest is underlined.

An example highlighting an instance with multiple implications is provided in Figure 1. Despite the challenges, SOTA LLMs like GPT-4 demonstrate a remarkable ability to implicitly identify patterns, surpassing traditional performance methods without requiring specific training.

4. Calibration

For a model's decisions to be considered trustworthy, they must be well-calibrated; this means that its confidence levels should accurately reflect the true likelihood of its predictions being correct. In this section, we will explore how various approaches influence models' calibration and accuracy.

4.1. Experimental Settings

Models Unfortunately, the *GPT-4* API does not provide enough information for the intended analysis. Therefore, we choose to use *Yi-Ko-6B* (Lee) and *EEVE-Korean-10.8B* (Kim et al., 2024) two pre-trained models with fewer than 14 billion parameters that demonstrate the highest performance on the KMMLU (Son et al., 2024) benchmark. See Appendix A for further details on the models.

Evaluation We evaluate ten distinct approaches, varying the number of in-context exemplars, the order of these exemplars, and the prompts themselves. See Appendix A for an explanation of each approach. For each approach, we append "The answer is" to a query and calculate the likelihood of each option following the query. Figure 2 provides an example of the query format.

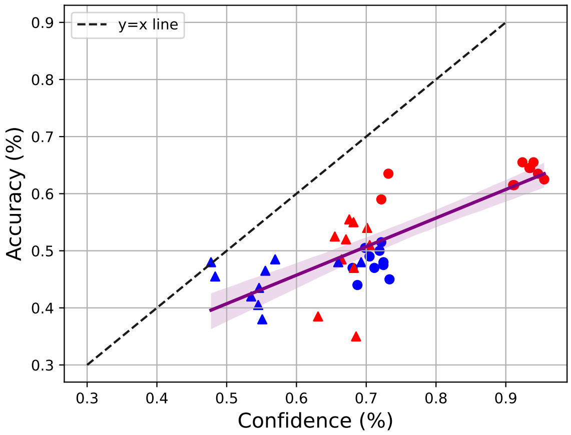


Figure 5: Relationship between accuracy and confidence of Yi-Ko-6B (circle) and EEVE-Korean-10.8B (triangle) for both subsets.(Red for 'Impact Type' and blue for 'Impact Duration'). Regression analysis exhibits a slope of 0.50.

4.2. Analysis

In Figure 5, we provide an overview of the calibration of models by testing how well the average confidence estimates the accuracy for each prompt. Surprisingly, both model appears to be well-calibrated, with a regression analysis exhibiting a slope of 0.5. In Table 3, we observe that *Yi-Ko-6B* outperforms *EEVE-Korean-10.8B* in both average and maximum scores. Additionally, *Yi-Ko-6B* exhibits a smaller delta, indicating greater robustness to prompt variations. This increased robustness may stem from extended continual pre-training, which is consistent with recent studies suggesting that the ICL capabilities of models are enhanced by encountering parallel structures in the training corpora (Chen et al., 2024b). Extended continual pre-training in Korean likely increases the model's exposure to parallel structures, thus improving its ability to capture implicit patterns robustly. Our analysis indicates that smaller, publicly available models can also effectively identify implicit patterns in ESG classification without prior training. Without needing task-specific fine-tuning, general pre-training seems to improve their robustness and overall performance.

5. Conclusion

In this work, we adopt multiple prompting, chain-of-thought reasoning, and in-context learning strategies to guide *GPT-4* in solving ESG classification tasks. We rank second in the Korean subset for Shared Task ML-ESG-3 in *Impact Type* prediction. Furthermore, we adopt open models to explain their calibration and robustness to different prompting strategies. The longer general pre-training correlates with enhanced performance in financial downstream tasks. While our work has been limited to the Korean language, we believe it will be equally applicable in different languages, especially in English, and leave for future works.

6. Bibliographical References

01-ai. 01-ai/yi-6b. https://huggingface.co/01-ai/Yi-6B. Accessed: 2024-03-08.

Alix Auzepy, Elena Tönjes, David Lenz, and Christoph Funk. 2023. Evaluating tcfd reporting: A new application of zero-shot analysis to climate-related financial disclosures. *arXiv preprint arXiv:2302.00326*.

Chang Heng Chen, Xiting Wang, Ting-En Lin, Ang Lv, Yuchuan Wu, Xin Gao, Ji-Rong Wen, Rui Yan, and Yongbin Li. 2024a. Masked thought: Simply masking partial reasoning steps can improve mathematical reasoning learning of language models.

Yanda Chen, Chen Zhao, Zhou Yu, Kathleen McKeown, and He He. 2024b. Parallel structures in pre-training data yield in-context learning. *arXiv preprint arXiv:2402.12530*.

Todd Cort and Daniel Esty. 2020. Esg standards: Looming challenges and pathways forward. *Organization & Environment*, 33(4):491–510.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, and Zhifang Sui. 2022. A survey for in-context learning. *arXiv preprint arXiv:2301.00234*.

Gunnar Friede, Timo Busch, and Alexander Bassen. 2015. Esg and financial performance: aggregated evidence from more than 2000 empirical studies. *Journal of sustainable finance & investment*, 5(4):210–233.

Parker Glenn, Alolika Gon, Nikhil Kohli, Sihan Zha, Parag Pravin Dakle, and Preethi Raghavan. 2023. Jetsons at the FinNLP-2023: Using synthetic data and transfer learning for multilingual ESG issue classification. In *Proceedings of the Fifth Workshop on Financial Technology and Natural Language Processing and the Second Multimodal AI For Financial Forecasting*, pages 133–139, Macao. -.

Dahyun Kim, Chanjun Park, Sanghoon Kim, Wonsung Lee, Wonho Song, Yunsu Kim, Hyeonwoo Kim, Yungi Kim, Hyeonju Lee, Jihoo Kim, et al. 2023. Solar 10.7 b: Scaling large language models with simple yet effective depth up-scaling. *arXiv preprint arXiv:2312.15166*.

Seungduk Kim, Seungtaek Choi, and Myeongho Jeong. 2024. Efficient and effective vocabulary expansion towards multilingual large language models. *arXiv preprint arXiv:2402.14714*.

Junbum Lee. beomi/yi-ko-6b. https://huggingface.co/beomi/Yi-Ko-6B. Accessed: 2024-03-08.

Mosh Levy, Alon Jacoby, and Yoav Goldberg. 2024. Same task, more tokens: the impact of input length on the reasoning performance of large language models. *arXiv preprint arXiv:2402.14848*.

Srishti Mehra, Robert Louka, and Yixun Zhang. 2022. Esgbert: Language model to help with classification tasks related to companies environmental, social, and governance practices. *arXiv preprint arXiv:2203.16788*.

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

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*.

Hariram Veeramani, Surendrabikram Thapa, and Usman Naseem. 2024. Enhancing esg impact type identification through early fusion and multilingual models. *arXiv preprint arXiv:2402.10772*.

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.

Jules White, Quchen Fu, Sam Hays, Michael Sandborn, Carlos Olea, Henry Gilbert, Ashraf Elnashar, Jesse Spencer-Smith, and Douglas C

Schmidt. 2023. A prompt pattern catalog to enhance prompt engineering with chatgpt. *arXiv preprint arXiv:2302.11382*.

A. Additional details for Section 4

A.1. Adopted Models

We adopt the following models with openly-available weights for analysis in Section 4. Due to hardware limitations, all models are used in 4-bit quantization.

1. *EEVE-Korean-10.8B* (Kim et al., 2024): A Korean vocabulary-extended ver sion of *SOLAR-10.7B* (Kim et al., 2023) that has undergone continual pre-training on a total of 3.2M documents (or, 3.2B tokens).

2. *Yi-Ko-6B* (Lee): A Korean vocabulary-extended version of Yi-6B (01-ai) that has undergone continual pre-training on 60B tokens.

A.2. Prompts

In Table 5, we provide an overview of the ten prompts used for analysis in Section 4.

Prompt Name	# of In-Context Exemplars	Order of Exemplars	Prompted to follow MSCI Guidelines
1-shot-standard_order-msci	1	Similar First	O
1-shot-standard_order-standard	1	Similar First	X
3-shot-reverse_order-msci	3	Similar Last	O
3-shot-reverse_order-standard	3	Similar Last	X
3-shot-standard_order-msci	3	Similar First	O
3-shot-standard_order-standard	3	Similar First	X
5-shot-reverse_order-msci	5	Similar Last	O
5-shot-reverse_order-standard	5	Similar Last	X
5-shot-standard_order-msci	5	Similar First	O
5-shot-standard_order-standard	5	Similar First	X

Table 5: Entire list of prompt settins used in Section 4.

A.3. Performance Details

In Tables 6 and 6 we present the detailed per prompt perfomrnace for each models.

Prompt	Accuracy	Confidence	Model	Task
1-shot-standard_order-msci_simple	0.635	0.731760	Yi-Ko-6B	Impact Type
1-shot-standard_order-standard	0.590	0.721608	Yi-Ko-6B	Impact Type
3-shot-reverse_order-msci_simple	0.625	0.955045	Yi-Ko-6B	Impact Type
3-shot-reverse_order-standard	0.635	0.946185	Yi-Ko-6B	Impact Type
3-shot-standard_order-msci_simple	0.645	0.933864	Yi-Ko-6B	Impact Type
3-shot-standard_order-standard	0.655	0.923851	Yi-Ko-6B	Impact Type
5-shot-reverse_order-msci_simple	0.645	0.934855	Yi-Ko-6B	Impact Type
5-shot-reverse_order-standard	0.655	0.939728	Yi-Ko-6B	Impact Type
5-shot-standard_order-msci_simple	0.615	0.910514	Yi-Ko-6B	Impact Type
5-shot-standard_order-standard	0.615	0.912037	Yi-Ko-6B	Impact Type
1-shot-standard_order-msci	0.505	0.698373	Yi-Ko-6B	Impact Duration
1-shot-standard_order-standard	0.500	0.719090	Yi-Ko-6B	Impact Duration
3-shot-reverse_order-msci	0.470	0.680418	Yi-Ko-6B	Impact Duration
3-shot-reverse_order-standard	0.490	0.704762	Yi-Ko-6B	Impact Duration
3-shot-standard_order-msci	0.475	0.724632	Yi-Ko-6B	Impact Duration
3-shot-standard_order-standard	0.515	0.721509	Yi-Ko-6B	Impact Duration
5-shot-reverse_order-msci	0.440	0.687383	Yi-Ko-6B	Impact Duration
5-shot-reverse_order-standard	0.470	0.711635	Yi-Ko-6B	Impact Duration
5-shot-standard_order-msci	0.450	0.733333	Yi-Ko-6B	Impact Duration
5-shot-standard_order-standard	0.480	0.724686	Yi-Ko-6B	Impact Duration

Table 6: Detailed performance of *Yi-Ko-6B* on different prompts.

Prompt	Accuracy	Confidence	Model	Task
1-shot-standard_order-msci_simple	0.35	0.685465	EEVE-Korean-10.8B	Impact Type
1-shot-standard_order-standard	0.385	0.630959	EEVE-Korean-10.8B	Impact Type
3-shot-reverse_order-msci_simple	0.525	0.654941	EEVE-Korean-10.8B	Impact Type
3-shot-reverse_order-standard	0.54	0.701319	EEVE-Korean-10.8B	Impact Type
3-shot-standard_order-msci_simple	0.485	0.664646	EEVE-Korean-10.8B	Impact Type
3-shot-standard_order-standard	0.55	0.681784	EEVE-Korean-10.8B	Impact Type
5-shot-reverse_order-msci_simple	0.51	0.704919	EEVE-Korean-10.8B	Impact Type
5-shot-reverse_order-standard	0.555	0.675689	EEVE-Korean-10.8B	Impact Type
5-shot-standard_order-msci_simple	0.47	0.682284	EEVE-Korean-10.8B	Impact Type
5-shot-standard_order-standard	0.52	0.670969	EEVE-Korean-10.8B	Impact Type
1-shot-standard_order-msci	0.48	0.659873	EEVE-Korean-10.8B	Impact Duration
1-shot-standard_order-standard	0.48	0.692712	EEVE-Korean-10.8B	Impact Duration
3-shot-reverse_order-msci	0.435	0.546392	EEVE-Korean-10.8B	Impact Duration
3-shot-reverse_order-standard	0.465	0.555405	EEVE-Korean-10.8B	Impact Duration
3-shot-standard_order-msci	0.42	0.535136	EEVE-Korean-10.8B	Impact Duration
3-shot-standard_order-standard	0.485	0.569464	EEVE-Korean-10.8B	Impact Duration
5-shot-reverse_order-msci	0.405	0.545175	EEVE-Korean-10.8B	Impact Duration
5-shot-reverse_order-standard	0.48	0.477536	EEVE-Korean-10.8B	Impact Duration
5-shot-standard_order-msci	0.38	0.55096	EEVE-Korean-10.8B	Impact Duration
5-shot-standard_order-standard	0.455	0.483521	EEVE-Korean-10.8B	Impact Duration

Table 7: Detailed performance of *EEVE-Korean-10.8B* on different prompts.