Benchmarking Foundation Models on Exceptional Cases: Dataset Creation and Validation

Anonymous ACL submission

Abstract

Foundation models (FMs) have achieved significant success across various tasks, leading to research on benchmarks for commonsense and reasoning abilities. However, there is a lack of studies on FMs performance in exceptional scenarios. This paper addresses these cases for the first time, developing a novel dataset for comprehensive FMs evaluation across multiple modalities, including graphic novels, calligraphy, news articles, and lyrics. It includes tasks for instance classification, character recognition, token prediction, and text generation. The paper also proposes prompt engineering techniques like Chain-of-Thought (CoT) and CoT+Few-Shot to enhance performance. Validation of FMs using various methods revealed improvements. The code repository is accessible at: https://anonymous.4open.science/r/Exceptional-Dataset-for-FMs/README.md

1 Introduction

In the real world, while everyday life may appear to consist solely of ordinary events, unusual situations occur more frequently than one might anticipate. Humans, even when confronted with such unexpected events within their repetitive routines, may initially be surprised but can subsequently analyze the causes and consequences of these events, thereby understanding the phenomenon. Recent studies have focused on assessing the commonsense reasoning capabilities of foundation models (FMs). These efforts aim to elevate the performance of FMs to human-like levels by employing human cognitive strategies. As a result, current FMs have achieved remarkable progress, demonstrating high performance across various tasks.(Sap et al., 2019), (Speer et al., 2017), (Jin et al., 2024). However, there are situations where FMs struggle to determine causal reasoning, especially when encountering non-ordinary circumstances. Despite the development of various datasets(Yue et al., 2023),(Zellers et al., 2019), there is a need for more diverse datasets that include less common scenarios. **Exceptional Cases** We define exceptional cases as situations that contravene commonsense knowledge.(Sap et al., 2019), (Speer et al., 2017) These scenarios are in-frequently encountered in typical contexts, are difficult to anticipate, and present significant challenges for reasoning. By creating datasets that encompass more exceptional cases rather than everyday situations, we can better evaluate the robustness of FMs. In this paper, we introduce the exceptional dataset for foundation models, to address various exceptional cases, we constructed a dataset divided into four different categories: graphic novels, calligraphy, news articles that report unusual incidents, and lyrics as shown in Figure 1. The main achievements outlined in this paper includes: (1) This paper initially addresses exceptional cases and develops a benchmark dataset for their evaluation using an automated pipeline. (2) We propose several prompt engineering techniques to enhance performance in exceptional cases, including Chain-of-Thought (CoT) and CoT+Few-Shot.(Wei et al., 2022) (3) In this paper, we validated FMs on the proposed exceptional cases dataset using various methods, such as Zero-Shot, CoT, and CoT+Few-Shot.

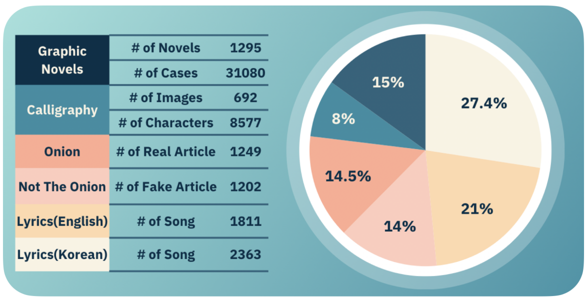

Figure 1: Distribution of Exceptional Cases Dataset and summary of four distinct datasets and their subsections, showcasing diverse characteristics.

2 Related Works

2.1 Foundation Model; Multi-Modal

With remarkable abilities of FMs(Devlin et al., 2018), (Brown et al., 2020), (Achiam et al., 2023), (Touvron et al., 2023a), (Touvron et al., 2023b), (Chowdhery et al., 2023), recent studies start to explore multi-modal data(Alayrac et al., 2022), (Li et al., 2022), (Li et al., 2023), (Team et al., 2023), (Liu et al., 2024), (OpenAI, 2023). Flamingo(Alayrac et al., 2022) was one of the earliest efforts to incorporate multiple modalities, using gated cross-attention to combine visual features. BLIP-2 (Li et al., 2023) employs a Q-Former between the image encoder and the FMs to bridge the modality gap. In order to enhance the understanding of human instructions, LLaVA (Liu et al., 2024) introduced visual instruction tuning. Recently, advanced FMs such as Gemini Pro (Team et al., 2023) and GPT-4V (OpenAI, 2023) have demonstrated outstanding performance on various multimodal tasks, including Visual Question Answering (VQA)(Mathew et al., 2021) and Optical Character Recognition (OCR). However, these models primarily focus on scenarios involving commonsense rather than exceptional cases. In this paper, we explore the understanding capabilities of FMs in exceptional situations.

2.2 Uni-Modal Commonsense Benchmark

There are many evaluations on the reasoning ability of FMs based on their knowledge on the wild world. Active research on benchmarks for text-based commonsense reasoning ability is a representative examples. Existing commonsense reasoning benchmarks are (1)composed by rule based methods as well as human expert-based methods and (2)have multi-choice problem structures in many cases (Talmor et al., 2018),(Sakaguchi et al., 2021). These benchmarks have evolved to include multiple-step problem and multi-task problems (Khot et al., 2020), (Lourie et al., 2021). Moreover, ongoing research efforts are actively focused on developing a variety of task types and uniformly structured datasets tailored to each task, with the ultimate goal of extending the overall dataset with high-quality data to enhance the capabilities of FMs.(Lin et al., 2024),(Dubois et al., 2024),(Li et al.). Our approach diverges by prioritizing critical thinking over the straightforward application of common or plausible external knowledge. Instead of evaluating understanding in routine human interactions, our tasks focus on comprehending exceptional cases and assessing the robustness of FMs in mimicking human reasoning.

2.3 Multi-Modal Commonsense Benchmark

Recently multi-modal investigations are drawn lots of attention. This leads arising benchmarks for various tasks like VQA(Mathew et al., 2021), scene understanding(Armitage et al., 2020). In addition to scene understanding, it necessitates the ability to identify connections between scenes and integrate sequences cohesively.(Wang et al., 2024). The integration of advanced Optical Character Recognition (OCR) capabilities with foundation models (FMs) can substantially enhance performance across a variety of tasks. Despite this potential, research on the capabilities of FMs remains limited. The few existing studies predominantly address English and Chinese text(Liu et al., 2023),(Shi et al., 2023), focusing on common elements such as handwriting and scene text. These studies suggest that FMs demonstrate inadequate recognition abilities for texts that are neither in English nor typographically standard. Within the entire OCR dataset, Korean OCR datasets are notably sparse. Additionally, datasets featuring art forms such as calligraphy are mainly available in Chinese, English, and Arabic(Liang et al., 2020),(Xie et al., 2022),(Alyafeai et al., 2022). To address these gaps, we conducted experiments using a Korean non-commonsense dataset and developed a multi-modal non-commonsense benchmark.

3 Methodology

3.1 Overview

It has been observed that in certain atypical scenarios(Chen et al., 2024), FMs exhibit errors in causal reasoning concerning the sequence of events. We designed experiments using four different datasets that feature various characters with multi types of tasks such as instance recognition, text generation, and character recognition, as described in Figure 2. In the experiments for all four datasets, we conducted all experimental tasks using GPT-4o, the most advanced model, which outperforms all previous GPT models.(Achiam et al., 2023) Also, we employed three prompt styles—Zero-Shot, Chain of Thought (CoT), and CoT+Few-Shot—to investigate how the accuracy of responses varies. The API temperature setting is regulated to 0 to ensure consistent results.

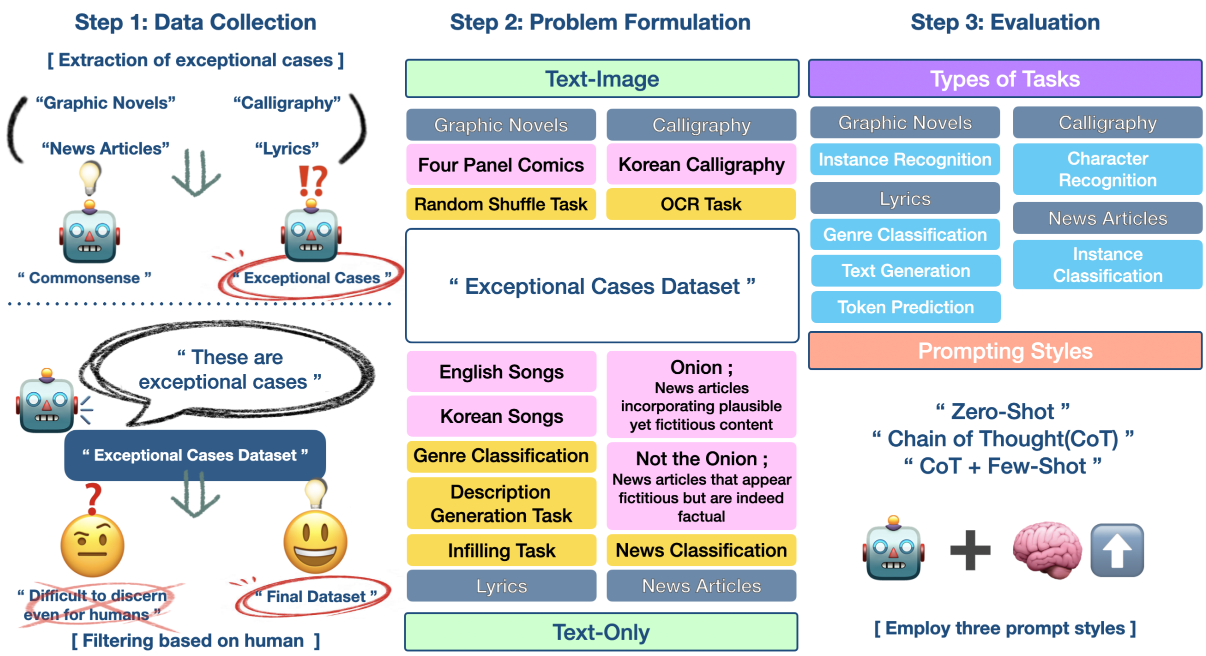

Figure 2: Three steps to construct a benchmark. Step 1: Collect data by extracting exceptional cases. After extraction, review all data to identify any ambiguous cases for reasoning. The dataset includes various modalities, such as text-only and text-image. Step 2: Define the characteristics of the dataset and the tasks that align with these characteristics. Step 3: The evaluation process encompasses a variety of tasks, including instance classification, recognition, token prediction, and text generation. To assess the diversity and accuracy of FMs' responses, Zero-Shot, CoT, and CoT+Few-Shot prompts were conducted for each task.

3.2 Graphic novels

3.2.1 Dataset Description

To evaluate the multi-modal causal reasoning capabilities of FMs, we constructed a dataset using graphic novels;'Old Master Q Comics' (Wong, 1973–1989). Written between 1973 and 1989, it used in this study consist of four panels, each forming a complete narrative with a beginning, development, climax, and conclusion. These narratives range from those easily understood through commonsense knowledge to those including exceptional cases. We hypothesize that if FMs can comprehend the story lines through causal reasoning, it is likely to be able to infer the correct sequence of panels when presented with a randomly shuffled input. Based on this hypothesis, we designed the random shuffle task as shown in Figure 3. The dataset comprises a total of 1,295 comics. When the four panels are randomly shuffled, it generates 31,080 possible number of cases. More details about the dataset can be found in App.A.1, App.A.2.

3.2.2 Experimental Design

The four input images are automatically shuffled in the code before being provided as a prompt to GPT-4o. We then measure the accuracy of all three prompt styles and analyze the results. More details about the experiment are provided in App.A.3.

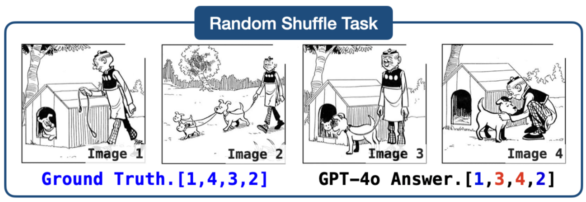

Figure 3: Example of the random shuffle task. The original sequence is [1, 4, 3, 2], but GPT-4o produce an incorrect result.

3.2.3 Evaluation Metrics

To ensure efficient and effective experimental evaluation, we developed a code that automatically prompts the input question and shuffled images, and calculates accuracy by comparing GPT-4o's responses with the ground truth order as shown in Figure 12 in App.A.3.

3.3 Calligraphy

3.3.1 Dataset Description

Given that calligraphy frequently utilizes distinctive writing styles not typically encountered in daily usage, this dataset provides an opportunity to assess

OCR performance in unusual scenarios as shown in Figure 4. Therefore, we selected 'Korean calligraphy' as our second dataset. We describe more detail in App.B.1. We gathered 692 calligraphy images through web crawling on Pinterest and labeled them using an OCR API. Typos from the OCR API were manually corrected. The preprocessing procedure is detailed in the App.B.2.

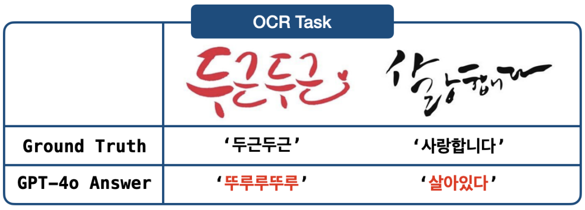

Figure 4: Examples of OCR tasks show that GPT-4o's performance results in various errors, including incorrect meanings.

3.3.2 Experimental Design

We experimented with GPT-4o to transcribe a Korean calligraphy piece using three different prompts, which are provided in the App.E.2. We used GPT-4o to transcribe a Korean calligraphy piece with three prompts in App.E.2. Before word-level evaluation, we removed punctuation and special symbols from GPT-4o's predictions and replaced '\n' with ' ' due to ambiguous line breaks in calligraphy.

3.3.3 Evaluation Metrics

We used Word-level Accuracy, CER, and WER, which are representative OCR metrics, to compare the performance of different OCR models. Additionally, we have included other metrics, such as sentence similarity (since FMs consider context in OCR tasks) and character-level evaluations, in the App.B.3.

3.4 Onion, Not The Onion

3.4.1 Dataset Description

In The Onion, Not The Onion dataset, we aim to evaluate whether FMs can judge if news that contain unpredictable event actually is occurred or fake news. This dataset contains satirical and exaggerated expressions, and deals with exceptional cases where it is difficult to distinguish between real and fake news with a naive application of common knowledge. Through this dataset, we ultimately aim to evaluate whether FMs can effectively engage in critical thinking in exceptional situations.

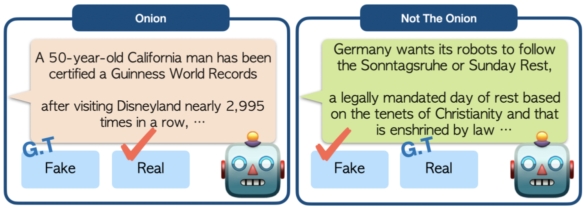

Figure 5: Examples of the fake and real news classification task include fake news that is plausible but exaggerated, and real news that features authentic yet seemingly unbelievable stories.

We collect datasets from The Onion, a website specializing in fake news contents, and Reddit's Not The Onion section, which features authentic but unbelievable news stories. It consists of 1,249 fake and 1,202 genuine news from each web source, covering news written from 2021 to May 2024.

3.4.2 Experimental Design

For our problem settings, we designed a '0' and '1' classification task, where each corresponds to fake and real news as shown in Figure 5. With the same setting as before, we concatenated for each category, followed by random shuffle and provided it to GPT-4o. Further details of prompt that we use are in Table 17 in App.E.3.

3.4.3 Evaluation Metrics

We assess the performance of FMs in classifying the authenticity of news articles using precision, recall, and F1-score. These evaluation metrics were applied to three prompting strategies to evaluate how different prompting methods affect the model's ability to identify fabricated news.

3.5 Lyrics

3.5.1 Dataset Description

Although lyrics often contain poetic licenses and uncommon expressions such as metaphors, song lyrics still allow for meaningful inference as one of the main literary genres. To evaluate the robustness of reasoning capabilities in FMs when dealing with exceptional data like lyrics, we constructed a dataset using song lyrics. We assess FMs' comprehension of song lyrics through three tasks: genre detection, song description generation, and infilling as shown in Figure 6. For more details on why we chose lyrics, refer to the App. D.1. For the English dataset, the Seen data was first collected from the Billboard Year-End Chart for 1990-2023, while the Unseen data was collected by combining data from the first week of January 2024(Considering the cut-off date for GPT-4o.) to the last week of April 2024.

The Korean dataset followed the same steps as the English dataset but was sourced from Melon. After preprocessing, the final dataset includes 1,811 English entries, 2,363 Korean entries. For more details about preprocessing steps, refer to App. D.2. Three tasks were performed separately. The first task was genre classification, which aimed to evaluate if GPT-4o can determine the genre based solely on lyrics without considering melody or musical elements. The second task was song description generation, designed to assess GPT-4o ability to understand and describe the context of lyrics. The third task was Infilling, which tested its capability to predict the masked word in a sentence. Further details of experiment that we used are in App. D.4.

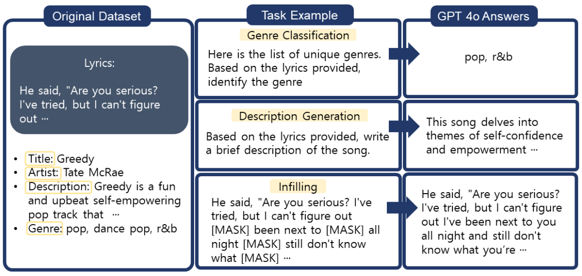

Figure 6: Overview of the lyrics dataset: an example of three different tasks and GPT-4o's responses.

3.5.2 Evaluation Metrics

The evaluation of genre detection was based on exact match scores, assigning a score of 1 if any genre exactly matched the ground truth genre, as well as overlap ratios. For the song description task, BERT F1 and ROUGE F1 scores were used for evaluation. For the third task, the evaluation used ROUGE recall scores and BERT F1 scores for the English dataset. For the Korean dataset, KoROUGE recall scores and cosine similarity between BERT embedding of text were used, as BERT score could not be performed. We provide additional details about the metrics in App. D.3.

4 Experiments Result

4.1 Graphic Novels

4.1.1 Quantitative Results

We evaluated GPT-4o's multi-modal causal reasoning capabilities using the random shuffle task. The performance was highest in the 'CoT+Few-Shot' condition, followed by the 'CoT' and 'Zero-Shot' conditions. Notably, the performance in the 'Zero-Shot' condition exceeded expectations, demonstrating an accuracy that was not significantly lower than that of the other prompting styles as shown in Table 1. Additional details about the experiment are in App.A.4.1.

	Zero-Shot	**CoT**	**CoT+Few-Shot**
Acc.	63.80	63.88	64.63

Table 1: Result(%) of the random shuffle task, We report accuracy by matching GPT-4o's responses to the ground truth order.

4.1.2 Qualitative Results

We evaluated whether GPT-4o comprehends the content of each image and performs causal inference during the random shuffle task by instructing it to generate descriptions for each image and assessing the accuracy of these descriptions in explaining the images. In this approach, we investigated how GPT-4o interprets the context within images and infers sequences based on this understanding in two scenarios: when the inferred order is completely correct, and when the inferred order is entirely incorrect.

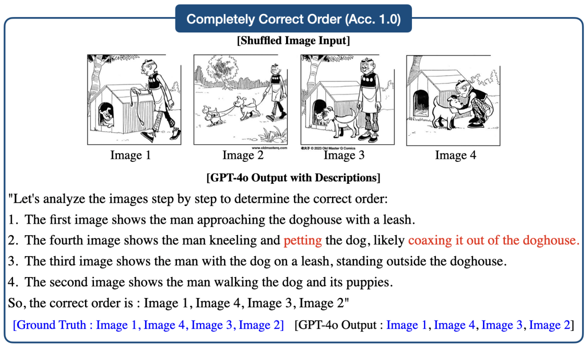

Figure 7: Correct Order Check: An example of the descriptions generated for each image shows that while GPT-4o can correctly order the images, it lacks the capability to fully understand the context. In some cases, it provides incorrect descriptions, such as using mismatched verbs (highlighted in red).

when the inferred order is completely correct: GPT-4o derives correct answers through in-context inference rather than fully understanding image contexts, because of inaccurate scene descriptions. For example, in Figure 7, it describes a man as 'kneeling and petting the dog, coaxing it out of the doghouse,' whereas the actual scene is 'squatting in front of the doghouse, putting a leash on the dog.' This indicates GPT-4o's poor instance recognition ability.

when the inferred order is completely incorrect: GPT-4o accurately describes only one of four images. It misidentifies objects or misunderstands emotions in the other three. For instance, it describes a man pulling a tiger's tail instead of removing an arrow from its paw, as shown in image 2 of Figure 8. Accurate instance recognition in a random shuffle task requires precise causal reasoning to sequence four scenes correctly. Thus, the graphic novels dataset serves as a robust benchmark for evaluating GPT-4o's multi-modal causal reasoning capabilities.

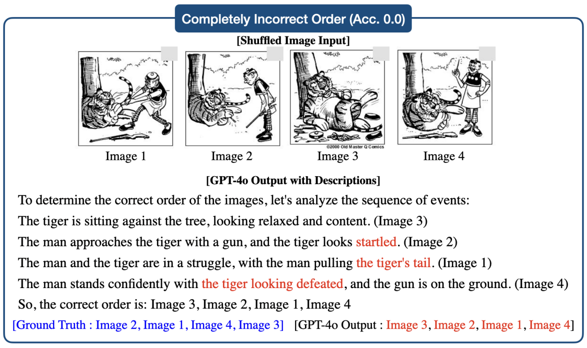

Figure 8: Incorrect Order Check: An example of descriptions presented by GPT-4o when it achieved an accuracy of 0.0. In three of four images, it provided incorrect character descriptions and showed poor object recognition (highlighted in red).

4.2 Calligraphy

4.2.1 Quantitative Results

To evaluate the multi-modal capability of GPT-4o on visual tasks related to text in exceptional cases, we devise an OCR task using a Calligraphy dataset. As shown in the Table 2, the overall results on our dataset indicate that GPT-4o performed inadequately on the OCR task. Depending on the prompt style, the CoT and CoT+Few-Shot approaches performed better compared to Zero-Shot, although the difference between the two CoT style prompts was negligible. To investigate the underlying reasons for these results, we conducted a qualitative evaluation.

	Zero-Shot	CoT	CoT+Few-Shot
Acc.(↑)	53.43	61.54	61.86
WER(↓)	64.41	45.81	45.39
CER(↓)	32.64	24.73	22.55

Table 2: Result(%) of Korean Calligraphy OCR task

4.2.2 Qualitative Results

Due to the unique characteristics of calligraphy as an art form, the dataset sometimes does not use exact sentences or words, such as expressing 'spring day' as 'spring d.' In these instances, GPT-4o tends to recognize 'd' not as a part of the word, but rather interprets only 'spring.' This tendency was more frequently observed in the CoT and CoT+Few-Shot prompts than Zero-Shot. This phenomenon arises due to the distinct characteristics of each prompt approach. In the Zero-Shot scenario, the OCR task tends to focus more on the appearance of individual words rather than the overall meaning conveyed by the calligraphy, leading to frequent typographical errors. In contrast, the CoT and CoT+Few-Shot approaches first grasp the overall meaning and then proceed with OCR based on the appropriate words. As a result, even when the output differs from the ground truth, it tends to produce semantically similar words or words that are more contextually appropriate than the ground truth. For instance, in the Figure 9, the first example of calligraphy means 'pray,' and the ground truth is '기도'. In Zero-Shot, GPT-4o recognizes it as '기드,' which is very similar in form but lacks meaning. The CoT approach recognized it as '기다,' which does not match the ground truth but has a meaning ('to crawl'). The CoT+Few-Shot approach correctly recognized it as '기도,' which matches the ground truth accurately. To accurately perform OCR on a Korean calligraphy dataset, FMs need precise skills in text detection, localization, and recognition, even for characters expressed artistically, which are considered exceptional cases. These aspects make the Korean calligraphy dataset an excellent benchmark for assessing FMs' recognition capabilities.

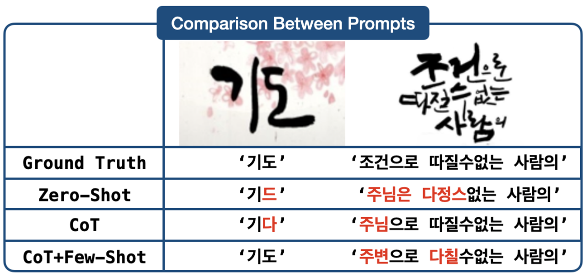

Figure 9: Examples of comparisons of OCR task results between prompts on Korean calligraphy data..

4.3 Onion, Not The Onion

4.3.1 Quantitative Results

Overall, the model demonstrates a commendable performance in fake news detection as shown in Table 3. However, its Zero-Shot capabilities are relatively low when compared to other prompting strategies CoT and CoT+Few-Shot. Both the Onion and Not The Onion categories consistently achieved high overall scores. In the qualitative analysis section, we investigated whether these predictions stemmed from inherent reasoning abilities or external influences.

		Zero-shot	CoT	CoT+Few-shot
	Acc.	80.70	89.88	94.74
Onion	Precision	78.70	86.14	92.49
	Recall	85.19	95.52	97.60
	F1-score	81.81	90.58	94.97
Not The Onion	Precision	83.17	94.75	97.35
	Recall	76.04	84.03	91.76
	F1-score	79.44	89.07	94.48

Table 3: Result(%) of news classification task. The overall scores were consistently high for both Onion and Not The Onion. In the qualitative results section, we examined whether these predictions were driven by genuine reasoning abilities or influenced by external factors.

Length of article		Q1	Q2	Q3	Q4	Q5
Not The Onion	Acc.(%)	84.23	90.42	91.25	96.25	96.68
Onion	Acc.(%)	91.20	96.80	100.00	100.00	100.00

Table 4: Longer articles improve the accuracy of the Not The Onion group. In contrast, The Onion group, consisting of shorter fake news articles, achieves higher accuracy, suggesting GPT-4o tends to classify shorter articles as fake news.

4.3.2 Qualitative Results

We observed that GPT-4o exhibits lower performance with relatively short articles. As illustrated in the Table 4, differences in accuracy based on article length indicate that as article length increases, accuracy tends to rise. Conversely, the Onion group, primarily composed of fake news articles, tends to feature shorter articles, maintaining consistently high accuracy across the dataset. This trend suggests that GPT-4o may be inclined to classify shorter articles as fake news, indicating that Not The Onion poses more challenges to GPT-4o for fake news classification. Additionally, We examine in more detail the rationale behind GPT-4o's decision-making and whether GPT-4o follows appropriate steps when it encounters relatively short articles, aiming to understand when and why GPT-4o reaches incorrect conclusions. In this approach, we observe that GPT-4o generally takes appropriate steps, many of which are plausible. However, it is notable that GPT-4o encounters difficulties with exceptional cases, as highlighted in Figure 10(marked in red). The article depicted in this figure includes several extraordinary claims, such as "Adidas urgently recalled the German national team jersey featuring the number 44 due to its resemblance to symbols used by the German SS division". To verify these claims, GPT-4o undergoes a validation process spanning from the second to the fourth step. Despite employing a search function in the fourth step, it fails to accurately determine the veracity of the article. Overall, to identify fake news, GPT-4o needs accurate causal reasoning to classify instances within an article. This makes the Onion, Not the Onion dataset a splendid benchmark for verifying their reasoning capabilities.

Reasoning Steps on Real News from Not The Onion (Article Length: Q1)

Step 1. **Source Evaluation**: The article mentions Adidas and the German Football Association (DFB), but it doesn't provide direct citations from official statements or reliable news outlets. The mention of Bild, a known tabloid, is notable but requires further verification from more credible sources.

Step 2. **Language Analysis**: The language used in the article is straightforward and factual. However, the claim about banning the number 44 and the specific names allowed and disallowed seems highly unusual and sensational.

Step 3. **Consistency and Accuracy**: The claim that Adidas would ban a number due to its resemblance to Nazi symbols is extraordinary and would likely be widely reported by major news outlets. The inconsistency in allowing certain names while banning others also seems improbable and lacks internal logic.

Step 4. **Cross-Verification**: A quick search reveals no major news outlets confirming this specific claim about Adidas and the number 44. Major changes to national team jerseys and policies would typically be well-documented and reported by multiple credible sources.

Step 5. **Content Credibility**: The mix of factual elements (such as the existence of the number 44 and customization options) with sensational and unverified claims (banning specific names) casts doubt on the article's overall credibility. The idea that Adidas would specifically allow certain controversial names while banning others is inconsistent with corporate policies aimed at avoiding controversy.

Based on these steps, the claim lacks credible verification and logical consistency, suggesting the article is likely fake. Inconsistent logic and lack of credible verification. **(0)** — GT (1)

Figure 10: An example where GPT-4o, despite following appropriate reasoning steps, produces an incorrect reasoning outcome due to an exceptional case(highlighted in red).

Comparing Song Description Generation Output

[Song Description: Ground Truth]
“Lose Control” is the second track on I've Tried Everything But Therapy by Teddy Swims. Swims uses this song to take everything off his shoulders- losing himself and all control when everything starts caving in; the only way to escape is being with that specific person.

➡️ **Song about Enduring difficult times with loved ones**

[Song Description: GPT-4o Output]
"Lose Control" by Nessa Barrett is about the intense emotional turmoil and dependency one feels in a relationship. The lyrics convey a sense of losing oneself and falling apart when the partner is not around, highlighting the addictive and problematic nature of their connection. The song captures the struggle of wanting to be independent but being irresistibly drawn to the other person.

➡️ **Song about Dealing with a problematic relationship and addictive emotions**

Figure 11: In the description generation task, It is evident that GPT-4o does not accurately comprehend song lyrics. Unlike the ground truth, which refers to 'enduring difficult times with loved ones,' GPT-4o generated content describing 'dealing with a problematic relationship and addictive emotions.'

4.4 Lyrics

4.4.1 Quantitative Results

In all three tasks, 'CoT+Few-Shot' achieved the highest performance, followed by 'CoT' and 'Zero-Shot'. In genre classification, the Korean dataset exceed the English dataset, as explained in Table 5. For the infilling task, we suggest adjusted result for accurate assessment. Initially, the overall score was high because GPT-4o was instructed to return the entire lyrics, not just 'masked part' as shown Table 6. So we get the adjusted result which only focuses on 'masked part' refer to Table 7. In the English description generation task, the overall scores were low, as shown in Table 8. Further details are discussed in the qualitative results section.

			Zero-Shot	CoT	CoT+Few-Shot
English	Seen	Overlap Ratio	0.594	0.610	0.620
		Exact Match	0.758	0.774	0.781
	Unseen	Overlap Ratio	0.474	0.497	0.509
		Exact Match	0.671	0.671	0.677
Korean	Seen	Overlap Ratio	0.642	0.665	0.733
		Exact Match	0.676	0.698	0.752
	Unseen	Overlap Ratio	0.668	0.690	0.750
		Exact Match	0.710	0.733	0.776

Table 5: Results of the genre classification task, the dataset was divided based on GPT-4o's training cutoff date. In the case of English data, it is noticeable that the performance of GPT-4o deteriorates after the cutoff. For the Korean data, the significantly lower quantity of unseen data compared to seen data appears to have a minimal impact on performance variation.

		masked	Zero-Shot	CoT	CoT+Few-Shot
English	ROUGE-1 (R)	0.853	0.928	0.934	0.939
	ROUGE-L (R)	0.853	0.923	0.930	0.935
	BERT Score (F1)	0.856	0.944	0.947	0.950
Korean	ROUGE-1 (R)	0.725	0.836	0.838	0.839
	ROUGE-L (R)	0.725	0.834	0.836	0.837
	BERT Score (F1)	0.877	0.926	0.932	0.934

Table 6: Result of infilling task using unseen dataset. Due to safety concerns, GPT-4o did not respond to seen data. Utilizing ROUGE recall for exact word matches and BERT scores for assessing semantic similarity. The 'masked' column represents the similarity between the masked input and the ground truth. For similarity on masked words, see the adjusted results in Table 7

4.4.2 Qualitative Results

In the genre classification task, the difference in the number of unique genres between the English and Korean datasets impacted the results: 11 in Korean, 58 in English, making more challenging for English, also the small number of unseen data compared to seen data resulted in only slight differences in the results. In the description generation task, the overall scores are poor, indicating that GPT-4o struggles to understand the meaning of song lyrics. As demonstrated in Figure 11, the song discusses 'enduring difficult times with loved ones,' whereas GPT-4o describes it as 'dealing with a problematic relationship and addictive emotions. In the infilling task, unmasked lyrics impacted the overall score as the score is calculated for the entire text. By applying the specified formula $\frac{\text{Prompt score}-\text{masked score}}{1-\text{masked score}}$, this influence can be mitigated, resulting in more accurate results. As demonstrated in the Table 7, the adjusted scores remain relatively low. Given the numerous repeated sections in song lyrics that GPT-4o could potentially exploit, it can be inferred that GPT-4o's performance on lyrics is suboptimal.

Adjusted Result		Zero-Shot	CoT	CoT+Few-Shot
English	ROUGE-1 (R)	0.510	0.551	0.585
	ROUGE-L (R)	0.476	0.524	0.558
	BERT Score (F1)	0.611	0.632	0.653
Korean	ROUGE-1 (R)	0.404	0.411	0.415
	ROUGE-L (R)	0.396	0.404	0.407
	BERT Score (F1)	0.398	0.447	0.463

Table 7: The adjusted results for the lyrics infilling task utilize the 'masked result'. The initially high scores required adjustment because the evaluation included the 'unmasked' parts, which inflated the overall score. After evaluating only the masked input, the overall scores decreased, indicating that GPT-4o struggles with token prediction in the lyrics dataset.

		Zero-Shot	CoT	CoT+Few-Shot
Seen	ROUGE-1 (P)	0.384	0.351	0.356
	ROUGE-1 (R)	0.073	0.142	0.148
	ROUGE-1 (F1)	0.151	0.247	0.251
	ROUGE-L (P)	0.274	0.232	0.227
	ROUGE-L (R)	0.073	0.142	0.148
	ROUGE-L (F1)	0.106	0.158	0.161
	BERT Score (P)	-0.091	-0.008	0.004
	BERT Score (R)	0.214	0.169	0.164
	BERT Score (F1)	0.057	0.080	0.084
Unseen	ROUGE-1 (P)	0.383	0.335	0.328
	ROUGE-1 (R)	0.117	0.240	0.259
	ROUGE-1 (F1)	0.163	0.252	0.262
	ROUGE-L (P)	0.270	0.212	0.202
	ROUGE-L (R)	0.082	0.160	0.166
	ROUGE-L (F1)	0.113	0.162	0.163
	BERT Score (P)	-0.034	0.050	0.062
	BERT Score (R)	0.241	0.181	0.174
	BERT Score (F1)	0.098	0.115	0.118

Table 8: Description generation task for English songs. The low overall score shows GPT-4o wrestle with understanding the meaning of lyrics.

5 Conclusion

We present an exceptional dataset, establishing a novel benchmark for the assessment of foundation models (FMs) across a wide range of scenarios, including those based on commonsense. The entire benchmark dataset is developed using an automated pipeline. Our dataset encompasses four distinct categories, each with its own characteristic dataset. These categories incorporate diverse tasks such as instance recognition, token prediction, instance classification, and text generation, designed to evaluate the capabilities of FMs across various modalities, including text-only, image-only, and image-text. We introduce several prompt engineering techniques, utilizing chain of thought (CoT) and CoT+Few-Shot, to enhance performance in exceptional cases. In this study, we validated FMs on the proposed exceptional cases dataset using various approaches, including Zero-Shot, CoT, and CoT+Few-Shot.

6 Limitation

This paper pioneers research into exceptional cases beyond commonsense knowledge, which has traditionally been the focus of benchmarking foundation models (FMs). It aims to explore how FMs, recognized for their high performance across various domains, can address situations they typically struggle with, thereby advancing towards human-like reasoning. To this end, the study develops datasets encompassing diverse modalities, including image-only, text-only, and multimodal combinations. However, current research still lacks coverage of exceptional cases involving graph interpretation(Tannert et al., 2023) and audio data(Yang et al., 2024). Future studies should establish benchmarks for exceptional cases in these and other unaddressed domains, defining appropriate tasks for their evaluation. We have only addressed English and Korean languages, leaving third-country languages unexplored and providing opportunities for further expansion. We utilized GPT-4o to ensure precise grammar and word usage.

References

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

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. 2022. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736.

Zaid Alyafeai, Maged S Al-shaibani, Mustafa Ghaleb, and Yousif Ahmed Al-Wajih. 2022. Calliar: an online handwritten dataset for arabic calligraphy. *Neural Computing and Applications*, 34(23):20701–20713.

Jason Armitage, Endri Kacupaj, Golsa Tahmasebzadeh, Swati, Maria Maleshkova, Ralph Ewerth, and Jens Lehmann. 2020. Mlm: A benchmark dataset for multitask learning with multiple languages and modalities. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pages 2967–2974.

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.

Yixin Chen, Shuai Zhang, Boran Han, Tong He, and Bo Li. 2024. Camml: Context-aware multimodal learner for large models. *arXiv preprint arXiv:2401.03149*.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2023. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113.

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*.

Yann Dubois, Chen Xuechen Li, Rohan Taori, Tianyi Zhang, Ishaan Gulrajani, Jimmy Ba, Carlos Guestrin, Percy S Liang, and Tatsunori B Hashimoto. 2024. Alpacafarm: A simulation framework for methods that learn from human feedback. *Advances in Neural Information Processing Systems*, 36.

Zhijing Jin, Yuen Chen, Felix Leeb, Luigi Gresele, Ojasv Kamal, Zhiheng Lyu, Kevin Blin, Fernando Gonzalez Adauto, Max Kleiman-Weiner, Mrinmaya Sachan, et al. 2024. Cladder: A benchmark to assess causal reasoning capabilities of language models. *Advances in Neural Information Processing Systems*, 36.

Tushar Khot, Peter Clark, Michal Guerquin, Peter Jansen, and Ashish Sabharwal. 2020. Qasc: A dataset for question answering via sentence composition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8082–8090.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR.

Tianle Li, Wei-Lin Chiang, Evan Frick, Lisa Dunlap, Banghua Zhu, Joseph E Gonzalez, and Ion Stoica. From live data to high-quality benchmarks: The arena-hard pipeline, april 2024. *URL https://lmsys.org/blog/2024-04-19-arena-hard*.

Jung Liang, Wen-Hung Liao, and Yi-Chieh Wu. 2020. Toward automatic recognition of cursive chinese calligraphy: An open dataset for cursive chinese calligraphy text. In *2020 14th International Conference* *on Ubiquitous Information Management and Communication (IMCOM)*, pages 1–5. IEEE.

Bill Yuchen Lin, Yuntian Deng, Khyathi Chandu, Faeze Brahman, Abhilasha Ravichander, Valentina Pyatkin, Nouha Dziri, Ronan Le Bras, and Yejin Choi. 2024. Wildbench: Benchmarking llms with challenging tasks from real users in the wild. *arXiv preprint arXiv:2406.04770*.

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

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024. Visual instruction tuning. *Advances in neural information processing systems*, 36.

Yuliang Liu, Zhang Li, Hongliang Li, Wenwen Yu, Mingxin Huang, Dezhi Peng, Mingyu Liu, Mingrui Chen, Chunyuan Li, Lianwen Jin, et al. 2023. On the hidden mystery of ocr in large multimodal models. *arXiv preprint arXiv:2305.07895*.

Nicholas Lourie, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2021. Unicorn on rainbow: A universal commonsense reasoning model on a new multitask benchmark. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13480–13488.

Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. 2021. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2200–2209.

OpenAI. 2023. Gpt-4v(ision) system card. https://cdn.openai.com/papers/GPTV_System_Card.pdf.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2021. Winogrande: An adversarial winograd schema challenge at scale. *Communications of the ACM*, 64(9):99–106.

Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A Smith, and Yejin Choi. 2019. Atomic: An atlas of machine commonsense for if-then reasoning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 3027–3035.

Yongxin Shi, Dezhi Peng, Wenhui Liao, Zening Lin, Xinhong Chen, Chongyu Liu, Yuyi Zhang, and Lianwen Jin. 2023. Exploring ocr capabilities of gpt-4v (ision): A quantitative and in-depth evaluation. *arXiv preprint arXiv:2310.16809*.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2018. Commonsenseqa: A question answering challenge targeting commonsense knowledge. *arXiv preprint arXiv:1811.00937*.

Simon Tannert, Marcelo G Feighelstein, Jasmina Bogojeska, Joseph Shtok, Assaf Arbelle, Peter WJ Staar, Anika Schumann, Jonas Kuhn, and Leonid Karlinsky. 2023. Flowchartqa: the first large-scale benchmark for reasoning over flowcharts. In *Proceedings of the 1st Workshop on Linguistic Insights from and for Multimodal Language Processing*, pages 34–46.

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, et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Xiyao Wang, Yuhang Zhou, Xiaoyu Liu, Hongjin Lu, Yuancheng Xu, Feihong He, Jaehong Yoon, Taixi Lu, Gedas Bertasius, Mohit Bansal, et al. 2024. Mementos: A comprehensive benchmark for multimodal large language model reasoning over image sequences. *arXiv preprint arXiv:2401.10529*.

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.

Alfonso Wong. 1973–1989. Old master q comics. *Old Master Q Comics*.

Xudong Xie, Ling Fu, Zhifei Zhang, Zhaowen Wang, and Xiang Bai. 2022. Toward understanding wordart: Corner-guided transformer for scene text recognition. In *European Conference on Computer Vision*, pages 303–321. Springer.

Qian Yang, Jin Xu, Wenrui Liu, Yunfei Chu, Ziyue Jiang, Xiaohuan Zhou, Yichong Leng, Yuanjun Lv, Zhou Zhao, Chang Zhou, et al. 2024. Air-bench: Benchmarking large audio-language models via generative comprehension. *arXiv preprint arXiv:2402.07729*.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. 2023. Mmmu:

A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. *arXiv preprint arXiv:2311.16502*.

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*.

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

A Graphic Novels

A.1 Why Choose Graphic Novels

We utilized graphic novels, which are rich in content and often depict exceptional cases, to test the FMs' understanding. The experiment involves short story graphic novels: four-panel graphic novels with shuffled sequences, where the task for the FMs is to rearrange the panels into the correct order. We selected 'Old Master Q Comics' (Wong, 1973–1989) for this purpose, as these graphic novels revolve around comedy and typically have short story lines. They depict exceptional cases that are unlikely to occur in everyday life but have solid plot lines. The probability of correctly ordering the four panels by chance alone is 1/24, making it highly unlikely to achieve the correct sequence purely by luck.

A.2 How to Collect the Dataset

We collected the graphic novels through web scraping and then segmented them panel by panel using automated Python scripts. We reviewed and excluded data entries that contained unevenly sized panels to maintain consistency in the dataset. This dataset allows us to evaluate the extent to which the FMs comprehends the story line. To ensure an accurate assessment, we eliminate all clues that provide information about the story line, including panel numbers and titles of the graphic novel as shown in Figure 13.

A.3 Experiments Details

The API temperature setting is regulated to 0 to ensure consistent results. In order to generate a concise answer, the model is instructed to output the answer solely in the format [1,2,3,4].(Blue letters in Figure 12 'Prompt').We fix the ground truth order as [1,3,2,4] to automate the task, as the input images are shuffled.((e) in Figure 12 'In the code') This predetermined order allows us to verify whether the GPT-4o provides the correct sequence. Additionally, we illustrate how we designed the prompts for each style in E.1.Table 15. We design the random shuffle experiment as follow, refer to Figure 12:

1. Inform the GPT-4o that the uploaded images represent parts of a story that have been shuffled and consist of four images.(Blue letters in Figure 12 'Prompt') Instruct it to analyze all the images and deduce the correct sequence.
2. Upload four images in a shuffled order, with each image assigned an ID number.((a),(b) in Figure 12 'In the code')
3. The uploaded images are indexed, and the GPT-4o infers the correct order, subsequently outputting the images in the proper indexed sequence.((c) in Figure 12 'In the code')
4. Using code, the indexed sequence is transformed into a sequence of image ID numbers to obtain the image order predicted by the GPT-4o.((d) in Figure 12 'In the code')
5. Compare the predicted image order with the ground truth order to determine accuracy.((e) in Figure 12 'In the code')

A.4 Experiments Result

A.4.1 Quantitative Result

We assessed the multimodal causal reasoning abilities of GPT-4o through a Random Shuffle task. The highest performance was observed in the CoT+Few-Shot condition, followed by CoT and then Zero-Shot.(Table 1.) Interestingly, the Zero-Shot performance exceeded expectations, displaying an accuracy that was not markedly lower than the other prompting styles. During the CoT style prompt experiments, we conducted various tests ranging from the very simple 'Let's think step by step' to more detailed descriptions of the reasoning sequence.(Table 9.)

Interestingly, the simplest 'Let's think step by step' prompt yielded the best performance. There was some variation depending on whether 'Let's think step by step' was prompted before or after the task images. In the case of CoT+Few-Shot, the number of Few-Shot examples impacted performance; with only one example, there was no difference compared to CoT, but increasing the examples to three resulted in a noticeable performance improvement.

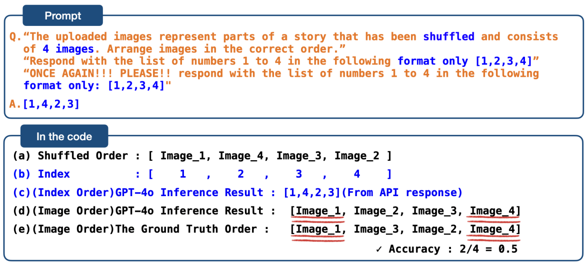

Figure 12: Description of the random shuffle experiment process: In the ’Prompt’, all essential information is provided, including the fact that all images are shuffled, that the four images are parts of a story, and the response format. The ’Code’ section illustrates the task sequence from (a) to (e). (a) shows the shuffled input image order, (b) is the index of the input image order, (c) is GPT-4o’s response which is the inferred result, (d) is the transformation from index order to image order, and (e) is the ground truth order used to calculate accuracy.

Graphic Novels	
Example	Prompt
CoT (Detailed Multi-Step Version)	Input : Q. “The uploaded images represent parts of a story that has been shuffled and consists of 4 images." "Arrange images in the correct order.” IMPORTANT: Respond ONLY with the list of numbers 1 to 4 in this format: [1, 2, 3, 4]. A. Let’s think step by step. 1. Initial Observation: Look at the comic image for a moment. What stands out to you immediately? 2. Setting Description: Describe the setting. Where does the scene take place? Include details about the background and environment. 3. Character Identification: Who are the characters in the image? Describe their appearance and any notable features. 4. Actions and Interactions: What are the characters doing? Describe their actions and how they interact with each other. 5. Text Elements: What text elements are present? What are the characters saying or thinking, and how does this contribute to the scene? 6. Emotional Tone and Atmosphere: What is the emotional tone of the scene? Describe the mood and emotions conveyed by the characters and setting. 7. Context and Story Progression: What do you think happened before this scene, and what might happen next? How does this image fit into the larger story? 8. Summary and Interpretation: Summarize your description. What is the key aspect of this comic image, and what theme or message does it convey? By these logical steps, the correct order of the images is: Output: A.

Table 9: We tried many other version of CoT to enhance capability of GPT-4o on Graphic Novels dataset such as the prompt in this table.

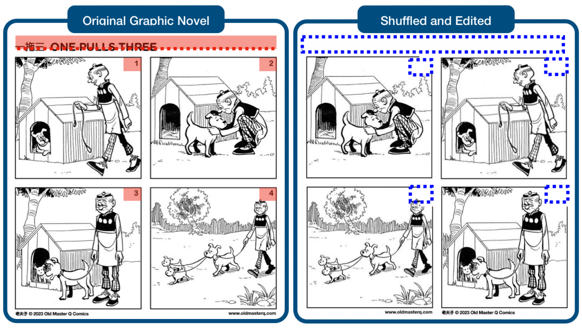

Figure 13: We remove clue-containing sections marked by red boxes that help determine the correct story line. These sections were removed as shown by the blue dotted line boxes in the ’Shuffled and Edited’ version.

B Calligraphy

B.1 The reason we select Korean calligraphy

The reason that we selected Korean calligraphy is three. First, OCR in Korean and artistic letters has not been researched a lot. Second, GPT-4o show Korean OCR as third lowest result in various languages about scene text OCR(Shi et al., 2023). Third, GPT-4o performed Word art(artistic English) as second lowest text recognition accuracy between various text datasets like IAM(English handwritten), ReCTS(Chinese scene text) etc in the (Liu et al., 2023).

B.2 How to preprocess Korean calligraphy

We preprocessed the dataset under three rules. First, delete images if the resolution of the image is too low or if the image has too many letters that even human cannot recognize. The criteria of the number is 35. As we can see in Figure 14, 35 is the irregularly large number in the data. And We found that images with longer than 35 characters are visually difficult for human to recognize. Thus we did not evaluate images longer than 35 characters. Second, separate the image by bounding box given from OCR API if image is overlapped with multiple calligraphy. Third, crop the typographies in the image such as sign, watermark that are considered to be irrelevant to the calligraphy. We also attached the example of preprocessed Korean calligraphy in Figure 15

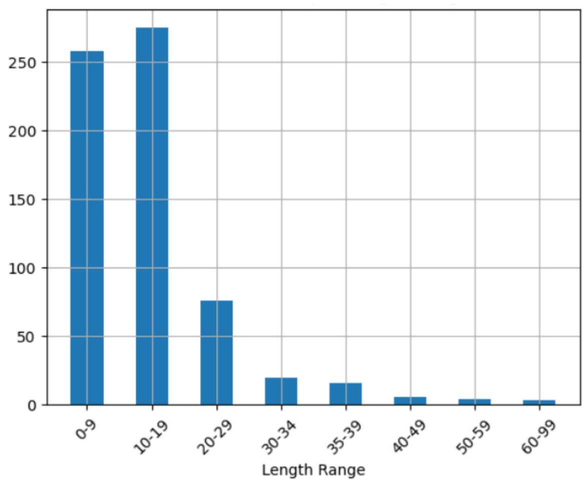

Figure 14: Length plot of Korean calligraphy images before preprocessing

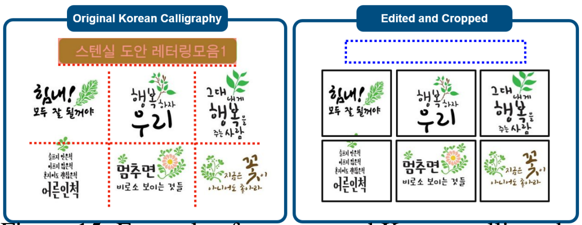

Figure 15: Example of prepocessed Korean calligraphy. We removed typography that isn't related to calligraphy. And we cropped the overlapped parts automatically by using bounding box that OCR API detected.

B.3 Evaluations with Various Metrics and Qualitative Results

For further exploration, we examine how the brevity of sentences affects the performance of GPT-4o, given that the calligraphy dataset contains sentences of varying lengths. To ensure an accurate examination, we evaluate various metrics such as precision, recall, and F1-Score. The denominators for recall and precision are determined by the word count in the predicted sentence and the number of accurately predicted characters divided by the word count in the ground truth sentence. Utilizing these metrics, we conduct experiments on our Korean calligraphy dataset. For precise observation, we scrutinize three different length splits: character-level, word-level, and the overall meaning of the calligraphy.

	Zero-Shot	CoT	CoT+Few-Shot
Precision	84.57	86.47	87.48
Recall	86.64	87.83	88.75
F1-score	85.59	87.14	88.11

Table 10: Result(%) of Korean Calligraphy. OCR task in word-level.

	Zero-Shot	CoT	CoT+Few-Shot
LD-Metric(↓)	4.06	3.08	2.81
Precision(↑)	75.46	76.27	77.25
Recall(↑)	74.82	75.50	76.17
F1-score(↑)	74.51	75.29	76.14

Table 11: Result(%) of Korean Calligraphy. OCR task in character-level.

	Zero-shot	CoT	CoT+Few-shot
Cosine Similarity	83.76	85.67	86.06

Table 12: Result(%) of Korean Calligraphy. OCR task in overall meaning of Korean Calligraphy.

C Onion,Not The Onion

C.1 How to Collect the Dataset

We performed web scraping on The Onion website and Reddit's Not The Onion section. Following the data collection, we implemented an additional filtering process to ensure the Sophistication of the dataset. Specifically, we removed instances where no content was collected, where the same content was repeated, and where advertisements were included.

During the preprocessing stage, we encountered valid data with either long or short lengths, which were indeed written by humans. However, these instances represent qualitative news articles; therefore we decided not to remove them to preserve the integrity of the dataset. Consequently, the mean and median text lengths are 2243 and 1433 respectively, resulting in a left-skewed distribution. A histogram illustrating the text lengths for each data example and category-specific statistics are presented in (Figure 16) and (Table **??**) below.

Through this process, we structured the dataset so that only the title and content of the original news articles influenced the FMs judgment during the fake news detection. This approach ensured that we had a reliable dataset, enabling us to evaluate the impact of textual data alone in fake news detection research.

C.2 Experiments Details

Recent studies have demonstrated that proper prompting can enhance the performance of FMs(Kojima et al., 2022). In this study, The default prompt simply asked the model to distinguish between fake news and real news. In contrast, the CoT prompts instructed the model to go through a step-by-step process of thinking to determine fake news(Wei et al., 2022). In this methodology, the model is instructed to take specific thought steps. Finally, we measured the performance of the model for the Fewshot and CoT prompts by providing examples of fake news and real news, as well as illustrating the judgment process. Through these comparisons, we evaluated the impact of various prompting methods on the model's ability to recognize fake news. The detailed prompts are provided in Table 17. By distinguishing between fake news and real news, we contribute to preventing social disruption and maintaining the credibility of information.

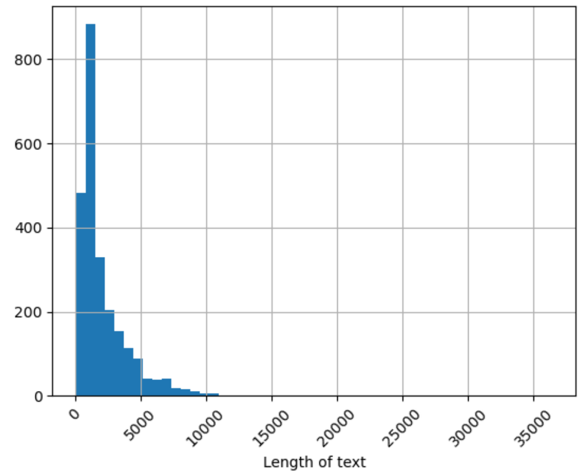

Figure 16: Length plot of the preprocessed Onion and Not the Onion news data.

D Lyrics

D.1 Why Choose Lyrics

We assess the ability of Foundation Models (FMs) to understand song lyrics, which often feature metaphorical and figurative language, slang, cultural references, and ambiguity, necessitating nuanced, human-like comprehension. We evaluate FMs' understanding of song lyrics through three tasks: genre detection, song description generation, and infilling.

The implications of these tests are significant and suggest several future research directions. Firstly, using only lyrics for genre detection is innovative and could lead to new methods for understanding the linguistic features that define musical genres. Additionally, evaluating GPT-4's comprehension of complex language features highlights the model's depth of language understanding, which is crucial for enhancing natural language processing tasks. Future research could expand into cross-linguistic analysis, incorporating more languages to examine how well GPT-4 and other language models handle multilingual datasets and cultural nuances.

D.2 How to Collect the Dataset

	English		Korean	
	Seen	**Unseen**	**Seen**	**Unseen**
Total	3400	1700	3400	1700
Delete duplicate songs	3112	353	2187	304
Lyrics and Description crawling	2435	246	2187	304
Genre crawling	1828	139	2187	304
Remove Multilingual	1803	131	X	X
Remove duplicate between yearly and weekly	X	121	X	176
Cleaning Genre	1703	108	X	X
Final	1703	108	2187	176

Table 13: During the collection of song data, various criteria were used to remove certain songs, as detailed in the first column of the table. Numbers in each blocks denotes the number of remaining data after each step. X indicates that the dataset did not go through that step.

For the English dataset, after collecting the title and artist of each song, we removed overlapping entries. A song was only removed if both the title and artist were the same, as different songs can share the same title. We then created links to the Genius site, where we obtained the lyrics and descriptions of the songs. This involved deleting strings following 'featuring' and modifying characters such as brackets and Latin alphabets. If it was impossible to crawl any of the descriptions, genre, or lyrics, due to link generation error or not being available on the site, we removed the song. Additionally, songs with non-English lyrics were also removed. If a song appeared in both the weekly and yearly data, it was excluded from the weekly data to ensure that the weekly dataset contained only data that GPT had not previously learned. For the genre detection task in English, we streamlined the genre list by removing infrequent genres. After combining all genre lists, we excluded genres with fewer than 10 occurrences. Songs that did not fall under these genres were also excluded, resulting in a unique genre list of 58 genres and a dataset of 1,811 songs. A similar process was applied to both the English and Korean datasets. However, for the Korean dataset, non-Korean lyrics were not removed due to their high frequency, and genre cleaning was not performed because there are not many kinds of genres in the dataset. Notably, no songs were removed during lyrics, description, or genre crawling for the Korean dataset because all song information was gathered from Melon, unlike the English dataset, which compiled data from multiple sites. The specific number of remaining data at each step is summarized in Table 13.

For the infilling task, only the 2024 unseen dataset was used due to GPT-4o's safety issues. In order to evaluate GPT-4o's understanding of exceptional data, we removed non-exceptional data by using a pre-trained BERT model to anticipate the masked part. We evaluated the prediction of BERT model on English dataset through BERT score, while the Korean dataset used cosine similarity between BERT-encoded sentences. Entries with scores exceeding a 0.9 threshold were removed, as high semantic similarity indicated non-exceptional data.

D.3 Evaluation metrics

		Genre Classification	Description Generation	Lyrics Infilling
Korean	Seen	- Overlap Ratio - Exact Match		
	Unseen	- Overlap Ratio - Exact Match		- ROUGE - Cosine Similarity
English	Seen	- Overlap Ratio - Exact Match	- ROUGE - BERT Score	
	Unseen	- Overlap Ratio - Exact Match	- ROUGE - BERT Score	- ROUGE - BERT Score

Table 14: Evaluation metric of each task using lyrics. Empty block denotes that we did not used the data for the corresponding task. For evaluation, We used BERT score(Zhang et al., 2019) and Rouge score(Lin, 2004) for lyrics infilling and description generation task, and Overlap Ratio and Exact Match score for genre classification task.

An exact match score assigns 1 if a predicted genre matches the original genre. The overlap ratio measures similarity based on shared elements.
The F1 score is calculated as the balanced average of precision and recall. Precision quantifies the number of tokens in the generated answer are in the ground truth, while recall measures how many of the ground truth tokens are present in the generated answer. Thus, the balanced average of precision and recall indicates the extent to which the generated answer and the ground truth overlap.
Recall scores were utilized to confirm if the original lyrics are present within the words generated by GPT-4o.

D.4 Experiments details

The API temperature setting is regulated to 0 to ensure consistent results. The model is instructed to generate answers in specific formats: for the Genre classification task, "Genre: the output"; for the song description generation task, "Description: the output"; and for the infilling task, GPT-4o should provide the complete lyrics, including the predicted masked part. Additional details about the prompts are in Appendix E.4

D.4.1 Genre classification

We design the genre classification task as follow:
1. A unique genre list was created by concatenating all possible genres and removing entries with fewer than 10 occurrences. This reduced the size of the genre lists and removed datasets with no genres.
2. We conducted separate experiments on the seen dataset, which includes data from 1990 to 2023, and the unseen dataset, covering January to April 2024. This was done to determine if there is a performance difference between the periods that GPT has been trained on and those it has not.
3. GPT-4o was then asked to select the most likely genre(s) based on the provided lyrics.
4. For the zero-shot approach, GPT-4o generated the output directly. For the CoT and CoT+Few-shot prompts, GPT-4o was instructed to think in alignment with the lyrics.

D.4.2 Description generation

We design the description generation task as follow:
1. GPT-4o was asked to generate a song description based on the provided lyrics.
2. We conducted separate experiments on the seen dataset, which includes data from 1990 to 2023, and the unseen dataset, covering January to April 2024. This was done to determine if there is a performance difference between the periods that GPT has been trained on and those it has not.
3. Since many ground truth song descriptions included additional information about the song (e.g., interviews, messages to fans, or musical features), for the CoT and CoT+Few-shot prompts, we included instructions for GPT-4o to add possible artist names, title names, and musical features.

D.4.3 Lyrics infilling

We design the lyrics infilling task as follow:
1. For the English seen and unseen datasets, masking was performed based on both word and token criteria to determine which masking technique would be more challenging.
2. Using BERT, we compared the two masking methods: the average score for word-based masking was lower, so we decided to use the word-based

masking dataset
3. The Korean unseen dataset was also masked based on words, without the process described in step 1. 4. The infilling task was performed on the Korean and English datasets using BERT.
5. The results from step 4 were evaluated using the BERT score for English and cosine similarity for Korean. Data with scores exceeding 0.9 were removed.
6. After step 5, the remaining data was used to perform the infilling task with GPT-4o. Due to GPT's safety issues, only the unseen dataset was used.

E Prompts

E.1 Graphic Novels

Graphic Novels	
Example	**Prompt**
Zero-Shot	Input : “The uploaded images represent parts of a story that has been shuffled and consists of 4 images." "Arrange images in the correct order.” “Respond with the list of numbers 1 to 4 in the following format only [1,2,3,4]” “ONCE AGAIN!!! PLEASE!! respond with the list of numbers 1 to 4 in the following format only: [1,2,3,4]" (Task Images) Output: A.
CoT + Zero-Shot	Input : Q. “The uploaded images represent parts of a story that has been shuffled and consists of 4 images." "Arrange images in the correct order.” IMPORTANT: Respond ONLY with the list of numbers 1 to 4 in this format: [1, 2, 3, 4]. (Task Images) Output: A. Let’s think step by step. The correct order is
CoT + Few-Shot	Input : Q. “The uploaded images represent parts of a story that has been shuffled and consists of 4 images." "Arrange images in the correct order.” IMPORTANT: Respond ONLY with the list of numbers 1 to 4 in this format: [1, 2, 3, 4]. **“The First, Example:”**: (1st Example Images) A. “Let’s think step by step. The correct order is [1,2,3,4]” **“The Second, Example:”**: (2nd Example Images) A. “Let’s think step by step. The correct order is [1,2,3,4]” **“The Third, Example:”**: (3rd Example Images) A. “Let’s think step by step. The correct order is [1,2,3,4]” Q. “The uploaded images represent parts of a story that has been shuffled and consists of 4 images." "Arrange images in the correct order.” IMPORTANT: Respond ONLY with the list of numbers 1 to 4 in this format: [1, 2, 3, 4]. (Task Images) Output: A. Let’s think step by step. The correct order is

Table 15: The description of each prompt style is provided. We assigned a response format to GPT-4o twice because, in Zero-Shot, the variation in responses is too broad, causing GPT-4o to occasionally break the response format rule. In CoT+Zero-Shot, we utilized the simplest CoT style because it achieved the best score compared to the more detailed CoT version (Table 9.). In CoT+Few-Shot, we used three different examples. The performance was insufficient when using only one or two examples.

E.2 Calligraphy

Dataset Name	
Example	**Prompt**
Zero-Shot	Input : One korean calligraphy image Prompt : "What are the all Korean characters in the image? Make sure that your answer only includes the result of the OCR without translating. You don’t need to describe the processing steps." Output: Only OCR result text
CoT + Zero-Shot	Input : One korean calligraphy image Prompt : "The image uploaded is Korean calligraphy with illustration. Transcribe the letters in the uploaded image. Solve it with following steps. 1. Identify the start and end of the sentence. Check if there are any line breaks in the middle of the sentence. 2. Split the recognized text into individual words. Combine the split words based on the context to form a coherent sentence. 3. Analyze the context to infer the meaning of the handwriting. Correct typos by comparing them with similar words and choosing the correct one. 4. Perform grammar and spelling checks to verify the recognized sentence. Ensure that the sentence flows naturally and makes sense. Don’t describe your steps. Just answer the result of the OCR without translating." Output: Only OCR Result text
CoT + Few-Shot	Input : One korean calligraphy image Prompt : "Below are examples of OCR task. I’ll show image first and explain step-by-step how to extract text from the image." **Example1**: example1 image "Step1: Identify the start and end of the sentence. Check if there are any line breaks in the middle of the sentence. Identify that the sentence starts with ’바라는게’ and ends with ’안그래?’ Step2: Split into words and translate each word in English and identify any typos based on the context.: 바라는게 (What I hope for) 무한정 (infinitely) 끝없이 (endlessly) 내리는 (falling) 게 (particle, indicating ’is’) 아닌게 (is not) 엄마나 (Typo: misidentified word, Correct: 얼마나, Translation: how much) 다행인지 (fortunately) 몰라 (I don’t know) 안그래? (isn’t it?) Step3: Correct the typos by comparing each word with similar words and combine the corrected words to form a coherent sentence.: ’엄마나’ should be ’얼마나’, ’알고 래?’ should be ’안그래?’ Step4: Combine based on context: ’바라는게 무한정 끝없이 내리는 게 아닌게 얼마나 다행인지 몰라 안그래?’ There is no weird word to use. Step5: Analyze the context to infer the meaning of the handwriting. Correct any misrecognized words by comparing them with similar words and choosing the correct one. Infer the context: The sentence talks about how fortunate it is that something is not happening endlessly. Correct any misrecognized words: ’얼 마나’ should be ’얼마나’ Step6: Perform grammar and spelling checks to verify the recognized sentence. Ensure that the sentence flows naturally and makes sense. Check grammar and spelling: Ensure ’바라는게 무한정 끝없이 내리는 게 아닌게 얼마나 다행인지 몰라 안그래?’ is grammatically correct and makes sense. Ensure the sentence flows naturally and the meaning is clear." prompt: "Now, please perform an OCR task on the following image like the example. The image is Korean calligraphy with an illustration. Transcribe the letters in the picture with a step-by-step explanation of your reasoning. But Don’t describe your steps. Just answer the result of the OCR without translating." Output: Only OCR Result text

Table 16: Korean Calligraphy Prompt: For the Cot+Few shot prompt, We utilized two examples but only one example is listed in the paper because it was too long to attach. The full prompt can be seen in GitHub.

E.3 Onion Not The Onion

Onion, Not The Onion	
Example	Prompt
Zero Shot	Input : A News article and Title Prompt: The uploaded text is one of the articles that may be real or fake. Please Answer whether below article is fake or real. Say nothing but **the number 0 or 1**. i.e. Answer 1 if you think the article is real, answer 0 if you think it is fake Output: (0 ‖ 1)
CoT + Zero Shot	Input : A News article and Title The uploaded text is one of the articles that may be real or fake.Please Answer whether below article is fake or real. Give a 20-character rationale for why you think that way, and output a 0 and 1 at the end of the sentence. To Solve this, You have to think step by step. **The first step** in identifying fake news is evaluating the reliability of the information source. Well-known and verified news organizations are generally more reliable, and their reports can be trusted more than unverified sources. **In addition** to source reliability, look at the language used in the content. Fake news often uses sensational or exaggerated language designed to elicit an emotional response. **It is also important to check** for consistency and accuracy in the information presented; fake news typically includes claims that are either unverified or clearly false. **Another critical step** is cross-verification,where check if the same claims are reported by multiple trusted sources. i.e. rationale + answer 1 if you think the article is real, rationale + answer 0 if you think it is fake. Must Keep in mind that the end of a sentence should end with either 0 or 1 Output: (rationales + (0‖1))
CoT + few Shot	Input : A News article and Title The uploaded text is one of the articles that may be real or fake.Please Answer whether below article is fake or real. Give a 20-character rationale for why you think that way, and output a 0 and 1 at the end of the sentence. To Solve this, You have to think step by step. **The first step** in identifying fake news is evaluating the reliability of the information source. Well-known and verified news organizations are generally more reliable, and their reports can be trusted more than unverified sources. **In addition** to source reliability, look at the language used in the content. Fake news often uses sensational or exaggerated language designed to elicit an emotional response. **It is also important to check** for consistency and accuracy in the information presented; fake news typically includes claims that are either unverified or clearly false. **Another critical step** is cross-verification,where check if the same claims are reported by multiple trusted sources. **See the example below**. i.e. rationale + answer 1 if you think the article is real, rationale + answer 0 if you think it is fake. Must Keep in mind that the end of a sentence should end with either 0 or 1 **Example**: we provided one fake news story from The Onion and one real news story from Reddit's Not the Onion. Additionally, rather than merely presenting the news, **we included examples of the rationales we derived for the two news stories, following the same prompting method**. Output: (rationales + (0‖1))

Table 17: We provided examples of prompts used to detect fake news, focusing on the implementation of CoT reasoning. We presented a structured approach that outlines the steps a FMs considers when analyzing and concluding whether a news story is fake or real. Lastly, this method involves a few-shot learning technique where examples of fake news and real news are given alongside rationales.

E.4 Lyrics

E.4.1 English Genre Classification

Lyrics	
Example	Prompt
Zero-Shot	Input : Lyrics Prompt : Here is a list of unique music genres: ['genre list str']. Say nothing but the Genre as Genre: the output. Output example: Genre: [pop, r&b, hip hop]. Lyrics: 'lyrics' Output: Genre: the output
CoT + Zero-Shot	Input : Lyrics Prompt : Here is a list of unique music genres: ['genre list str']. Based on the lyrics provided, identify the genres. Say nothing but the Genre as Genre: the output. Output example: Genre: [pop, r&b, hip hop]. Lyrics: 'lyrics' Output: Genre: the output
CoT + Few-Shot	Input : Lyrics Prompt : Here is a list of unique music genres: ['genre list str']. **Example Lyrics**: And she spoke words that would melt in your hands And she spoke words of wisdom To the basement, people, to the basement Many surprises await you In the basement, people, in the basement You hid there last time, you know we're gonna find you Sick in the car seat, 'cause you're not up to going Out on the main streets, completing your mission You hid there last time, you know we're gonna find you Sick in the car seat, 'cause you're not up to going Out on the main streets, completing your mission **Example Description**: indie pop Now, based on the lyrics provided, identify the genres. Say nothing but the Genre as Genre: the output. Output example: Genre: [pop, r&b, hip hop]. Lyrics: 'lyrics' Output: Genre: the output

Table 18: Prompt for English genre classification task

E.4.2 Korean Genre Classification

Lyrics	
Example	Prompt
Zero-Shot	Input : Lyrics Prompt : Here is a list of unique music genres: ['genre list str']. Say nothing but the Genre as Genre: the output. Output example: Genre: [발라드, 댄스, 랩/힙합]. Lyrics: 'lyrics' Output: Genre: the output
CoT + Zero-Shot	Input : Lyrics Prompt : Here is a list of unique music genres: ['genre list str']. Based on the lyrics provided, identify the genres. Say nothing but the Genre as Genre: the output. Output example: Genre: [발라드, 댄스, 랩/힙합]. Lyrics: 'lyrics' Output: Genre: the output
CoT + Few-Shot	Input : Lyrics Prompt : Here is a list of unique music genres: ['genre list str']. **Example Lyrics**: 처음 그대 내게로 오던 그날에 잠시 동안 적시는 그런 비가 아니길 간절히 난 바래왔었죠 그대도 내 맘 아나요 매일 그대만 그려왔던 나를 오늘도 내 맘에 스며들죠 그대는 선물입니다 하늘이 내려준 홀로 선 세상 속에 그댈 지켜줄게요 어느 날 문득 소나기처럼 내린 그대지만 오늘도 불러 봅니다 내겐 소중한 사람 Oh 떨어지는 빗물이 어느새 날 깨우고 그대 생각에 잠겨요 이제는 내게로 와요 언제나처럼 기다리고 있죠 그대 손을 꼭 잡아줄게요' **Example Description**: 발라드, 국내드라마 Now, based on the lyrics provided, identify the genres. Say nothing but the Genre as Genre: the output. Output example: Genre: [발라드, 댄스, 랩/힙합]. Lyrics: 'lyrics' Output: Genre: the output

Table 19: Prompt for Korean genre classification task

E.4.3 English Song Description Generation

Lyrics	
Example	Prompt
Zero-Shot	Input : Lyrics Prompt : Say nothing but the Description as Description: the output Output example: Description: The song explores themes of love and heartbreak. Lyrics: ’lyrics’ Output: Description: the output
CoT + Zero-Shot	Input : Lyrics Prompt : Based on the provided lyrics, write a brief description of the song. Include the possible song title and artist name in the description. Say nothing but the Description as Description: the output Output example: Description: Honeymoon Avenue by Ariana Grande is about knowing you are at the end of a relationship and wishing it could not be the end and go back to the beginning and start over. Output: Description: the output
CoT + Few-Shot	Input : Lyrics Prompt : **Example Lyrics**: I’d like to say we gave it a try I’d like to blame it all on life Maybe we just weren’t right But that’s a lie, that’s a lie And we can deny it as much as we want But in time, our feelings will show ’Cause sooner or later, we’ll wonder why we gave up The truth is everyone knows, oh Almost, almost is never enough So close to being in love If I would have known that you wanted me the way I wanted you Then maybe we wouldn’t be two worlds apart (Ah) But right here in each other’s arms And we almost, we almost knew what love was But almost is never enough (Ah) If I could change the world overnight (Ah) There’d be no such thing as goodbye (Ah) You’d be standing right where you were (Ah) And we’d get the chance we deserve, oh (Ah) See upcoming pop shows Get tickets for your favorite artists Try to deny it as much as you want But in time, our feelings will show (Ah) ’Cause sooner or later, we’ll wonder why we gave up The truth is everyone knows (Ah) **Example Description**: On the collaborative track “Almost Is Never Enough,” Ariana Grande & Nathan Sykes play a couple who had a relationship that hadn’t gone right. Ariana would like to say things were going well but she knows that’s a lie and like the title states, almost is never enough to make the relationship work; you need to put full effort in. Both of them state that they didn’t feel the relationship while in it, but the mood of the song and lyrics suggest that they both want to either reconnect or they simply just miss better times. At the time of the song’s release, Nathan and Ariana were dating. Unfortunately, their relationship ended a few months later. Now, based on the provided lyrics, write a brief description of the song. Include the possible song title and artist name in the description. Say nothing but the Description as Description: the output Output example: Description: Honeymoon Avenue by Ariana Grande is about knowing you are at the end of a relationship and wishing it could not be the end and go back to the beginning and start over. Output: Description: the output

Table 20: Prompt for English song description generation task

E.4.4 English Song Infilling

Lyrics Infilling Task	
Example	Prompt
Zero-Shot	Input : Masked lyrics Prompt : You are a powerful language model. Fill in the blanks in the following text with appropriate words. The text is a part of a song with certain words masked by [MASK]. Lyrics: 'lyrics Say nothing but the filled lyrics as 'Filled lyrics: the output'. Output example: Filled lyrics: 'I know this pain (I know this pain) why do you lock yourself up in these chains? (these chains). . . Output: Filled lyrics: the output
CoT + Zero-Shot	Input : Lyrics Prompt : You are a powerful language model. Fill in the blanks in the following text with appropriate words.The text is a part of a song with certain words masked by [MASK]. For each blank, think step by step about the context and meaning of the surrounding text before choosing the word. To do this, follow these steps: a. Carefully read and analysis the lyrics. b-1. Check the entire lyrics to see if there are any repeating parts. b-2. If repeating parts exist, replace the [MASK] with the corresponding word. c-1. Make the list of possible words for the masked part. c-2. Select a suitable word from the candidate list. c-3. Replace [MASK] with the word that you selected. Lyrics: 'lyrics Step-by-step reasoning and filled lyrics as 'Filled lyrics: the output'. Say nothing but the filled lyrics as 'Filled lyrics: the output'. Output example: Filled lyrics: 'I know this pain (I know this pain) why do you lock yourself up in these chains? (these chains). . . Output: Filled lyrics: the output
CoT + Few-Shot	Input : Lyrics Prompt : You are a powerful language model. Fill in the blanks in the following text with appropriate words. The text is a part of a song with certain words masked by [MASK]. For each blank, think step by step about the context and meaning of the surrounding text before choosing the word. To do this, follow these steps: a. Carefully read and analysis the lyrics. b-1. Check the entire lyrics to see if there are any repeating parts. b-2. If repeating parts exist, replace the [MASK] with the corresponding word. c-1. Make the list of possible words for the masked part. c-2. Select a suitable word from the candidate list. c-3. Replace [MASK] with the word that you selected. Example: Lyrics: Rotgut whiskey's gonna ease my mind Beach [MASK] rests on the dryin' line Do I remind you of your daddy in his '88 Ford? Labrador [MASK] out the passenger door The sand from your hair is blowin' in my eyes [MASK] it on [MASK] [MASK] grown men don't cry [MASK] [MASK] remember that beat down basement couch? I'd sing [MASK] my love songs [MASK] you'd tell me about How your mama [MASK] off and pawned her ring [MASK] remember, I remember everything Filled lyrics: Rotgut whiskey's gonna ease my mind Beach towel rests on the dryin' line Do I remind you of your daddy in his '88 Ford? Labrador hangin' out the passenger door The sand from your hair is blowin' in my eyes Blame it on the beach, grown men don't cry Do you remember that beat down basement couch? I'd sing you my love songs and you'd tell me about How your mama ran off and pawned her ring I remember, I remember everything Now, based on the provided lyrics, fill in the blanks with appropriate words. Lyrics: 'lyrics Step-by-step reasoning and filled lyrics as 'Filled lyrics: the output'. Say nothing but the filled lyrics as 'Filled lyrics: the output'. Output example: Filled lyrics: 'I know this pain (I know this pain) why do you lock yourself up in these chains? (these chains). . . Output: Filled lyrics: the output

Table 21: Prompt for English lyrics infilling task. Examples in CoT+Few-shot are composed of data removed during BERT testing.

E.4.5 Korean Song Infilling task

Lyrics	
Example	**Prompt**
Zero-Shot	Input : Masked lyrics Prompt : You are a powerful language model. Fill in the blanks in the following text with appropriate words. The text is a part of a song with certain words masked by [MASK]. Lyrics: 'lyrics Say nothing but the filled lyrics as 'Filled lyrics: the output'. Output example: Filled lyrics: 'I know this pain (I know this pain) why do you lock yourself up in these chains? (these chains)... Output: Filled lyrics: the output
CoT + Zero-Shot	Input : Lyrics Prompt : You are a powerful language model. Fill in the blanks in the following text with appropriate words. The text is a part of a song with certain words masked by [MASK]. For each blank, think step by step about the context and meaning of the surrounding text before choosing the word. To do this, follow these steps: a. Carefully read and analysis the lyrics. b-1. Check the entire lyrics to see if there are any repeating parts. b-2. If repeating parts exist, replace the [MASK] with the corresponding word. c-1. Make the list of possible words for the masked part. c-2. Select a suitable word from the candidate list. c-3. Replace [MASK] with the word that you selected. Lyrics: 'lyrics Step-by-step reasoning and filled lyrics as 'Filled lyrics: the output'. Say nothing but the filled lyrics as 'Filled lyrics: the output'. Output example: Filled lyrics: 'I know this pain (I know this pain) why do you lock yourself up in these chains? (these chains)... Output: Filled lyrics: the output
CoT + Few-Shot	Input : Lyrics Prompt : You are a powerful language model. Fill in the blanks in the following text with appropriate words. The text is a part of a song with certain words masked by [MASK]. For each blank, think step by step about the context and meaning of the surrounding text before choosing the word. To do this, follow these steps: a. Carefully read and analysis the lyrics. b-1. Check the entire lyrics to see if there are any repeating parts. b-2. If repeating parts exist, replace the [MASK] with the corresponding word. c-1. Make the list of possible words for the masked part. c-2. Select a suitable word from the candidate list. c-3. Replace [MASK] with the word that you selected. Example: Lyrics: 세상에 음악의 신이 있다면 고맙다고 안아주고 싶어 전 세계 공통의 Language 자음과 모음이 달라도 상관없는 건 Music 말이 안 통해도 [MASK] 있다면 [MASK] 지금부터는 아주 친한 친구 너와 내가 모르는 사이여도 춤출 [MASK] 있어 We [MASK] mix it up right Sugar and spice Brass sound and guitar 네 [MASK] 다 내 [MASK] 쿵치팍치 또한 내 이름인가 이것 또한 나를 위한 소린가 [MASK] [MASK] Drum bass Piano [MASK] Filled lyrics: 세상에 음악의 신이 있다면 고맙다고 안아주고 싶어 전 세계 공통의 Language 자음과 모음이 달라도 상관없는 건 Music 말이 안 통해도 음악이 있다면 우리는 지금부터는 아주 친한 친구 너와 내가 모르는 사이여도 춤출 수 있어 We can mix it up right Sugar and spice Brass sound and guitar 네 글자면 다 내 이름이래 쿵치팍치 또한 내 이름인가 이것 또한 나를 위한 소린가 Kick snare Drum bass Piano Bassline Lyrics: 'lyrics Step-by-step reasoning and filled lyrics as 'Filled lyrics: the output'. Say nothing but the filled lyrics as 'Filled lyrics: the output'. Output example: Filled lyrics: 'I know this pain (I know this pain) why do you lock yourself up in these chains? (these chains)... Output: Filled lyrics: the output

Table 22: Prompt for Korean lyrics infilling task. Examples in CoT+Few-shot are composed of data removed during BERT testing.