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 reasoning abilities. However, there is a lack of studies on FMs performance in exceptional scenarios, which we define as out-of-distribution (OOD) reasoning tasks. This paper is the first to address these cases, developing a novel dataset for evaluation of FMs 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 introduces prompt engineering techniques, Out-of-distribution Reasoning Chain-of-Thought (ORCoT) and ORCoT+Few-Shot, to improve performance. Validation of FMs using various methods revealed improvements. The code repository contains all relevant code and supplementary materials, including prompts such as ORCoT. It is accessible at: https://github.com/Code4PaperBlind/ExceptionalBenchmark

1 Introduction

Recent studies (Joshi et al., 2025; Bandyopadhyay et al., 2025) have focused on assessing general-purpose reasoning capabilities of FMs (Achiam et al., 2023; Team et al., 2023; anthropic, 2024). As a result, current FMs have achieved remarkable progress, demonstrating high performance across various tasks (Cherian et al., 2023; Wang et al., 2019). However, there are situations where FMs struggle to determine reasoning. Despite the development of several datasets (Xiong et al., 2025; Arora et al., 2023; Zellers et al., 2019), there remains a need for benchmarks that assess performance in OOD scenarios—what we refer to as **exceptional cases**. In our work, an exceptional case is defined as an OOD scenario where the test data differs significantly from the pre-training data.

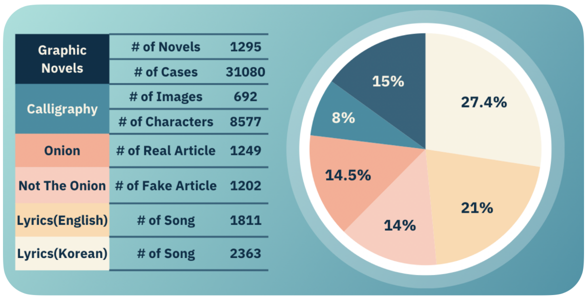


Figure 1: Distribution of Exceptional Cases Dataset and summary of four distinct datasets and their subsections, showcasing diverse characteristics.

While such cases may exist within the pre-training data, they reside in the tail of the distribution, making them rare and challenging for the model to generalize. This rarity means that even if such cases are included in the pre-training data, they are insufficiently represented to allow the models to learn robust generalizable patterns, as evidenced by consistent performance drops across multiple FMs. To formalize this notion, we first introduce the key probability distributions used in our analysis. Accordingly, we define an exceptional case in a reasoning task as one that is OOD.

Notation. Let $x \in \mathcal{X}$ denote an input (image, text, etc.) and $y \in \mathcal{Y}$ its label. Table 1 lists all probability distributions used in this paper.

Symbol	Description
$P_{tr}(x, y)$	Joint distribution in **training** data
$P_{te}(x, y)$	Joint distribution in **test** (exceptional) data
$P_{tr}(x)$, $P_{te}(x)$	Marginal over inputs x (train / test)
$P_{tr}(y)$, $P_{te}(y)$	Marginal over labels y (train / test)
$P_{tr}(y\|x)$, $P_{te}(y\|x)$	Conditional label distributions (train / test)

Table 1: Training vs. exceptional-case distributions

Under the ideal *independent and identically distributed (IID)* assumption, the joint distribution of training and test data is equal, i.e., $P_{\text{tr}}(x, y) = P_{\text{te}}(x, y)$. However, *exceptional cases* inherently violate the IID assumption, as they exhibit distributional shifts: $P_{tr}(x, y) \neq P_{te}(x, y)$, because at least one of the following holds: $P_{tr}(x) \neq P_{te}(x)$, $P_{tr}(y) \neq P_{te}(y)$, $P_{tr}(y|x) \neq P_{te}(y|x)$ (Yang et al., 2024). Based on which of the three distributional shifts occurs, each dataset can be associated with a specific type of exceptional case:

$$P_{tr}(x) \neq P_{te}(x) \text{ (Graphic Novels, Calligraphy)} \tag{1}$$

The graphic novel presents bold, cartoonish storylines ($P_{te}(x)$). Additionally, the calligraphy is artistically stylized, which makes $P_{te}(x)$ significantly different from $P_{tr}(x)$. Both elements are types of content that FMs have rarely encountered in their pre-training datasets ($P_{tr}(x)$).

$$P_{tr}(y) \neq P_{te}(y) \quad \text{(Lyrics)} \tag{2}$$

The task involving lyrics assesses whether FMs can accurately complete masked segments ($P_{te}(y)$), which are designated as exceptional cases, representing scenarios that FMs rarely encounter during training ($P_{tr}(y)$).

$$P_{tr}(y|x) \neq P_{te}(y|x) \text{ (Onion, Not The Onion)} \tag{3}$$

In the case of Onion-style plausible fake news and Not The Onion's real news that appear fake, the label distributions $P_{te}(y|x)$ diverge from those seen during training ($P_{tr}(y|x)$). Despite syntactic similarity to typical news, their semantic-label mappings are flipped or ambiguous, making them challenging for FMs to classify correctly. Figure 1 illustrates the distribution of these four distinct datasets, highlighting the diversity and complexity of the exceptional cases we investigate. Building on these insights, we summarize the key contributions of this paper as follows.

First Multimodal Benchmark for Evaluating FMs on Exceptional Cases: We introduce the first benchmark explicitly designed to evaluate FMs in OOD reasoning scenarios—exceptional cases that have been largely overlooked in prior research. Our contribution includes the construction of a novel multimodal dataset comprising graphic novels, Korean calligraphy, news articles (Onion, Not The Onion), and song lyrics, spanning diverse tasks such as classification, recognition, and generation across text, image, and hybrid modalities.

Comprehensive Validation and Analysis: We provide a thorough validation of FMs using various methods, demonstrating improvements with the proposed techniques. The code repository containing all relevant code and supplementary materials, including ORCoT prompts, is publicly available.

2 Related Work

2.1 Out Of Distribution

OOD(Hendrycks and Gimpel, 2016; Yang et al., 2024) refers to samples that fall outside the statistical distribution of the training data used to develop the model. Researching and enhancing OOD detection capabilities is crucial, as it ensures the reliability and safety of machine learning systems, particularly in applications where decision-making is dependent on reasoning processes. OOD detection presents several significant challenges. Firstly, there is the lack of guidance from unknown data during the training process, as models are typically trained exclusively on in-distribution (ID) data. Secondly, anticipating the locations of OOD data is inherently challenging due to the expansive and intricate nature of the unknown space in high-dimensional environments. Thirdly, these challenges intensify the tendency of large-scale neural networks to produce overly confident predictions(Kang et al., 2023). Fourthly, real-world images consist of various objects and elements. To address these challenges, numerous investigations(Liu et al., 2020; Osada et al., 2023) have been conducted to explore effective methods.

2.2 Foundation Models

Recent advancements in FMs (Achiam et al., 2023; Team et al., 2023; anthropic, 2024) have spurred interest in the integration of multimodal data. To improve the comprehension of human instructions, LLaVA (Liu et al., 2024) proposed visual instruction tuning, enhancing multimodal interaction. More recently, cutting-edge FMs, including Gemini Pro, Claude 3.5 Sonnet, and GPT-4o, have exhibited remarkable performance across a range of multimodal tasks, such as Visual Question Answering (VQA) (Mathew et al., 2021).

2.3 Benchmarks

Evaluating the capabilities of FMs is essential, as it supports the further development and refinement of these models. This has led to the creation of benchmarks across various domains, including reasoning, question answering, coding, and mathematics (Arora et al., 2023; Zheng et al., 2023). Additionally, assessing FMs' abilities with multimodal data—where language is combined with another modality such as images—is critical for expanding the applicability of FMs in diverse fields. As FMs continue to advance, there is increasing interest

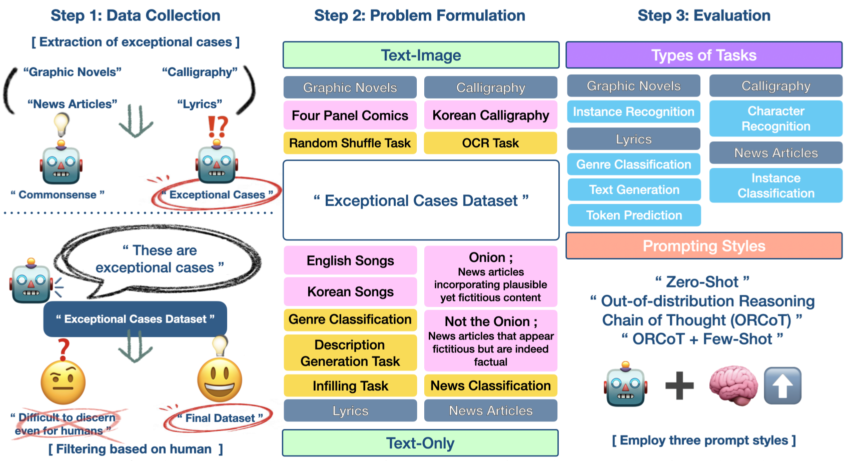


Figure 2: Three-step process for benchmark construction: Step 1 involves collecting data by extracting exceptional cases and reviewing for ambiguous reasoning instances, covering modalities like text-only and text-image. Step 2 defines dataset characteristics and corresponding tasks. Step 3 evaluates FMs through tasks such as instance classification, recognition, token prediction, and text generation, using Zero-Shot, ORCoT, and ORCoT+Few-Shot prompts to assess diversity and accuracy.

in developing benchmarks grounded in real-world data to enhance their reliability and address challenges such as hallucination (Lu et al., 2024).

3 Methodology

3.1 Overview

It has been observed that in certain atypical scenarios (Chen et al., 2024), FMs exhibit errors in reasoning concerning the sequence of events. We designed experiments using four different datasets that feature various characters with multiple types of tasks such as instance recognition, text generation, token prediction, and character recognition, as described in Figure 2. We also propose ORCoT, a revised version of CoT (Wei et al., 2022), to achieve enhanced performance. We employed three prompt styles—Zero-Shot (Kojima et al., 2022), ORCoT, and ORCoT+Few-Shot (Brown et al., 2020)—to investigate how the accuracy of responses varies.

3.2 Out-of-Distribution Chain-of-Thought

The ORCoT prompting strategy extends standard CoT reasoning to enhance robustness in OOD scenarios. Unlike general CoT, which relies on learned patterns, ORCoT incorporates explicit strategies to navigate unfamiliar domains. For story ordering, our strategy uses a two-phase process of analysis and synthesis to identify the underlying causal narrative. A key adaptation for OOD performance is that the model is explicitly guided to follow this narrative causality even when it defies real-world physics (Table 2). This emphasis on structured reasoning ensures the model remains effective with unfamiliar data, and this approach can be adapted to other tasks.

Prompt
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."
Example Analysis Output (Cartoon Logic): Image 1: (a) A cat is standing on a branch, holding a large magnet connected to a battery. Below, a dog with a metal collar is walking by. (b) End State: The cat has a mischievous look and is about to flip the switch on the battery. Image 2: (a) The dog is flying vertically up towards the magnet, looking surprised. The magnet is now glowing. (b) End State: The dog is in mid-air, having been pulled off the ground. Image 3: (a) The dog is walking peacefully, unaware of the cat above. (b) End State: The scene is calm, setting up the situation. Image 4: (a) The dog is stuck to the magnet by its collar. The cat is laughing. (b) End State: The cat's plan has succeeded, and the dog is trapped.
Phase 2: Synthesize and determine the story order. Now, compare the descriptions and end states for all four images. Find the most logical narrative sequence by looking for cause and effect, **even if it defies real-world physics.** Identify the beginning: Which panel sets up the initial scene? It should have no obvious preceding cause within the other panels. Find the causal links: How does the end state of one panel become the start state or cause of another? Explain your reasoning. (Example reasoning: "The dog walking peacefully (Image 3) is the calm beginning. The cat preparing its trap (Image 1) is the next logical step, as it introduces the conflict. The dog being pulled into the air (Image 2) is the direct result of the magnet being activated. The dog being stuck to the magnet (Image 4) is the final outcome of the sequence. Therefore, the internal story logic dictates the order is [3, 1, 2, 4].")
Phase 3: Construct the final order.
Output: A.

Table 2: We explored multiple variations of ORCoT to enhance the capabilities of FMs. The prompt shown here is one such example, tailored for the story ordering task on the Graphic Novels dataset, which uses 'analyze-synthesize' structure to improve OOD reasoning.

3.3 Model Selection

General-Purpose Reasoning Capability Unlike task-specific models such as BERT (Devlin et al., 2018), T5 (Raffel et al., 2023), or Vision Transformers (Dosovitskiy et al., 2021), the selected FMs possess general-purpose reasoning capabilities without requiring task-specific pre-training or fine-tuning. These models are extensively used in real-world applications, providing a balance between high performance, multimodal reasoning capabilities, and accessibility. Given their rising adoption as general-purpose assistants, evaluating their robustness in handling exceptional cases (OOD reasoning tasks) is critical to understanding their practical limitations and strengths.

Multimodal Proficiency The benchmark tasks developed in this study explicitly involve both textual and visual understanding and reasoning. The latest generation of multimodal models demonstrates state-of-the-art performance in integrating image and text within unified architectures. In contrast, open-source alternatives such as OLMo (Groeneveld et al., 2024) still show limited capabilities in multimodal reasoning, as seen in recent benchmarks, making them less suitable for evaluating complex OOD scenarios.

3.4 Graphic Novels

Dataset Description To evaluate the multimodal reasoning capabilities of FMs, we constructed a dataset from the graphic novel series 'Old Master Q Comics' (Wong, 1973–1989). Each comic consists of four-panel narratives presenting a complete story, including a beginning, development, climax, and conclusion, with instances of exceptional cases. We hypothesize that if FMs can reason through storylines, they should infer the correct sequence of panels from a randomly shuffled input. Based on this, we designed a random shuffle task. The dataset includes 1,295 comics, generating 31,080 possible permutations when the four panels are shuffled.

Experimental Design The four images are automatically shuffled within the code before being presented as a prompt to the FMs. We subsequently measure the accuracy of all three prompt styles and conduct an analysis of the results.

Evaluation Metrics To facilitate efficient and accurate experimental evaluation, we developed a code that automatically generates prompts consisting of the input question and shuffled images. We calculate accuracy by comparing model responses to the ground truth (Figure 3).

Image Order Accuracy

```
(Image Order)                                  ✓ Accuracy : 2/4 = 0.5
FMs Inference Result :  [Image_1, Image_2, Image_3, Image_4]
Ground Truth Order :    [Image_1, Image_3, Image_2, Image_4]
```

Figure 3: To evaluate accuracy in the Graphic Novel task, we compare the inferred image sequence with the ground truth and measure the number of correctly ordered images.

3.5 Calligraphy

Dataset Description Unlike the standard text that conventional Optical Character Recognition (OCR) technology primarily handles (Ruder et al., 2023; Kausadikar et al., 2025), calligraphy uses unique writing styles not commonly seen in everyday life. Due to this characteristic, the relevant dataset holds significant value for testing OCR performance in atypical environments. The dataset focuses on OCR tasks specifically for transcribing Korean calligraphy. Initially, we considered developing an English calligraphy dataset; however, this was deemed unnecessary due to the high accuracy already achieved by FMs in this domain. For example, in the WordArt dataset (Shi et al., 2023), an English calligraphy dataset, GPT-4's accuracy is 60.20%, which improves significantly to 77.61% when evaluated with GPT-4o.

Experimental Design We conducted experiments using FMs to transcribe a Korean calligraphy piece. Prior to performing word-level evaluations, we removed punctuation and special symbols from the FMs' predictions, and replaced '\n' with a space (' ') to address ambiguous line breaks inherent in calligraphy.

Evaluation Metrics We employed Word-level Accuracy, Character Error Rate (CER), and Word Error Rate (WER) as standard OCR evaluation metrics, widely used across various OCR models.

3.6 Onion, Not The Onion

Dataset Description The Onion, Not The Onion dataset assesses whether FMs can differentiate between real and fabricated news stories involving unpredictable events. Although parts of this dataset may appear in FMs' training data, it likely resides in the distribution's tail, making it a valuable resource for investigation. Featuring satirical and exaggerated expressions, the dataset focuses on challenging cases where distinguishing real news from fake has become increasingly difficult. Its primary objective is to evaluate whether FMs can demonstrate critical thinking in exceptional scenarios. The dataset is sourced from The Onion, a satirical fake news website, and Reddit's Not The Onion, which highlights real yet seemingly unbelievable stories.

Experimental Design We designed a binary classification task, where '0' corresponds to fake news and '1' to real news. We concatenated examples from each category, applied a random shuffle, and then provided them to FMs.

Evaluation Metrics We evaluate FMs performance in classifying news authenticity using accuracy, precision, and recall. We applied these evaluation metrics to three prompting strategies to assess how variations in prompting affect the model's ability to detect fabricated news.

3.7 Lyrics

Dataset Description Song lyrics, extensively studied in prior research (Jamdar et al., 2015; Tsaptsinos, 2017; Barman et al., 2019), often feature poetic license and literary expressions uncommon in everyday language. Building on this, we designed tasks leveraging BERT (Devlin et al., 2018) to identify exceptional elements within lyrics. These elements were masked, and FMs were tasked with predicting the masked tokens. The dataset also includes genre detection and song description generation tasks.

Experimental Design The three tasks were conducted independently. The first task, Infilling, evaluated the model's ability to predict words masked by BERT within a sentence. The second task, Song Description Generation, assessed the FMs' ability to comprehend and describe the context of song lyrics. The third task, Genre Classification, aimed to determine whether FMs could classify the genre based solely on lyrics.

Evaluation Metrics We evaluated the FMs' responses in the infilling task using BERT F1 scores (Zhang et al., 2019). The evaluation of genre detection was based on exact match scores, assigning a score of 1 if the predicted genre exactly matched the ground truth, along with overlap ratios. For the song description task, BERT F1 and ROUGE F1 scores were used as evaluation metrics.

4 Experiment Result

4.1 Impact of Distribution Shift on Model Reasoning

Distribution shifts critically challenge multi-step reasoning, particularly with OOD inputs like cartoons that can cause perceptual errors to cascade through the reasoning chain. To investigate this, we tasked models with ordering an ID story (real-life photos) and an OOD story (cartoons). Our findings reveal two distinct failure modes. For the ID task, models generated accurate scene descriptions but struggled with the ordering task, producing highly plausible alternative sequences—though they were incorrect. In stark contrast, the OOD task prompted a more fundamental failure at the perceptual level. The models' inability to interpret scenes violating real-world priors led to poor descriptions from the outset, causing a complete collapse in reasoning that yielded logically incoherent sequences. This distinction highlights a key vulnerability: a distribution shift can trigger a systemic failure where initial perceptual errors shatter the entire reasoning process. Further details are provided in Appendix B.1 and B.3.

4.2 Graphic Novels

Quantitative Results We evaluated the multimodal reasoning capabilities of FMs using the random shuffle task. The ORCoT+Few-Shot condition demonstrated superior performance compared to the ORCoT and Zero-Shot conditions. Although the Zero-Shot setting showed slightly lower accuracy compared to ORCoT-based prompts, its performance still exceeded expectations. Among the models, GPT-4o achieved significantly higher accuracy than the other two. Gemini-1.5-Pro showed similar performance levels in the ORCoT and ORCoT+Few-Shot conditions, while Claude-3.5-Sonnet exhibited a substantial improvement in the ORCoT+Few-Shot condition, as shown in Table 3.

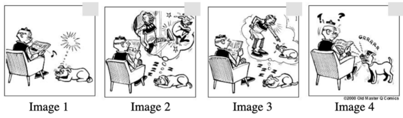
Acc.(%)	Zero-Shot	ORCoT	ORCoT+Few-Shot
Claude-3.5-Sonnet	44.69	44.75	49.92
Gemini-1.5-Pro	51.41	52.45	52.51
GPT-4o	**63.80**	**63.88**	**64.63**

Table 3: Result(%) of random shuffle task. The overall scores are low, indicating that FMs struggle to correctly reason the order of shuffled images.

Qualitative Results To investigate whether image style was impeding reasoning, we prompted FMs to describe single images. The models were generally able to generate detailed descriptions, identifying characters, actions, and even hypothesizing thoughts. Additionally, we assessed their ability to understand image content in the random shuffle task by instructing them to generate descriptions for each image. The accuracy of these descriptions in capturing the visual content was evaluated using the SBERT (Reimers and Gurevych, 2019). We evaluated the FMs' generated descriptions by comparing them to manually crafted ground truth sentences for the example images using cosine similarity. This approach allowed us to investigate how FMs interpret the context within images and infer sequences based on their understanding. In Figure 4, none of the FMs inferred the correct answer. When SBERT scores were relatively low, the

Analysis of Example Input Results

[Shuffled Image Example Input]



[Random Shuffle Task Accuracy For The Example Input]

	GPT-4o	Gemini-1.5-Pro	Claude-3.5-Sonnet
Shuffle Task Acc.(%)	50.00%	50.00%	50.00%

[SBERT Scores (Cosine-Similarity) For The Example Input]

	GPT-4o	Gemini-1.5-Pro	Claude-3.5-Sonnet
Image 1	0.666	0.742	0.519
Image 2	0.663	0.615	0.428
Image 3	0.597	0.572	0.511
Image 4	0.837	0.688	0.786

[Description For Image 3]

Ground Truth : The dog is dreaming about being scolded by its owner, while the man continues to sit in the chair reading the newspaper.

GPT-4o : The dog dreams of being punished and looks remorseful.

Gemini-1.5-Pro : The man continues to sleep, dreaming of his dog fetching a ball.

Claude-3.5-Sonnet : The person imagines the dog chasing them while it continues to sleep.

Figure 4: Cosine similarity scores were used to evaluate FM-generated descriptions. Low SBERT scores often reflected inaccuracies, such as misidentifying character counts or using incorrect verbs (red), likely contributing to errors in image ordering.

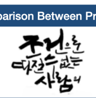
models tended to generate inappropriate words, as illustrated in Figure 4 (with improper words highlighted in red). In *Image 3*, FMs used inaccurate verbs, nouns, and descriptions, leading to poor inference. GPT-4o, while successfully describing the dog's status, produced insufficient information by omitting any mention of the man. Gemini-1.5-Pro generated an inaccurate sentence, misidentifying the characters' statuses and incorrectly interpreting their actions. For example, it depicted the man as sleeping, confusing his status with the dog's, while he was actually reading the newspaper. Similarly, Claude-3.5-Sonnet accurately identified the characters but produced an incorrect scene interpretation for *Image3*.

4.3 Calligraphy

Quantitative Results As shown in the Table 4, the overall results indicate that FMs performed inadequately on the OCR task. Among the different prompt styles, the ORCoT and ORCoT+Few-Shot approaches outperformed the Zero-Shot approach, although the difference between the two ORCoT-based methods was negligible. Of the three models, GPT-4o achieved the best performance, mainly because of its improved ability to detect spacing (' ') more accurately than the other models.

Qualitative Results Due to the unique characteristics of calligraphy, the dataset occasionally includes abbreviated or non-standard forms, such as 'spring day' written as 'spring d.' In such cases, FMs often misinterpret 'd' as a separate element, recognizing only 'spring.' This issue was more prevalent with ORCoT and ORCoT+Few-Shot prompts compared to Zero-Shot. In Zero-Shot scenarios, OCR tasks focus on individual word appearances, leading to frequent typographical errors, as demonstrated by Gemini-1.5-Pro's Zero-Shot result in Figure 5. In contrast, ORCoT and ORCoT+Few-Shot approaches prioritize the overall meaning, generating contextually appropriate outputs even when deviating from the ground truth. For instance, in Figure 5, the calligraphy translates to "A person who cannot be judged by conditions" (ground truth: '조건으로 따질 수 없는 사람의'). In the Zero-Shot setting, GPT-4o and Claude-3.5-Sonnet produced structurally similar but semantically incorrect sentences, while Gemini-1.5-Pro recognized isolated words. Using ORCoT and ORCoT+Few-Shot, all FMs interpreted the input as sentences. Although none perfectly matched the ground truth, they conveyed partial meaning. For example, in ORCoT, Claude-3.5-Sonnet generated '주거를 다시 시작하는 사람의,' meaning "A person starting to live in a residence again," which, while coherent, deviates from the intended meaning. In ORCoT+Few-Shot, Claude-3.5-Sonnet produced '주어로 다짐하는 사람의,' which lacks cohesiveness but includes meaningful tokens like '주어로' ("as the subject"), '다짐하는' ("making a resolution"), and '사람의' ("person's").

Comparison Between Prompts



Base Models	GPT-4o	Gemini-1.5-Pro	Claude-3.5-Sonnet
Ground Truth	'조건으로 따질수없는 사람의'	'조건으로 따질수없는 사람의'	'조건으로 따질수없는 사람의'
Zero-Shot	'주견으로 타협할 수 없는 사람의'	'저,쓴,대,쯘,없,는,사,람'	'주거의 대전을 만드는 사람들'
ORCoT	'주견으로 타협할 수 없는 사람의'	'저 사랑 대전 없는 사람'	'주거를 다시 시작하는 사람의'
ORCoT+Few-Shot	'주견으로 다정스없는 사람의'	'죄는 대전으로 사람의'	'주어로 다짐하는 사람의'

Figure 5: In the Zero-Shot approach, OCR emphasizes individual word appearance, as seen with Gemini-1.5-Pro. Other methods prioritize overall meaning, producing contextually or semantically similar words even when deviating from the ground truth.

4.4 Onion, Not The Onion

Quantitative Results Overall, the models demonstrates notable behavior in fake news detection, as shown in Table 5. All FMs illustrate major increase of performance from Zero-Shot to ORCoT+Few-Shot. However, Claude-3.5-Sonnet exhibits relatively lower performance across all three prompt styles. Unlike other FMs, it shows comparable capabilities in both the Zero-Shot and ORCoT.

Qualitative Results We observed that FMs exhibit

Additional Experiment For Onion, Not The Onion

[GPT-4o]

Step 5. **Plausibility**: The idea that a single surname would take over an entire country's population in 500 years is implausible based on current demographic trends. Even though it references a law on surnames in Japan, the conclusions drawn from this are unrealistic. (0) — GT (1)

[Claude-3.5-Sonnet]

Step 5. **Plausibility**: The idea that everyone in Japan would have the same surname by 2531 seems highly implausible and exaggerated. Given these factors, especially the mention of April Fool's and the highly improbable prediction, I believe this article is likely fake. April Fool's joke. (0) — GT (1)

[Gemini-1.5-Pro]

Step 5. **Plausibility**: The claim that everyone in Japan will have the same surname within a few centuries is highly implausible. It's statistically improbable for a single surname to completely dominate an entire population within such a timeframe. The lack of a credible source, sensational language, difficulty verifying the study's accuracy, and the absence of corroborating reports from other reliable sources — strongly suggests that the article is fake. (0) — GT (1)

Figure 6: This example is sourced from *Not The Onion*. All FMs misclassified the story as fake, basing their decision on the implausible claim that "within the next 500 years, nearly all people in Japan will share the same surname." This unlikely statement led the models to conclude that the news was false.

	Model	Zero-Shot	ORCoT	ORCoT+Few-Shot
Acc.(%)(↑)	Claude-3.5-Sonnet	14.42	26.40	32.20
	Gemini-1.5-Pro	17.55	18.50	20.20
	GPT-4o	**53.43**	**61.54**	**61.86**
WER(%)(↓)	Claude-3.5-Sonnet	81.39	74.00	65.62
	Gemini-1.5-Pro	**90.52**	**89.55**	**88.45**
	GPT-4o	64.41	45.81	45.39
CER(%)(↓)	Claude-3.5-Sonnet	**77.65**	**77.21**	**76.15**
	Gemini-1.5-Pro	74.04	71.85	69.55
	GPT-4o	32.64	24.73	22.55

Table 4: The results (%) of the Korean Calligraphy OCR task indicate that the overall OCR capabilities of FMs are limited. GPT-4o exhibited superior performance, largely due to its enhanced ability to accurately detect spacing (' ') compared to other models.

lower performance on relatively short articles. As shown in Table 6, accuracy tends to drop as article length decreases. For a more detailed analysis, we divided the dataset into five sections based on article length, with Q1 representing the shortest and Q5 the longest articles. The *Onion* group, primarily consisting of fake news articles, tends to feature shorter articles while maintaining consistently high accuracy across the dataset. This trend suggests that FMs may be more inclined to classify shorter articles as fake news, indicating that *Not The Onion* articles pose greater challenges for fake news classification. We analyzed FMs' decision-making rationale to assess if they follow proper reasoning steps with shorter articles, aiming to understand when and why they make incorrect conclusions. In this analysis, we observed that while FMs generally take plausible steps, they encounter difficulties with exceptional cases, as highlighted in Figure 6 (highlighted in red, all models received the same input for evaluation). The example is a seemingly fake but real news story sourced from *Not The Onion*. The article discusses the possibility that, in the distant future, most people in Japan will have the surname 'Sato'. All FMs incorrectly classified the news as fake, basing their judgment on the unbelievable claim that "within the next 500 years, nearly all people in Japan will share the same surname." This implausible fact led them to categorize the news as fake.

	Model	Zero-Shot	ORCoT	ORCoT+Few-Shot
Acc.(%)	Claude-3.5-Sonnet	71.00	69.58	85.93
	Gemini-1.5-Pro	**83.97**	87.81	91.91
	GPT-4o	80.70	**89.88**	**94.74**
Precision(%)	Claude-3.5-Sonnet	67.21	70.21	88.85
	Gemini-1.5-Pro	**83.78**	92.35	**97.82**
	GPT-4o	83.17	**94.75**	97.35
Recall(%)	Claude-3.5-Sonnet	68.68	57.27	78.42
	Gemini-1.5-Pro	**79.96**	79.94	84.18
	GPT-4o	76.04	**84.03**	**91.76**

Table 5: Results (%) of the news classification task. Overall, the models exhibit strong performance in detecting fake news. However, Claude-3.5-Sonnet performs comparatively worse across all three prompt styles. Unlike other FMs, it demonstrates similar effectiveness in both Zero-Shot and ORCoT prompting styles.

Length of Article		Model	Q1	Q2	Q3	Q4	Q5
Not The Onion	Acc.(%)	Claude-3.5-Sonnet	***69.30***	80.59	73.63	79.60	89.05
		Gemini-1.5-Pro	***67.35***	80.20	89.58	92.18	91.66
		GPT-4o	***84.23***	90.42	91.25	96.25	96.68
Onion	Acc.(%)	Claude-3.5-Sonnet	78.22	83.87	98.79	99.59	99.59
		Gemini-1.5-Pro	92.98	99.12	100.00	100.00	100.00
		GPT-4o	91.20	96.80	100.00	100.00	100.00

Table 6: Accuracy tended to decline with shorter articles. To investigate this trend in more detail, we divided the dataset into five sections based on article length, with Q1 representing the shortest and Q5 the longest articles. In *Not The Onion*, FMs often misclassified real short articles as fake due to the common association between shorter length and fake news. This tendency led to a noticeable drop in performance. Despite the task being a binary classification, Claude-3.5-Sonnet achieved only 69.30% accuracy.

4.5 Lyrics

Quantitative Results In the infilling task, as shown in Table 7, FMs demonstrated poor performance. It is clear that the FMs struggle to predict the masked portions of lyrics classified as exceptional cases by BERT. Notably, Claude-3.5-Sonnet refused to respond due to safety concerns related to copyright

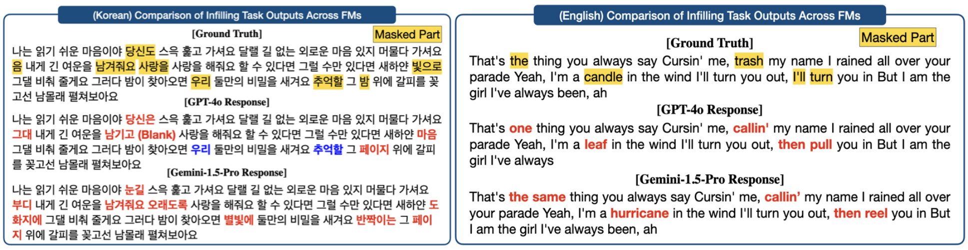


Figure 7: Examples from the infilling task. In the Korean dataset (left), GPT-4o correctly predicts some masked sections (blue) but misses others (red), while Gemini-1.5-Pro generates plausible yet often inaccurate interpretations (red). In the English dataset (right), both FMs frequently produce suboptimal responses, relying on common expressions (red). Notably, Gemini-1.5-Pro occasionally outputs two tokens for a single masked section (e.g., predicting "the same" for "the"), compounding errors.

issues, making it impossible to include in the evaluation. Additionally, within the Korean dataset, we observed a significant performance degradation in Gemini-1.5-Pro compared to GPT-4o. In the English description generation task, the overall scores were low, as shown in Table 8. In the genre classification task, all models demonstrated stronger performance in Korean than in English in the After Cut-Off data, as shown in Table 9.

Infilling Result		Baseline Model	Zero-shot	ORCoT	ORCoT+Few-shot
English	BERT Score(F1)	Gemini-1.5-Pro	**0.613**	0.616	0.643
		GPT-4o	0.611	**0.632**	**0.653**
Korean	BERT Score(F1)	Gemini-1.5-Pro	0.032	0.155	0.324
		GPT-4o	**0.398**	**0.447**	**0.463**

Table 7: The poor performance in the lyrics infilling task suggests that FMs struggle to predict tokens involving irregular and complex sentence structures, as well as uncommon words. In the Korean dataset, there is a significant performance drop in both FMs, further supporting the observation that FMs face greater challenges with Korean songs compared to English ones. Claude-3.5-Sonnet, however, declined to participate in the evaluation due to safety concerns regarding copyright issues, thus excluding it from the analysis.

Qualitative Results In the infilling task, both FMs underperform on the English dataset, often filling masked sections with common phrases, leading to suboptimal results. For example, Gemini-1.5-Pro generated *the same* instead of the masked token *the*, occasionally producing two tokens for a single mask (see Figure 7). On the Korean dataset, FMs also show subpar inference performance. While GPT-4o successfully predicts some masked sections, it sometimes overlooks others entirely. Similarly, Gemini-1.5-Pro provides contextually plausible interpretations but deviates significantly from the ground truth. In the description generation task, FMs misinterpret song lyrics, producing responses that summarize or repeat the input rather than offering deeper insights. In the genre classification task, the disparity in unique genres—11 in Korean and 58 in English—makes classification more challenging for the English dataset.

Description Generation Result		Baseline Model	Zero-Shot	ORCoT	ORCoT+Few-Shot
Before Cut-Off	ROUGE-L (F1)	Claude-3.5-Sonnet	0.107	0.135	0.157
		Gemini-1.5-Pro	0.100	0.113	0.102
		GPT-4o	0.106	0.158	0.161
	BERT Score (F1)	Claude-3.5-Sonnet	0.049	0.041	0.071
		Gemini-1.5-Pro	-0.049	-0.039	-0.055
		GPT-4o	0.057	0.080	0.084
After Cut-Off	ROUGE-L (F1)	Claude-3.5-Sonnet	0.105	0.129	0.149
		Gemini-1.5-Pro	0.104	0.119	0.112
		GPT-4o	0.113	0.162	0.163
	BERT Score (F1)	Claude-3.5-Sonnet	-0.052	-0.066	-0.048
		Gemini-1.5-Pro	-0.066	-0.049	-0.058
		GPT-4o	0.098	0.115	0.118

Table 8: In the song description generation task, low scores show FMs struggle to interpret lyrics. No significant differences in were found between "Before" and "After" Cut-Off (Training cut-off); datasets. The Korean task was not performed, as descriptions are only given for album concepts, not individual songs.

Genre Classification Result			Baseline Model	Zero-Shot	ORCoT	ORCoT+Few-Shot
English	Before Cut-Off	Overlap Ratio	Claude-3.5-Sonnet	**0.660**	**0.690**	**0.703**
			Gemini-1.5-Pro	0.214	0.218	0.306
			GPT-4o	0.594	0.610	0.620
	After Cut-Off	Overlap Ratio	Claude-3.5-Sonnet	**0.538**	**0.580**	**0.590**
			Gemini-1.5-Pro	0.405	0.429	0.550
			GPT-4o	0.474	0.497	0.509
Korean	Before Cut-Off	Overlap Ratio	Claude-3.5-Sonnet	**0.695**	**0.727**	**0.766**
			Gemini-1.5-Pro	0.619	0.581	0.609
			GPT-4o	0.642	0.665	0.733
	After Cut-Off	Overlap Ratio	Claude-3.5-Sonnet	**0.692**	**0.708**	**0.754**
			Gemini-1.5-Pro	0.503	0.665	0.673
			GPT-4o	0.668	0.690	0.750

Table 9: In the genre classification task, Claude-3.5-Sonnet achieved the highest performance, followed by GPT-4o and Gemini-1.5-Pro. After Cutoff, all models performed better on Korean data than English, with Claude and GPT showing significant differences between "Before" and "After" Cutoff (Training cut-off) results in the English dataset.

5 Conclusion

This paper introduces a benchmark for evaluating FMs on OOD reasoning tasks. Our comprehensive evaluation across text-only, image-only, and multimodal tasks reveals a key failure mode: models over-rely on learned statistical priors, causing initial perceptual errors that shatter the entire reasoning chain. We demonstrate that our proposed ORCoT prompting strategy mitigates this by enforcing a causal analysis that prioritizes the data's internal logic, validating our approach against Zero-Shot and ORCoT+Few-Shot techniques.

6 Limitation

This paper pioneers research into exceptional cases, examining how FMs handle scenarios where they typically underperform, thereby advancing toward robust reasoning. We introduce benchmark datasets across multiple modalities and tasks, focusing on instance classification, ordering, text infilling, and short-text generation. However, our benchmark does not cover more complex tasks such as long-form continuation, dialogue grounding, or interactive reasoning, which we leave as future work. AI tools were used only to assist with language editing; all scientific content was developed by the authors.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

anthropic. 2024. Claude 3.5 sonnet model card addendum. https://www-cdn.anthropic.com/fed9cc193a14b84131812372d8d5857f8f304c52/Model_Card_Claude_3_Addendum.pdf.

Daman Arora, Himanshu Gaurav Singh, and Mausam. 2023. Have llms advanced enough? a challenging problem solving benchmark for large language models. *Preprint*, arXiv:2305.15074.

Dibyanayan Bandyopadhyay, Soham Bhattacharjee, and Asif Ekbal. 2025. Thinking machines: A survey of llm based reasoning strategies. *Preprint*, arXiv:2503.10814.

Manash Pratim Barman, Amit Awekar, and Sambhav Kothari. 2019. Decoding the style and bias of song lyrics. *Preprint*, arXiv:1907.07818.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 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*.

Anoop Cherian, Kuan-Chuan Peng, Suhas Lohit, Kevin A. Smith, and Joshua B. Tenenbaum. 2023. Are deep neural networks smarter than second graders? *Preprint*, arXiv:2212.09993.

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*.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. *Preprint*, arXiv:2010.11929.

Dirk Groeneveld, Iz Beltagy, Pete Walsh, Akshita Bhagia, Rodney Kinney, Oyvind Tafjord, Ananya Harsh Jha, Hamish Ivison, Ian Magnusson, Yizhong Wang, Shane Arora, David Atkinson, Russell Authur, Khyathi Raghavi Chandu, Arman Cohan, Jennifer Dumas, Yanai Elazar, Yuling Gu, Jack Hessel, and 24 others. 2024. Olmo: Accelerating the science of language models. *Preprint*, arXiv:2402.00838.

Dan Hendrycks and Kevin Gimpel. 2016. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136*.

Adit Jamdar, Jessica Abraham, Karishma Khanna, and Rahul Dubey. 2015. Emotion analysis of songs based on lyrical and audio features. *International Journal of Artificial Intelligence amp; Applications*, 6(3):35–50.

Abhinav Joshi, Areeb Ahmad, Divyaksh Shukla, and Ashutosh Modi. 2025. Towards quantifying commonsense reasoning with mechanistic insights. *Preprint*, arXiv:2504.10077.

Katie Kang, Amrith Setlur, Claire Tomlin, and Sergey Levine. 2023. Deep neural networks tend to extrapolate predictably. *arXiv preprint arXiv:2310.00873*.

Harshal Kausadikar, Tanvi Kale, Onkar Susladkar, and Sparsh Mittal. 2025. Historic scripts to modern vision: A novel dataset and a vlm framework for transliteration of modi script to devanagari. *Preprint*, arXiv:2503.13060.

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.

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

Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. 2020. Energy-based out-of-distribution detection. *Advances in neural information processing systems*, 33:21464–21475.

Yujie Lu, Dongfu Jiang, Wenhu Chen, William Yang Wang, Yejin Choi, and Bill Yuchen Lin. 2024. Wildvision: Evaluating vision-language models in the wild with human preferences. *arXiv preprint arXiv:2406.11069*.

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.

Genki Osada, Tsubasa Takahashi, Budrul Ahsan, and Takashi Nishide. 2023. Out-of-distribution detection with reconstruction error and typicality-based penalty. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5551–5563.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2023. Exploring the limits of transfer learning with a unified text-to-text transformer. *Preprint*, arXiv:1910.10683.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Sebastian Ruder, Jonathan Clark, Alexander Gutkin, Mihir Kale, Min Ma, Massimo Nicosia, Shruti Rijhwani, Parker Riley, Jean-Michel Sarr, Xinyi Wang, John Wieting, Nitish Gupta, Anna Katanova, Christo Kirov, Dana Dickinson, Brian Roark, Bidisha Samanta, Connie Tao, David Adelani, and 8 others. 2023. Xtreme-up: A user-centric scarce-data benchmark for under-represented languages. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, page 1856–1884. Association for Computational Linguistics.

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. *Preprint*, arXiv:2310.16809.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, and 1 others. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Alexandros Tsaptsinos. 2017. Lyrics-based music genre classification using a hierarchical attention network. *Preprint*, arXiv:1707.04678.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. *Advances in neural information processing systems*, 32.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 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*.

Kai Xiong, Xiao Ding, Yixin Cao, Yuxiong Yan, Li Du, Yufei Zhang, Jinglong Gao, Jiaqian Liu, Bing Qin, and Ting Liu. 2025. Com^2: A causal-guided benchmark for exploring complex commonsense reasoning in large language models. *Preprint*, arXiv:2506.07064.

Jingkang Yang, Kaiyang Zhou, Yixuan Li, and Ziwei Liu. 2024. Generalized out-of-distribution detection: A survey. *International Journal of Computer Vision*, pages 1–28.

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*.

Qinkai Zheng, Xiao Xia, Xu Zou, Yuxiao Dong, Shan Wang, Yufei Xue, Lei Shen, Zihan Wang, Andi Wang, Yang Li, and 1 others. 2023. Codegeex: A pre-trained model for code generation with multilingual benchmarking on humaneval-x. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 5673–5684.

Supplementary Material for Benchmarking Foundation Models on Exceptional Cases: Dataset Creation and Validation

A Experimental Setting

It has been observed that in certain atypical scenarios(Chen et al., 2024), FMs exhibit errors in reasoning concerning the sequence of events. We designed experiments using four different datasets that feature various characters with multiple types of tasks such as instance recognition, text generation, token prediction, and character recognition, as described in Figure 2. In the experiments for all four datasets, we conducted all experimental tasks using GPT-4o, Gemini-1.5-pro, and Claude-3.5-Sonnet (Achiam et al., 2023; Team et al., 2023; anthropic, 2024). We also propose ORCoT, a revised version of CoT (Wei et al., 2022), to achieve enhanced performance. We employed three prompt styles—Zero-Shot (Kojima et al., 2022), ORCoT, and ORCoT+Few-Shot (Brown et al., 2020)—to investigate how the accuracy of responses varies. The API temperature setting is regulated to 0 in GPT-4o, Claude-3.5-Sonnet, and 0.01 in Gemini-1.5-pro to ensure consistent results.

B Graphic Novels

B.1 Experiments on Visual Storytelling Images

Figure 8: Five randomly shuffled panels of real-world photos depicting a short, real-life story.

In this task, the model was provided with shuffled real-life story images (Figure 8) and instructed to reconstruct the original sequence. While the model successfully generated accurate descriptions for each scene, it failed to determine the correct order. As shown in Table 10, the resulting sequence was often plausible and internally coherent, but did not match the ground truth.

Image Descriptions for ID Images
Image 1: A blue-and-white soccer ball is suspended mid-air above the roof's edge. Below it lie a black frisbee and a decorative golden plate. The clear sky suggests daylight. This moment captures an object seemingly falling or being dislodged from the roof it signals that some retrieval action is underway.
Image 2: A person stands barefoot in front of a house with a red door, arms extended upward. Above them, a black-and-white soccer ball is descending. The person appears to be attempting a catch. This scene reflects a mid-action moment where the object freed from the roof is about to be retrieved.
Image 3: A close-up of the roof shows only the black frisbee remaining. The other items—blue-and-white ball, black-and-white ball, and golden plate—are no longer present. This indicates that the retrieval was successful and we're seeing the post-action state.
Image 4: The same person is now walking on the lawn, barefoot, next to the black-and-white soccer ball. They appear relaxed and ready to play, with a red truck and garage in the background. This signals the end of the sequence: the ball is recovered and back in use.
Image 5: Both soccer balls (black-and-white and blue-and-white), the frisbee, and the golden plate are all stationary on the roof. This is a clear setup image, establishing the initial conflict where multiple items are stuck and need to be retrieved.
Correct Order: [5, 1, 2, 3, 4]

Table 10: Detailed visual descriptions for the ID story ordering task. The model's predicted sequence, [5, 1, 2, 3, 4], deviates from the ground truth order of [1, 2, 3, 4, 5]. Notably, while the predicted sequence is incorrect, **it represents a plausible and internally coherent narrative**, highlighting a failure in identifying the specific ground truth rather than a general inability to construct a logical story.

B.2 Task Details

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 storylines. These present vividly exaggerated storylines that are seldom encountered by FMs.

Data Details: 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 comprehend the storyline. To ensure an accurate assessment, we eliminate all clues that provide information about the storyline, including panel numbers and titles of the graphic novel as shown in Figure 9.

Experiments Details: The API temperature setting is adjusted to 0.01 for Gemini-1.5-Pro and 0 for GPT-4 to ensure consistent results. To generate a concise answer, the model is instructed to output the response solely in the format [1,2,3,4], as shown in the blue text in Figure 10 ('Prompt'). We set the ground truth order as [1,3,2,4] to automate the task, given that the input images are shuffled, as shown in (e) in Figure 10 ('In the code'). This predetermined order allows us to verify whether FMs produce the correct sequence. Additionally, we demonstrate how the prompts were designed for each style in E.1. Table 16. We design the random shuffle experiment as follows.

1. Inform the FMs that the uploaded images represent parts of a story that have been shuffled and

consist of four images as shown in the blue letters in Figure 10 ('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 as shown in (a), (b) in Figure 10 ('In the code').
3. The uploaded images are indexed, and the FMs infers the correct order, subsequently outputting the images in the proper indexed sequence as shown in (c) in Figure 10 ('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 FMs as shown in (d) in Figure 10 ('In the code').
5. Compare the predicted image order with the ground truth order to determine accuracy as shown in (e) in Figure 10 ('In the code').

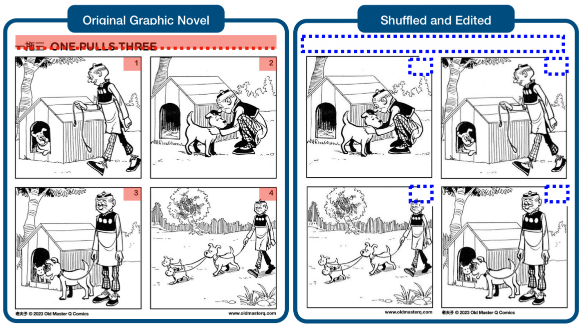


Figure 9: We remove clue-containing sections marked by red boxes that help determine the correct storyline. These sections were removed as shown by the blue dotted line boxes in the 'Shuffled and Edited' version. The original Graphic Novel is shown on the left, with clue-containing sections marked by red boxes—these sections provide hints for determining the correct storyline. On the right, the 'Shuffled and Edited' version displays the result after removing these sections, as indicated by the blue dotted line boxes.

B.3 Task Result

We assessed the multimodal causal reasoning abilities of FMs through a Random Shuffle task. 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 11. The highest performance was observed in the ORCoT+Few-Shot condition, followed by ORCoT and then Zero-Shot. Interestingly, the Zero-Shot performance exceeded expectations, displaying an accuracy that was not markedly lower than the other prompting styles. 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 ORCoT+Few-Shot, the number of Few-Shot examples impacted performance; with only one example, there was no difference compared to ORCoT, but increasing the examples to three resulted in a noticeable performance improvement.

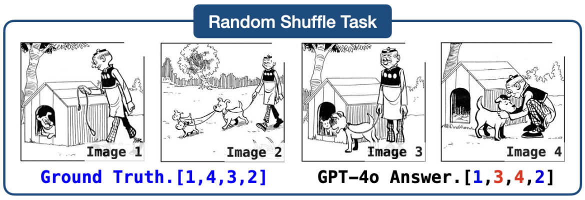


Figure 11: Example of the random shuffle task. The original sequence is [1, 4, 3, 2], but GPT-4o produces an incorrect result.

When the inferred order is completely correct: FMs occasionally makes mistakes in scene descriptions, even when it derives correct answers. For example, in Figure 12, GPT-4o 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.'
When the inferred order is completely incorrect: FMs sometimes misidentify objects or misunderstand emotions. For instance, GPT-4o describes a man pulling a tiger's tail instead of removing an arrow from its paw, refer to image 2 of Figure 13.

C Calligraphy

C.1 Task Details

Data Details: We preprocessed the dataset according to three rules. First, we deleted images if their resolution was too low or if they contained too many letters that even a human could not recognize. We set the threshold at 35 characters, as shown in Figure 14, where 35 is an irregularly large number in the dataset. We observed that images with more than 35 characters are visually challenging for humans to recognize, so we excluded such images from evaluation. Second, we separated overlapping calligraphy in an image by applying bounding boxes provided by the OCR API. Third, we cropped out typographic elements such as signs and watermarks that were deemed irrelevant to the

Prompt

```
Q."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]"
A.[1,4,2,3]
```

In the code

```
(a) Shuffled Order : [ Image_1, Image_4, Image_3, Image_2 ]
(b) Index          : [    1   ,    2   ,    3   ,    4    ]
(c)(Index Order)GPT-4o Inference Result : [1,4,2,3](From API response)
(d)(Image Order)GPT-4o Inference Result :  [Image_1, Image_2, Image_3, Image_4]
(e)(Image Order)The Ground Truth Order :   [Image_1, Image_3, Image_2, Image_4]
                                                      ✓ Accuracy : 2/4 = 0.5
```

Figure 10: 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.

Completely Correct Order (Acc. 1.0)

[Shuffled Image Input]

Image 1 Image 2 Image 3 Image 4

[GPT-4o Output with Descriptions]

"Let's analyze the images step by step to determine the correct order:
1. The first image shows the man approaching the doghouse with a leash.
2. The fourth image shows the man kneeling and petting the dog, likely coaxing it out of the doghouse.
3. The third image shows the man with the dog on a leash, standing outside the doghouse.
4. The second image shows the man walking the dog and its puppies.

So, the correct order is : Image 1, Image 4, Image 3, Image 2"

[Ground Truth : Image 1, Image 4, Image 3, Image 2] [GPT-4o Output : Image 1, Image 4, Image 3, Image 2]

Figure 12: Correct Order Check: This example shows that while GPT-4o can correctly order the images, it sometimes lacks in scene description such as using mismatched verbs (highlighted in red).

Completely Incorrect Order (Acc. 0.0)

[Shuffled Image Input]

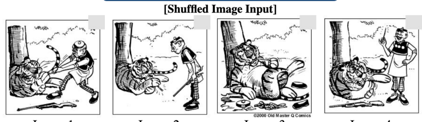


Image 1 Image 2 Image 3 Image 4

[GPT-4o Output with Descriptions]

To determine the correct order of the images, let's analyze the sequence of events:
The tiger is sitting against the tree, looking relaxed and content. (Image 3)
The man approaches the tiger with a gun, and the tiger looks startled. (Image 2)
The man and the tiger are in a struggle, with the man pulling the tiger's tail. (Image 1)
The man stands confidently with the tiger looking defeated, and the gun is on the ground. (Image 4)
So, the correct order is: Image 3, Image 2, Image 1, Image 4

[Ground Truth : Image 2, Image 1, Image 4, Image 3] [GPT-4o Output : Image 3, Image 2, Image 1, Image 4]

Figure 13: Incorrect Order Check: In three of four images, GPT-4o provided incorrect character descriptions and showed poor object recognition (highlighted in red).

calligraphy. An example of the preprocessed Korean calligraphy is shown in Figure 15. 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.

Experiments Details: The API temperature setting is adjusted to 0.01 for Gemini-1.5-Pro and 0 for GPT-4 to ensure consistent results. Before word-level evaluation, we removed punctuation and special symbols from FM predictions and replaced '\n' with ' ' due to ambiguous line breaks in the calligraphy. We used Word-level Accuracy, CER, and WER, which are representative OCR metrics.

C.2 Task Result

The artistic nature of calligraphy sometimes leads to unconventional representations in the dataset, such as abbreviating 'spring day' to 'spring d.' In these cases, FMs tend to process 'd' as a separate element rather than part of the word, recognizing only 'spring.' This tendency was more pronounced in the ORCoT and ORCoT+Few-Shot prompts compared to Zero-Shot. In the Zero-Shot scenario, the OCR task tends to prioritize the visual recognition of individual words over the holistic meaning conveyed by the calligraphy, resulting in a higher frequency of typographical errors. Conversely, the ORCoT and ORCoT+Few-Shot approaches first interpret the overall meaning and then perform OCR based on contextually relevant words. Consequently, even when the output deviates from the ground truth, it tends to generate semantically similar words or words that are more contextually fitting than the ground truth. As illustrated in Figure 16, the first calligraphy example signifies 'pray,' with the ground truth being '기도.' In the Zero-Shot scenario, GPT-4o recognizes it as '기드,' which bears a close visual resemblance but lacks semantic meaning. The ORCoT approach interprets it as '기다,' which, although not aligning with the ground truth, at least carries the meaning 'to crawl.' Notably, the ORCoT+Few-Shot approach accurately identifies it as '기도,' precisely matching the ground truth.

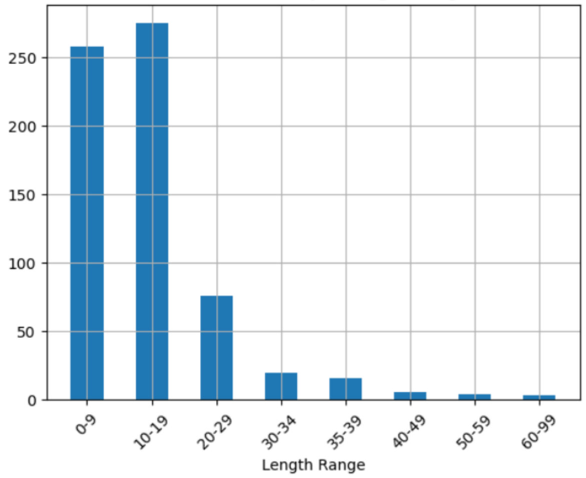


Figure 14: Length plot of Korean calligraphy images. We determined that images with over 35 characters presented considerable visual recognition difficulties, even for humans, prompting their exclusion from our evaluation.

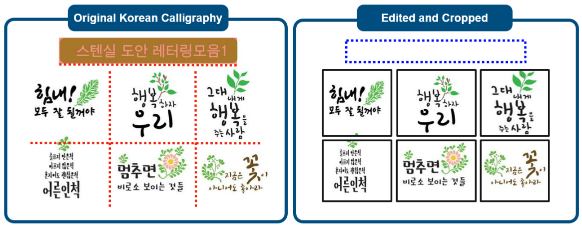


Figure 15: Example of preprocessed Korean calligraphy. We removed typographic elements unrelated to the calligraphy and automatically cropped overlapping sections using bounding boxes detected by the OCR API.

This figure illustrates the preprocessing of Korean calligraphy. On the left, the original calligraphy is shown with typographic elements unrelated to the calligraphy removed. On the right, the edited and cropped version is displayed, processed automatically using bounding boxes detected by the OCR API.

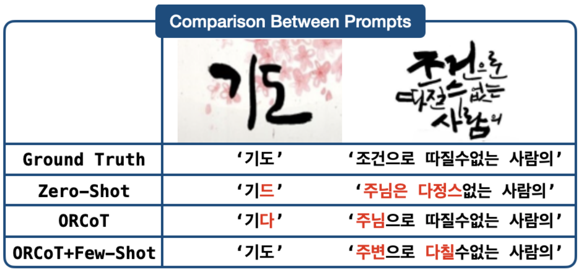


Figure 16: Examples of comparisons of OCR task results between prompts on Korean calligraphy data in GPT-4o.

D Onion, Not The Onion

D.1 Task Details

Data Details: We performed web scraping on The Onion website and Reddit's Not The Onion section. Following data collection, we implemented an additional filtering process using Python scripts to enhance the dataset's sophistication. Specifically, we automated the removal of instances where no content was collected, where content was duplicated, and where advertisements were included. During preprocessing, we encountered valid data with varying lengths, both long and short, that were indeed written by humans. These instances represent qualitative news articles, so we chose not to remove them to preserve the dataset's integrity. As a result, the mean and median text lengths are 2243 and 1433, respectively, leading to a left-skewed distribution. A histogram illustrating text lengths and category-specific statistics is presented in Figure 17. Through this process, we ensured that only the title and content of the original news articles influenced the FMs' judgment during fake news detection. This approach provided a reliable dataset, allowing us to evaluate the impact of textual data alone in fake news detection research. It contains 1,249 fake and 1,202 genuine articles from Jan 2021 to May 2024.

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 ORCoT 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 Few-shot and ORCoT 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 18. By distinguishing between fake news and real news, we contribute to preventing social disruption and maintaining the credibility of information.

D.2 Task Result

Overall, FMs exhibit high performance on Onion, Not The Onion dataset as shown in Table 11, but we observed a reduction in performance with relatively short articles. As shown in Table 5, accuracy differ-

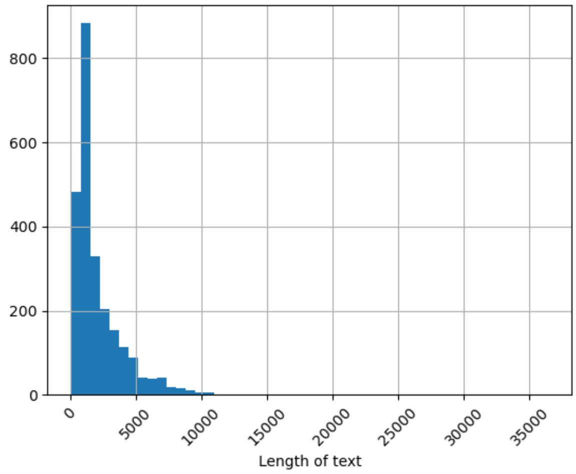


Figure 17: Length plot of the preprocessed Onion and Not the Onion news data.

Table 11: Comparison of performance metrics between Gemini-1.5-Pro and GPT-4o across different settings.

Metric		Model	Zero-Shot	ORCoT	ORCoT+Few-Shot
Acc.		Gemini-1.5-Pro	83.97	87.81	91.91
		GPT-4o	80.70	89.88	94.74
Onion	Precision	Gemini-1.5-Pro	84.12	84.84	88.07
		GPT-4o	78.70	86.14	92.49
	Recall	Gemini-1.5-Pro	87.27	94.43	98.42
		GPT-4o	85.19	95.52	97.60
	F1-score	Gemini-1.5-Pro	85.67	89.38	92.96
		GPT-4o	81.81	90.58	94.97
Not The Onion	Precision	Gemini-1.5-Pro	83.78	92.35	97.82
		GPT-4o	83.17	94.75	97.35
	Recall	Gemini-1.5-Pro	79.96	79.94	84.18
		GPT-4o	76.04	84.03	91.76
	F1-score	Gemini-1.5-Pro	81.82	85.70	90.49
		GPT-4o	79.44	89.07	94.48

ences based on article length reveal that accuracy generally improves as article length increases. In contrast, the Onion group, predominantly consisting of fake news articles, typically features shorter articles and maintains consistently high accuracy across the dataset. This pattern suggests that FMs may have a tendency to classify shorter articles as fake news, highlighting the greater challenge posed by Not The Onion in fake news classification. Furthermore, we delve deeper into the rationale behind FMs's decision-making process, particularly when encountering relatively short articles, to better understand the circumstances under which FMs arrives at incorrect conclusions and whether it follows appropriate steps in such cases. In this approach, we observe that FMs generally takes appropriate steps, many of which are plausible. However, it is notable that FMs encounters difficulties with exceptional cases, as highlighted in Figure 18(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.

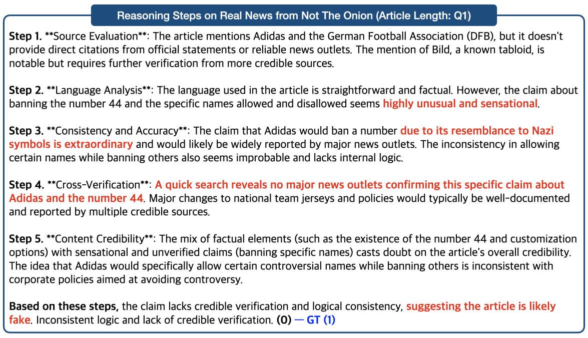


Figure 18: An example where GPT-4o, despite following appropriate reasoning steps, produces an incorrect reasoning outcome due to an exceptional case(highlighted in red).

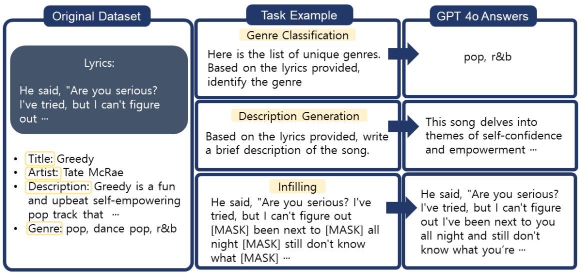


Figure 19: Overview of the lyrics dataset: an example of three different tasks and GPT-4o's responses.

E Lyrics

E.1 Task Details

Data Details: 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. The English dataset was sourced from the Billboard Year-End Chart (1990–2023) for the Before Cut-Off period and entries from January to April 2024 for the After Cut-Off period, reflecting the FMs training cutoff date. Similarly, the Korean dataset was sourced from Melon. After preprocessing, the dataset contained 1,811 English and 2,363 Korean entries. We assess FMs' comprehension of song lyrics through three tasks: genre detection, song description generation, and infilling as shown in Figure 19. For the infilling task, we used a pre-trained BERT model to anticipate the masked parts and removed non-exceptional data. Entries with BERT scores exceeding a 0.9 threshold were excluded, as high semantic similarity indicated non-exceptional content. When collecting the dataset, we divided it into two parts: 'yearly' and 'weekly.' The yearly dataset comprises data from before the FMs cut-off date (before the end of 2023), while the weekly dataset includes data from after the cut-off date (after the end of 2023). For the English dataset, after collecting the title and artist of each song, we removed duplicate entries—only removing a song if both the title and artist were identical, as different songs can share the same title. We then generated links to the Genius site to obtain the lyrics and descriptions of the songs. This process involved removing strings following 'featuring' and modifying characters such as brackets and Latin alphabets. If it was impossible to retrieve any of the descriptions, genre, or lyrics due to link generation errors or unavailability on the site, we excluded the song. Additionally, songs with non-English lyrics were also removed. To ensure that the weekly dataset contained only data that the FMs had not previously encountered, any song appearing in both the weekly and yearly data was excluded from the weekly dataset. For the genre detection task in English, we streamlined the genre list by removing infrequent genres. After consolidating all genre lists, we excluded genres with fewer than 10 occurrences, resulting in a final list of 58 unique 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 the dataset contained fewer genre categories. Notably, no songs were excluded during the crawling of lyrics, descriptions, or genres in the Korean dataset, as all song information was sourced 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 12.

Experiments Details: We employed several metrics for precise testing, including BERT Score and ROUGE, which are well-known, as well as Exact Match and Overlap Ratio, specifically utilized for this dataset as shown in Table 13. An Exact Match score assigns a value of 1 if the predicted genre matches the original genre. The Overlap Ratio measures similarity based on shared elements. The F1 score reflects the extent of overlap between the generated answer and the ground truth. Recall scores were used to confirm whether the original lyrics were present within the words generated by the FMs. 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, FMs should provide the complete lyrics, including the predicted masked part. Additional details about the prompts are in Appendix F.4

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 Before Cut-Off dataset, which includes data from 1990 to 2023, and the After Cut-Off dataset, covering January to April 2024. This was done to determine if there is a performance difference between the periods that FMs has been trained on and those it has not.

Table 12: 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.

	English		Korean	
	Before Cut-Off	After Cut-Off	Before Cut-Off	After Cut-Off
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: Evaluation metric of each task using lyrics. Empty block denotes that we did not used the data for the corresponding task.

		Genre Classification	Description Generation	Lyrics Infilling
Korean	**Before Cut-Off**	- Overlap Ratio - Exact Match		
	After Cut-Off	- Overlap Ratio - Exact Match		- ROUGE - BERT Score
English	**Before Cut-Off**	- Overlap Ratio - Exact Match	- ROUGE - BERT Score	
	After Cut-Off	- Overlap Ratio - Exact Match	- ROUGE - BERT Score	- ROUGE - BERT Score

3. FMs was then asked to select the most likely genre(s) based on the provided lyrics.
4. For the zero-shot approach, FMs generated the output directly. For the ORCoT and ORCoT+Few-shot prompts, FMs was instructed to think in alignment with the lyrics.

Description generation: We design the description generation task as follow:
1. FMs 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 FMs 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 ORCoT and ORCoT+Few-shot prompts, we included instructions for FMs to add possible artist names, title names, and musical features.

Infilling: We design the 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. Data with scores exceeding 0.9 were removed.
6. After step 5, the remaining data was used to perform the infilling task with FMs. Due to FMs's safety issues, only the After Cut-Off dataset was used.

E.2 Task Result

In 2.Experiments and Results, we discussed the infilling task. Here, we focus on the genre classification and song description generation tasks.

Genre Classification: In the genre classification task, the difference in the number of unique genres between the English and Korean datasets influenced the results: 11 genres in Korean and 58 in English. This made the task more challenging for the English dataset, leading to FMs struggling more with the English data than the Korean data, as shown in Table 14.

Table 14: Results of the genre classification task, GPT-4o generally outperforms Gemini-1.5-Pro across the entire dataset. Interestingly, after the cut-off, both baseline models showed better performance in Korean than in English.

			Model	Zero-Shot	ORCoT	ORCoT+Few-Shot
English	Before Cut-Off	Overlap Ratio	Gemini-1.5-Pro	0.214	0.218	0.306
			GPT-4o	0.594	0.610	0.620
		Exact Match	Gemini-1.5-Pro	0.306	0.316	0.434
			GPT-4o	0.758	0.774	0.781
	After Cut-Off	Overlap Ratio	Gemini-1.5-Pro	0.405	0.429	0.550
			GPT-4o	0.474	0.497	0.509
		Exact Match	Gemini-1.5-Pro	0.486	0.514	0.657
			GPT-4o	0.671	0.671	0.677
Korean	Before Cut-Off	Overlap Ratio	Gemini-1.5-Pro	0.619	0.581	0.609
			GPT-4o	0.642	0.665	0.733
		Exact Match	Gemini-1.5-Pro	0.652	0.615	0.642
			GPT-4o	0.676	0.698	0.752
	After Cut-Off	Overlap Ratio	Gemini-1.5-Pro	0.503	0.665	0.673
			GPT-4o	0.668	0.690	0.750
		Exact Match	Gemini-1.5-Pro	0.538	0.710	0.713
			GPT-4o	0.710	0.733	0.776

Description Generation: In the description generation task, the overall scores are poor, indicating that FMs struggle to accurately understand the meaning of song lyrics as shown in Table 15. As illustrated in Figure 20, the song discusses 'enduring difficult times with loved ones,' while GPT-4o describes it as 'dealing with a problematic relationship and addictive emotions.

Table 15: Description generation task for English songs. The low overall score shows FMs wrestle with understanding the meaning of lyrics.

			Zero-Shot	ORCoT	ORCoT+Few-Shot
Before Cut-Off	ROUGE-1 (P)	Gemini-1.5-Pro	0.347	0.322	0.321
		GPT-4o	0.384	0.351	0.356
	ROUGE-1 (R)	Gemini-1.5-Pro	0.108	0.140	0.117
		GPT-4o	0.073	0.142	0.148
	ROUGE-1 (F1)	Gemini-1.5-Pro	0.148	0.175	0.154
		GPT-4o	0.151	0.247	0.251
	ROUGE-L (P)	Gemini-1.5-Pro	0.239	0.209	0.216
		GPT-4o	0.274	0.232	0.227
	ROUGE-L (R)	Gemini-1.5-Pro	0.073	0.091	0.078
		GPT-4o	0.073	0.142	0.148
	ROUGE-L (F1)	Gemini-1.5-Pro	0.100	0.113	0.102
		GPT-4o	0.106	0.158	0.161
	BERT Score (P)	Gemini-1.5-Pro	-0.127	-0.111	0.120
		GPT-4o	-0.091	-0.008	0.004
	BERT Score (R)	Gemini-1.5-Pro	0.028	0.032	0.010
		GPT-4o	0.214	0.169	0.164
	BERT Score (F1)	Gemini-1.5-Pro	-0.049	-0.039	-0.055
		GPT-4o	0.057	0.080	0.084
After Cut-Off	ROUGE-1 (P)	Gemini-1.5-Pro	0.298	0.289	0.280
		GPT-4o	0.383	0.335	0.328
	ROUGE-1 (R)	Gemini-1.5-Pro	0.126	0.154	0.140
		GPT-4o	0.117	0.240	0.259
	ROUGE-1 (F1)	Gemini-1.5-Pro	0.161	0.185	0.171
		GPT-4o	0.163	0.252	0.262
	ROUGE-L (P)	Gemini-1.5-Pro	0.201	0.187	0.185
		GPT-4o	0.270	0.212	0.202
	ROUGE-L (R)	Gemini-1.5-Pro	0.080	0.101	0.092
		GPT-4o	0.082	0.160	0.166
	ROUGE-L (F1)	Gemini-1.5-Pro	0.104	0.119	0.112
		GPT-4o	0.113	0.162	0.163
	BERT Score (P)	Gemini-1.5-Pro	-0.087	-0.069	-0.067
		GPT-4o	-0.034	0.050	0.062
	BERT Score (R)	Gemini-1.5-Pro	-0.465	-0.031	-0.051
		GPT-4o	0.241	0.181	0.174
	BERT Score (F1)	Gemini-1.5-Pro	-0.066	-0.049	-0.058
		GPT-4o	0.098	0.115	0.118

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 20: In the description generation task, It is evident that FMs does not accurately comprehend song lyrics. In the example, 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.'

F Prompts

F.1 Graphic Novels

Table 16: The description of each prompt style is provided. We assigned a response format to FMs twice because, in Zero-Shot, the variation in responses is too broad, causing FMs to occasionally break the response format rule. In ORCoT+Zero-Shot, we utilized the simplest ORCoT style because it achieved the best score compared to the more detailed ORCoT version (Table 9.). In ORCoT+Few-Shot, we used three different examples. The performance was insufficient when using only one or two examples.

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.
ORCoT + 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) 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. The correct order is
ORCoT + 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. (Same as ORCoT+Zero-Shot) Output: A.The correct order is [1,2,3,4]” **“The Second, Example:”**: (2nd Example Images) A. “Let’s think step by step. (Same as 1st) The correct order is [1,2,3,4]” **“The Third, Example:”**: (3rd Example Images) A. “Let’s think step by step. (Same as 1st) 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. (Same as 1st) The correct order is

F.2 Calligraphy

Table 17: Korean Calligraphy Prompt: For the ORCoT+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.

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
ORCoT + 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
ORCoT + 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

F.3 Onion Not The Onion

Table 18: We provided examples of prompts used to detect fake news, focusing on the implementation of ORCoT 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.

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)
ORCoT + 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))
ORCoT + 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))

F.4 Lyrics

English Genre Classification

Table 19: Prompt for English genre classification task

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
ORCoT + Zero-Shot	Input : Lyrics Prompt : You are a music genre classifier that analyzes lyrics by reasoning about their thematic content, word choice, rhythm, and stylistic elements. Given a list of unique music genres: ['genre list str'], infer the most appropriate genre(s) based on the provided lyrics. Carefully consider the tone, vocabulary, flow, and subject matter. 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
ORCoT + Few-Shot	Input : Lyrics Prompt : (Same as ORCoT+Zero-Shot) 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

Korean Genre Classification

Table 20: Prompt for Korean genre classification task

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
ORCoT + Zero-Shot	Input : Lyrics Prompt : You are a music genre classifier that analyzes lyrics by reasoning about their thematic content, word choice, rhythm, and stylistic elements. Given a list of unique music genres: ['genre list str'], infer the most appropriate genre(s) based on the provided lyrics. Carefully consider the tone, vocabulary, flow, and subject matter. 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
ORCoT + Few-Shot	Input : Lyrics Prompt : (Same as ORCoT+Zero-Shot) 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

English Song Description Generation

Table 21: Prompt for English song description generation task

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
ORCoT + 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
ORCoT + 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

English Song Infilling

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

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
ORCoT + 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
ORCoT + 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

Korean Song Infilling task

Table 23: ORCoT+Few-shot Prompt for Korean 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
ORCoT + 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
ORCoT + 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