



Let's Think Outside the Box: Exploring Leap-of-Thought in Large Language Models with Creative Humor Generation

Shanshan Zhong^{1,2*} Zhongzhan Huang^{1,2*} Shanghua Gao³ Wushao Wen² Liang Lin²

Marinka Zitnik³ Pan Zhou^{1†}

¹Sea AI Lab ²Sun Yat-sen University ³Harvard University

*Co-first author: {zhongshsh5, huangzhzh23}@mail2.sysu.edu.cn

†Corresponding author: zhoupan@sea.com

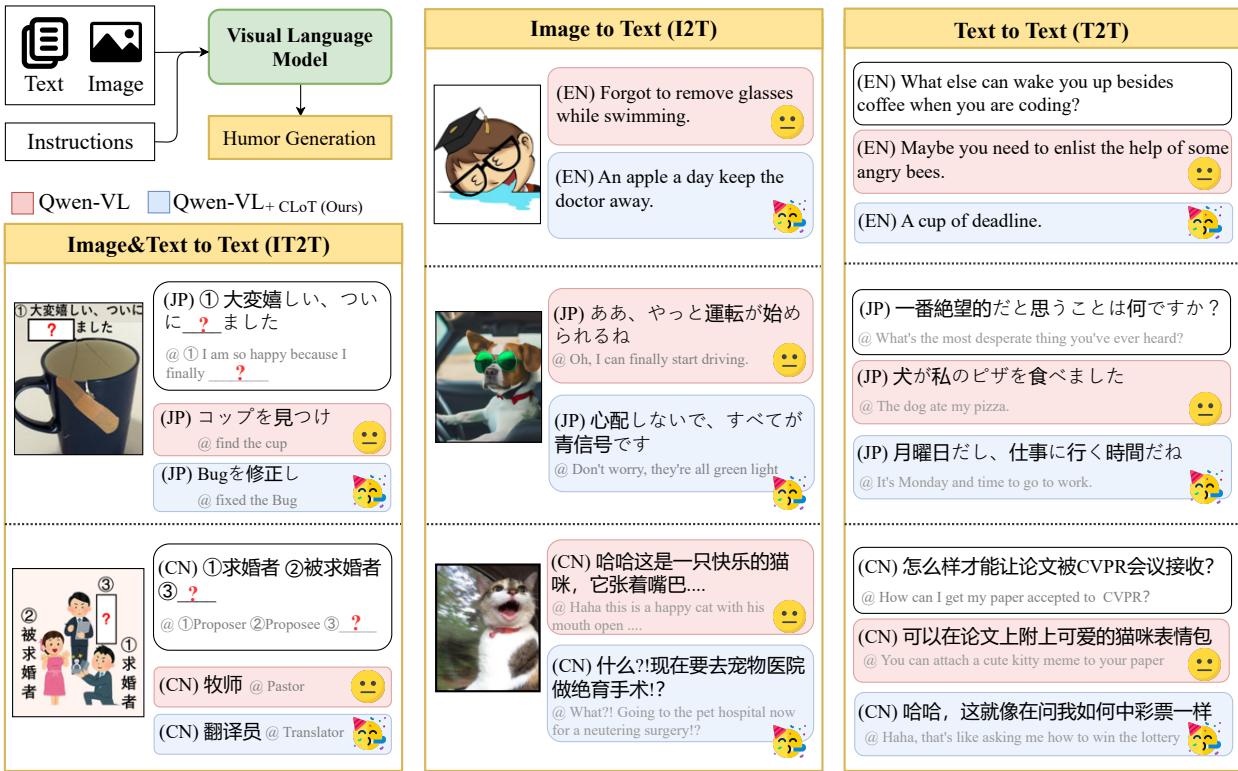


Figure 1. Comaprison between (multimodal) large language model (LLM, ■ red) and its CLoT-integrated version (■ blue) for Oogiri-style multimodal humor generation. According to the model input that can be image, text or both, there are three Oogiri tasks, “Image&Text to Text (IT2T)”, “Image to Text (I2T)”, and “Text to Text (T2T)”, where text can be English (EN), Chinese (CN), and Japanese (JP). “@” denotes translations. The baseline LLM is Qwen-VL [1]. While humor is subjective, these examples demonstrate CLoT’s leap-of-thought capacity of using excellent creative thinking to produce high-quality humor responses. See more examples in Appendix.

Abstract

Chain-of-Thought (CoT) [2, 3] guides large language models (LLMs) to reason step-by-step, and can motivate their logical reasoning ability. While effective for logical tasks, CoT is not conducive to creative problem-solving which often requires out-of-box thoughts and is crucial for

innovation advancements. In this paper, we explore the Leap-of-Thought (LoT) abilities within LLMs — a non-sequential, creative paradigm involving strong associations and knowledge leaps. To this end, we study LLMs on the popular Oogiri game which needs participants to have good creativity and strong associative thinking for responding unexpectedly and humorously to the given image, text, or

both, and thus is suitable for LoT study. Then to investigate LLMs’ LoT ability in the Oogiri game, we first build a multimodal and multilingual Oogiri-GO dataset which contains over 130,000 samples from the Oogiri game, and observe the insufficient LoT ability or failures of most existing LLMs on the Oogiri game. Accordingly, we introduce a creative Leap-of-Thought (CLoT) paradigm to improve LLM’s LoT ability. CLoT first formulates the Oogiri-GO dataset into LoT-oriented instruction tuning data to train pretrained LLM for achieving certain LoT humor generation and discrimination abilities. Then CLoT designs an explorative self-refinement that encourages the LLM to generate more creative LoT data via exploring parallels between seemingly unrelated concepts and selects high-quality data to train itself for self-refinement. CLoT not only excels in humor generation in the Oogiri game as shown in Fig. 1 but also boosts creative abilities in various tasks like “cloud guessing game” and “divergent association task”. These findings advance our understanding and offer a pathway to improve LLMs’ creative capacities for innovative applications across domains. The dataset, code, and models will be released online.

1. Introduction

Large language models (LLMs) [4–13] have catalyzed a transformative era in problem-solving abilities, revolutionizing various domains within artificial intelligence. The advent of the Chain-of-Thought (CoT) paradigm [3] and its further enhancements [2, 14–16] have equipped these LLMs with a human-like step-by-step reasoning capacity. This augmentation has enabled LLMs to excel in intricate reasoning tasks spanning from language comprehension to visual understanding. As shown in Fig. 2 (Left), CoT instills LLMs with a sequential thinking process wherein each subsequent thought builds upon the previous one. This paradigm enhances the precision and rigor in logical processing, making it exceedingly effective for problems that demand closely linked logical reasoning.

However, the sequential nature of CoT might fall short in nurturing creativity and innovation, potentially limiting solutions in creative problem-solving scenarios [17, 18]. For instance, proving an algebraic inequality often follows a step-by-step CoT process that progresses from one inequality to the next. Yet, an intuitive flash, e.g., a geometric interpretation, can yield a more creative solution. This type of insight, known as “Leap-of-Thought” (LoT) [19, 20], a.k.a. mental leap [21–24]—the art of non-sequential thinking by association, drawing parallels between seemingly unrelated concepts, and facilitating a “leap” of knowledge transfer. In contrast to CoT reasoning, LoT as depicted in Fig. 2 (Right), fosters associative reasoning and encourages thinking outside the box, which bridges disparate ideas and fa-

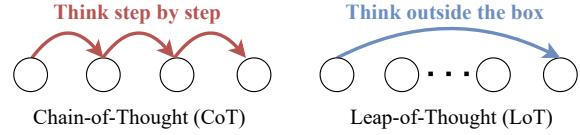


Figure 2. Comparison of CoT and LoT. “○” denotes the thought and “→” represents the connection between two thoughts.

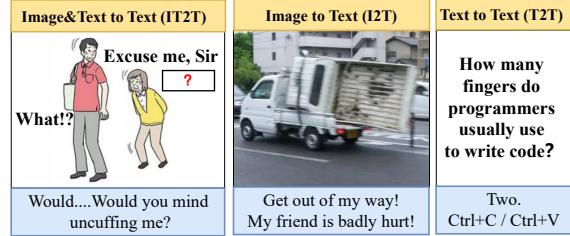


Figure 3. Examples of the three types of LoT-based Oogiri games. Players are required to make surprising and creative humorous responses (blue box) to the given multimodal information e.g., images, text, or both.

cilitates conceptual leaps. Embracing LLMs with a strong LoT ability can unlock significant potential for innovation, contributing to advancements in creative applications.

In this paper, we aim to initially explore and enhance the LoT ability of LLMs. However, thoroughly assessing LoT is challenging due to the complexity of measuring creative thinking [25–27] and the difficulty in gathering pertinent data, since generating novel ideas is challenging, even for humans [17]. Given these constraints, we propose studying LoT in LLMs through the lens of Oogiri-style humor generation. Oogiri, a traditional Japanese creative game [28], requires participants to provide unexpected and humorous responses to prompts in the form of images, text, or a combination of both, as shown in Fig. 3. This game challenges LLMs to demonstrate a sudden burst of insight and strong associative thinking, presenting a unique challenge for CoT-based methods, making it an ideal testbed for assessing the leap-of-thought abilities of LLMs. Moreover, the extensive online presence of Oogiri guarantees a wealth of human-generated creative content, ideal for compiling an expansive leap-of-thought dataset.

To investigate the LoT ability of LLMs in the Oogiri game, we present the multilingual and multimodal Oogiri-GO dataset which comprises more than 130,000 high-quality Oogiri samples in English, Chinese, and Japanese, and curated to prompt textual humor in response to inputs that can be images, text, or both. Through extensive experiments, we discover that even the advanced LLMs and reasoning frameworks [2, 4, 6, 29], such as GPT-4 and CoT, despite their exceptional reasoning capabilities, possessing a rich prior knowledge of diverse forms of humor [2], still struggle to exhibit sufficient LoT ability for creative humor generation. Moreover, directly fine-tuning LLMs on

the Oogiri-GO is not easy to improve the LoT ability. The more efficient utilization of humorous knowledge is needed to help LLM elicit creative responses.

Motivated by the human mental leap exercise process of “remote association & self-refinement” [30], to enable LLMs with strong LoT ability for creation, we propose the Creative Leap-of-Thought (CLoT) paradigm which relies on two LoT-boosting stages. The first one is the associative instruction tuning stage which designs an associative instruction template to formulate the Oogiri-GO dataset into instruction data and trains an LLM to improve its LoT ability. The core here is the instruction template with a dual purpose: it randomly provides LLM with clues to establish connections between game inputs and creative responses, while also introducing empty clues to encourage LLM for unrestrained exploration and remote association thinking.

The second stage is explorative self-refinement which encourages the LLM to generate more creative LoT data via exploring parallels between seemingly unrelated concepts under weakly-associated conditions, and selects high-quality data to train itself for self-refinement. These weakly-associated conditions can either be empty, or randomly sampled from an object noun set collected from the Oogiri-GO dataset. The former empty conditions to allow LLM to operate freely, and the latter ones help the LLM to link seemingly-unrelated and weakly-related concepts, and encourage the LLM to explore knowledge outside of traditional cognitive limitations. This exploration strategy can help generate diverse high-quality data for self-refinement.

Experimental results show that CLoT can greatly enhance the LoT ability of LLMs like Qwen [1] and CogVLM [29] across several types of Oogiri games. Specifically, CLoT can help LLMs to generate much better humors in Fig. 1. Moreover, CLoT-integrated LLMs achieve higher quantitative performance than the corresponding vanilla and CoT-integrated LLMs across the multiple-choice and ranking questions in the Oogiri game. Also, CLoT can boost creative abilities on other tasks like “cloud guessing game” and “divergent association task” [31–33], showing its remarkable generalization ability.

2. Related Works

(1) Oogiri game (大喜利) is a general term for a series of traditional Japanese comedy games. In ancient times, there were different types of Oogiri, such as actors performing sumo wrestling, telling ghost stories, etc. The modern Oogiri game mainly refers to one specific type known as Tonchi (頓智), typically presented in the format of game shows or intellectual quiz programs [28]. Players are provided with various multimodal contents, which can be simple questions, random images, etc., and are then prompted to come up with humorous, creative responses to achieve surprising comedic effects, as the examples are shown in

Fig. 3. It is worth noting that the character “頓” in both Japanese and Chinese denote “sudden”, while “智” means “intelligence, insight or intuition”. This highlights the connection between the Oogiri game and the requirement for strong associative abilities in LoT, making Oogiri an ideal platform for exploring LoT capabilities within LLMs.

(2) Multimodal LLMs and their creativity. Recently, multimodal Language Models [1, 29, 34, 35] have garnered significant attention, particularly due to their impressive reasoning abilities [7–12, 36]. Moreover, there is a growing focus on exploring the creativity [37–40] of LLMs for applications such as scientific discovery [18, 41–44], creative writing [45–49], etc.

(3) Computational humor is a branch of computational linguistics and artificial intelligence that uses computers in humor research [50], which encompasses various tasks, including humor detection [51–58], humor interpretation [58–61], and humor generation [62–66], etc. With the advancement of generative LLMs [1, 4, 29], humor generation has become a popular focus while humor generation still faces challenges such as insufficient punchlines [67] and limited in multimodal contexts [68, 69].

(4) Chain-of-Thought based Methods provide the models with “chain of thoughts” [2, 3, 14–16], i.e., reasoning exemplars [3], or a simple prompt “Let’s think step by step” [2], to encourage LLMs to engage in reasoning rather than simply providing answers directly [70].

Category	English	Chinese	Japanese	Total
I2T	17, 336	32, 130	40, 278	89, 744
T2T	6, 433	15, 797	11, 842	34, 072
IT2T	—	912	9, 420	10, 332

Table 1. Data distribution of the Oogiri-GO dataset. For the IT2T task, its English version is not available due to cultural preference.

3. Oogiri-GO Dataset

As introduced in Sec. 2, in the Oogiri game, the participants need to unexpectedly and humorously respond to the given images, text, or both. See three types of examples in Fig. 3. This game requests a sudden burst of insight and strong associative thinking to the given context, and provides an ideal platform to assess the leap-of-thought (LoT) ability of LLMs. Accordingly, we collect Oogiri game data to build a large-scale Oogiri-GO dataset which serves as a benchmark to evaluate and improve LoT ability.

Specifically, Oogiri-GO is a multimodal and multilingual humor dataset, and contains more than 130,000 Oogiri samples in English, Chinese, and Japanese. Notably, in Oogiri-GO, 77.95% of samples are annotated with human preferences, namely the number of likes, indicating the popularity of a response. As illustrated in Fig. 3, Oogiri-GO contains three types of Oogiri games according to the input that can be images, text, or both, and are respectively called “Text

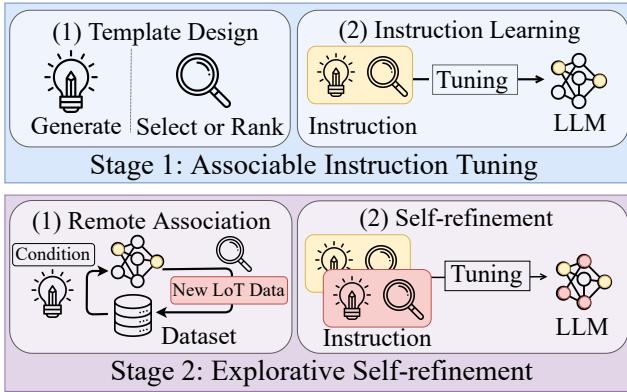


Figure 4. The overview of proposed Creative Leap-of-Thought.

to Text” (T2T), “Image to Text” (I2T), and “Image & Text to Text” (IT2T) for brevity. See more examples in Fig. 1. Table 1 summarizes the distribution of these game types. For training purposes, 95% of the samples are randomly selected to construct the training dataset, while the remaining 5% form the test dataset for validation and analysis.

To create the Oogiri-GO dataset, there are three main steps, including online data collection, machine filtering by LLM, and manual screening. Firstly, to collect sufficient data, we source Oogiri game data from the official Oogiri game platform, Bokete, and other popular platforms, such as Twitter and Weibo which also host some Oogiri-game-like data. Then, to guard against the inclusion of bias, violence, explicit content, offensive language, etc., we have placed a strong emphasis on rigorous safety checks during both machine and manual screening. We first use the multimodal LLM Qwen-14B [5] to do the initial screening of the raw data by constructing safety-checking prompts. Then, manual checking is performed on the remaining data. See more details about the dataset creation in the Appendix.

4. Creative Leap-of-Thought (CLoT)

To augment the Leap-of-Thought (LoT) ability in (multimodal) Large Language Models (LLMs) for creative generation, we propose a novel Creative LoT framework (CLoT). As shown in Fig. 4, CLoT relies on two LoT-boosting stages. The first one is associative instruction tuning that formulates the Oogiri-GO dataset into instruction tuning data for training an LLM to improve its LoT ability (Sec. 4.1). The second one is explorative self-refinement that encourages the LLM to generate more creative LoT data via exploring parallels between seemingly unrelated concepts, and selects high-quality data to train itself for self-refinement (Sec. 4.2). Finally, we present the CLoT inference to induce the LoT ability of the trained LLM (Sec. 4.3).

4.1. Associable Instruction Tuning

LoT ability mainly includes associative generation and discrimination ability [30]. Given an input, associative generation draws its parallels with seemingly unrelated concepts via remote association and then generates innovative responses, e.g., the unexpected humor for the Oogiri input. Associable discrimination is to judge the matchiness among input and responses though they are seemingly unrelated, and then to select the most creative response.

Unfortunately, both associative generation and discrimination are not present in current LLMs, e.g., poor performance of GPT4v [71] in the Oogiri game observed in Sec. 5. Moreover, it is hard to improve these two LoT abilities via popular CoT-like prompt techniques. Indeed, as shown in Sec. 5, CoT even sometimes impairs the LoT performance of the LLMs like Qwen-VL [1] in the Oogiri game.

To address this issue, we propose associative instruction tuning which trains LoRA [72] for LLMs on the Oogiri-GO dataset to achieve certain associative generation and discrimination abilities. It has two steps, including instruction generation and discrimination template design, and associative instruction learning.

(1) Instruction Generation & Discrimination Templates. We design LoT-oriented instruction templates to transform the Oogiri-GO dataset into instruction tuning data, and then train LLM to achieve associative generation and discrimination abilities. Our templates primarily comprise two components in Fig. 5: task-specific prompt and response. For different abilities, the templates need some special design.

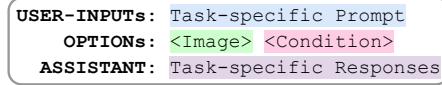


Figure 5. The LoT-oriented instruction templates.

For associative generation, “USER-INPUTs” contains “Task-specific Prompt” along with two optional conditions, “Image” and “Condition”. For “Task-specific Prompt”, we elaborately design several templates for different types of Oogiri game. See the Appendix for details and there is an image-2-text (I2T) Oogiri example in Fig. 6. For “Image” condition, it relies on the type of Oogiri game, e.g., being the image embeddings in I2T game and empty in T2T type. For the “condition” option, it’s set to empty with a probability of ρ_c , and otherwise is randomly set as one noun in “task-specific responses”. This design gives the LLM a clue to connect the game input and the correct responses while also encouraging LLM to explore and unleash its creative thinking with probability ρ_c . Finally, “Task-specific Responses” are the ground truth responses of an Oogiri-GO data, and need to be predicted by LLM during training. This task enforces the LLM to draw parallels between seemingly unrelated concepts in inputs and responses for giving innovative responses, e.g., the humor for the Oogiri input. This

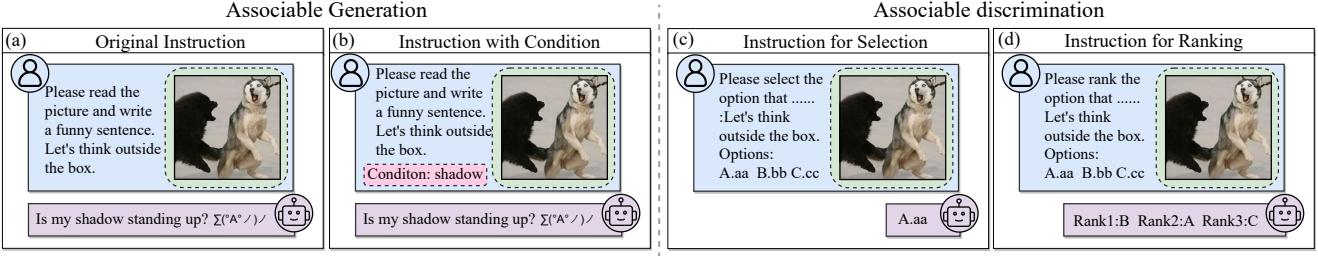


Figure 6. The details of LoT-oriented instructions templates. We take “Image to Text” as an example, see the Appendix for the details of other categories’ instructions. (a) and (b) are the instruction templates with/without conditions for associative generation. (c) and (d) are the two instructions about the selection and ranking of associative discrimination. All templates follow the formats in Fig. 5.

associable generation ability can assist the LLM to think outside the box and learn remote association thinking.

Regarding associative discrimination, we aim to develop fundamental LoT discrimination skills for LLM. Based on the Oogiri-GO data, we design choice questions to enhance LLM’s LoT discrimination ability, i.e., **selection** skill. Besides, as 77.95% of the Oogiri-GO data have human preference annotations, i.e., the number of likes of several responses (see Sec. 3), we design ranking questions to improve another discrimination skill, i.e., **ranking** ability.

For a choice question, as shown in Fig. 6 (c), the options in “Task-specific Prompt” contain the random permutations of ground truth response (GTR), image captions generated by BLIP2 [73], GTR from other images, rewrites of GTR by Qwen-14B [5]. See details in Appendix. For “task-specific responses”, it is the GTR. This design is to train LLM to improve its LoT selection ability. For a ranking question, as shown in Fig. 6 (d), it is to enforce LLM to rank multiple distinct responses of a given input to match their human preferences. By training on the choice and ranking questions, LLM is encouraged to distinguish LoT responses and align human creative preferences, improving its LoT discriminative selection and ranking abilities.

(2) Associable Instruction Learning. By using the above instruction templates, we augment the 130,000 samples in the Oogiri-GO dataset to more than 500,000 instructions whose formulation is in Fig. 5. During training, LLM is required to predict the “task-specific responses” according to the “USER-INPUTs” which include “Task-specific Prompt” and two additional optional conditions like image and text condition. To avoid over-fitting, we only train standard LoRA [72] for the LLM with the associative instruction data. See more details in Appendix.

4.2. Explorative Self-Refinement

After associative instruction tuning, we aim to generate more high-quality creative data by LLM which are then used to train LLM for self-refinement. To this end, we introduce an innovative stage called explorative self-refinement,

inspired by human LoT exercise process of “remote association & self-refinement”, also known as mental leap [21, 24, 30]. The remote association process refers to generating new ideas by associating remote concepts or thoughts, and self-refinement uses the generated data to enhance one’s own LoT ability. In the following, we design two similar LoT exercise processes for LLM to improve its LoT ability.

(1) Explorative Remote Association. The core here is to prompt the LLM to generate a diverse array of creative responses under weakly-associated conditions. To implement this, we extract a set of object nouns, denoted as \mathcal{S} , from the text in the Oogiri-GO training data. See details in Appendix. Then, for each user-input I (see Fig. 5), we generate n weakly-associated conditions $\{C_i\}_{i=1}^n$. These conditions can either be empty with a probability $\rho \in (0, 1)$ to give freedom to LLM, or uniformly randomly sampled from the noun set \mathcal{S} to enforce LLM to build connections between different concepts. Next, we add the condition C_i into user-input I , and feed I into the LLM to generate a humor candidate R_i . Repeating this process with different conditions C_i can generate a total of n candidates $\{R_i\}_{i=1}^n$.

Then the LLM ranks these candidates by its discriminative ranking ability learned in Sec. 4.1. Next, it mixes the top-2 candidates with the ground truth responses (GTR), and selects the top-1 as the final response. Finally, if the selected top-1 response is the GTR, we discard this sample. Here first filtering out low-quality responses can improve the accuracy of subsequent top-1 selection, since $(n + 1)$ -choice problem is often more challenging than 3-choice problem as shown in Sec. 5. By repeating this process, we progressively gather sufficient new high-quality data.

The core of this approach is the weakly-associated conditions $\{C_i\}_{i=1}^n$ which can encourage the LLM to engage in remote associations. This is because the empty conditions allow LLM to operate freely, while the object noun conditions compel the LLM to draw connections between seemingly unrelated concepts. This mechanism facilitates the establishment of links between seemingly-unrelated and weakly-related concepts, encouraging the LLM to explore

Model	Size	Image&Text to Text (IT2T)					Image to Text (I2T)					Text to Text (T2T)				
		3T1	4T1	5T2	Rank	Avg.	3T1	4T1	5T2	Rank	Avg.	3T1	4T1	5T2	Rank	Avg.
GPT4v [71]	-	19.3	14.9	3.2	56.7	23.5	29.1	15.1	3.9	60.4	27.1	27.1	16.8	6.8	53.5	26.1
LLaVA-1.5 [34]	13B	13.2	13.7	13.9	68.1	27.2	29.3	22.7	3.9	60.9	29.2	33.8	25.2	4.0	62.6	31.4
MiniGPT-v2 [35]	7B	6.1	3.4	4.0	60.7	18.6	5.3	4.0	3.8	60.5	18.4	10.8	7.3	3.5	59.4	20.3
mPLUG-Owl _{Multilingual} [12]	7B	28.1	26.0	10.5	64.4	32.2	19.2	18.6	6.0	60.5	26.1	24.4	22.2	10.7	60.1	29.4
VisualGLM-6B [74]	6B	24.1	22.5	9.7	67.4	30.9	14.3	20.4	8.8	61.9	26.4	13.1	20.2	7.1	61.3	25.4
Qwen-VL [1]	7B	30.2	26.0	10.4	67.7	33.6	23.2	23.1	11.9	62.2	30.1	23.4	25.0	13.3	59.6	30.3
Qwen-VL+AIT (Ours)	7B	39.7	38.9	15.7	67.3	40.4 _{+ 6.8}	38.8	30.5	15.7	62.3	36.8 _{+ 6.7}	30.6	28.7	16.7	62.6	34.6 _{+ 4.3}
Qwen-VL+CLoT (Ours)	7B	41.8	38.7	21.6	68.5	42.7_{+ 9.1}	39.8	35.1	22.7	64.4	40.5_{+ 10.4}	38.8	29.4	21.0	64.7	38.5_{+ 8.2}

Table 2. The accuracy (%) of choice questions and the NDCG (%) of ranking questions on **multimodal multilingual models**. mTn choice question selects n correct answers from m options. “Avg.” is the average of all metrics. “AIT” denotes associative instruction tuning.

Algorithm 1 Inference Step of CLoT

Input: Input I , CLoT-trained LLM \mathcal{A} , response number n
Output: Creative response R .

- 1: ▷ Creating the candidate responses
- 2: construct n weakly-associated conditions $\{C_i\}_{i=1}^n$
- 3: $\{\mathbf{R}_i\}_{i=1}^n \leftarrow \mathcal{A}([I, \{C_i\}_{i=1}^n])$
- 4: ▷ Choosing most creative response
- 5: Top-2 $\mathbf{R}'_1, \mathbf{R}'_2 \leftarrow \mathcal{A}([I, \{\mathbf{R}_i\}_{i=1}^n])$ with ranking ability
- 6: Best $\mathbf{R} \leftarrow \mathcal{A}([I, \mathbf{R}'_1, \mathbf{R}'_2])$ with selection ability
- 7: **return** Best response \mathbf{R} .

knowledge outside of traditional cognitive limitations. The exploration ability distinguishes our CLoT from CoT which primarily guides the LLM to exploit its inherent reasoning ability without emphasizing knowledge exploration.

(2) Self-refinement. Here we combine the above generated instructions with vanilla instruction tuning samples in Sec. 4.1 to form a dataset with more than 550,000 samples to train our LLM again. Since the above generated data is of high diversity because of its exploration strategy, they prevent performance collapse [75, 76] during self-refinement phase, and can improve the LoT performance across several creative tasks as shown in Sec. 5. See the ablation study and more discussions in Sec. 5.5.

4.3. CLoT Inference

After the two LoT-boosting phases in Sec. 4.1 and 4.2, the LLM acquires sufficient LoT ability. Now we introduce the inference steps of LLM to release its LoT ability. Formally, given an Oogiri user-input I of the formation in Fig. 5, LLM first uses explorative remote association in Sec. 4.2 to construct n weakly-associated conditions, and then follows Sec. 4.2 to generate n responses $\{\mathbf{R}_i\}_{i=1}^n$. Next, LLM ranks these responses by using its learned ranking skill and finally selects the top-1 one from the ranked top-2 response via its selection skill. The reason to first use ranking before selection is that as shown by experimental results in Sec. 5, directly choosing the best one from a large number of options has poor accuracy, and ranking can filter out low-quality candidates to improve the selection accuracy. See Algorithm 1 for an overview of CLoT inference steps.

5. Experiments

5.1. Evaluation Questions and Metrics

Inspired by the humor benchmarks in [80], we first develop choice and ranking questions as introduced in Sec. 4.1 (see examples in Fig. 6 (c-d)), and then quantitatively evaluate the LoT ability of LLMs on the Oogiri-GO test dataset. For the *choice questions*, mTn for short, they need LLMs to choose n “leap-of-thought” humor responses from m options given the input. Here we build four types of mTn questions, including 2T1, 3T1, 4T1, and 5T2. 2T1 means two options, the ground-truth response (GTR) and an image caption generated by BLIP2 [73]. 3T1 adds unrelated answers, e.g., other image captions. 4T1 further adds the GTR rewrite by Qwen-14B [5]. 5T2 has an extra GTR. For these questions, their difficulty increases progressively, and is diverse to ensure comprehensive evaluation. For choice questions, we use accuracy as the evaluation metric. Additionally, for the questions in test set whose responses have ground-truth human preference, e.g., the number of likes, we develop the *ranking questions* that always rank five candidates. For evaluation, we adopt the top-1 accuracy and the widely used ranking metric, i.e., Normalized Discounted Cumulative Gain (NDCG) [81, 82]. We provide more experimental details in the Appendix.

5.2. Evaluation by Choice and Ranking Questions

Evaluation on Multimodal Multilingual LLMs. We plug our associative instruction tuning (AIT) and our CLoT into the SoTA open-source multimodal multilingual model Qwen-VL [1] to obtain Qwen-VL_{+AIT} and Qwen-VL_{+CLoT}, respectively. Table 2 shows that, on three tasks (IT2T, I2T and T2T) which include English, Chinese and Japanese questions, Qwen-VL achieves the best LoT performance among all baselines in most cases. In comparison, Qwen-VL_{+AIT} achieves a noticeable improvement on the SoTA Qwen with average accuracy enhancements of 6.8%, 6.7%, and 4.3% on the three tasks, respectively. Importantly, Qwen-VL_{+CLoT} further enhances Qwen-VL, showing improvements of 9.1%, 10.4%, and 8.2% in accuracy across these tasks. These results demonstrate the efficacy of the

Model	Size	Image to Text (I2T)						Text to Text (T2T)					
		2T1	3T1	4T1	5T2	Rank	Avg.	2T1	3T1	4T1	5T2	Rank	Avg.
InstructionBLIP [77]	13B	19.8	13.7	15.5	1.1	65.5	23.1	22.3	16.0	17.0	0.7	59.5	23.1
mPLUG-Owl _{LLaMA2} [12]	7B	22.3	12.7	15.0	4.2	59.9	22.8	24.2	13.7	12.6	3.1	59.2	22.6
Otter [78]	7B	15.8	9.9	8.5	7.1	61.3	20.5	3.8	3.3	4.8	5.4	58.5	15.1
CogVLM-17B [29]	7B	37.6	26.4	18.3	2.5	64.6	29.9	35.1	27.8	24.8	7.5	64.1	31.9
CogVLM-17B _{+AIT} (Ours)	7B	57.4	37.4	33.5	21.8	64.6	42.9 _{+13.1}	55.4	46.5	26.4	18.2	64.4	42.2 _{+10.3}
CogVLM-17B _{+CLoT} (Ours)	7B	66.9	47.6	43.4	30.7	69.4	51.6_{+21.7}	64.8	52.9	33.6	21.8	68.6	48.3_{+16.4}

Table 3. The accuracy (%) of choice questions and the NDCG (%) of ranking questions on various **multimodal non-multilingual models** (English). See notations in Table 2. We only consider I2T and T2T since English IT2T is not available due to cultural preference.

Model	Size	3T1	4T1	5T2	Rank	Avg.
GPT-3.5 [71]	-	45.3	30.4	6.7	61.6	36.0
GPT-4 [71]	-	49.2	20.4	3.6	54.7	32.0
LLAMA2 [4]	7B	18.9	13.5	1.1	60.4	23.5
	13B	15.6	20.0	1.8	60.5	24.5
	70B	27.8	16.1	3.8	62.0	27.4
Baichuan2 [79]	7B	28.3	22.6	11.6	64.6	31.8
	13B	21.7	18.3	8.9	61.5	27.6
Qwen [5]	7B	23.1	20.4	8.0	61.4	28.2
	14B	27.4	22.2	12.3	59.5	30.3
ChatGLM3 [74]	6B	15.6	17.0	5.4	59.4	24.3
Vicuna-v1.5 [6]	7B	32.6	23.5	0.0	63.0	29.8
	13B	30.2	23.0	2.7	62.2	29.5
Qwen-VL _{+CLoT} (Ours)	7B	51.7	32.3	24.8	65.0	43.4
CogVLM-17B _{+CLoT} (Ours)	7B	52.9	33.6	21.8	68.6	44.2

Table 4. The accuracy (%) of choice questions and the NDCG (%) of ranking questions on various **large language models**. Here we use English T2T task for test. See notations in Table 2.

two stages in CLoT, i.e., associative instruction tuning and explorative self-refinement.

Evaluation on Multimodal Non-multilingual LLMs. Here we integrate our CLoT with the SoTA multimodal non-multilingual model, CogVLM-17B [29], and evaluate it on the English I2T and T2T tasks. Table 3 shows that CogVLM-17B_{+AIT} achieves remarkable improvements over the standard CogVLM-17B, and CogVLM-17B_{+CLoT} consistently demonstrates significantly superior performance compared to CogVLM-17B.

Evaluation on Single-Modal LLMs. Now we test LLMs that can handle only pure texts, using the English T2T task for evaluation. Table 4 also indicates the insufficient LoT ability within existing LLMs, ranging from small to large models. Fortunately, our CLoT significantly improves the LoT ability of these LLMs, as demonstrated by the notable improvement in accuracy.

Comparison with CoT-alike Reasoning Frameworks.

We also find that existing reasoning frameworks are not as effective as CLoT in enhancing LoT ability. Fig. 8 compares CLoT with CoT [2, 3], CoT-SC [83], and prompted-based LoT (PLoT) with the prompt “let’s think outside the box”. The results reveal that CoT-alike frameworks do not enhance LoT performance of LLMs, while CLoT demon-

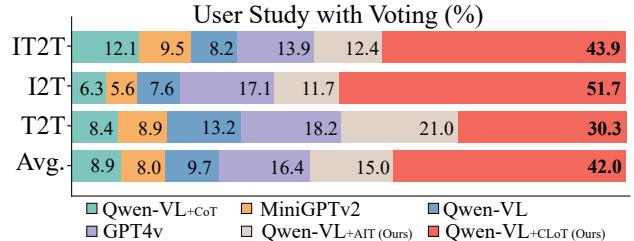


Figure 7. User study with voting (%) for Oogiri-style creative responses by different models and improved methods.

strates the ability to consistently enhance LLMs.

Our experiments and analysis reveal that, unlike CoT-based methods, LoT cannot be directly achieved by prompting alone. This is because the inherent reasoning capabilities and extensive knowledge of LLMs are not sufficient to enable LoT ability. However, when trained with our proposed CLoT method, LLMs can effectively engage in a range of creative tasks. Additionally, the use of specific prompting techniques can enhance the LoT ability of CLoT-trained LLMs. These findings suggest that LoT could potentially be considered an additional general reasoning ability for LLMs that is not contained in current LLMs.

5.3. Human Evaluation

We conduct a user preference study to test creativity of LLMs. Here we select six LLMs to generate responses for a total of eighteen questions across three tasks (IT2T, I2T and T2T). We use choice questions, and ask users to choose the most creative and humorous responses. Fig. 7 summarizes the statistical analysis of 154 valid surveys. The results show that users have a strong inclination towards selecting the results of CLoT across three tasks, highlighting the high-quality creative content generated by CLoT. See more user study details in Appendix.

5.4. Evaluation on Other Creative Tasks

To evaluate the generalization ability of CLoT, we test CLoT on another two creative tasks, including Cloud Guessing Game (CGG) and Divergent Association Task (DAT). In CGG, the LLM is to identify the shape of white clouds, and then to select the corresponding shapes from given options. For instance, the white clouds in Fig. 9 (c)

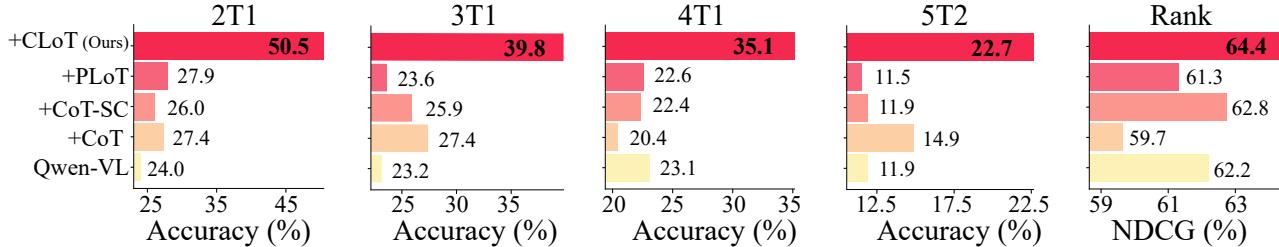


Figure 8. The accuracy (%) of choice questions and the NDCG (%) of ranking questions on our CLoT and various **reasoning frameworks**. The baseline is Qwen-VL on multilingual I2T task. For mTn choice questions, one needs to select n correct answers from m options.

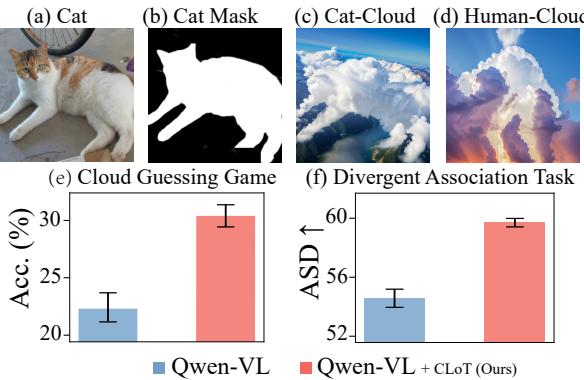


Figure 9. Evaluation of CLoT on the creative CGG (e) and DAT (f) tasks. (c-d): examples of cloud guessing games. (b): conditional masks of image (a) for generating cloud images.

has a shape of a cat, and the one in Fig. 9 (d) is similar to a human. These white cloud images are generated by a control diffusion model [31, 32, 84, 85], guided by masks shown in Fig. 9 (b). We use top-1 accuracy as metric. See more details in Appendix. For DAT, it is a classic creativity test [33, 86] which needs participants to choose words with larger semantic distances among 10 unrelated nouns. Here for test easily, we transfer the DAT benchmark [33] to a series of choice questions and take the standard average semantic distance (ASD) as a metric. These questions can challenge the LLM to select the one word from nine options that differs from the given word most. See more details in Appendix. CGG and DAT can test the LoT ability of LLMs, specifically their remote association thinking ability, and provide quite different evaluation platforms. As shown in Fig. 9 (e-f), CLoT can also significantly enhance the performance of the SoTA Qwen-VL on both CGG and DAT tasks. Specifically, CLoT-integrated Qwen-VL improves the vanilla Qwen-VL by about 8% on the CGG task and 5% on the DAT task. These results well demonstrate the good generalization and transferability of CLoT.

5.5. Ablation Study

Weakly-associated Conditions. By default, to encourage remote association, we use weakly-associated conditions randomly sampled from the noun set on the whole dataset in Sec. 4.2.

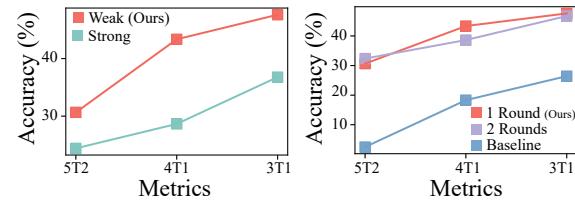


Figure 10. The ablation study of CLoT. We use CogVLM as baseline on the English I2T task. **Left:** weakly-associated condition v.s. strongly-associated condition during explorative remote association. **Right:** The effect of rounds of self-refinement.

To verify the effectiveness of weakly-associated conditions, now we resort to strongly-associated conditions sampled from the noun set of the current image caption. Results in Fig. 10 (Left) show that using weakly-associated conditions is superior and more conducive to fostering the creativity of LLMs. The weakly-associated conditions enable the LLM to generate more diverse LoT responses, while the strong clue from the strongly-associated conditions limit the diversity of LoT generations.

Round of Self-refinement. By default, we run one-round self-refinement for the Oogiri game. Here we explore whether more rounds of the self-refinement can further improve the LoT ability. Fig. 10 (Right) shows that a single round of self-refinement already yields promising performance, whereas additional rounds do not yield significant further improvements. As shown in Fig. 10 (Left), the diversity of the condition set is crucial to self-refinement, since it decides whether the associative remote stage can generate high quality and diverse data. However, the condition set is not expanded during the second-round self-refinement, which consequently limits further improvements in performance. Effectively increasing the scale of the condition set is an effective way for further improvement. See more discussion in Appendix. But its exploration falls outside the scope of this work and is left for our future research.

6. Conclusion

In this paper, we propose a Creative Leap-of-Thought (CLoT) paradigm to improve LLM’s leap-of-thought (LoT) ability. CLoT first collects a multimodal Oogiri-GO dataset, and formulates it into instruction tuning data to train LLM

to improve its LoT ability. Then CLoT designs an explorative self-refinement that lets LLM generate more creative LoT data via exploring parallels among different concepts and selects high-quality data to train itself for self-refinement. Experimental results show the effectiveness and generalization ability of CLoT across several creative tasks.

References

- [1] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023. [1](#), [3](#), [4](#), [6](#), [15](#), [16](#), [20](#), [29](#)
- [2] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213, 2022. [1](#), [2](#), [3](#), [7](#)
- [3] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837, 2022. [1](#), [2](#), [3](#), [7](#)
- [4] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023. [2](#), [3](#), [7](#)
- [5] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023. [4](#), [5](#), [6](#), [7](#), [18](#)
- [6] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhang-hao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, March 2023. [2](#), [7](#)
- [7] Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*, 2021. [3](#)
- [8] Abulhair Saparov and He He. Language models are greedy reasoners: A systematic formal analysis of chain-of-thought. *arXiv preprint arXiv:2210.01240*, 2022.
- [9] Andy Zeng, Maria Attarian, Brian Ichter, Krzysztof Choromanski, Adrian Wong, Stefan Welker, Federico Tombari, Aveek Purohit, Michael Ryoo, Vikas Sindhwani, et al. Socratic models: Composing zero-shot multimodal reasoning with language. *arXiv preprint arXiv:2204.00598*, 2022.
- [10] Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al. Palm-e: An embodied multimodal language model. *arXiv preprint arXiv:2303.03378*, 2023.
- [11] Runpei Dong, Chunrui Han, Yuang Peng, Zekun Qi, Zheng Ge, Jinrong Yang, Liang Zhao, Jianjian Sun, Hongyu Zhou, Haoran Wei, et al. Dreamllm: Synergistic multimodal comprehension and creation. *arXiv preprint arXiv:2309.11499*, 2023.
- [12] Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*, 2023. [3](#), [6](#), [7](#)
- [13] Shanshan Zhong, Zhongzhan Huang, Weushao Wen, Jinghui Qin, and Liang Lin. Sur-adapter: Enhancing text-to-image pre-trained diffusion models with large language models. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 567–578, 2023. [2](#)
- [14] Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. Automatic chain of thought prompting in large language models. *arXiv preprint arXiv:2210.03493*, 2022. [2](#), [3](#)
- [15] Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models, may 2023. *arXiv preprint arXiv:2305.10601*, 2023.
- [16] Jieyi Long. Large language model guided tree-of-thought. *arXiv preprint arXiv:2305.08291*, 2023. [2](#), [3](#)
- [17] Daniel Kahneman. *Thinking, fast and slow*. macmillan, 2011. [2](#)
- [18] Michael Park, Erin Leahy, and Russell J Funk. Papers and patents are becoming less disruptive over time. *Nature*, 613(7942):138–144, 2023. [2](#), [3](#)
- [19] Alon Talmor, Oyvind Tafjord, Peter Clark, Yoav Goldberg, and Jonathan Berant. Leap-of-thought: Teaching pre-trained models to systematically reason over implicit knowledge. *Advances in Neural Information Processing Systems*, 33:20227–20237, 2020. [2](#)
- [20] Ewen Callaway. Cognitive science: Leap of thought, 2013. [2](#)
- [21] Keith J Holyoak, Paul Thagard, and Stuart Sutherland. Mental leaps: analogy in creative thought. *Nature*, 373(6515):572–572, 1995. [2](#), [5](#)
- [22] Carl Olson. The leap of thinking: A comparison of heidegger and the zen master dogen. *Philosophy Today*, 25(1):55, 1981.
- [23] Douglas Hofstadter. A review of mental leaps: analogy in creative thought. *AI Magazine*, 16(3):75–75, 1995.
- [24] Keith J Holyoak and Paul Thagard. *Mental leaps: Analogy in creative thought*. MIT press, 1996. [2](#), [5](#)
- [25] Joanna Kitto, David Lok, and Elizabeth Rudowicz. Measuring creative thinking: An activity-based approach. *Creativity Research Journal*, 7(1):59–69, 1994. [2](#)
- [26] Matthias Mölle, Lisa Marshall, Britta Wolf, Horst L Fehm, and Jan Born. Eeg complexity and performance measures of creative thinking. *Psychophysiology*, 36(1):95–104, 1999.
- [27] Hui Jiang and Qing-pu Zhang. Development and validation of team creativity measures: A complex systems perspective. *Creativity and Innovation Management*, 23(3):264–275, 2014. [2](#)
- [28] Wikimedia. Glossary of owarai terms. https://en.wikipedia.org/wiki/Glossary_of_owarai_terms, 2023. [2](#), [3](#), [28](#)

- [29] Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, Jiazheng Xu, Bin Xu, Juanzi Li, Yuxiao Dong, Ming Ding, and Jie Tang. Cogvlm: Visual expert for pretrained language models. *arXiv preprint arXiv: 2311.03079*, 2023. 2, 3, 7, 16
- [30] JungMi Lee. Mental leap. In Norbert M. Seel, editor, *Encyclopedia of the Sciences of Learning*, pages 2194–2194, Boston, MA, 2012. Springer US. 3, 4, 5
- [31] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 3, 8, 24
- [32] Lvmian Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023. 8, 24
- [33] Jay A Olson, Johnny Nahas, Denis Chmoulevitch, Simon J Cropper, and Margaret E Webb. Naming unrelated words predicts creativity. *Proceedings of the National Academy of Sciences*, 118(25):e2022340118, 2021. 3, 8, 2, 25
- [34] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning, 2023. 3, 6
- [35] Jun Chen, Deyao Zhu, Xiaoqian Shen, Xiang Li, Zechu Liu, Pengchuan Zhang, Raghuraman Krishnamoorthi, Vikas Chandra, Yunyang Xiong, and Mohamed Elhoseiny. Minigpt-v2: large language model as a unified interface for vision-language multi-task learning. *arXiv preprint arXiv:2310.09478*, 2023. 3, 6
- [36] Xiaoying Xing, Mingfu Liang, and Ying Wu. TOA: Task-oriented active VQA. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. 3
- [37] Zhan Ling, Yunhao Fang, Xuanlin Li, Tongzhou Mu, Mingu Lee, Reza Pourreza, Roland Memisevic, and Hao Su. Unleashing the creative mind: Language model as hierarchical policy for improved exploration on challenging problem solving. *arXiv preprint arXiv:2311.00694*, 2023. 3
- [38] Douglas Summers-Stay, Clare R Voss, and Stephanie M Lukin. Brainstorm, then select: a generative language model improves its creativity score. In *The AAAI-23 Workshop on Creative AI Across Modalities*, 2023.
- [39] Yuqian Sun, Xingyu Li, Jun Peng, and Ze Gao. Inspire creativity with oriba: Transform artists' original characters into chatbots through large language model. In *Adjunct Proceedings of the 2023 ACM International Joint Conference on Pervasive and Ubiquitous Computing & the 2023 ACM International Symposium on Wearable Computing*, pages 78–82, 2023.
- [40] Bhavya Bhavya, Jinjun Xiong, and Chengxiang Zhai. Cam: A large language model-based creative analogy mining framework. In *Proceedings of the ACM Web Conference 2023*, pages 3903–3914, 2023. 3
- [41] Hyeonsu B Kang, Xin Qian, Tom Hope, Dafna Shahaf, Joel Chan, and Aniket Kittur. Augmenting scientific creativity with an analogical search engine. *ACM Transactions on Computer-Human Interaction*, 29(6):1–36, 2022. 3
- [42] Tom Hope, Ronen Tamari, Daniel Hershcovich, Hyeonsu B Kang, Joel Chan, Aniket Kittur, and Dafna Shahaf. Scaling creative inspiration with fine-grained functional aspects of ideas. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pages 1–15, 2022.
- [43] Senwei Liang, Zhongzhan Huang, and Hong Zhang. Stiffness-aware neural network for learning hamiltonian systems. In *International Conference on Learning Representations*, 2021.
- [44] Zhongzhan Huang, Senwei Liang, Hong Zhang, Haizhao Yang, and Liang Lin. On fast simulation of dynamical system with neural vector enhanced numerical solver. *Scientific Reports*, 13(1):15254, 2023. 3
- [45] Ben Swanson, Kory Mathewson, Ben Pietrzak, Sherol Chen, and Monica Dinalescu. Story centaur: Large language model few shot learning as a creative writing tool. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 244–256, 2021. 3
- [46] Tuhin Chakrabarty, Vishakh Padmakumar, and He He. Help me write a poem: Instruction tuning as a vehicle for collaborative poetry writing. *arXiv preprint arXiv:2210.13669*, 2022.
- [47] Tongshuang Wu, Ellen Jiang, Aaron Donsbach, Jeff Gray, Alejandra Molina, Michael Terry, and Carrie J Cai. Promptchainer: Chaining large language model prompts through visual programming. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts*, pages 1–10, 2022.
- [48] Piotr Mirowski, Kory W Mathewson, Jaylen Pittman, and Richard Evans. Co-writing screenplays and theatre scripts with language models: Evaluation by industry professionals. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–34, 2023.
- [49] Hai Dang, Sven Goller, Florian Lehmann, and Daniel Buschek. Choice over control: How users write with large language models using diegetic and non-diegetic prompting. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–17, 2023. 3
- [50] Kim Binsted, Anton Nijholt, Oliviero Stock, Carlo Strapparava, G Ritchie, R Manurung, H Pain, Annalu Waller, and D O'Mara. Computational humor. *IEEE intelligent systems*, 21(2):59–69, 2006. 3
- [51] Dafna Shahaf, Eric Horvitz, and Robert Mankoff. Inside jokes: Identifying humorous cartoon captions. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1065–1074, 2015. 3
- [52] Kohtaro Tanaka, Hiroaki Yamane, Yusuke Mori, Yusuke Mukuta, and Tatsuya Harada. Learning to evaluate humor in memes based on the incongruity theory. In *Proceedings of the Second Workshop on When Creative AI Meets Conversational AI*, pages 81–93, 2022.
- [53] Haojie Xu, Weifeng Liu, Jiangwei Liu, Mingzheng Li, Yu Feng, Yasi Peng, Yunwei Shi, Xiao Sun, and Meng

- Wang. Hybrid multimodal fusion for humor detection. In *Proceedings of the 3rd International on Multimodal Sentiment Analysis Workshop and Challenge*, pages 15–21, 2022.
- [54] Chengxin Chen and Pengyuan Zhang. Integrating cross-modal interactions via latent representation shift for multimodal humor detection. In *Proceedings of the 3rd International on Multimodal Sentiment Analysis Workshop and Challenge*, pages 23–28, 2022.
- [55] Vijay Kumar, Ranjeet Walia, and Shivam Sharma. Deephumor: a novel deep learning framework for humor detection. *Multimedia Tools and Applications*, 81(12):16797–16812, 2022.
- [56] Jiaming Wu, Hongfei Lin, Liang Yang, and Bo Xu. Mumor: A multimodal dataset for humor detection in conversations. In *Natural Language Processing and Chinese Computing: 10th CCF International Conference, NLPCC 2021, Qingdao, China, October 13–17, 2021, Proceedings, Part I* 10, pages 619–627. Springer, 2021.
- [57] Dan Ofer and Dafna Shahaf. Cards against ai: Predicting humor in a fill-in-the-blank party game. *arXiv preprint arXiv:2210.13016*, 2022.
- [58] Binzhu Xie, Sicheng Zhang, Zitang Zhou, Bo Li, Yuanhan Zhang, Jack Hessel, Jingkang Yang, and Ziwei Liu. Funqa: Towards surprising video comprehension. *arXiv preprint arXiv:2306.14899*, 2023. 3
- [59] EunJeong Hwang and Vered Shwartz. Memecap: A dataset for captioning and interpreting memes. *arXiv preprint arXiv:2305.13703*, 2023.
- [60] Jonathan B Evans, Jerel E Slaughter, Aleksander PJ Ellis, and Jessi M Rivin. Gender and the evaluation of humor at work. *Journal of Applied Psychology*, 104(8):1077, 2019.
- [61] Camilla Vásquez and Erhan Aslan. “cats be outside, how about meow”: Multimodal humor and creativity in an internet meme. *Journal of Pragmatics*, 171:101–117, 2021. 3
- [62] Miriam Amin and Manuel Burghardt. A survey on approaches to computational humor generation. In *Proceedings of the The 4th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 29–41, 2020. 3
- [63] Hang Zhang, Dayiheng Liu, Jiancheng Lv, and Cheng Luo. Let’s be humorous: Knowledge enhanced humor generation. *arXiv preprint arXiv:2004.13317*, 2020.
- [64] Nabil Hossain, John Krumm, Tanvir Sajed, and Henry Kautz. Stimulating creativity with funlines: A case study of humor generation in headlines. *arXiv preprint arXiv:2002.02031*, 2020.
- [65] Alessandro Valitutti, Hannu Toivonen, Antoine Doucet, and Jukka M Toivanen. “let everything turn well in your wife”: generation of adult humor using lexical constraints. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 243–248, 2013.
- [66] Tanishq Chaudhary, Mayank Goel, and Radhika Mamidi. Towards conversational humor analysis and design. *arXiv preprint arXiv:2103.00536*, 2021. 3
- [67] Olga Popova and Petra Dadić. Does ai have a sense of humor? clef 2023 joker tasks 1, 2 and 3: using bloom, gpt, simplet5, and more for pun detection, location, interpretation and translation. *Proceedings of the Working Notes of CLEF*, 2023. 3
- [68] Dushyant Singh Chauhan, Gopendra Vikram Singh, Asif Ekbal, and Pushpak Bhattacharyya. Mhadig: A multilingual humor-aided multiparty dialogue generation in multimodal conversational setting. *Knowledge-Based Systems*, 278:110840, 2023. 3
- [69] Shraman Pramanick, Aniket Roy, and Vishal M Patel. Multimodal learning using optimal transport for sarcasm and humor detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3930–3940, 2022. 3
- [70] Jie Huang and Kevin Chen-Chuan Chang. Towards reasoning in large language models: A survey. *arXiv preprint arXiv:2212.10403*, 2022. 3
- [71] OpenAI. Gpt-4 technical report, 2023. 4, 6, 7
- [72] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. 4, 5, 20, 29
- [73] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023. 5, 6
- [74] Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. Glm: General language model pretraining with autoregressive blank infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 320–335, 2022. 6, 7
- [75] Ryuichiro Hataya, Han Bao, and Hiromi Arai. Will large-scale generative models corrupt future datasets? In *ICCV*, 2023. 6, 27
- [76] Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Yarin Gal, Nicolas Papernot, and Ross Anderson. Model dementia: Generated data makes models forget. *arXiv preprint arXiv:2305.17493*, 2023. 6, 27
- [77] Wenliang Dai, Junnan Li, and et al. Instructblip: Towards general-purpose vision-language models with instruction tuning. *arXiv preprint arXiv:2305.06500*, 2023. 7
- [78] Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Jingkang Yang, and Ziwei Liu. Otter: A multi-modal model with in-context instruction tuning. *arXiv preprint arXiv:2305.03726*, 2023. 7
- [79] Baichuan. Baichuan 2: Open large-scale language models. *arXiv preprint arXiv:2309.10305*, 2023. 7
- [80] Jack Hessel, Ana Marasović, Jena D. Hwang, Lillian Lee, Jeff Da, Rowan Zellers, Robert Mankoff, and Yejin Choi. Do androids laugh at electric sheep? Humor “understanding” benchmarks from The New Yorker Caption Contest. In *Proceedings of the ACL*, 2023. 6, 16
- [81] Kalervo Järvelin and Jaana Kekäläinen. Cumulated gain-based evaluation of ir techniques. *ACM Transactions on Information Systems (TOIS)*, 20(4):422–446, 2002. 6

- [82] Filip Radlinski and Nick Craswell. Comparing the sensitivity of information retrieval metrics. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, pages 667–674, 2010. [6](#) [16](#)
- [83] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2022. [7](#)
- [84] Zhongzhan Huang, Pan Zhou, Shuicheng YAN, and Liang Lin. Scalelong: Towards more stable training of diffusion model via scaling network long skip connection. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. [8](#) [24](#)
- [85] Changhao Shi, Haomiao Ni, Kai Li, Shaobo Han, Mingfu Liang, and Martin Renqiang Min. Exploring compositional visual generation with latent classifier guidance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 853–862, 2023. [8](#) [24](#)
- [86] Kenes Beketayev and Mark A Runco. Scoring divergent thinking tests by computer with a semantics-based algorithm. *Europe's journal of psychology*, 12(2):210, 2016. [8](#) [2](#) [25](#)
- [87] Jiaxi Tang and Ke Wang. Ranking distillation: Learning compact ranking models with high performance for recommender system. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 2289–2298, 2018. [16](#)
- [88] Koustuv Sinha, Robin Jia, Dieuwke Hupkes, Joelle Pineau, Adina Williams, and Douwe Kiela. Masked language modeling and the distributional hypothesis: Order word matters pre-training for little. *arXiv preprint arXiv:2104.06644*, 2021. [20](#)
- [89] Julian Salazar, Davis Liang, Toan Q Nguyen, and Katrin Kirchhoff. Masked language model scoring. *arXiv preprint arXiv:1910.14659*, 2019.
- [90] Yonatan Bitton, Gabriel Stanovsky, Michael Elhadad, and Roy Schwartz. Data efficient masked language modeling for vision and language. *arXiv preprint arXiv:2109.02040*, 2021. [20](#)
- [91] Steven Bird, Ewan Klein, and Edward Loper. *Natural language processing with Python: analyzing text with the natural language toolkit.* ” O'Reilly Media, Inc.”, 2009. [21](#)
- [92] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014. [25](#)
- [93] Jamshed Memon, Maira Sami, Rizwan Ahmed Khan, and Mueen Uddin. Handwritten optical character recognition (ocr): A comprehensive systematic literature review (slr). *IEEE Access*, 8:142642–142668, 2020. [28](#)
- [94] Minghao Li, Tengchao Lv, Jingye Chen, Lei Cui, Yijuan Lu, Dinei Florencio, Cha Zhang, Zhoujun Li, and Furu Wei. Trocr: Transformer-based optical character recognition with pre-trained models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 13094–13102, 2023. [28](#)
- [95] Patrick Pérez, Michel Gangnet, and Andrew Blake. Poisson image editing. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, pages 577–582. 2023. [28](#)
- [96] Felix Stahlberg. Neural machine translation: A review. *Journal of Artificial Intelligence Research*, 69:343–418, 2020. [28](#)
- [97] Shuheng Yang, Yuxin Wang, and Xiaowen Chu. A survey of deep learning techniques for neural machine translation. *arXiv preprint arXiv:2002.07526*, 2020. [28](#)
- [98] Yuexiang Zhai, Shengbang Tong, Xiao Li, Mu Cai, Qing Qu, Yong Jae Lee, and Yi Ma. Investigating the catastrophic forgetting in multimodal large language models. *arXiv preprint arXiv:2309.10313*, 2023. [29](#)
- [99] Guile Wu, Shaogang Gong, and Pan Li. Striking a balance between stability and plasticity for class-incremental learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1124–1133, 2021. [29](#)
- [100] Zhongzhan Huang, Mingfu Liang, Senwei Liang, and Wei He. Altersgd: Finding flat minima for continual learning by alternative training. *arXiv preprint arXiv:2107.05804*, 2021.
- [101] Mingfu Liang, Jiahuan Zhou, Wei Wei, and Ying Wu. Balancing between forgetting and acquisition in incremental subpopulation learning. In *European Conference on Computer Vision*, pages 364–380. Springer, 2022. [29](#)

Let's Think **Outside** the Box: Exploring Leap-of-Thought in Large Language Models with Multimodal Humor Generation

—Appendix—



Contents

A Introduction of Appendix	2
B Highlight Perspective	2
B.1. The Novelty of Our Paper	2
B.2. The Proposed CLoT is not Tailored for Humor Generation	2
C More Experimental Results	3
C.1. The Performance in Various Languages	3
C.2. The Diversity Responses of Oogiri Game by LLMs	4
C.3. More Examples for Oogiri-style Humor Generation	5
D The Construction of Oogiri-GO dataset	13
D.1. Online Data Collection	13
D.2 Machine Screening by LLM	15
D.3 Manual Screening	15
E Experimental Details	16
E.1. The Details of Implementation Details	16
E.2. The Details of Instruction Templates	16
E.3. The Details of Tuning by LoRA	20
E.4. The Extraction of Weakly-associated Conditions	21
E.5. The Details of User Study	23
F. The Details of other Creative Tasks	24
F.1. The Details of Cloud Guessing Game (CGG)	24
F.2. The Details of Divergent Association Task (DAT)	25
G The Analysis for Self-Refinement	26
G.1. The Discussion for the Round of Self-Refinement	26
G.2 Self-Refinement doesn't Cause a Performance Collapse in CLoT	27
H Further Discussions	28
H.1. The Oogiri-GO Dataset doesn't Include English IT2T Type Data	28
H.2 Fine-tuning Directly on Oogiri-GO is Hard to Achieve Good LoT Ability	28
H.3 How to Further Enhance CLoT?	29

A. Introduction of Appendix

The appendix is structured as follows. In Appendix B, we initially provide a detailed summary of the novelty in our paper and emphasize that our proposed CLoT is not tailored for humor generation. Instead, it focuses on the Leaf-of-Thought capability of large language models. In Appendix C, we further demonstrate the superiority of CLoT through various aspects. Firstly, we showcase the performance of CLoT in each language, i.e., English, Chinese, and Japanese respectively, emphasizing its versatility across languages. Following that, to illustrate CLoT's impact on enhancing creativity, we present its ability to generate diverse creative solutions for the same Oogiri game data sample. Lastly, we provide additional generated humor responses of various types of Oogiri games for different LLMs. Appendix D outlines the construction details of the Oogiri-GO dataset, encompassing the data collection process and both machine and human-driven filtering processes. In Appendix E, we meticulously detail main experiments presented in this paper, while Appendix F provides a comprehensive description of other creative tasks. Furthermore, Appendix G discusses two pivotal issues during the self-refinement phase, namely the round of refinement and potential performance collapse. Finally, in Appendix H, we address noteworthy concerns about the leaf-of-thought through a series of discussions.

B. Highlight Perspective

B.1. The Novelty of Our Paper.

The novelty of this paper can be summarized as follows:

- (**Pioneering**) To the best of our knowledge, we are the first to profoundly explore the Leap-of-Thought (LoT) capability in multimodal large language models (LLMs). This involves challenging LLMs to think outside the box, a non-sequential thinking skill equally crucial alongside popular sequential thinking abilities, such as Chain-of-Thought based methods [2, 3, 14–16]. The LoT ability serves as a cornerstone for creative exploration and discovery in LLMs.
- (**Large-scale Creative Dataset**) Given the scarcity of large-scale datasets for investigating creativity in the current community and the inherent challenges in collecting creative data (refer to Appendix G.1), in this paper, we identify the Oogiri game as an ideal platform for exploring the LoT ability of LLMs (refer to Appendix B.2), and gather over 130,000 creative data samples about Oogiri game, forming a large-scale creative dataset named Oogiri-GO.
- (**Novel Paradigm for Improving LoT**) Our experiments and analysis reveal that existing LLMs struggle to evoke LoT ability solely relying on their intrinsic reasoning abilities and extensive prior knowledge. Therefore, we propose a novel Creative Leap-of-Thought (CLoT) paradigm, employing associative instruction tuning and explorative self-refinement to significantly enhance LLMs' LoT ability. Further experiments demonstrate the effectiveness and versatility of the proposed CLoT across various creative tasks.

B.2. The Proposed CLoT is not Tailored for Humor Generation

In this paper, our primary focus is on exploring the Leap-of-Thought (LoT) capability of large language models, a crucial cognitive skill akin to Chain-of-Thought [2, 3, 14–16], rather than humor generation per se. The selection of the Oogiri game as the humor generation task in this study is justified on following three main grounds:

(1) **The Oogiri game serves as an ideal platform for investigating the LLM's LoT ability.** As discussed in the related works section, the Oogiri game aligns well with the characteristics of LoT, demanding players to think creatively outside the box in response to multimodal information. Moreover, the three primary types within the Oogiri game (including I2T, T2T, and IT2T) also align with the input and output types of multimodal LLMs. Hence, the Oogiri game proves highly suitable for exploring the LoT capability of LLMs;

(2) **The Oogiri game boasts a substantial corpus of manually annotated creative data.** Due to its widespread popularity on the Internet, the game attracts a large user base generating creative human responses which can constitute an extensive dataset for LoT exploration;

(3) **The Oogiri game facilitates visualization for measuring LoT ability.** Unlike most association-related cognitive tests [33, 86], the Oogiri game, being inherently a text and image multimodal task, lends itself readily to visualizing LoT ability in a clear and interesting format. Furthermore, this method of showcasing LoT ability, coupled with the amusement factor of the Oogiri game, encourages a broader audience to participate in LoT measurement, ensuring the accuracy of LoT analysis.

C. More Experimental Results

In this section, we present additional experimental results to demonstrate the effectiveness of CLoT in enhancing LLM’s LoT ability. This includes CLoT’s performance across distinct languages (Appendix C.1), diverse responses to single images (Appendix C.2), and increased Oogiri-style humor generation (Appendix C.3).

C.1. The Performance in Various Languages

The results shown in Table 2 in the main text consider all languages together. In this section, we present the performance of all multimodal LLMs individually across different languages, including english (EN), chinese (CN) and japanese (JP). The results are illustrated in Fig. 11, with Qwen-VL serving as the primary baseline model. It is evident that our proposed CLoT not only significantly enhances the performance of the baseline model but also surpasses other advanced LLMs across various metrics in different languages. This experiment once again underscores the effectiveness of the proposed CLoT.

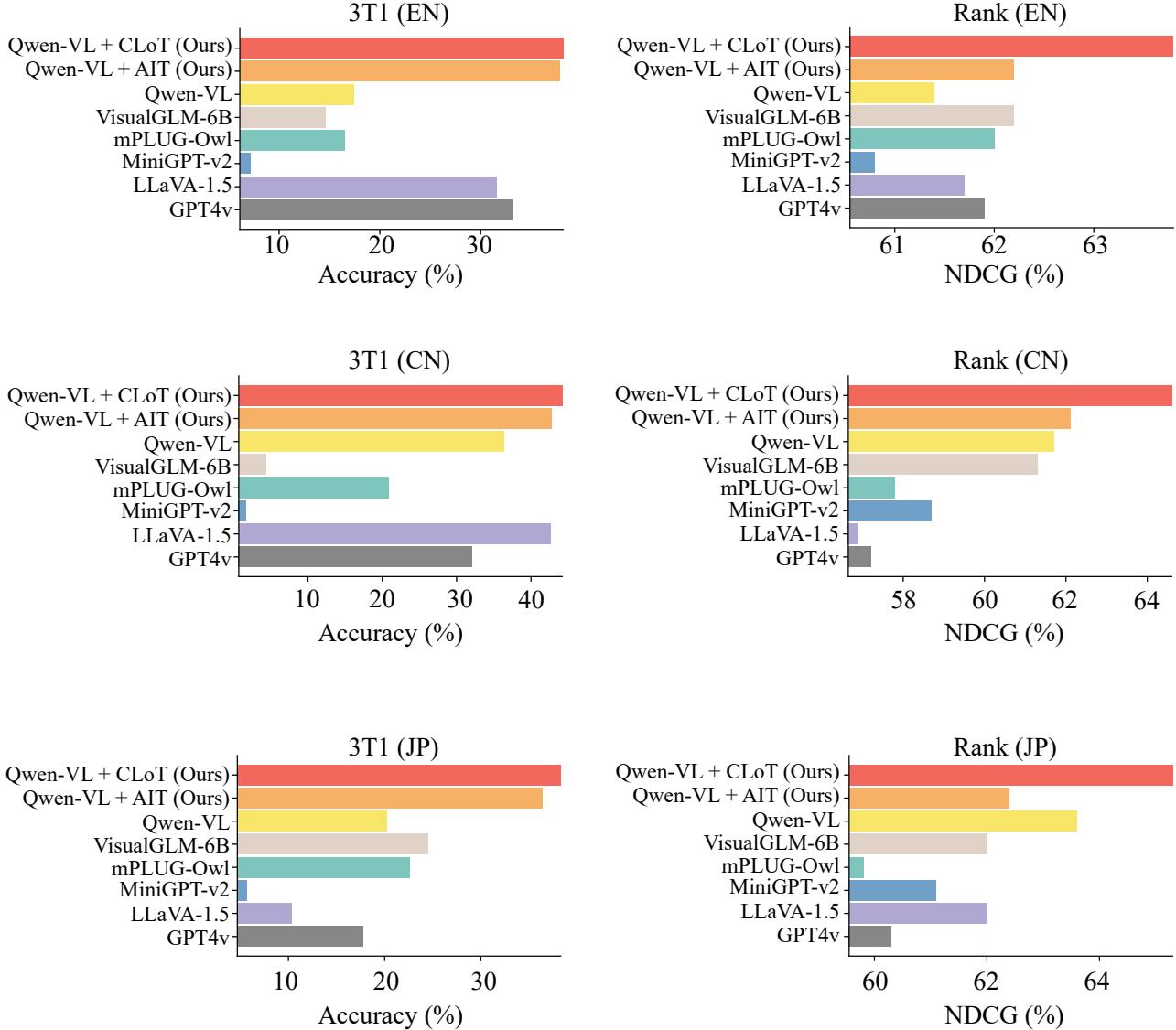


Figure 11. The performance of different LLM for Oogiri game in various languages.

C.2. The Diversity Responses of Oogiri Game by LLMs

To illustrate the creativity of CLoT, we present the diverse responses of Qwen-VL₊CLoT on the same Oogiri game samples, as depicted in Fig. 12. In order to impartially showcase CLoT’s creativity, we opt for Chinese Oogiri which has a moderate dataset size in Oogiri-GO, as the limited dataset of English Oogiri may fail to fully capture the advantages of CLoT, and the largest dataset of Japanese Oogiri may overly emphasize CLoT’s strengths. To facilitate comprehension for readers of different languages, Fig. 12 simultaneously displays the English translations of the Chinese Oogiri responses. However, due to cultural factors and other constraints, the translated content may not entirely convey the intended meaning of the Chinese Oogiri responses. Nevertheless, the diversity of responses in Fig. 12 underscores CLoT’s ability to engage in divergent thinking and approach challenges from multiple perspectives, showcasing its capacity to think outside the box.

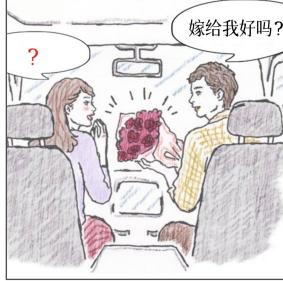
	(1) 你又想偷吃我的猫粮! @ You're trying to sneak a bite of my cat food again, aren't you? (2) 在路边被一车子压扁之后..... @ After getting squashed by a car on the roadside... (3) 猫版大鹏展翅。 @ The cat's version of a majestic spread of wings, like a giant roc taking flight. (4) 你这猫也太会摆姿势了吧! @ Oh, your cat really knows how to strike a pose! (5) 早上好! 今天我来教你如何用爪子打招呼。 @ Good morning! Today, I'll teach you how to greet with a pawshake.
	(1) 猎豹扑向鹰的瞬间被鹰带飞。 @ The moment the cheetah pounced, it got scooped up by the eagle and taken for a ride. (2) 别追了，我已经把你孩子叼走了。 @ Stop chasing me. I've already grabbed your kid. (3) 你先别飞，我还有话要说。 @ Don't take off just yet; I've got something to say. (4) 奢官你还没付钱！ @ Hey, you haven't paid yet, sir! (5) 每次做坏事都要替你擦屁股。 @ I always have to clean up after you every time you mess up.
	(1) 终于知道为什么他们不说话了。 @ Finally figured out why those two aren't barking. (2) 你们是不是又把家具都拆了。 @ Did you two go and take apart all the furniture again? (3) 不让你们吃巧克力，就只能这样了。 @ If I can't let you have chocolate, this is the only option. (4) 最近又胖了。 @ We've packed on some pounds again lately. (5) 说好的一起藏猫猫呢？ @ We're all hidden and waiting, but you haven't come to find us yet!
	(1) 什么?! 现在要去宠物医院做绝育手术!? @ What?! Going to the pet hospital now for a neutering surgery!? (2) 哇，真的有条大鱼！ @ Wow, that's a really big fish! (3) 司机师傅，快看路啊！ @ Hey driver, watch the road! Hurry up! (4) 安全带是系在脖子上的吗？ @ Is the seatbelt supposed to be fastened around the neck? (5) 快！追上前面那辆车！ @ Hurry! Catch up with the car in front!

Figure 12. The diversity responses of proposed Creative Leap-of-Thought. “@” denotes English translations.

C.3. More Examples for Oogiri-style Humor Generation

We provide additional examples of humor generation for the multimodal multilingual LLMs mentioned in Table 2 (main text) to illustrate the effectiveness of CLoT. Fig. 13, 14 showcase responses on the task of Image&Text to Text in Chinese and Japanese, respectively. As English Oogiri data lacks Image&Text to Text samples, we do not display responses in English for this task (see Appendix H.1 for details). Fig. 15, 16, and 17 exhibit responses on Image to Text, while Fig. 18, 19, and 20 display responses on Text to Text. Finally, we supplement more Qwen-VL+CLoT responses on Chinese Oogiri samples in Fig. 21.

Image&Text to Text (CN)



> Man: Will you marry me?
> Woman: _____ ?

- > GPT4v: 你是指这束巧克力花束吗？因为我闻不出它有任何玫瑰的香味。 @ Are you referring to this bouquet of chocolate? Because I can't smell any rose fragrance from it.
- > LLaVA-1.5: 这位女士惊讶地发现它们实际上是伪装的胡萝卜。
@ The woman is surprised to see that they are actually carrots in disguise.
- > MiniGPT-v2: 你太擅长了！
@ You're so good at it!
- > mPLUG-Owl: 他居然能从一只小猫咪的嘴巴里吃下一块巧克力棒。
@ He can actually eat a chocolate bar from a kitten's mouth.
- > VisualGLM-6B: 这是什么？
@ What is this?
- > Qwen-VL: 哟，这个花是给我的吗？
@ Oh, is this flower for me?
- > Qwen-VL+CLoT (Ours): 老公别闹了。
@ Don't tease, hubby.

Image&Text to Text (CN)

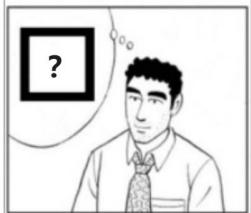


> Nobita Nobi: The island is right in front of us!
> Nobita Nobi: Hey! We've arrived at Treasure Island, you know!
> Takeshi Goda, Suneo Honekawa: _____

- > GPT4v: 这个漫画中的人物误将游乐场中的恐龙雕塑当成了真的恐龙，他紧张地大喊：“救命啊！恐龙来了！”然而在下一格里我们看到一位工作人员平静地回应：“别慌，那是假的。”
@ The character in this comic mistakenly took the dinosaur sculpture in the amusement park for a real dinosaur. Nervously, he shouted, "Help! The dinosaur is coming!" However, in the next panel, we see a staff member calmly responding, "Don't panic, it's fake."
- > LLaVA-1.5: 一个男人站在船上，指着坐在椅子上的女人。 @ A man is standing on a boat and pointing at a woman who is sitting in a chair.
- > MiniGPT-v2: 一个男人正在演奏陶笛，但它实际上是一个玩具，面前的一块假石头后面藏着一个怪物！ @ A man is playing an ocarina, but it's actually a toy with a monster hiding behind a fake rock in front of him!
- > mPLUG-Owl: 一个人走进一家酒吧，看到一个人在桌子上吃东西，他走过去问那个人：“为什么你的食物是放在这里的？”那人笑着回答：“因为我喜欢吃这里的食物。” @ A person walks into a bar and sees someone eating at a table. He approaches and asks, "Why is your food placed here?" The person laughs and replies, "Because I like the food here."
- > VisualGLM-6B: 本图是《名侦探柯南》中毛利小五郎和工藤新一的第一次对话。 @ This scene depicts the first conversation between Kogoro Mori and Shinichi Kudo in "Detective Conan".
- > Qwen-VL: 哟，我的眼镜呢？原来它被我丢在这里，让我来取回它吧！ @ Oh shoot, where are my glasses? Turns out I left them here. Let me grab 'em real quick!
- > Qwen-VL+CLoT (Ours): 好家伙！
@ Goodness gracious!

Figure 13. The responses of LLMs in Chinese Oogiri IT2T samples. “@” denotes English translations.

Image&Text to Text (JP)



> **Woman:** You've been waiting for a while, my apologies. Please, take your time and enjoy!

> **Man:** _____ ?



> **GPT4v:** この紳士はおそらく「サプライズセット」を注文したようで、彼の前には疑問符のついた枠が現れました。彼自身も考えているように、これは新しいトレンドの見えない美食なのかしら？

@ The gentleman seems to have ordered a 'surprise package,' as there is a box with a question mark in front of him. Even he is wondering if this is the latest trend in invisible cuisine.



> **LLaVA-1.5:** 男性が机の前に座っており、彼の後ろには女性が立っています。

@ A man is sitting at a desk with a woman standing behind him.



> **MiniGPT-v2:** お互いに訴え合いながらも、彼女たちはお互いに惹かれていて、肉の塩だけを食べるしかありません。@ They accuse each other but still enjoy, and all they can do is eat humble pie.



> **mPLUG-Owl:** もともと私は慈善晩会に参加しに来たのですが、結果的にはお見合いパーティーに招待されていることに気づきました！@ I came to attend a charity event, only to discover that I've been invited to a blind date party.



> **VisualGLM-6B:** 男性が皿の前に立ち、飲み物とサンドイッチを持っている様子で、お昼を楽しんでいるようです。

@ A man stands before a plate, holding a drink and a sandwich. It seems he is enjoying his lunch.



> **Qwen-VL:** 不思議ですね、なぜこのレストランの皿が四角いのでしょうか？

@ Weird, right? Why are the plates at this joint all square-shaped?



> **Qwen-VL+CLoT (Ours):** もし私がこれ以上食べ続けたら、本当におかがいいっぱいになってしまいます。

@ If I have another bite, I swear I'll pop.

Image&Text to Text (JP)



> **Woman:** _____ ?



> **GPT4v:** ああ、どうやら誰かが「父親を背負う」という成語をあまりにも文字通りに理解してしまったようですね！@ Oh, it looks like someone took the idiom 'carry your father on your back' a bit too literally!



> **LLaVA-1.5:** 一人の男性がトイレを抱えており、その後ろには女性が立っています。トイレには会話の吹き出しがあります。

@ A man is hugging a toilet with a woman standing behind him, and the toilet has a speech bubble.



> **MiniGPT-v2:** 放してあげるわよ、老頭子！

@ I will let you go, old man!



> **mPLUG-Owl:** 毎回スーパーでパンを見ると、なぜかパンが私を見ている気がします。@ Why is it that every time I see bread at the supermarket, it feels like the bread is staring back at me?



> **VisualGLM-6B:** トイレに座りながら抱き合う老夫婦のシーン。@ A senior couple sharing a tender moment, sitting together in the bathroom.



> **Qwen-VL:** 彼は老いたパートナーをぎゅっと抱きしめ、まるでトイレの中のように自由でした。

@ He holds his spouse close, as freely as if they are in the bathroom.



> **Qwen-VL+CLoT (Ours):** これは確かに大きな問題ですね！なぜあなたはお尻を下ろせないのか？

@ This is a big deal! Why can't you just sit your butt down?

Figure 14. The responses of LLMs in Japanese Oogiri IT2T samples. “@” denotes translations.

Image to Text (EN) 	<ul style="list-style-type: none"> > GPT4v: This cat seems to be practicing the ancient art of the innocent cat face—perfectly concealing how it just turned the vase into a modern art masterpiece in pieces. > LLaVA-1.5: A white cat sits on the ground, with fragments scattered around, as if it's in the process of breaking a toy. > MiniGPT-v2: This is not a toilet brush, it's actually a hamster cage! > mPLUG-Owl: This morning, I woke up to find myself with a pale complexion, yellowed teeth, and disheveled hair. I didn't recognize myself at all. Turns out, I was abducted by aliens! > VisualGLM-6B: The cat looks a bit lost and dejected, as if it has misplaced one of its toy balls. > Qwen-VL: Why is the cat wearing a flower on its head and why is there a broken vase on the floor? Because it's the culprit of that vase! > Qwen-VL+CLoT (Ours): Why did you break my cat's nest?
Image to Text (EN) 	<ul style="list-style-type: none"> > GPT4v: It's really a bath class taught by a chef that turns soaking into an artistic delight. > LLaVA-1.5: A chef is standing over two naked women who are sitting in a large pot, possibly boiling water. > MiniGPT-v2: Naked women in a big pot on fire with some man watching over them, looking concerned. > mPLUG-Owl: "Why can't pigs go online?" "Because they'll get caught by the cats on the internet!" > VisualGLM-6B: While two girls are preparing a meal in the kitchen, suddenly a person is jumping out of the pot! > Qwen-VL: Why does the chef look so solemn? > Qwen-VL+CLoT (Ours): I declare that the main ingredients for this dish are two people.

Figure 15. The responses of LLMs in English Oogiri I2T samples.

Image to Text (CN)



> GPT4v: 这辆车看起来像是误解了“下水道”和“停车道”的区别，结果选择了个全新的“泊车方式”。
@ This car seems to have misunderstood the difference between "sewer" and "parking," opting for a whole new "parking style."



> LLaVA-1.5: 一辆白色货车被困在沟里，这似乎是动画片中货车试图逃离大勺子的场景。
@ A white van is stuck in a ditch, and it appears to be a scene from a cartoon where the van is trying to escape from a giant spoon.



> MiniGPT-v2: 谁能想到一辆白色货车停在水坑中央会引起如此大的骚动？@ Who would have thought a white van parked in the middle of a puddle could cause such a big commotion?



> mPLUG-Owl: 今天我穿着一只鞋，两只袜子，一双拖鞋，还有一只雨鞋。@ Today, I'm wearing one shoe, two socks, a pair of slippers, and even a rain boot.



> VisualGLM-6B: 一辆白色面包车在池塘里翻倒，似乎正在被水淹没。@ A white van flipped over in the pond, seemingly getting submerged in the water.



> Qwen-VL: 为什么司机不停车，而是把车开进水里呢？因为他想要一辆全新的丰田海拉克斯！
@ Why didn't the driver stop but instead drove the car into the water?
Because he wanted a brand new Toyota Hilux!



> Qwen-VL+CLoT (Ours): 停车不规范，亲人两行泪。
@ Improper parking, brings tears to loved ones.

Image to Text (CN)



> GPT4v: 在日常生活中，我们以为互相拧耳朵是孩子们的游戏，但这两位先生显然决定上一堂成人版“捉迷藏”的高级课程。
@ In everyday life, we often think that twisting each other's ears is a children's game, but these two gentlemen have clearly decided to attend an advanced course in adult hide-and-seek.



> LLaVA-1.5: 在这个黑色礼服和白色礼服的两位亚洲男子面对面的场景中，他们的嘴巴竟然变成了一个巨大的钥匙，让人不禁感到出乎意料和滑稽。@ In this scene of two Asian men facing each other in black and white formal attire, their mouths have unexpectedly transformed into a giant key, creating a surprising and amusing sight.



> MiniGPT-v2: 成龙是中国古代的间谍，他伪装成年长、睿智的自己，为正义而战，而甄子丹则伪装成年轻、缺乏经验的间谍，不小心给两人带来了麻烦。@ Jackie Chan was a spy in ancient china who disguised himself as his older, wiser self to fight for justice, while Donnie Yen pretended to be a younger, more inexperienced spy and accidentally caused trouble for both of them.



> mPLUG-Owl: 一个男人穿着女人的衣服在公园里散步，引来了围观者。@ A man walking in the park wearing women's clothing attracts onlookers.



> VisualGLM-6B: 两个男人面对面站着，穿着西装打领带，其中一个人在摸另一个的鼻子。@ Two men standing face to face in suits with ties, and one of them is touching the other's nose.



> Qwen-VL: 哎，你听说了吗？其实黑社会片里的大佬都是富二代。
@ Hey, did you hear? Turns out, the big shots in gangster movies are actually rich second-generation kids.



> Qwen-VL+CLoT (Ours): 别忘了，今天是疯狂星期四。
@ Don't forget, today is Crazy Thursday!

Figure 16. The responses of LLMs in Chinese Oogiri I2T samples. “@” denotes English translations.

Image to Text (JP)

> GPT4v: 人類の「牽引競争」は本当に体力を使いますね！ @ It seems like the human "pulling a car race" is really physically demanding!

> LLaVA-1.5: 一台警察車が草地に停まっており、2匹の猫がその警察車を監視しています。 @ A police car is parked on the grass, and two cats are keeping a watchful eye on it.

> MiniGPT-v2: 猫たちが窓辺に座って、電車がレールを走るのを眺めています。 @ Cats sitting on window sills watching the train move down the tracks.

> mPLUG-Owl: 昨夜、スイカを食べたら、今朝おなかが痛い結果になりました。 @ I had watermelon last night, and now my stomach hurts this morning.

> VisualGLM-6B: 一匹猫がパトカーのそばに立っていて、もう一匹は窓の上でそれを見て座っている。 @ One cat is standing next to a police car, while another cat is sitting on the windowsill, looking at it.

> Qwen-VL: 警察のおじさんがパトカーの後を追いかけると、2匹の猫は横で冷静に見ていて、まるで「もっと遅く、もっと遅く、私たちがついていけないよ！」と言っているかのようです。 @ As the police officer chased the police car, two cats watched calmly from the side, as if saying, "Take it slow, take it slow, we can't keep up!"

> Qwen-VL+CLoT (Ours): これって新しい耕地のやり方？ @ Is this a new form of fieldwork?

Image to Text (JP)

> GPT4v: 老紳士が請求書を読んでいると、彼は若い頃に隠していたチョコレートクッキー缶の秘密の貯金口座を見つけました。2人は口角が合わないほど笑い、彼らが始める予定のチョコレートの冒険の旅行を驚きで発見しました。 @ While reading the bills, the old man discovers the secret savings account hidden in the chocolate cookie jar from his youth. Both of them burst into laughter, delighted to embark on the chocolate adventure that awaits them.

> LLaVA-1.5: 一人のおじいさんと一人のおばあさんが一緒に携帯の画面を見ています。 @ An old man and an old woman are looking at a mobile phone screen together.

> MiniGPT-v2: ああ、あなたは私の財布の中身を知りたいのですね？わかりました、教えてあげましょう。 @ Oh, you want to know what's in my purse? Well, let me show you.

> mPLUG-Owl: 今朝目が覚めてみると、自分が一匹の小さな子猫に身体の下で圧迫されていることに気づきました！ @ This morning, I woke up and found myself being squished under a little kitty!

> VisualGLM-6B: この老夫婦はテーブルに座り、新聞を見ながら微笑んでいます。 @ The elderly couple is sitting at a table, looking at a paper, and smiling.

> Qwen-VL: おばあちゃんの歯はなぜあんなに白いのですか？ なぜなら、彼女は毎日歯を磨くからです。そして、おじいちゃんの歯は毎日皿を洗うのでとても黄色いからです。 @ Why are grandma's teeth so white? Because she brushes them every day. And why are grandpa's teeth so yellow? Because he brushes the dishes every day!

> Qwen-VL+CLoT (Ours): 退職金もなければ、年金もない、医療保険もないし、福祉も何もない、福祉も何もない… @ No pension, no retirement fund, no health insurance, no benefits, no welfare, no welfare, no welfare...

Figure 17. The responses of LLMs in Japanese Oogiri I2T samples. “@” denotes English translations.

<p>What are some magical rewards you've received after completing a survey?</p> <p>> GPT4v: A coupon that allows me to get a free piece of cheese on the moon!</p> <p>> LLaVA-1.5: A little teddy bear, a tiny chocolate book, or an adorable puppy toy.</p> <p>> MiniGPT-v2: The rewards could be some welcoming or favorite gifts, or even a small bonus.</p> <p>> mPLUG-Owl: A mysterious candy, a free little gift, or a discount coupon.</p> <p>> VisualGLM-6B: Various shapes, sizes, and colors of chocolate pieces, candy bars, or other small treats.</p> <p>> Qwen-VL: Flying unicorns, free tickets to Hawaii, a personal chef for a day, and a lifetime supply of chocolate.</p> <p>> Qwen-VL+CLoT (Ours): The survey reward is your own personal information.</p>	<p>What job has a 99% mortality rate if you stick with it?</p> <p>> GPT4v: A professional food taster for explosive flavors.</p> <p>> LLaVA-1.5: Place the cup on the table.</p> <p>> MiniGPT-v2: Bankers.</p> <p>> mPLUG-Owl: The answer isn't unique, but it could be understood as referring to a highly dangerous or risky profession, such as a soldier, firefighter, police officer, doctor, and so on.</p> <p>> VisualGLM-6B: Pilots.</p> <p>> Qwen-VL: Firefighters</p> <p>> Qwen-VL+CLoT (Ours): 99% of jobs.</p>
---	---

Figure 18. The responses of LLMs in English Oogiri T2T samples.

<p>她忍不住放了个屁，现在尴尬地沉默着，给她一句贴心的话吧。@ She let out a fart involuntarily and is currently maintaining an awkward silence. Share a comforting word with her.</p> <p>> GPT4v: 别担心，人人都放屁，就当是给房间免费增加了一点自然香气吧！ @ Don't worry, everyone farts. Just consider it as a free natural air freshener for the room!</p> <p>> LLaVA-1.5: 哇，你真是个屁糗大师！ @ Wow, you're a master of awkward farts!</p> <p>> MiniGPT-v2: 她不知道我喜欢吃蛋糕，如果你喜欢蛋糕，我会继续帮助你。 @ She doesn't know I like cakes. If you're into cakes, I'll keep assisting you.</p> <p>> mPLUG-Owl: 没关系，放屁是人类的本能，不用太自责。 @ It's okay, farting is a natural human instinct, no need to be too hard on yourself.</p> <p>> VisualGLM-6B: 你放了一个可爱的屁！这很自然、有趣和令人愉悦。 @ You let out a cute fart! It's natural, amusing, and quite delightful.</p> <p>> Qwen-VL: 哎呀，看来你的屁比你的声音还要响啊，哈哈！ @ Oh dear, it seems like your fart is even louder than your voice, haha!</p> <p>> Qwen-VL+CLoT (Ours): 你这屁，屁到我心里去了。 @ Your fart, it really struck a chord with me.</p>	<p>当被高高举起时，婴儿可能会想的事情。 @ What a baby might be thinking when lifted up high.</p> <p>> GPT4v: 嘿，这里的风景不错，但我希望他们不会忘记把我放下来。 @ Hey, the scenery here is nice, but I hope they don't forget to set me down.</p> <p>> LLaVA-1.5: 哇，我在空中！ @ Wow, I'm in the air!</p> <p>> MiniGPT-v2: 互动和玩耍。 @ Interaction and playfulness.</p> <p>> mPLUG-Owl: 想象自己在飞翔，或者想象自己在玩捉迷藏，或者想象自己在骑着滑板车。 @ Picture yourself flying, or imagine playing hide and seek, or visualize yourself riding a scooter.</p> <p>> VisualGLM-6B: 想到自己正在飞起来或飞行。 @ Imagine yourself lifting off and flying.</p> <p>> Qwen-VL: 哇，我怎么这么轻？我是不是变成了一只小鸟？ @ Wow, why am I so light? Did I turn into a little bird?</p> <p>> Qwen-VL+CLoT (Ours): 我想尿尿。 @ I have the urge to pee.</p>
---	--

Figure 19. The responses of LLMs in Chinese Oogiri T2T samples. “@” denotes translations.

<p>若者にとって、幸福とは何でしょうか？ @ For young people, what is happiness?</p> <p> > GPT4v: 若者にとって、幸福とは携帯電話にまだ50%の電力が残っていることを発見することであり、0%ではないことです。 @ For young people, happiness is finding that there's still 50% battery left on their phone, not 0%.</p> <p> > LLaVA-1.5: 若者にとって、幸福はアイスクリーミングと可愛らしい犬がそばにいることかもしれません。@ For young people, happiness might be a cup of ice cream and a cute puppy.</p> <p> > MiniGPT-v2: 私たちがこの問題を多角的に考えることができます。 @ We can consider this issue from multiple perspectives.</p> <p> > mPLUG-Owl: 若者にとって、幸福はおいしい夕食、素敵な映画、楽しいパーティー、または興奮する旅行かもしれません。 @ For young people, happiness could be a delicious dinner, a good movie, a fun party, or an exciting trip.</p> <p> > VisualGLM-6B: 幸福は、車、家、アパートまたは別荘を所有すること、あるいは安定した仕事、収入、そして良好な福祉を持つことを意味するかもしれません。@ Happiness might mean owning a car, a house, and a set of apartments or a villa, or having a stable job, income, and good benefits.</p> <p> > Qwen-VL: 若者にとって、幸福とは、まるでパンダのように黒と白が交互になった髪型を持っているかもしれません！@ Happiness for young people might be having hair that's black and white like a panda!</p> <p> > Qwen-VL+CLoT (Ours): 親からの「もう相手を見つけるべきだ」「もう結婚するべきだ」「もう子供を産むべきだ」という言葉を聞かなくともよくなりますが... @ No more hearing parents say, "It's time to find a partner"; "It's time to get married"; "It's time to have children"...</p>	<p>あなたが聞いた中で最も絶望的だと感じた一言は何ですか？ @ What's the most despairing statement you've ever heard?</p> <p> > GPT4v: 私の携帯に入っている全ての写真がなくなりました。エイリアンのセルфиーも含めてです。@ All the photos on my phone are gone, including the selfie I took with the alien.</p> <p> > LLaVA-1.5: 私のコンピュータがフリーズしてしまっています。うちの犬もどうすれば助けられるかわからないようです！@ My computer is frozen, and my dog has no idea how to rescue it!</p> <p> > MiniGPT-v2: 私は今、どこに走ってしまったのでしょうか。@ I have no idea where I've ended up now.</p> <p> > mPLUG-Owl: 私はすべて試してみましたが、どれも失敗しました。 @ I've tried everything, but it all failed.</p> <p> > VisualGLM-6B: なんでライオンはりんご好きじゃないんだろう？@ Why don't lions like eating apples?</p> <p> > Qwen-VL: 犬が私のピザを食べました。 @ The dog ate my pizza.</p> <p> > Qwen-VL+CLoT (Ours): 月曜日だし、仕事に行く時間だね。@ It's Monday and time to go to work.</p>
---	--

Figure 20. The responses of LLMs in Japanese Oogiri T2T samples. “@” denotes English translations.



Figure 21. The responses of Qwen-VL_{+CL_{oT}} in Chinese Oogiri I2T samples. “@” denotes English translations.

D. The Construction of Oogiri-GO dataset

In this section, we delve into the introduction of data collection and screening for Oogiri-GO. In Appendix D.1, we elucidate the origins of our dataset. Using Bokete as a case study, we expound on the rationale and essential code employed in data crawling. Additionally, Appendices D.2 and D.3 provide a detailed breakdown of the procedural steps for machine screening and manual screening, respectively.

D.1. Online Data Collection

We source Oogiri game data from the official Oogiri game platform, Bokete (<https://bokete.jp>), and other popular platforms, such as Twitter (<https://twitter.com>) and Weibo (<https://m.weibo.cn>) which also host some Oogiri-game-alike data. Through extensive data collection from different platforms, we gather over 200,000 unfiltered raw samples. Notably, the Bokete website stands out as the preeminent Oogiri game-dedicated platform on the Internet, characterized by the highest data volume and user engagement. Consequently, we select it as a representative case study, providing a comprehensive account of our data acquisition methodology.

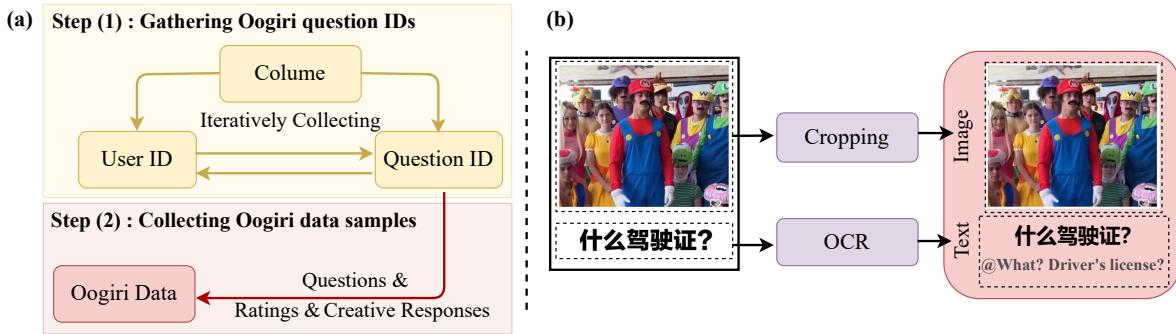


Figure 22. The processing of online data collection. (a) The crawling flow of Bokete website. (b) The image processing of Oogiri data.

Specifically, as illustrated in Fig. 22 (a), the primary approach to crawl the Bokete website involves two key steps:

(1) **Gathering Oogiri question IDs.** On the Bokete website, an Oogiri question refers to content uploaded by either the official site or users. It exists in the form of images, and even for T2T types, the textual content is embedded in the pictures. Players are tasked with generating creative responses based on these images. Initially, question IDs on the homepage and all corresponding user IDs under each question are preliminarily obtained from various columns, including “Best”, “Rising”, “Popular” and others. Notably, on each user’s homepage, their history of creative responses and ratings is available, allowing us to expand our pool of question IDs from these records. Through this iterative process, we progressively enlarge the pool of both question IDs and user IDs.

(2) **Collecting Oogiri data samples.** Subsequently, utilizing the gathered question IDs, all creative responses (answers) under a specific question are crawled to compile the Oogiri data. Simultaneously, we record their rating information for the subsequent training of the LLM’s discrimination ability in the CLoT framework.

To show the process of online data collection more clearly, we print the core code for both steps below.

```

1 def processing_url(url, page):
2     ''' The core code for step (1) Gathering Oogiri question IDs
3     Args:
4         url (str): basic URL of Bokete, e.g. https://bokete.jp/boke/legend
5         page (int): page number of basic URL, e.g. 1
6     '''
7     url = f'{url}?page={page}'
8     print('processing', url)
9
10    # get content of url
11    r = requests.get(url)
12    r.raise_for_status()
13

```

```

14 # parse the content and find all hyperlinks <a></a>
15 soup = BeautifulSoup(r.text, 'html.parser')
16 links = soup.find_all('a', href=True)
17 for link in links:
18     # find user id
19     if "/user/" in link['href']:
20         with open('user.txt', 'a') as f:
21             f.write(link['href'].split('/')[-1] + '\n')
22     # find Oogiri question id
23     if "/odai/" in link['href']:
24         with open('question.txt', 'a') as f:
25             f.write(link['href'].split('/')[-1] + '\n')

```

```

1 def processing_odai(odai, page):
2     ''' The core code for step (2) Collecting Oogiri data samples
3     Args:
4         odai (str): question ID, e.g. 6902364
5         page (int): page number of question URL, e.g. 1
6     '''
7     url = f'https://bokete.jp/odai/{odai}?page={page}'
8     print('processing', url)
9
10    # get content of url
11    r = requests.get(url)
12    r.raise_for_status()
13
14    # parse the content
15    soup = BeautifulSoup(r.text, 'html.parser')
16
17    # find the image url of the question
18    img = soup.find('a', href=f"/odai/{odai}").find('img')
19    link = 'https:' + img.get('src')
20
21    # find user id
22    links = soup.find_all('a', href=True)
23    for link in links:
24        if "/user/" in link['href']:
25            with open('user.txt', 'a') as f:
26                f.write(link['href'].split('/')[-1] + '\n')
27
28    # find all answers
29    texts = soup.find_all('a', class_='boke-text')
30    stars = soup.find_all('div', class_='boke-stars')
31    times = soup.find_all('div', class_='boke-information-label')
32    for text, star, t in zip(texts, stars, times):
33        with open('data.jsonl', 'a') as f:
34            f.write(json.dumps({
35                'id': text['href'].split('/')[-1], # id
36                'text': text.text, # content
37                'attitudes_count': star.text, # rate
38                'created_at': t.text, # creation time
39                'pics': { # image information
40                    'pid': odai, # question id
41                    'url': link, # image link
42                }
43            }, ensure_ascii=False) + '\n')

```

It's worth noting that, unlike the Bokete website, where questions and responses are distinct, on other platforms, data may have questions and responses combined in a single image, as illustrated in Fig. 22 (b)¹. In such cases, it is necessary to separate them to construct Oogiri data with a consistent format. Specifically, we utilize PaddleOCR² to recognize text within the main image. Subsequently, leveraging the positional information of the text, we employ image cropping to distinguish the image, resulting in an Oogiri sample with separated questions and responses.

D.2. Machine Screening by LLM

After collecting raw Oogiri data as outlined in Appendix D.1, it is important to acknowledge that the Oogiri game, being a comedy game, may involve responses with biases or other offensive humor. Additionally, since the Oogiri game allows participation from any Internet user, the potential for encountering such issues grows with the game's Internet dissemination. Therefore, effective filtering of the raw data becomes essential. Specifically, to prevent the inclusion of bias, violence, explicit content, offensive language, etc., we employ the multimodal language model Qwen-VL [1] as a checker for the initial screening of the raw data. This screening is performed by constructing safety-checking prompts. The design of the screening template for Qwen-VL is outlined as follows:

Does the image or text contain content related to <Label>? Or the combination of image and text shows the metaphor related to <Label>? If so, kindly respond with “Yes”; otherwise, respond with “No.”

Here is the text: <Text>

where the tag <Label> represents the keyword (e.g., violence, explicit content, offensive language, etc.) and <Text> denotes the responses from per collected Oogiri sample. To further enhance the effectiveness of safety-checking, we additionally employ the <Label> utilized by NudeNet³, which includes a substantial number of keywords associated with Not Safe For Work (NSFW) content. Finally, after machine screening, the number of samples is reduced to about 160,000.

D.3. Manual Screening

Although the majority of inappropriate content is successfully detected from the dataset through machine screening, some subtly metaphorical inappropriate content proves challenging to eliminate entirely. Consequently, aided by translation software, we conducted manual screening to further enhance the quality of the dataset. The criteria for manual screening are consistent with those used in machine screening, involving the removal of content related to the specific <Label>. The detailed process of manual screening is outlined as two parts.

(1) **Manual inspection.** Each sample in the dataset is meticulously examined to determine whether it contains content related to the specified keywords. The inspection involves a careful examination of both images and text to ensure accurate identification and labeling of inappropriate content.

(2) **Iterative screening.** To ensure accuracy and consistency in manual screening, we conduct two rounds of iterative manual screening. Each round involves different individuals to minimize the impact of subjective judgments and enhance the reliability of the dataset. Following manual screening, we successfully further reduce the presence of inappropriate content in the dataset, refining the sample count to more than 130,000. The introduction of manual screening contributes to ensuring a high-quality dataset and more sensitive detection of inappropriate content.

¹<https://m.weibo.cn/detail/4909366778531862>

²<https://github.com/PaddlePaddle/PaddleOCR>

³<https://github.com/notAI-tech/NudeNet>

E. Experimental Details

E.1. The Details of Implementation Details

Metrics. In this paper, we delve into experiments that include choice and ranking questions inspired by the humor benchmarks in [80]. Additionally, we conduct a user study to directly evaluate the effectiveness of humor generation, along with other creative tasks such as the Cloud Guessing Game (CGG) and the Divergent Association Task (DAT). Subsequently, individual metrics for each of these experiments will be provided.

(1) For the *choice questions*, we utilize classification accuracy as the evaluation metric. Specifically, the accuracy of LLMs is calculated by dividing the number of correctly answered questions by the total number of questions.

(2) For the *ranking questions*, we adopt the widely used ranking metric, i.e., Normalized Discounted Cumulative Gain (NDCG) [82]. We adopt top-1 accuracy as the positions at the top of rank lists are more significant in ranking scenarios [87].

(3) For the *user study*, we conduct a user survey, tallying the total number of votes received by various LLMs across different categories of Oogiri. Subsequently, we calculate the percentage of votes each LLM garnered in relation to its overall vote count for different types. see Appendix E.5 for more details.

(4) For the *other creative tasks*, we employed classification accuracy and average semantic distance (ASD) as metrics for the CGG and DAT tasks, respectively. ASD represents the average semantic distance of all test examples, where the semantic distance for each test example is calculated based on the ten words following the completion of each choice question. see Appendix F for more details.

Hyperparameters of Associable Instruction Tuning. For “Image” condition, it relies on the type of Oogiri game, e.g., being the image embeddings in I2T game and empty in T2T type. For the “condition” option, it’s set to empty with a probability of ρ_c , and otherwise is randomly set as one noun in “task-specific responses”. We set the value of ρ_c to 0.50. This setting is driven by the fact that training LLMs to perform associative generation assists in the remote association of self-refinement, and unconditionally controlling leap-of-thought generation is the capability we aim for the model to acquire.

Hyperparameters of Explorative Self-Refinement. During explorative remote association, we generate n weakly-associated conditions $\{C_i\}_{i=1}^n$. These conditions can either be empty with a probability $\rho = 0.5$ to give freedom to LLM, or uniformly randomly sampled from the noun set \mathcal{S} to enforce LLM to build connections between different concepts. Next, we add the condition C_i into user-input I , and feed I into the LLM to generate a humor candidate R_i . Repeating this process with different conditions C_i can generate a total of n candidates $\{R_i\}_{i=1}^n$. We set the value of n to 5, aiming not only to control the difficulty of ranking for reliable results but also to align with the number of options in the discrimination during associative instruction tuning. Then the LLM ranks these candidates by its discriminative ranking ability learned in Sec. 4.1 (main text). Next, it mixes the top-2 candidates with the ground truth responses, and selects the top-1 as the final response.

Here we not only incorporate ranking but also introduce a selection process to achieve explorative remote association. This decision is based on our experimental results, as demonstrated in the experimental section, indicating that the accuracy of LLMs tends to increase with a decrease in the number of choices for choice questions. Directly having LLMs choose an option from a pool of n candidates poses a significant challenge. Hence, we design a two-step process involving ranking followed by selection. Furthermore, the choice of selecting the Top-2 candidates from the ranking results is intended to ensure the accuracy of LLMs in completing choice questions.

Hyperparameters of Training. We use the official code of Qwen-VL [1] and CogVLM [29] for implementation and training. All models are trained utilizing 8 Nvidia A100 (40G) GPUs. The training and hyperparameters for each model are specified as follows.

(1) Qwen-VL+CLoT is trained using AdamW optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.95$, $eps = 1e^{-8}$. We set the learning rate to $1e^{-5}$ and use a weight decay of $1e^{-1}$. The training process uses a batch size of 64. LoRA in Qwen-VL has a rank of 64, a normalization parameter of 16, and a dropout rate of 0.05.

(2) CogVLM-17B+CLoT is trained using AdamW optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.95$, $eps = 1e^{-8}$. We set the learning rate to $1e^{-5}$ and use a weight decay of $5e^{-2}$. The training process uses a batch size of 128. LoRA in CogVLM-17B has a rank of 10, a normalization parameter of 1, and a dropout rate of 0.00.

E.2. The Details of Instruction Templates

After completing data collection and screening, the next step is to transform the collected Oogiri data into instruction-tuning data that will be utilized to train models. We design some LoT-oriented instruction templates to transform the Oogiri-GO dataset into instruction tuning data, and then train LLM to achieve associative generation and discrimination abilities. Our templates primarily comprise two components in Fig. 23: task-specific prompt and response. For different abilities, the templates need some special design. In this section, we will elaborate on the details of instruction templates for each task.

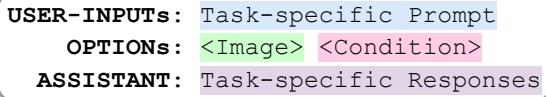


Figure 23. The LoT-oriented instruction templates.

Instruction Templates of Image to Text. Based on Fig. 23, we can categorize the instruction templates for Image to Text into the following four types:

Original Instruction

Based on the image, think of a sentence that is unexpected and humorous. Let's think outside the box. A satisfactory response is

Image: <Image>
<Response>

Instruction with Condition

Please carefully understand the image and give an answer that contains conditional words and is surprising and funny. Let's think outside the box. A surprising and funny answer containing conditional word is

Condition: <Condition>
Image: <Image>
<Response>

Instruction for Ranking

Please evaluate the degree of unexpected and humorous effect when each of the option contents is combined with the image.

Options:
A. <Content A>
B. <Content B>
C. <Content C>
D. <Content D>
E. <Content E>

Response Format: Please respond in the format of ranking the humorlessness of the options from high to low, for example, “1. A. xxx. 2. B. xxx. 3. C. xxx. 4. D. xxx. 5. E. xxx.”. Be sure to rank all five options.

Let's think outside the box. The result of ranking the options from most surprising and funny to least is

Image: <Image>
<Response>

Instruction for 3T1 Selection

Please select the option that, when combined with the image, creates an unexpected and humorous effect. Only one option meets the requirements.

Options:
A. <Content A>
B. <Content B>
C. <Content C>

Response Format: Please respond in the format of “Option id. Option content”, for example, “A. xxx”.

Let's think outside the box. The satisfactory option is

Image: <Image>
<Response>

where the tags <Image>, <Response>, <Condition> and <Content X> serve as placeholders for inserting the embeddings of visual image, the text response, the text condition, and the text option content. The condition of instruction with condition

is from the nouns of ground truth response, and the candidate options of instruction for ranking are from the Oogiri data with multiple answers. Besides, we illustrate the instructions for selection taking 3T1 selection as an example. For other types of selection instructions, only minor modifications to the number of options and quantifiers are needed.

We set the number of original instruction templates for each task to three. This decision is rooted in the precise requirements of the three tasks in Oogiri-GO, which is that we aim for LLMs to generate unexpected and humorous content based on given images or texts. Consequently, a large number of prompts is unnecessary for robust generalization across these tasks and our experiments support this observation. In the Image to Text task, we compare the effects of using three and fifty prompt templates and the fifty prompt templates are obtained by rewriting the original three templates using Qwen-14B [5]. The experimental results demonstrate that the number of prompt templates does not significantly impact the performance of LLMs.

Instruction Templates of Text to Text. Instruction templates for Text to Text are very similar to those for Image to Text, and can also be categorized into the following four types:

Original Instruction

Please carefully understand the provided question and come up with a surprising and humorous response.

Question: <Question>

Let's think outside the box. A satisfactory response is

<Response>

Instruction with Condition

Please carefully understand the question and give an answer that contains conditional words and is surprising and funny.

Question: <Question>

Let's think outside the box. A surprising and funny answer containing conditional word is

Condition: <Condition>

<Response>

Instruction for Ranking

Please evaluate the degree of unexpected and humorous effect when each of the option contents is combined with the question.

Question: <Question>

Options:

- A. <Content A>
- B. <Content B>
- C. <Content C>
- D. <Content D>
- E. <Content E>

Response Format: Please respond in the format of ranking the humorosity of the options from high to low, for example, “1. A. xxx. 2. B. xxx. 3. C. xxx. 4. D. xxx. 5. E. xxx.”. Be sure to rank all five options.

Let's think outside the box. The result of ranking the options from most surprising and funny to least is

<Response>

Instruction for 3T1 Selection

Please select the option that, when combined with the question, creates an unexpected and humorous effect. Only one option meets the requirements.

Question: <Question>

Options:

- A. <Content A>
- B. <Content B>
- C. <Content C>

Response Format: Please respond in the format of “Option id. Option content”, for example, “A. xxx”.

Let’s think outside the box. The satisfactory option is

<Response>

where the tag <Question> denotes the text question of Oogiri data.

Instruction Templates of Image&Text to Text. The instruction templates for Image&Text to Text are similar to those of the other two tasks, but due to the unique nature of Image&Text to Text, we incorporate a special character [MASK] into the templates. The instruction templates for Image&Text to Text are as follows:

Original Instruction

In this image, there are sections of text that need to be completed, and the content to fill in is denoted by [MASK]. Let’s think outside the box and complete the [MASK] to make the response unexpectedly funny. A satisfactory response is

Image: <Image>

<Response>

Instruction with Condition

In this image, there are sections of text that need to be completed, and the content to fill in is denoted by [MASK]. Let’s think outside the box and complete the [MASK] with a response that contains conditional words and is surprising and funny. A surprising and funny response containing conditional word is

Condition: <Condition>

Image: <Image>

<Response>

Instruction for Ranking

In this image, there are sections of text that need to be completed, and the content to fill in is denoted by [MASK]. Please evaluate the degree of unexpected and humorous effect when the options are the content of the [MASK].

Options:

- A. <Content A>
- B. <Content B>
- C. <Content C>
- D. <Content D>
- E. <Content E>

Response Format: Please respond in the format of ranking the humorlessness of the options from high to low, for example, “1. A. xxx. 2. B. xxx. 3. C. xxx. 4. D. xxx. 5. E. xxx.”. Be sure to rank all five options.

Let’s think outside the box. The result of ranking the options from most surprising and funny to least is

Image: <Image>

<Response>

Instruction for 3T1 Selection

In this image, there are sections of text that need to be completed, and the content to fill in is denoted by [MASK]. Please select the option that, creates an unexpected and humorous effect when being the content of the [MASK]. Only one option meets the requirements.

Options:

- A. <Content A>
- B. <Content B>
- C. <Content C>

Response Format: Please respond in the format of “Option id. Option content”, for example, “A. xxx”.

Let’s think outside the box. The satisfactory option is

Image: <Image>

<Response>

From Table 1 in the main text, it is evident that the data volume for the IT2T category is significantly lower compared to

the other two Oogiri game types. Consequently, there is a need to augment more data to enhance instruction tuning for IT2T. It is noteworthy that IT2T can be regarded as a form of masked language modeling (MLM) task [88–90]. Therefore, we contemplate achieving this objective by constructing MLM tasks for the I2T and T2T data types. The details are as follows:

Mask Instruction for I2T

Please carefully understand the provided image and complete the answer by replacing the [MASK] part to make the answer unexpectedly funny.

Answer: <Answer with [MASK]>

Let's think outside the box. The content of [MASK] is

Image: <Image>

<Response>

Mask Instruction for T2T

Please carefully understand the provided question and complete the answer by replacing the [MASK] part to make the answer unexpectedly funny.

Question: <Question>

Answer: <Answer with [MASK]>

Let's think outside the box. The content of [MASK] is

<Response>

Here, we probabilistically replace nouns or verb phrases from the Oogiri answers with [MASK], and use the replaced Oogiri answer as <Answer with [MASK]>.

E.3. The Details of Tuning by LoRA

LoRA [72] is a widely employed method for fine-tuning LLMs. It effectively reduces the number of trainable parameters by learning pairs of rank-decomposition matrices while maintaining the original weights in a frozen state. LoRA currently stands out as a superior adaptation method. Hence, we train LoRA for the LLMs with the associative instruction data.

The code snippets below illustrate the insertion points for LoRA during the training of Qwen-VL [1]. The first code snippet demonstrates how to insert LoRA into the textual module of Qwen-VL, while the second code snippet shows how to insert LoRA into the visual module of Qwen-VL.

```

1 # add LoRA to the textual module of Qwen-VL
2 QWenLMHeadModel(
3     transformer: QWenModel(
4         (wte): Embedding(151936, 4096)
5         (drop): Dropout(p=0.0, inplace=False)
6         (rotary_emb): RotaryEmbedding()
7         (h): ModuleList(
8             (0-31): 32 x QWenBlock(
9                 (ln_1): RMSNorm()
10                (attn): QWenAttention(
11                    (c_attn): Linear(in_features=4096, out_features=12288, bias=True) # + LoRA
12                    (c_proj): Linear(in_features=4096, out_features=4096, bias=False) # + LoRA
13                    (attn_dropout): Dropout(p=0.0, inplace=False)
14                )
15                (ln_2): RMSNorm()
16                (mlp): QWenMLP(
17                    (w1): Linear(in_features=4096, out_features=11008, bias=False)
18                    (w2): Linear(in_features=4096, out_features=11008, bias=False)
19                    (c_proj): Linear(in_features=11008, out_features=4096, bias=False)
20                )
21            )
22        )
23    .....

```

```

1 # add LoRA to the visual module of Qwen-VL
2 QWenLMHeadModel(
3     transformer): QWenModel(
4     .....
5     (visual): VisionTransformer(
6         (conv1): Conv2d(3, 1664, kernel_size=(14, 14), stride=(14, 14), bias=False)
7         (ln_pre): LayerNorm((1664,), eps=1e-06, elementwise_affine=True)
8         (transformer): TransformerBlock(
9             (resblocks): ModuleList(
10                 (0-47): 48 x VisualAttentionBlock(
11                     (ln_1): LayerNorm((1664,), eps=1e-06, elementwise_affine=True)
12                     (ln_2): LayerNorm((1664,), eps=1e-06, elementwise_affine=True)
13                     (attn): VisualAttention(
14                         (in_proj): Linear(in_features=1664, out_features=4992, \
15                             bias=True) # + LoRA
16                         (out_proj): Linear(in_features=1664, out_features=1664, \
17                             bias=True) # + LoRA
18                     )
19                     (mlp): Sequential(
20                         (c_fc): Linear(in_features=1664, out_features=8192, bias=True) # + LoRA
21                         (gelu): GELU(approximate='none')
22                         (c_proj): Linear(in_features=8192, out_features=1664, bias=True)
23                     )
24                 )
25             )
26         )
27         (attn_pool): Resampler(
28             (kv_proj): Linear(in_features=1664, out_features=4096, bias=False)
29             (attn): MultiheadAttention(
30                 (out_proj): NonDynamicallyQuantizableLinear(in_features=4096, \
31                     out_features=4096, bias=True) # + LoRA
32             )
33             (ln_q): LayerNorm((4096,), eps=1e-06, elementwise_affine=True)
34             (ln_kv): LayerNorm((4096,), eps=1e-06, elementwise_affine=True)
35         )
36         (ln_post): LayerNorm((4096,), eps=1e-06, elementwise_affine=True)
37     )
38 )
39 (lm_head): Linear(in_features=4096, out_features=151936, bias=False)
40 )

```

To investigate the appropriate insertion strategy for LoRA, we conduct three sets of associative instruction tuning experiments using Oogiri-GO I2T data. LoRA is inserted separately into the textual, visual, and both textual and visual modules of Qwen-VL. Experimental results indicate that, based on the 3T1 metric, the accuracy of LoRA insertion solely into the textual module (38.8) surpasses the performance of simultaneous insertion into both textual and visual modules (37.0), while the accuracy is lowest when LoRA is inserted only into the visual module (25.4). Therefore, we exclusively train LoRA in the textual module of Qwen-VL.

E.4. The Extraction of Weakly-associated Conditions

The core of explorative remote association is to prompt the LLM to generate a diverse array of creative responses under weakly-associated conditions. To implement this, we extract a set of object nouns from the text in the Oogiri-GO training data. In this section, we provide the details of the extraction of weakly-associated conditions.

We initiate by extracting a set of nouns from the responses in data of Oogiri-GO. For this, we employ NLTK [91], Jieba⁴, and Janome⁵ for various languages, considering the performance differences of different part-of-speech analysis tools across

⁴<https://github.com/fxsjy/jieba>

⁵<https://github.com/mocobeta/janome>

languages.

Specifically, we utilize NLTK, a suite of open source Python modules, data sets, and tutorials supporting research and development in natural language processing, to extract nouns from English text and the core code snippet is presented below:

```
1 from nltk import word_tokenize, pos_tag
2
3 def extract_nouns(text):
4     tokens = word_tokenize(text)
5     tagged_words = pos_tag(tokens)
6     conditions = [word for word, pos in tagged_words if pos.startswith('N')]
7     return conditions
```

For Chinese text, we utilize the widely adopted Chinese segmentation tool, Jieba, for part-of-speech analysis. The core implementation code is provided below:

```
1 import jieba.posseg as psg
2
3 def extract_nouns(text):
4     tagged_words = psg.cut(text)
5     conditions = [x.word for x in tagged_words if 'n' in x.flag]
6     return conditions
```

For Japanese text, we use the specialized Japanese morphological analysis engine, Janome, to extract nouns from the Oogiri-GO corpus. The core implementation code is as follows, where tag is 名詞 (nouns).

```
1 from janome.tokenizer import Tokenizer
2
3 tokenizer = Tokenizer()
4
5 def extract_nouns(text):
6     tokens = tokenizer.tokenize(text)
7     conditions = []
8     for token in tokens:
9         if tag == token.part_of_speech.split(',')[0]:
10             conditions.append(token.surface)
11     return conditions
```

Leveraging the three aforementioned part-of-speech analysis tools, we extract nouns from Oogiri-GO. Subsequently, we conducted deduplication on these nouns and performed manual quality screening to form the final noun set \mathcal{S} for weakly-associated condition sampling in CLoT.

E.5. The Details of User Study

We conduct a user preference study to directly verify the creativity of LLMs. Fig. 24 is the questionnaire homepage of user study where users can select the preferred language of questionnaire. Subsequently, we present choice questions in the preferred language based on user selections, and ask users to choose the most creative and humorous responses as shown in Fig. 25. Here we select six advanced LLMs to generate responses for a total of eighteen questions across three tasks (IT2T, I2T and T2T), and the six responses from six distinct LLMs are randomly permuted in options.

We conduct an extensive survey through the online survey platform⁶, ultimately collecting 154 valid questionnaires with 2772 votes. Within these collected questionnaires, we can calculate the proportion of times each LLM is selected for each question, as illustrated in Fig. 25 (bottom). Finally, we aggregate the total number of times each LLM is chosen across all Oogiri samples. The ratio of this sum to the overall number of selections among all LLMs signifies the user preference for each LLM.

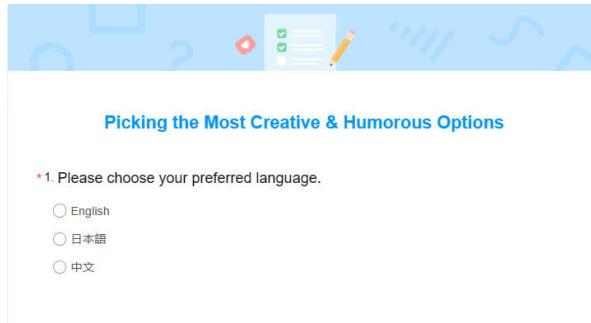


Figure 24. The questionnaire homepage of user study.

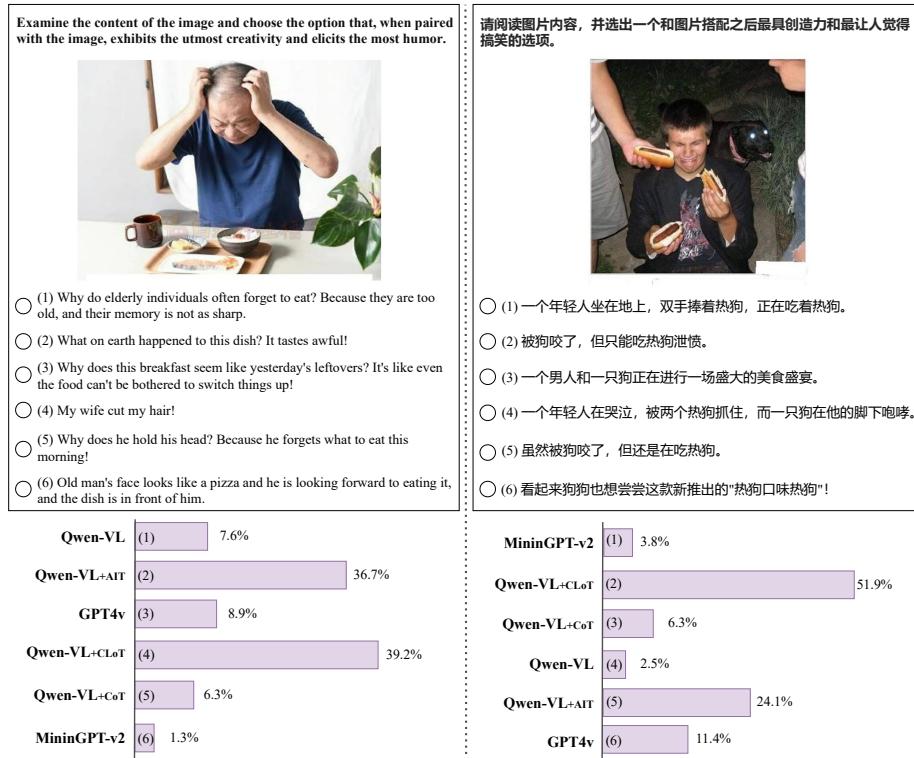


Figure 25. The questionnaire examples of user study.

⁶<https://www.wjx.cn>

F. The Details of other Creative Tasks

In this section, we provide experimental details and more examples of other creative tasks, including Cloud Guessing Game (CGG), Divergent Association Task (DAT).

F.1. The Details of Cloud Guessing Game (CGG)

The Cloud Guessing Game (CGG) is a task that requires LLM to identify the shapes of white clouds and then select the corresponding shapes from given options. The motivation behind CGG lies in the ever-changing shapes of white clouds and creative thinking can associate specific shapes with different cloud formations, as illustrated in Fig. 9 (c-d) in the main text. Therefore, the classification performance of cloud shapes in CGG can, to some extent, analyze the LLM’s LoT ability.

Specifically, the data construction process of CGG is as follows: First, we consider four categories—cat, human, and giraffe. We selected unambiguous and categorically distinct images for each category from the Internet. Taking the example of a cat, we chose the image shown in Fig. 9 (a) in main text, then use Photoshop carefully to create a mask of the cat as shown in Fig. 9 (b) in main text. Subsequently, employing the control diffusion model [31, 32, 84, 85] with the following prompt, we generated a cat-like white cloud as depicted in Fig. 9 (c) in main text.

(Prompt for white cloud generation) masterpiece, best quality, white cloud++, Stratus cloud, Altostratus cloud, Cirrus cloud, blue sky, light rays, sharp focus, HDR, UHD, 8K, masterpiece, Highly detailed, extreme detail detail, reality, realistic light, real, physics, reality, photo reality, Deconstruction

Continuously generated through the control diffusion model, we manually screen until we identify 30 unambiguous and difficulty challenging white cloud images for each category. The difficulty is adjusted by the “controlnet_scale,” a coefficient used to control the intensity of mask control. A higher value implies a stronger correlation between the generated images and masks, resulting in lower difficulty. Finally, employing the instruction template from Fig. 6 (a) in the main text, we construct choice questions for 4T1, with options randomly arranged from both the ground truth category and three words sampled randomly from the unrelated word set [‘chair’, ‘cup’, ‘sing’, ‘jump’, ‘rap’, ‘basketball’, ‘computer’, ‘egg’, ‘phone’, ‘house’, ‘lamp’, ‘shoes’], ensuring each question’s validity and clarity through manual verification.

Finally, we constructed three choice questions for each white cloud image. The various LLMs are instructed to choose the option containing the word that best resembled the shape of the given white cloud. In the experimental setup depicted in Fig. 9 (c) of the main text, CLoT refers to the Qwen-VL+CLoT model trained as outlined in Table 2. Additionally, we used classification accuracy as a metric. The results presented in Fig. 9 (c) demonstrate that the proposed CLoT can further enhance the performance of the CGG task. This to some extent validates the versatility and effectiveness of CLoT.



Figure 26. The examples of the generated data in CGG.

F.2. The Details of Divergent Association Task (DAT)

For DAT, it is a classic creativity test [33, 86] which needs participants to choose words with larger semantic distances among 10 unrelated nouns. Building on existing research findings that suggest participants with the ability to select unrelated nouns with large semantic distances tend to have stronger associative ability, we leverage this insight to analyze the LoT ability of LLMs using the DAT benchmark [33]. To streamline the analysis, we adapt the DAT benchmark into a series of choice questions, with the standard average semantic distance (ASD) measured by GloVe [92] serving as the metric. These questions challenge LLMs to choose the word from a set of nine options that differs the most from the given word.

Please carefully understand the provided question and select the option that satisfies the problem. Only one option meets the requirements. Question: Please select the option least relevant to the current set of words.

Words: <Words>

Options: <Options>

Answer Format: Please respond in the format of 'Option id. Option content,' for example, 'A. xxx.' Response: Satisfactory option is

<Response>

Specifically we use the instruction template above for the DAT task on LLM. Below we provide some examples of words and options:

Example 1:

<Words>: Guitar Amplifier Strings Pick Melody Chord Song Musician Concert

<Options>: A.studio B.hat C.piano D.umbrella

Example 2:

<Words>: Guitar Amplifier Strings Pick Melody Chord Song Musician Concert

<Options>: A.flame B.orange C.diamond D.earth

Example 3:

<Words>: Soccer Amplifier Marathon Surfing Volleyball Basketball Carrot Running Yoga

<Options>: A.canvas B.wire C.volcano D.bracelet

Example 4:

<Words>: Pepper Zucchini Eggplant Surfing Garlic Potato Carrot Bean Gymnastics

<Options>: A.drill B.bee C.hourglass D.brick

Example 5:

<Words>: Decaf Pastry Brew Roast Forest Outdoors Compass Bean Backpack

<Options>: A.cake B.whip C.space D.river

G. The Analysis for Self-Refinement

In this section, we will further analyze why we employ a one-round self-refinement in CLoT on Oogiri-GO. Additionally, we will discuss the reasons why CLoT does not induce performance collapse during “Explorative Self-Refinement” stage.

G.1. The Discussion for the Round of Self-Refinement

In Section 5 of the main text, we empirically demonstrated that additional rounds do not significantly enhance the LoT ability. Consequently, we default to a one-round setting when performing self-refinement on the Oogiri-GO dataset. In this section, we delve into a detailed analysis of the underlying reasons for this phenomenon. Indeed, to achieve efficacy with multiple rounds of self-refinement, two strategies can be employed.

Firstly, by expanding diversity in creative data (strategy 1). This involves providing the model with a more diverse set of creative data during the “Associable Instruction Tuning” stage to enhance initial LoT capabilities. This diversity enables the LLM to continuously generate novel data effectively even under various weakly-associated conditions;

Secondly, by ensuring diversity in the noun set \mathcal{S} (strategy 2). This implies using a more diverse and effective set of nouns in \mathcal{S} for sampling weakly-associated conditions, thereby facilitating better associative capabilities in LLM and ensuring the quality of newly generated data.

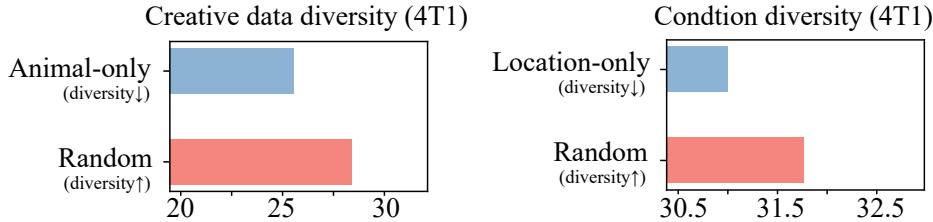


Figure 27. The impact of data diversity on CLoT performance. The baseline is Qwen-VL on I2T type Oogiri game. **Left:** For diversity of Oogiri-GO . **Right:** For diversity of weakly-associated conditions in noun set \mathcal{S} .

For strategy 1. We first substantiate the impact of creative data diversity on model performance. We conduct experiments using two subsets from the Oogiri-GO dataset: an “animal-only” subset and a “random” subset. The former consisted of 10,000 randomly sampled data containing animal-related nouns in responses, while the latter comprised 10,000 samples randomly drawn from the Oogiri-GO dataset. As depicted in Fig. 27 (Left), despite having an equal data size, the diverse “random” subset exhibited significant performance advantages. Therefore, for strategy 1, collecting a wide range of potential and diverse creative data is crucial to enhance LLM performance in generating effective new data under various weakly associated conditions. However, the inherent scarcity of high-quality creative data poses a challenge, as continuous production of innovative data by humans is not easily sustained. Additionally, the Oogiri-GO dataset already encompasses responses from a substantial portion of online Oogiri games, making it difficult to obtain a large-scale collection of new data. Hence, the inherent scarcity of innovative data constrains the further expansion of creative dataset diversity.

For strategy 2. Similarly, we conduct a simple experiment to illustrate the significant impact of \mathcal{S} diversity on model performance. We randomly sampled 5% of nouns from \mathcal{S} to create a “random” subset and extracted an equal number of location-related nouns to form a “location-only” subset. The results in Fig. 27 (Right) revealed the importance of \mathcal{S} diversity for LLM’s LoT ability. Therefore, for strategy 2, expanding the diversity of effective nouns in \mathcal{S} is essential. However, when utilizing nouns from \mathcal{S} for conditional generation by LLM, the number of newly generated effective nouns is limited. This limitation arises because the quantity of nouns in the response text is limited, and most nouns depend on those sampled in \mathcal{S} . Consequently, existing paradigms encounter difficulty in diversifying \mathcal{S} due to the constraints on the number of effective nouns generated. Moreover, introducing new nouns, and potentially verbs, from external knowledge bases into \mathcal{S} poses a challenge, as it necessitates ensuring that the newly added vocabulary provides sufficient clues to guide creative responses. For instance, external conditions may be entirely unrelated to the current task, making it challenging for LLM to draw upon existing knowledge for generating effective new creative data through associative thinking. The ideal scenario involves conditions that have a certain distance from the current knowledge domain but are not entirely irrelevant—termed as weakly-associated conditions. Therefore, the intrinsic difficulty in expanding the diversity of \mathcal{S} hinders the augmentation of \mathcal{S} diversity.

In conclusion, **due to inherent constraints on expanding the diversity of \mathcal{S} and creative data**, existing paradigms struggle to provide sufficient diversity for multi-round self-refinement. The experiments in Section 5 of the main text indicate

that one-round self-refinement effectively utilizes the existing diversity in \mathcal{S} and creative data. Consequently, multiple rounds of self-refinement do not yield a significant performance boost, as one-round already achieves satisfactory performance.

G.2. Self-Refinement doesn't Cause a Performance Collapse in CLoT

During the training of large language models, there is a phenomenon known as “Performance Collapse” [75, 76] while using the LLM-generated data. Specifically, due to the impressive performance of these models and their widespread use by various users, the Internet is now flooded with a vast amount of text generated by large language models, including answers, conversations, chat records, and more. Despite the generated text appearing close to those generated by human, it has irreversibly polluted Internet text data [75, 76]. This pollution is expected to result in a performance decline when large language models are retrained in the future to update parameters. The generated data from the Internet often exhibits similar patterns or implicit characteristics. Continuously feeding self-generated data to large language models [75, 76], i.e., self-refinement, leads to an accumulation of similar data during training, restricting the diversity of model outputs and ultimately causing “Performance Collapse”.

However, for CLoT, the proposed **“Explorative Self-Refinement” stage does not lead to “Performance Collapse”**. This is because, (1) during this stage, the generated data is produced under the constraints of various weak-associated conditions, ensuring diversity and alleviating the issue of similar patterns; (2) in the “Explorative Self-Refinement” stage, the generated data undergoes rigorous filtering through the discrimination ability by tuned LLM during the “Associable Instruction Tuning” stage. This process ensures that the generated data is of high quality and mitigates the potential risk of “Performance Collapse”. These two mechanisms provide effective safeguards for CLoT when enhancing the LoT capability of LLM. Experimental results in the main text empirically demonstrate that “Explorative Self-Refinement” does not impact model performance and significantly promotes the Leap-of-Thought ability of the model.

H. Further Discussions

In this section, we further discuss CLoT from other perspectives.

H.1. The Oogiri-GO Dataset doesn't Include English IT2T Type Data

Tables 1 and 3 in the main text reveal the absence of English IT2T type data in the Oogiri-GO dataset. This can be attributed to two main reasons.

(1) **Cultural difference.** Firstly, Oogiri games are traditional comedy games in Japan [28], with their content heavily influenced by the cultural context of players. Currently, Oogiri games are predominantly popular in countries with similar cultures, such as China and Japan. Due to cultural differences, the more intricate IT2T type of Oogiri games is not widely embraced in English-speaking countries' online communities, where participation primarily revolves around I2T or T2T formats.

(2) **Complex processes.** Secondly, a considerable portion of Oogiri game data on the Internet relies on bloggers and website operators who disseminate the Oogiri games through translation in their respective countries. The creation of IT2T-type Oogiri game data requires specific tools for Optical Character Recognition (OCR) [93, 94], image editing [95] and machine translation [96, 97], while other types of data are very simple to migrate to English. The Complex process for IT2T type data presents certain barriers, leading to a scarcity of IT2T content in English on the Internet.

In summary, given the rarity of English IT2T type data and the challenges associated with its creation, including the editing and translation of image text, our Oogiri-GO dataset does not encompass this type of data. Conversely, the dataset includes an ample amount of other types of data, adequately fulfilling the requirements for validating and analyzing the proposed CLoT.

H.2. Fine-tuning Directly on Oogiri-GO is Hard to Achieve Good LoT Ability

In the main text, we substantiate the efficacy of CLoT's "Associable Instruction Tuning" and "Explorative Self-Refinement" stages in enhancing the LoT capabilities of LLM through extensive experiments and analyses. This results in the impressive generation of humor. In this section, we revisit these two stages, asserting that a simple direct fine-tuning approach falls short in achieving sufficiently robust LoT ability.

To illustrate this conclusion, we conduct an experiment wherein the model underwent direct fine-tuning based on the instruction template of the "Associable Instruction Tuning" stage, as outlined in the main text. Specifically, the model is fine-tuned using the template depicted in only Fig. 6 (a). The results, presented in Fig. 28, reveal a significant performance gap between direct fine-tuning on Oogiri-GO and CLoT. Two primary reasons account for this phenomenon:

(1) **A lack of targeted associative exercises for LoT.** In fact, directly fine-tuning on the given creative data merely amounts to a rigorous fitting of the data. This fitting process only captures the inherent creative patterns within the data, failing to stimulate "thinking outside the box" for generating novel ideas. Furthermore, as mentioned in Appendix G.1, creative data is inherently scarce, and relying solely on dataset fitting easily leads to being trapped in local patterns. Hence, there is a pressing need for associative exercises to foster a departure from conventional thinking.

(2) **The creativity is uneven.** While Oogiri-GO responses stem from human creativity, the creativity in these responses varies widely. Some are highly imaginative, while others are mundane. The inherent difficulty in generating creative responses, even for humans, leads to uneven quality in the dataset, with a scarcity of exceptionally high-quality creative instances. Intuitively, only such instances have the potential to stimulate the LoT capabilities of the model. Learning from general creative responses is insufficient to foster strong LoT abilities.

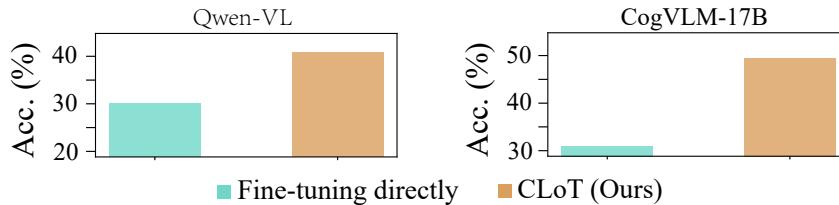


Figure 28. The performance of LLM with direct fine-tuning under 3T1 and I2T settings.

Fortunately, CLoT not only introduces associative exercises for LLM but also leverages the varied quality of creative data, enabling LLM to discern and generate exceptionally high-quality creative responses. This distinctive approach ultimately yields performance beyond what direct fine-tuning can achieve.

H.3. How to Further Enhance CLoT?

Although CLoT has demonstrated strong efficacy in enhancing the Leap-of-Thought capability of LLM, there are still notable areas for improvement in the future. For instance, as mentioned in Appendix H.2, the creative quality in data such as Oogiri-GO is uneven, and CLoT has leveraged this diversity to enhance LLM’s discrimination ability towards creative data, thereby aiding in the generation of high-quality creative content. Furthermore, these human-annotated data, featuring human rankings, can be utilized to construct evaluators, which employ the innovative Reinforcement Learning from Human Feedback (RLHF) technique to further boost CLoT’s performance, a pivotal approach for enhancing large language models.

Additionally, within this study, we reveal that prompting alone is insufficient to stimulate LLM’s LoT ability. Despite LLM possessing rich prior knowledge and excellent reasoning capabilities, additional training is currently necessary to activate LoT. Therefore, exploring ways to maximize LoT activation through prompting or minimizing LLM training is a meaningful research direction. Moreover, the tuning of LLM’s instructions, as seen in methods like LoRA, inevitably results in partial forgetting [1, 72, 98] of its inherent knowledge. For creative tasks, preserving as much of the original knowledge as possible is valuable. Hence, future work should focus on continuous learning approaches [99–101] to ensure the model retains existing knowledge to the greatest extent possible.