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

Shanshan Zhong[1,3*] **Zhongzhan Huang**[1,3*] **Shanghua Gao**[4] **Wushao Wen**[1]
Liang Lin[1] **Marinka Zitnik**[4] **Pan Zhou**[2,3†]
[1] Sun Yat-Sen University, [2] Singapore Management University, [3] Sea AI Lab, [4] Harvard University
*Co-first author: {zhongshsh5,huangzhzh23}@mail2.sysu.edu.cn
†Corresponding author: panzhou@smu.edu.sg

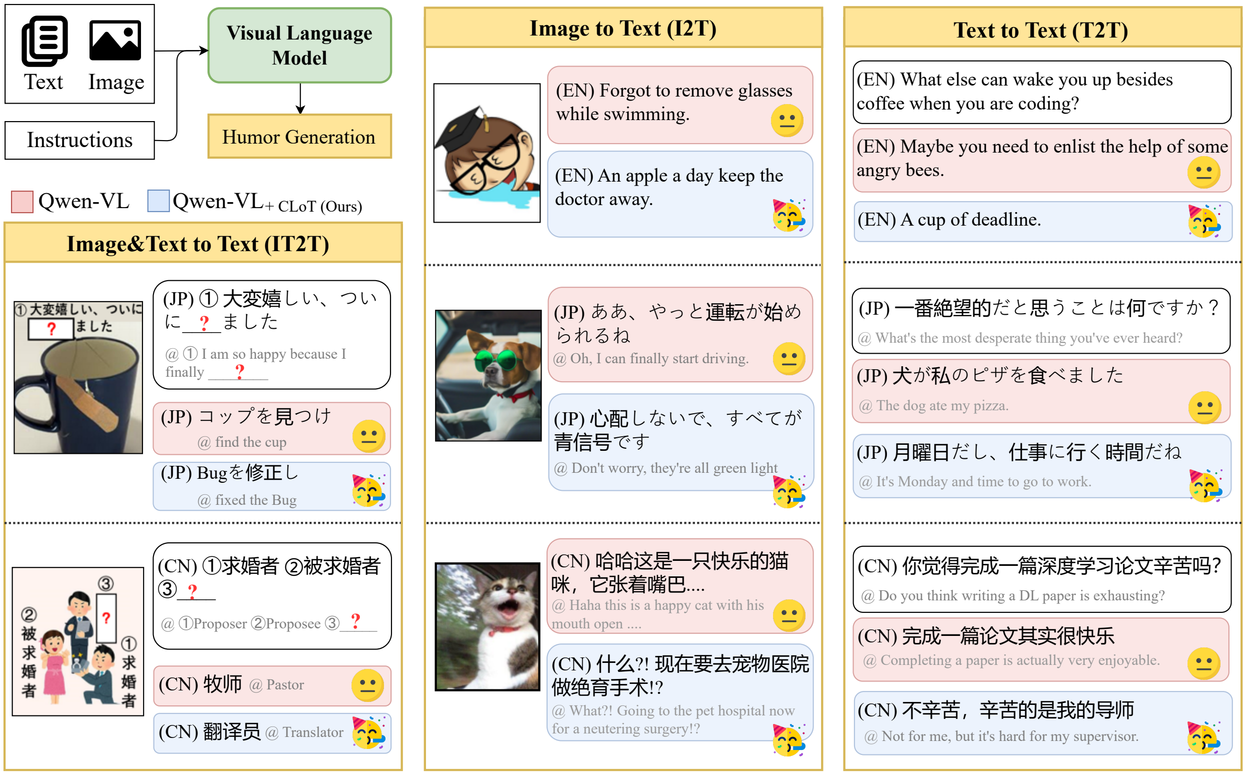

Text
Image
Instructions
Visual Language Model
Humor Generation
🟥 Qwen-VL
🟦 Qwen-VL+ CLoT (Ours)
Image&Text to Text (IT2T)
①大変嬉しい、ついに
?
ました
(JP) ① 大変嬉しい、ついに＿?＿ました
@ ① I am so happy because I finally ＿?＿
(JP) コップを見つけ
@ find the cup
(JP) Bugを修正し
@ fixed the Bug
③
?
②被求婚者
①求婚者
(CN) ①求婚者 ②被求婚者 ③＿?＿
@ ①Proposer ②Proposee ③＿?＿
(CN) 牧师 @ Pastor
(CN) 翻译员 @ Translator
Image to Text (I2T)
(EN) Forgot to remove glasses while swimming.
(EN) An apple a day keep the doctor away.
(JP) ああ、やっと運転が始められるね
@ Oh, I can finally start driving.
(JP) 心配しないで、すべてが青信号です
@ Don't worry, they're all green light
(CN) 哈哈这是一只快乐的猫咪，它张着嘴巴....
@ Haha this is a happy cat with his mouth open ....
(CN) 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!?
Text to Text (T2T)
(EN) What else can wake you up besides coffee when you are coding?
(EN) Maybe you need to enlist the help of some angry bees.
(EN) A cup of deadline.
(JP) 一番絶望的だと思うことは何ですか？
@ What's the most desperate thing you've ever heard?
(JP) 犬が私のピザを食べました
@ The dog ate my pizza.
(JP) 月曜日だし、仕事に行く時間だね
@ It's Monday and time to go to work.
(CN) 你觉得完成一篇深度学习论文辛苦吗?
@ Do you think writing a DL paper is exhausting?
(CN) 完成一篇论文其实很快乐
@ Completing a paper is actually very enjoyable.
(CN) 不辛苦，辛苦的是我的导师
@ Not for me, but it's hard for my supervisor.

Figure 1. Comparison 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 con-

tains 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 have been released online: https://zhongshsh.github.io/CLoT.

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

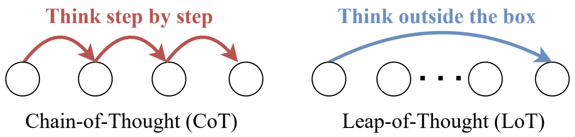


Figure 2. Comparison of CoT and LoT. "○" denotes the thought and "→" represents the connection between two thoughts.

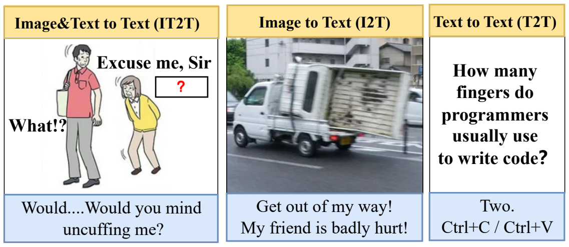

Image&Text to Text (IT2T)
Excuse me, Sir
?
What!?
Would....Would you mind uncuffing me?
Image to Text (I2T)
Get out of my way! My friend is badly hurt!
Text to Text (T2T)
How many fingers do programmers usually use to write code?
Two. Ctrl+C / Ctrl+V

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.

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 associable instruction tuning stage which designs an associable 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 [68–70].

(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 [71].

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 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.

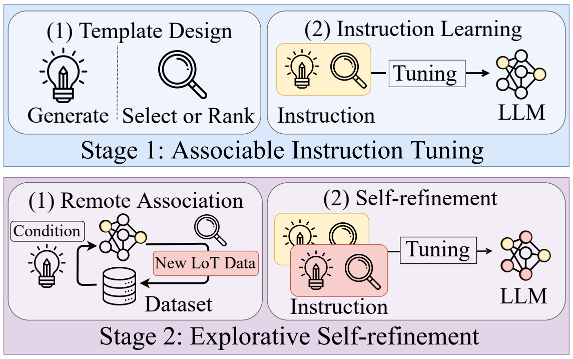


Figure 4. The overview of proposed Creative Leap-of-Thought.

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-alike 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-VL [1] 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 associable 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).

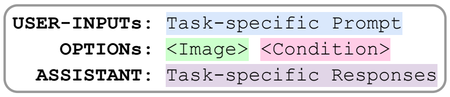
4.1. Associable Instruction Tuning

LoT ability mainly includes associable generation and discrimination ability [30]. Given an input, associable 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 associable generation and discrimination are not present in current LLMs, e.g., poor performance of GPT4v [72] 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 associable instruction tuning which trains LoRA [73] for LLMs on the Oogiri-GO dataset to achieve certain associable generation and discrimination abilities. It has two steps, including instruction generation and discrimination template design, and associable 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 associable 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.



```
USER-INPUTs: Task-specific Prompt
   OPTIONs: <Image> <Condition>
 ASSISTANT: Task-specific Responses
```

Figure 5. The LoT-oriented instruction templates.

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

Regarding associable 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 re-

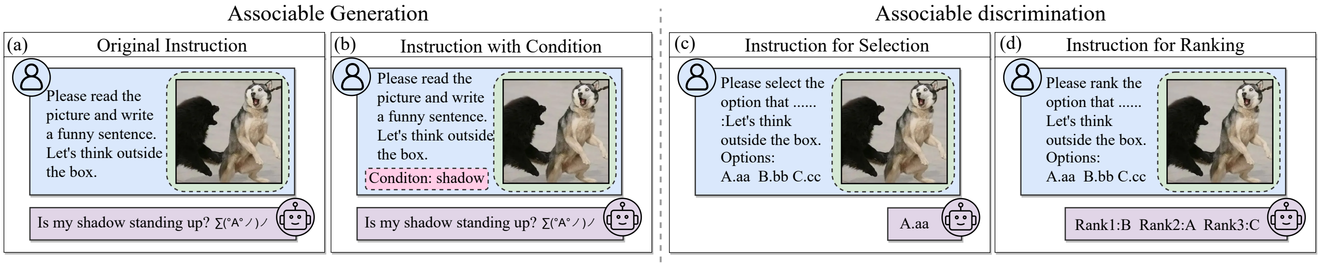

Associable Generation
Associable discrimination
(a) Original Instruction
Please read the picture and write a funny sentence. Let's think outside the box.
Is my shadow standing up? Σ(°A°ノ)ノ
(b) Instruction with Condition
Please read the picture and write a funny sentence. Let's think outside the box.
Conditon: shadow
Is my shadow standing up? Σ(°A°ノ)ノ
(c) Instruction for Selection
Please select the option that ...... :Let's think outside the box. Options: A.aa B.bb C.cc
A.aa
(d) Instruction for Ranking
Please rank the option that ...... Let's think outside the box. Options: A.aa B.bb C.cc
Rank1:B Rank2:A Rank3:C

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 associable generation. (c) and (d) are the two instructions about the selection and ranking of associable discrimination. All templates follow the formats in Fig. 5.

sponses (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 [74], 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 [73] for the LLM with the associable instruction data. See more details in Appendix.

4.2. Explorative Self-Refinement

After associable 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 $\boldsymbol{S}$, from the text in the Oogiri-GO training data. See details in Appendix. Then, for each user-input $\boldsymbol{I}$ (see Fig. 5), we generate n weakly-associated conditions $\{\boldsymbol{C}_i\}_{i=1}^{n}$. 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 $\boldsymbol{S}$ to enforce LLM to build connections between different concepts. Next, we add the condition $\boldsymbol{C}_i$ into user-input $\boldsymbol{I}$, and feed $\boldsymbol{I}$ into the LLM to generate a humor candidate $\boldsymbol{R}_i$. Repeating this process with different conditions $\boldsymbol{C}_i$ can generate a total of n candidates $\{\boldsymbol{R}_i\}_{i=1}^{n}$.

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 $\{\boldsymbol{C}_i\}_{i=1}^{n}$ 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 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 [76, 77] during self-refinement phase, and can improve the LoT performance across several creative tasks as shown in Sec. 5. See the ablation study and

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 [72]	-	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$_{\text{Multilingual}}$ [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 [75]	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$_{\text{+AIT (Ours)}}$	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$_{\text{+CLoT (Ours)}}$	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 **mutilmodal multilingual models**. mTn choice question selects n correct answers from m options. "Avg." is the average of all metrics. "AIT" denotes associable instruction tuning.

Algorithm 1 Inference Step of CLoT

```
Input: Input $I$, CLoT-trained LLM $\mathcal{A}$, response number $n$
Output: Creative response $R$.
                                   ▷ Creating the candidate responses
construct $n$ weakly-associated conditions $\{C_i\}_{i=1}^n$
$\{R_i\}_{i=1}^n \leftarrow \mathcal{A}([I, \{C_i\}_{i=1}^n])$
                                   ▷ Choosing most creative response
Top-2 $R_1', R_2' \leftarrow \mathcal{A}([I, \{R_i\}_{i=1}^n])$ with ranking ability
Best $R \leftarrow \mathcal{A}([I, R_1', R_2'])$ with selection ability
return Best response $R$.
```

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 $\boldsymbol{I}$ of the formation in Fig. 5, LLM first uses explorative remote association in Sec. 4.2 to construct n weakly-associated conditions (these conditions can be empty since the training paradigm in Section 4), and then follows Sec. 4.2 to generate n responses $\{\boldsymbol{R}_i\}_{i=1}^n$. 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 [81], 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 [74]. 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) [82, 83]. We provide more experimental details in the Appendix.

5.2. Evaluation by Choice and Ranking Questions

Evaluation on Multimodal Multilingual LLMs. We plug our associable instruction tuning (AIT) and our CLoT into the SoTA open-source multimodal multilingual model Qwen-VL [1] to obtain Qwen-VL$_{\text{+AIT}}$ and Qwen-VL$_{\text{+CLoT}}$, 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$_{\text{+AIT}}$ 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$_{\text{+CLoT}}$ 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 two stages in CLoT, i.e., associable 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$_{\text{+AIT}}$ achieves remarkable improvements over the standard CogVLM-17B, and CogVLM-17B$_{\text{+CLoT}}$ 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

Model	Size	Image to Text (I2T)						Text to Text (T2T)					
		2T1	3T1	4T1	5T2	Rank	Avg.	2T1	3T1	4T1	5T2	Rank	Avg.
InstructionBLIP [78]	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$_{\text{LLaMA2}}$ [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 [79]	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$_{\text{+AIT (Ours)}}$	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$_{\text{+CLoT (Ours)}}$	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 **mutilmodal 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 [72]	-	45.3	30.4	6.7	61.6	36.0
GPT-4 [72]	-	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 [80]	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 [75]	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$_{\text{+CLoT (Ours)}}$	7B	51.7	32.3	**24.8**	65.0	43.4
CogVLM-17B$_{\text{+CLoT (Ours)}}$	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.

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 [84], 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 demonstrates 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.

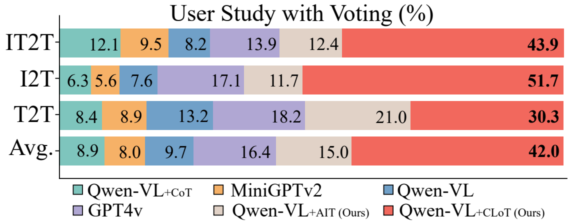


Figure 7. User study with voting (%) for Oogiri-style creative responses by different models and improved methods.

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) 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, 85, 86], 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, 87] 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,

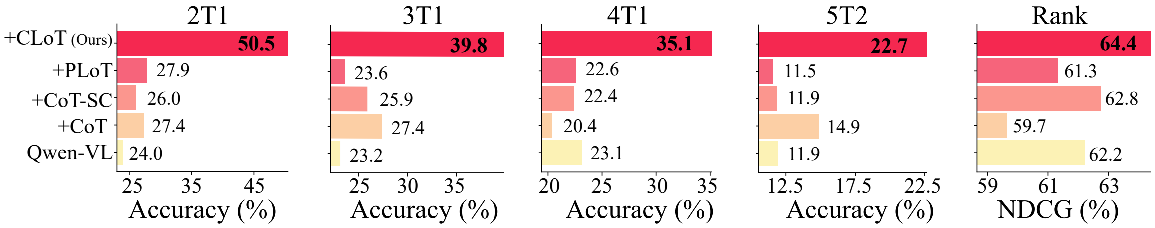


Figure 8. The accuracy (%) of choice questions and the NDCG (%) of ranking questions on our CLoT and various **reasoning frameworks**. The baseline is Qwen-VL on multilingual I2T task. For mTn choice questions, one needs to select n correct answers from m options.

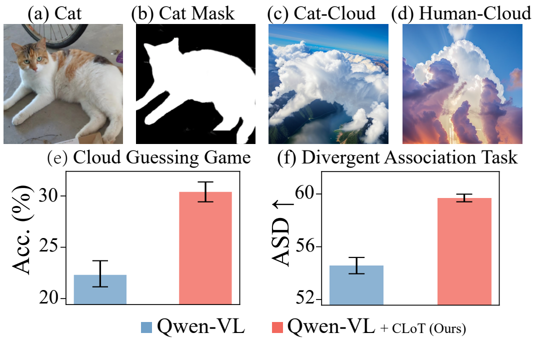


Figure 9. Evaluation of CLoT on the creative CGG (e) and DAT (f) tasks. (c-d): examples of cloud guessing games. (b): conditional masks of image (a) for generating cloud images.

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. 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 $\mathcal{S}$ is crucial to self-refinement, since it decides whether the associable remote stage can generate high quality and diverse data. However, the condition set $\mathcal{S}$ 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.

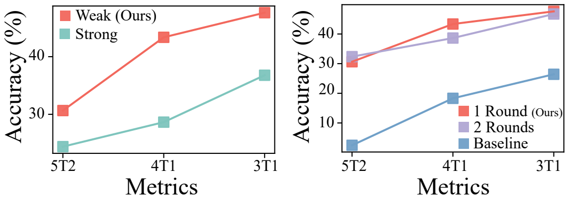


Figure 10. The ablation study of CLoT. We use CogVLM as baseline on the English I2T task. **Left:** weakly-associated condition v.s. strongly-associated condition during explorative remote association. **Right:** The effect of rounds of self-refinement.

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.

7. Acknowledgments

This work was supported by National Science and Technology Major Project (No.2021ZD0111601), National Natural Science Foundation of China (No.62325605, U1711264), and Guangdong Basic and Applied Basic Research Foundation (No.2023A1515011374), and Guangzhou Science and Technology Program (No.2024A04J6365), and Guangdong Province Key Laboratory of Information Security Technology. Pan Zhou acknowledges support from the Singapore Ministry of Education (MOE) Academic Research Fund (AcRF) Tier 1 grant.

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

[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, Prajjwal 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. 5, 6, 7

[6] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao 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 Leahey, 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

[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, 6, 7

[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, 7

[32] Lvmin 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. 7

[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, 7

[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 multi-modal 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.

[70] Shanshan Zhong, Zhongzhan Huang, Daifeng Li, Wushao Wen, Jinghui Qin, and Liang Lin. Mirror gradient: Towards robust multimodal recommender systems via exploring flat local minima. *arXiv preprint arXiv:2402.11262*, 2024. 3

[71] Jie Huang and Kevin Chen-Chuan Chang. Towards reasoning in large language models: A survey. *arXiv preprint arXiv:2212.10403*, 2022. 3

[72] OpenAI. Gpt-4 technical report, 2023. 4, 6, 7

[73] 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

[74] 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

[75] 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

[76] Ryuichiro Hataya, Han Bao, and Hiromi Arai. Will large-scale generative models corrupt future datasets? In *ICCV*, 2023. 5

[77] 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. 5

[78] Wenliang Dai, Junnan Li, and et al. Instructblip: Towards general-purpose vision-language models with instruction tuning. *arXiv preprint arXiv:2305.06500*, 2023. 7

[79] 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

[80] Baichuan. Baichuan 2: Open large-scale language models. *arXiv preprint arXiv:2309.10305*, 2023. 7

[81] 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

[82] Kalervo Järvelin and Jaana Kekäläinen. Cumulated gain-based evaluation of ir techniques. *ACM Transactions on Information Systems (TOIS)*, 20(4):422–446, 2002. 6

[83] 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

[84] 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

[85] 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. 7

[86] 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. 7

[87] 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. 7