

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

—Appendix—

Shanshan Zhong Zhongzhan Huang Shanghua Gao Wushao Wen Liang Lin
Marinka Zitnik Pan Zhou



Creative LoT

Let’s think **outside** the **box!**

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 [1–5]. 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 associable 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 [1–5], 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 [6, 7], 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). While humor is **subjective**, the additional results demonstrate CLoT’s LoT capacity of using excellent creative thinking to produce high-quality humor responses.

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. 1, 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.

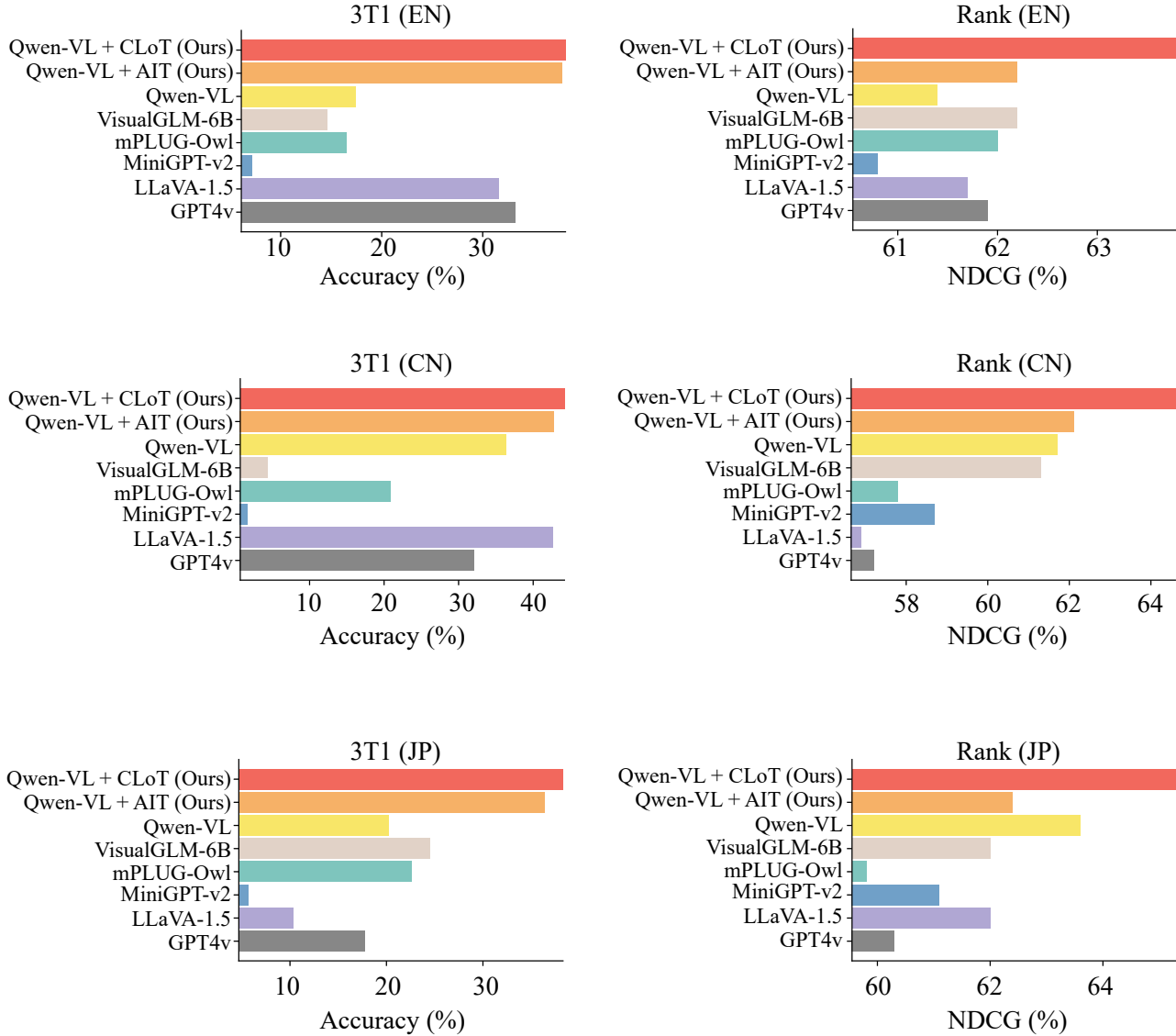


Figure 1. The performance of different LLM for Oogiri game in various languages.

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. 2. 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. 2 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. 2 underscores CLoT’s ability to engage in divergent thinking and approach challenges from multiple perspectives, showcasing its capacity to think outside the box.

	<p>(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?</p> <p>(2) 在路边被一车子压扁之后..... @ After getting squashed by a car on the roadside...</p> <p>(3) 猫版大鹏展翅。 @ The cat's version of a majestic spread of wings, like a giant roc taking flight.</p> <p>(4) 你这猫也太会摆姿势了吧! @ Oh, your cat really knows how to strike a pose!</p> <p>(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.</p>
	<p>(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.</p> <p>(2) 别追了, 我已经把你的孩子叼走了。 @ Stop chasing me. I've already grabbed your kid.</p> <p>(3) 你先别飞, 我还有话要说。 @ Don't take off just yet; I've got something to say.</p> <p>(4) 客官你还没付钱! @ Hey, you haven't paid yet, sir!</p> <p>(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.</p>
	<p>(1) 终于知道为什么他俩不说话了。 @ Finally figured out why those two aren't barking.</p> <p>(2) 你俩是不是又把家具都拆了。 @ Did you two go and take apart all the furniture again?</p> <p>(3) 不让你吃巧克力, 就只能这样了。 @ If I can't let you have chocolate, this is the only option.</p> <p>(4) 最近又胖了。 @ We've packed on some pounds again lately.</p> <p>(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!</p>
	<p>(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!?</p> <p>(2) 哇, 真的有条大鱼! @ Wow, that's a really big fish!</p> <p>(3) 司机师傅, 快看路啊! @ Hey driver, watch the road! Hurry up!</p> <p>(4) 安全带是系在脖子上的吗? @ Is the seatbelt supposed to be fastened around the neck?</p> <p>(5) 快! 追上前面那辆车! @ Hurry! Catch up with the car in front!</p>

Figure 2. 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. 3, 4 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. 5, 6, and 7 exhibit responses on Image to Text, while Fig. 8, 9, and 10 display responses on Text to Text. Finally, we supplement more Qwen-VL+CLoT responses on Chinese Oogiri samples in Fig. 11.

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:** 噢, 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:** 哎呀, 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 3. 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、本当におなかが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 4. The responses of LLMs in Japanese Oogiri IT2T samples. “@” denotes translations.

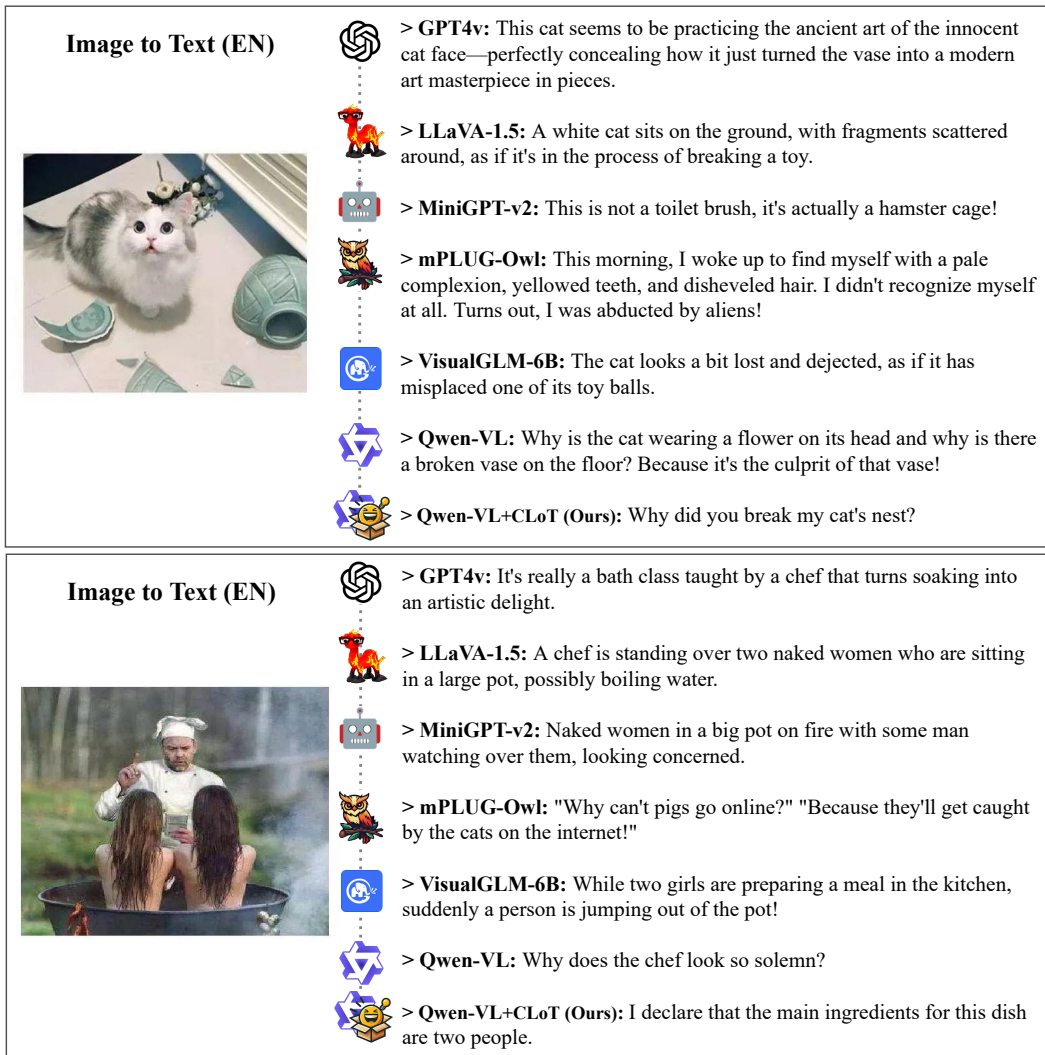


Figure 5. The responses of LLMs in English Oogiri I2T samples.

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，他伪装成年老、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 6. 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よね? 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 7. The responses of LLMs in Japanese Oogiri I2T samples. “@” denotes English translations.

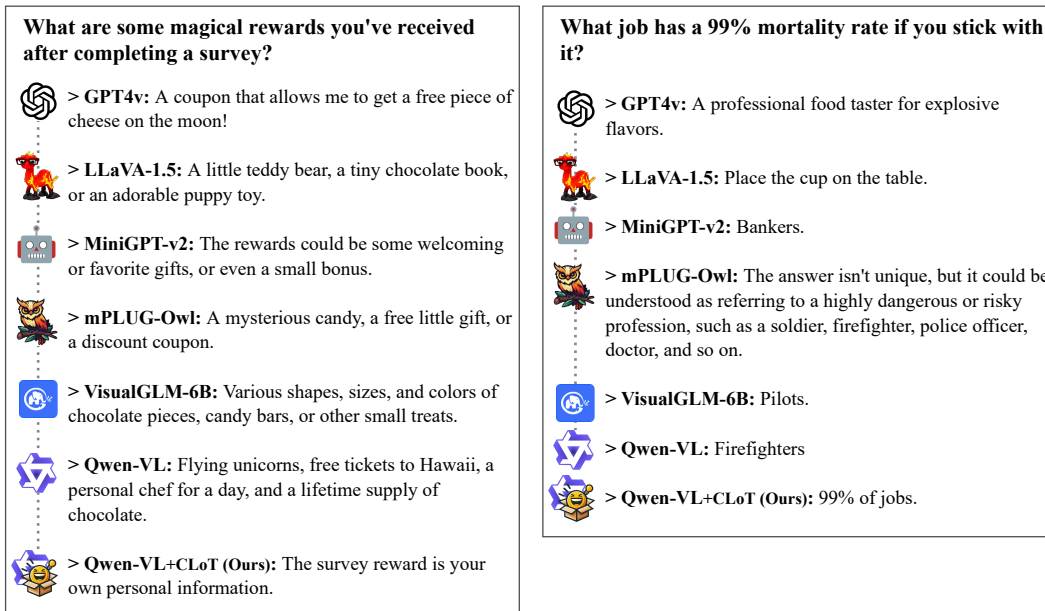


Figure 8. The responses of LLMs in English Oogiri T2T samples.



Figure 9. 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ムと可愛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ことかもしれません! @ 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「もう相手を見つけたらいい」「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てもよくなります... @ 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。エイリアンのセルフィーも含めてです。 @ 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: なぜライオンはりんごが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 10. The responses of LLMs in Japanese Oogiri T2T samples. “@” denotes English translations.

		
<p>今天是情人节..... @ Today is Valentine's Day...</p>	<p>抹茶奥利奥 @ Matcha Oreo</p>	<p>你是不是在外面有狗了?! @ Is there another dog in your life?!</p>
		
<p>狗界最好的演唱会! @ The ultimate dog concert!</p>	<p>如果我这么做的时候, 会被别人误会成小偷吗? @ Would people mistake me for a thief if I do this?</p>	<p>我今天有作业要交, 可以借我一下吗? @ I have an assignment due today. Can I borrow yours for a bit?</p>
		
<p>你可真会抢镜 @ You sure know how to steal the spotlight.</p>	<p>突然发现野餐地点是墓地 @ Just realized the picnic spot is a graveyard all of a sudden.</p>	<p>又到了给主人买饭的时间 @ It's time to grab some food for the owner again.</p>
		
<p>宇宙旅行前的最后祈祷 The final prayer before a journey through the cosmos.</p>	<p>别以为我不知道你作弊 @ Don't think I don't know you're cheating.</p>	<p>情侣车出门约会 @ Couple's car heading out for a date</p>

Figure 11. The responses of Qwen-VL+CLoT 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.

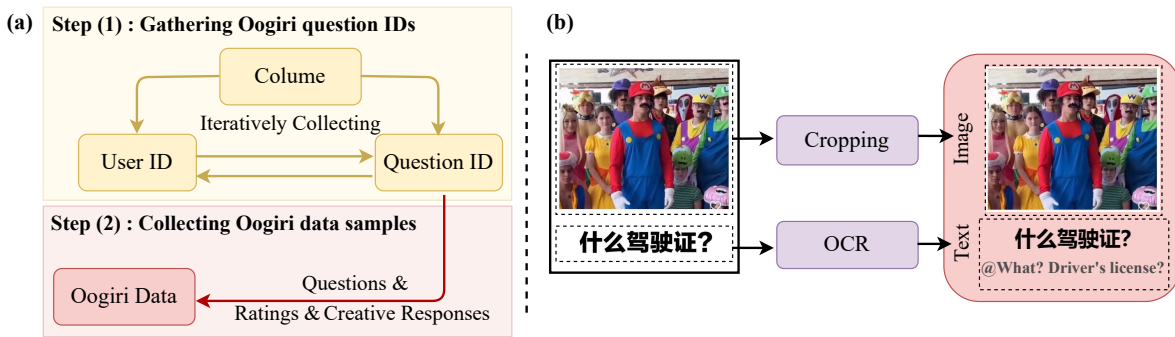


Figure 12. The processing of online data collection. (a) The crawling flow of Bokete website. (b) The image processing of Oogiri data.

Specifically, as illustrated in Fig. 12 (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. 12 (b) ¹. In such cases, it is necessary to separate them to construct Oogiri data with a consistent format. Specifically, we utilize PaddleOCR ² 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 [8] 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 NudeNet ³, 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 [9]. 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) [10]. We adopt top-1 accuracy as the positions at the top of rank lists are more significant in ranking scenarios [11].

(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 associable 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 associable 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 [8] and CogVLM [12] 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 associable generation and discrimination abilities. Our templates primarily comprise two components in Fig. 13: 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.

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

Figure 13. The LoT-oriented instruction templates.

Instruction Templates of Image to Text. Based on Fig. 13, 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 humorousness 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 [13]. 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 humorousness 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 humorousness 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 [14–16]. 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 [17] 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 associable instruction data.

The code snippets below illustrate the insertion points for LoRA during the training of Qwen-VL [8]. 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 associable 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 [18], 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. 14 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. 15. 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. 15 (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.

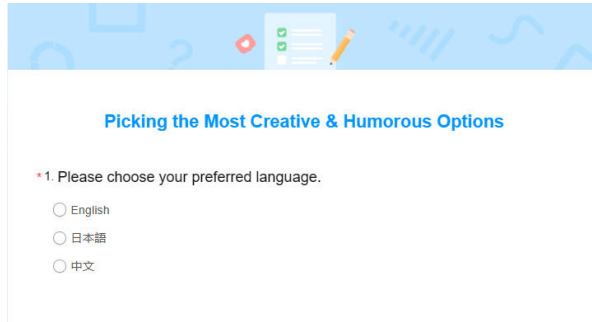


Figure 14. The questionnaire homepage of user study.

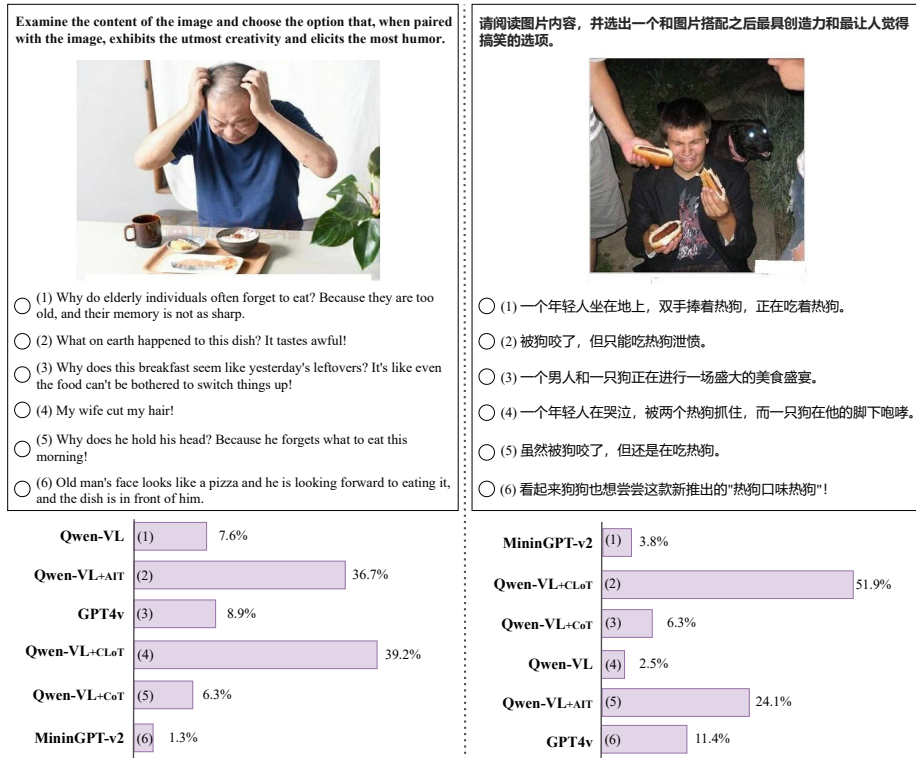


Figure 15. The questionnaire examples of user study.

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 [19–23] 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 16. 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 [6, 7] 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 [7]. 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 [24] 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 “Associative 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.

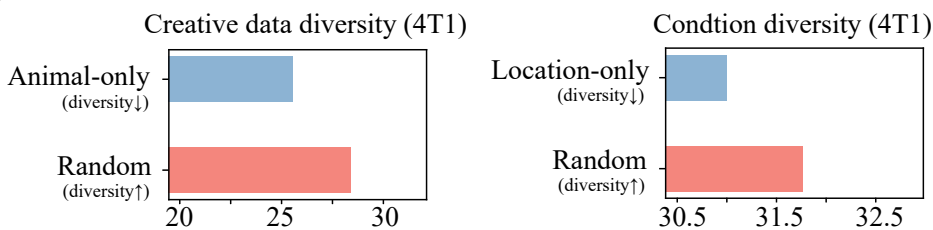


Figure 17. The impact of data diversity on CLoT performance. The baseline is Qwen-VL on I2T type Oogiri game. **Left:** For diversity of Oogiri-GO. **Right:** For diversity of weakly-associated conditions in noun set \mathcal{S} .

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. 17 (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. 17 (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” [25, 26] 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 [25, 26]. 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 [25, 26], 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 [27], 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) [28, 29], image editing [30] and machine translation [31, 32], 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. 18, 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.

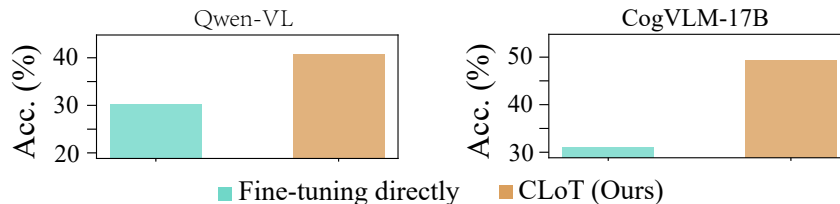


Figure 18. The performance of LLM with direct fine-tuning under 3T1 and I2T settings.

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 [8, 17, 33] 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 [34–36] to ensure the model retains existing knowledge to the greatest extent possible.

References

- [1] 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. [2](#)
- [2] Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. Automatic chain of thought prompting in large language models. *arXiv preprint arXiv:2210.03493*, 2022.
- [3] 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.
- [4] 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.
- [5] Jieyi Long. Large language model guided tree-of-thought. *arXiv preprint arXiv:2305.08291*, 2023. [2](#)
- [6] 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. [2](#), [25](#)
- [7] 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. [2](#), [25](#)
- [8] 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. [15](#), [16](#), [20](#), [29](#)
- [9] 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. [16](#)
- [10] 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. [16](#)
- [11] Jiayi 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](#)
- [12] Weihang 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. [16](#)
- [13] 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. [18](#)
- [14] 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](#)
- [15] Julian Salazar, Davis Liang, Toan Q Nguyen, and Katrin Kirchhoff. Masked language model scoring. *arXiv preprint arXiv:1910.14659*, 2019.
- [16] 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](#)
- [17] 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. [20](#), [29](#)

- [18] 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
- [19] 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. 24
- [20] 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.
- [21] 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.
- [22] 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.
- [23] 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. 24
- [24] 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
- [25] Ryuichiro Hataya, Han Bao, and Hiromi Arai. Will large-scale generative models corrupt future datasets? In *ICCV*, 2023. 27
- [26] Iliia 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. 27
- [27] Wikimedia. Glossary of owarai terms. https://en.wikiped ia.org/wiki/Glossary_of_owarai_terms, 2023. 28
- [28] 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
- [29] 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
- [30] Patrick Pérez, Michel Gangnet, and Andrew Blake. Poisson image editing. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, pages 577–582. 2023. 28
- [31] Felix Stahlberg. Neural machine translation: A review. *Journal of Artificial Intelligence Research*, 69:343–418, 2020. 28
- [32] Shuoheng Yang, Yuxin Wang, and Xiaowen Chu. A survey of deep learning techniques for neural machine translation. *arXiv preprint arXiv:2002.07526*, 2020. 28
- [33] 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
- [34] 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
- [35] 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.
- [36] 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