Metaphor Reasoning is Meta-reasoning

Anonymous ACL submission

Abstract

Metaphor reasoning is an essential cognitive ability that maps knowledge from familiar domains to more abstract domains. This ability functions as a meta-ability underlying many types of reasoning. However, existing work rarely investigates how metaphor reasoning affects other reasoning abilities. To bridge this gap, we systematically study how metaphor reasoning, particularly through metaphorical riddles, can enhance broader reasoning abilities in large language models. We propose METAR, an automated system for synthesizing metaphorical riddles that satisfy five quality dimensions: *diverse*, *balanced*, *reasoning-oriented*, *challenging*, and *verifiable*. Leveraging that answer categories determine riddle categories, we employ a hierarchical answer taxonomy for the former three criteria and a multi-agent refinement framework for the latter two, generating a high-quality dataset. Training with reinforcement learning on verifiable rewards using only thousands of metaphorical riddles, we demonstrate improvements across six out-of-distribution reasoning domains. Analysis reveals transfer effectiveness depends on model scale and pattern-target domain alignment. We will release our code and data.

1 Introduction

Metaphor reasoning is an essential cognitive ability (Lakoff, 1993). Through metaphors, humans are able to understand abstract concepts by mapping knowledge from familiar domains (i.e., source domains) to more abstract domains (i.e., target domains) (Lakoff and Johnson, 2024). As illustrated in Figure 1, comparing *ducks* that handle pests in rice fields to *wardens* conveys their role with only one word. Hence, metaphor reasoning can be considered as a meta-ability underlying many types of reasoning, such as abstract and analogical reasoning (Thibodeau and Boroditsky, 2011; Khatin-Zadeh et al., 2022).

Figure 1: An example of a *metaphorical riddle*.

However, existing work rarely investigates metaphor reasoning's impact on reasoning abilities. Current research on metaphor reasoning focuses on detection (Tian et al., 2024; Chen et al., 2024), interpretation (Tong et al., 2024; Sanchez-Bayona and Agerri, 2025), creative generation (Chakrabarty et al., 2021; He et al., 2023a,b), or downstream tasks such as sentiment analysis (Li et al., 2022a), translation (Wang et al., 2024), jailbreaking (Yan et al., 2025), and word-games (Xu and Zhong, 2025). While recent work demonstrates that metaphor reasoning can enhance reasoning abilities (Kramer, 2025), this work lacks evaluation on public reasoning datasets. Therefore, metaphor reasoning's impact on broader reasoning domains remains under-explored.

To investigate how metaphor reasoning affects reasoning abilities in broader domains, we first need to identify an appropriate task form. *Riddles* serve as a typical task form for metaphor reasoning (Panagiotopoulos et al., 2025; Le et al., 2025):

> *For the essence of a riddle is to express true facts under impossible combinations. Now this cannot be done by any arrangement of ordinary words, but by the use of metaphor it can.*
>
> *—Poetics, Chapter XXII, by Aristotle*

Specifically, riddles use ingenious and enigmatic language to describe a hidden answer, requiring

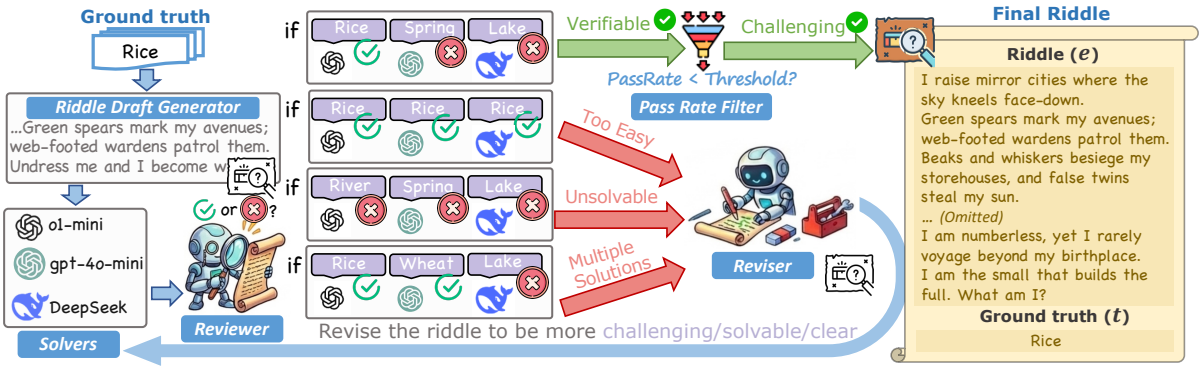

Figure 2: Multi-agent iterative refinement framework for riddle generation: (1) *Riddle Draft Generation* constructs initial riddles from metaphors given a ground truth answer; (2) *Solver Answer Collection* collects candidate answers from diverse solvers $\mathcal{M}$; (3) *Reviewer Assessment* verifies *Verifiability* and *Challengingness* through quality assessment; (4) *Reviser Refinement* iteratively improves problematic riddles based on diagnostic outcomes; (5) *Pass Rate Filter* applies a final quality gate to ensure challengingness requirement.

solvers to infer the answer through deliberate reasoning. On one hand, *figurative language is common in riddles*, with 37.5% of riddles containing figurative language, of which 24% contain metaphors (Zhang and Wan, 2022). In our paper, we refer to riddles that contain metaphors as *metaphorical riddles*. On the other hand, *metaphorical riddle reasoning inherently involves multiple traditional metaphor reasoning tasks*, such as component identification (Li et al., 2022b) and metaphor interpretation (Tong et al., 2024). As shown in Figure 1, this metaphorical riddle contains multiple metaphorical clues that require mapping through metaphor reasoning to arrive at the final answer. Hence, we adopt metaphorical riddles as our task form to study how metaphor reasoning affects reasoning abilities in broader domains.

However, using metaphorical riddles to enhance models' reasoning abilities faces several challenges: (1) *Scalable dataset construction*: Previous work typically collects metaphorical riddles through web scraping (Lin et al., 2021; Zhang and Wan, 2022), which suffers from limited scalability. (2) *Effective training recipe*: Existing methods either rely on prompting without improving model capabilities (Panagiotopoulos et al., 2025) or fine-tune small models on Question-Answer pairs, but fail to evaluate generalization to broader reasoning domains (Lin et al., 2021).

To bridge the gap, we systematically study how metaphor reasoning affects reasoning abilities in other domains using riddles as our task form. We propose METAR, an automated system for synthesizing metaphorical riddles with five key quality dimensions: *diverse*, *balanced*, *reasoning-oriented*, *challenging*, and *verifiable*. Since each riddle describes a specific answer, the answer's semantic category determines the riddle's thematic category, enabling us to achieve these criteria through two mechanisms: a *riddle answer taxonomy* (3 hierarchical levels, 5,466 answers) ensuring the first three dimensions via category selection and popularity-based entity selection; and a *multi-agent iterative refinement framework* (Figure 2) ensuring the latter two dimensions. We adopt reinforcement learning with verifiable rewards (RLVR) (Guo et al., 2025; Yu et al., 2025) to enhance models' metaphor reasoning abilities. Experiments show that training on only 3,444 metaphorical riddles improves reasoning in six out-of-distribution domains, demonstrating metaphor reasoning as a meta-reasoning ability and revealing the importance of model scale and domain alignment.

Overall, our contributions are as follows: (1) We are the first to systematically study whether metaphor reasoning can enhance models' broader reasoning abilities. (2) We propose a scalable automated system for synthesizing metaphorical riddles, enabling large-scale generation of high-quality training data. (3) Through extensive experiments, we demonstrate that metaphor reasoning functions as a meta-ability and provide interpretable analysis of the underlying mechanisms.

2 Related Work

2.1 Metaphor Reasoning

Research on metaphor reasoning encompasses detection (Li et al., 2024; Tian et al., 2024; Chen et al., 2024), generation (Chakrabarty et al., 2021; He

et al., 2023a,b), and interpretation (He et al., 2022; Tong et al., 2024; Sanchez-Bayona and Agerri, 2025), with applications in sentiment analysis (Li et al., 2022a), translation (Wang et al., 2024), jailbreaking (Yan et al., 2025), and word games (Xu and Zhong, 2025). Metaphor reasoning may enhance general reasoning capabilities, as metaphors form the foundation of abstract reasoning (Lakoff and Johnson, 2024; Thibodeau and Boroditsky, 2011; Khatin-Zadeh et al., 2022). While conceptual metaphor-based prompting has shown promise to improve reasoning capabilities (Kramer, 2025), no prior work has investigated whether metaphor reasoning enhances model performance on established reasoning benchmarks.

2.2 Metaphor and Riddle Reasoning

Riddle reasoning datasets are typically collected from online sources and reformatted as question-answer (QA) questions (Lin et al., 2021; Jiang et al., 2024). Researchers have improved model performance through fine-tuning on QA datasets (Lin et al., 2021) and prompting techniques (Panagiotopoulos et al., 2025), but have not examined whether riddle reasoning training enhances reasoning capabilities in other domains.

Figurative language is common in riddles (Zhang and Wan, 2022). Panagiotopoulos et al. (2025) highlights metaphors when reconstructing in-context riddles, improving riddle-solving performance. Le et al. (2025) incorporates novel metaphors in riddle generation to increase difficulty. However, no prior work has used riddles as a medium to study how metaphor reasoning affects other reasoning capabilities.

2.3 Reinforcement Learning with Verifiable Rewards

Reinforcement Learning with Verifiable Rewards (RLVR) leverages objective, verifiable metrics as reward signals, enabling models to exhibit advanced deliberative thinking capabilities (Guo et al., 2025; Yu et al., 2025; Hu et al., 2025), such as planning and reflection, through extended reasoning chains. This approach has proven effective for complex reasoning tasks (Seed et al., 2025; Team et al., 2025; Chen et al., 2025). However, no prior work has applied RLVR to metaphor reasoning.

3 Task Formulation

We formulate metaphorical riddle reasoning as a question-answering (QA) task. Formally, let

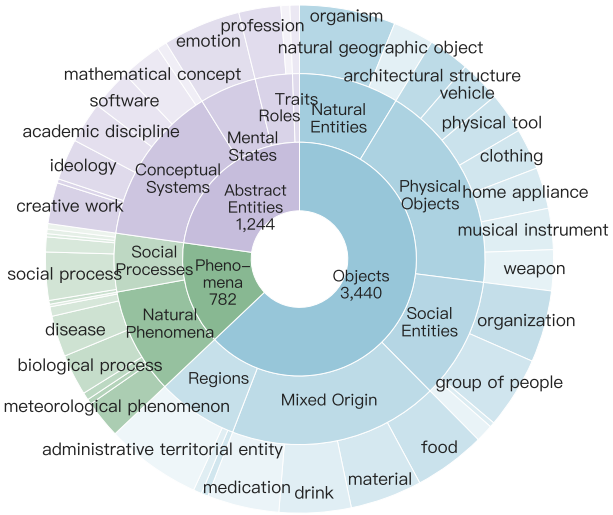

Figure 3: The statistics of the Riddle Answer Taxonomy.

$\mathcal{D} = \{(e_i, t_i)\}_{i=1}^{|\mathcal{D}|}$ denote a dataset of metaphorical riddles, where each pair (e, t) consists of a riddle e (which serves as the question) and its corresponding ground truth answer t. Given a riddle e, the task is to predict the answer $\hat{t}$ that matches the ground truth t. An example is shown in Figure 2.

4 METAR

We propose METAR, an automated system for synthesizing metaphorical riddles. We establish design criteria for high-quality riddle datasets and present mechanisms to achieve them.

4.1 Design Criteria

Our riddle dataset satisfies five quality dimensions: (1) *diverse*, ensuring model generalization across different domains; (2) *balanced*, avoiding bias toward specific entity types; (3) *reasoning-oriented*, requiring metaphorical reasoning rather than knowledge retrieval; (4) *challenging*, incentivizing models' elaborate reasoning processes; (5) *verifiable*, providing reliable supervision signals for subsequent training.

Following §3, the *answer* t is the target entity, and the *question* (riddle) e is its metaphorical description. Since each riddle describes a specific answer, the answer's category determines the riddle's category. Hence, we achieve these criteria via two mechanisms: the first three through answer taxonomy design with category and popularity-based entity selection; the latter two through a multi-agent iterative refinement framework.

H_1	H_2	H_3	**Example Answers**
Objects	Mixed Origin	drink	coffee, tea, milk, beer
Objects	Natural Entities	natural geographic object	mountain, ocean, valley, volcano
Phenomena	Natural Phenomena	astronomical phenomenon	eclipse, meteor, sunset
Phenomena	Social Processes	historical event	battle, armistice, ceasefire
Abstract Entities	Conceptual Systems	ideology	nationalism, conservatism, fascism, feminism
Abstract Entities	Roles	social status	citizen, prisoner, slave, reputation

Table 1: Examples of riddle answers t organized by taxonomic hierarchy.

4.2 Riddle Answer Taxonomy

To achieve the criteria of *diverse*, *balanced*, and *reasoning-oriented*, we construct a three-level answer taxonomy (H_i for $i \in \{1, 2, 3\}$), with statistics in Figure 3 and examples in Table 1[1].

Diverse Coverage. To partition entities into fundamental categories at the top level H_1, we adopt UFO (Guizzardi et al., 2022)'s foundational categories: *Material Object*, *Phenomenon*, *Abstract Entities*. Since manually enumerating all relevant domain-specific subcategories would be impractical, we use GPT-5 (OpenAI, 2025b)[2] to generate second-level H_2 categories. To ensure comprehensive coverage, we leverage Wikidata's *subclass-of* relationship: for each $h_2 \in H_2$, we generate seeds $\mathcal{S}_{h_2} = \{s_1, s_2, \ldots, s_{n_{h_2}}\}$ where each seed s corresponds to a Wikidata Q-id Q_s.

Common Answers. To ensure *reasoning-oriented* criterion, we select *common* well-known entities, avoiding obscure answers that test knowledge rather than metaphor reasoning. For each seed s with Q-id Q_s, we employ two-stage sorting: (1) retrieve candidates via *subclass-of*, sort by *sitelinks* (cross-language Wikipedia links), select top candidates as C_s; (2) then sort C_s by *popularity* using QRank[3](Arora et al., 2024), ranking entities by Wikimedia page views.

Quality Control and Taxonomic Balance. To ensure answer quality, we apply two filters: (1) *Conciseness*: remove entities with labels exceeding 2 words, as overly complex answers make it difficult to satisfy the *verifiable* criterion; (2) *Quality threshold*: filter entities with sitelinks below threshold τ, as they tend to be obscure. An entity x is valid, Valid(x), if its label ≤ 2 words and sitelinks $> \tau$. To ensure *balanced* criterion, we allocate fixed quota K to each $h_2 \in H_2$, uniformly distributed among seeds $\mathcal{S}_{h_2}$. For h_2 with n_{h_2} seeds, quota k_s per seed $s \in \mathcal{S}_{h_2}$ is:

$$k_s = \min\left(\left\lfloor \frac{K}{n_{h_2}} \right\rfloor, |\{x \in C_s : \text{Valid}(x)\}|.\right) \quad (1)$$

4.3 Riddle Generation

We design a multi-agent iterative refinement framework to achieve *challenging* and *verifiable* criteria. Figure 2 illustrates the framework. It takes ground truth answers $t \in \mathcal{T}$ from the Riddle Answer Taxonomy and refines riddles through five stages. The pseudocode can refer to Algorithm 1 in Appendix A.1, and examples of generated riddles are provided in Appendix A.1.4.

Stage 1: Riddle Draft Generation. To obtain initial drafts, we require: (1) *metaphorical* language obscuring the answer; (2) sufficient *challenging* through multi-layered metaphors. Our prompt template (Table 4) includes: (1) example e_t demonstrating metaphorical reasoning patterns; (2) Wikidata descriptions d_t for t, providing contextual information. Formally, for $t \in \mathcal{T}$, we generate draft $e^{(0)}$ using model M_g: $e^{(0)} = M_g(t, e_t, d_t)$.

Stage 2: Solvers' Answer Collection. To prepare for subsequent assessment stages, we collect answers from a diverse set of solvers. We employ solver set $\mathcal{M} = \{M_s^{(1)}, M_s^{(2)}, \ldots, M_s^{(k)}\}$, where each $M_s^{(i)} \in \mathcal{M}$ represents different *architectural origins* (model families, training paradigms) and *capability levels*. To prevent bias, solvers must be *distinct* from generator, reviewer, and reviser models. Each $M_s^{(i)} \in \mathcal{M}$ independently generates answer $a_i \leftarrow M_s^{(i)}.\text{Solve}(e^{(r)})$ for draft $e^{(r)}$ at round r, forming $\mathcal{A}^{(r)} = \{a_1, a_2, \ldots, a_k\}$.

Stage 3: Reviewer Assessment. To assess riddle quality along *verifiability* and *challengingness*, we employ two steps: (1) reviewer independently evaluates each solver's answer based solely on the riddle, without ground truth access; (2) then, independent judgments are cross-validated against ground truth to diagnose issues (excessive simplicity, unsolvability, multiple solutions).

[1]Please refer to Appendix A.1.4 for more examples.

[2]Specifically, the version of the adopted model is high version of GPT-5-2025-08-17.

[3]https://qrank.toolforge.org/

Reviewer Independent Assessment. Reviewer M_r receives riddle $e^{(r)}$ and answer set $\mathcal{A}^{(r)} = \{a_1, a_2, \ldots, a_k\}$ from Stage 2. For each $a_i \in \mathcal{A}^{(r)}$, reviewer evaluates correctness given only $e^{(r)}$, producing $c_i = M_r(a_i, e^{(r)}) \in \{0, 1\}$, yielding $\mathcal{C}^{(r)} = \{c_1, c_2, \ldots, c_k\}$.

Reviewer Cross-Validation. A riddle is *solvable* if at least one answer is judged correct and matches ground truth:

$$\text{Solvable}(e^{(r)}, t) \iff \exists i : (c_i = 1) \wedge (a_i = t) \quad (2)$$

For *Pass (PASS)* status, a riddle must satisfy: (1) *Verifiable* and (2) *Challenging*:

$$\begin{cases} \text{Verifiable}(e^{(r)}, t) \iff \text{Solvable}(e^{(r)}, t) \wedge \\ \qquad\qquad \neg(\exists i : (c_i = 1) \wedge (a_i \neq t)) \\ \text{Challenging}(e^{(r)}, t) \iff \text{PassRate}(e^{(r)}, t) \geq \theta_p \end{cases} \quad (3)$$

where $\text{PassRate}(e^{(r)}, t)$ is the rate of answers both judged correct and matching ground truth:

$$\text{PassRate}(e^{(r)}, t) = \frac{|\{i : (c_i = 1) \wedge (a_i = t)\}|}{k} \quad (4)$$

Cross-validation yields four mutually exclusive outcomes: (1) *Multiple Solutions (MULTI)*: at least one answer judged correct but $\neq t$, indicating ambiguity; (2) *Too Easy (EASY)*: all answers $= t$, indicating insufficient challenge; (3) *Unsolvable (UNSOLV)*: no answer judged correct or no correct answer matches t, indicating excessive obscurity; (4) *Pass (PASS)*: solvable and challenging, indicating single unambiguous answer requiring genuine reasoning. Formally:

$$\begin{cases} \text{MULTI}(e^{(r)}, t) \iff \exists i : (c_i = 1) \wedge (a_i \neq t) \\ \text{EASY}(e^{(r)}, t) \iff \forall i : a_i = t \\ \text{UNSOLV}(e^{(r)}, t) \iff \neg\text{Solvable}(e^{(r)}, t) \\ \text{PASS}(e^{(r)}, t) \iff \text{Verifiable}(e^{(r)}, t) \wedge \\ \qquad\qquad \text{Challenging}(e^{(r)}, t) \end{cases} \quad (5)$$

These four outcomes provide clear signals for the next refinement stage.

Stage 4: Reviser Refinement. Reviser M_v refines problematic riddles based on Stage 3 outcomes, addressing ambiguity, insufficient challenge, or unsolvability. Formally, given $e^{(r)}$ and outcome $d^{(r)} \in \{\text{MULTI}, \text{EASY}, \text{UNSOLV}\}$, reviser generates $e^{(r+1)} = M_v(e^{(r)}, d^{(r)})$. Then, the refined riddle $e^{(r+1)}$ returns to Stage 2, establishing an iterative loop until PASS status or exceeding $R_{\max}$ rounds.

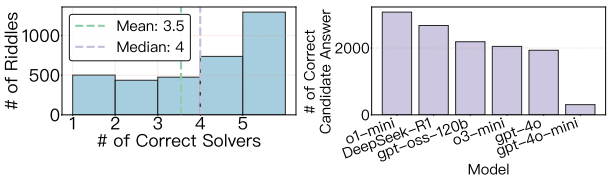

Figure 4: The statistics of valid generated riddles. The distribution of correct solver answer counts per riddle (left) and the correctness of each riddle across different solvers (right).

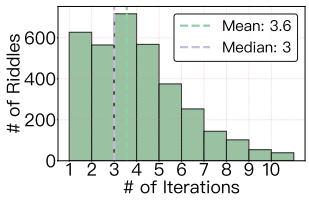

Figure 5: The distribution of iteration counts per riddle during generation.

Stage 5: Pass Rate Filter. We apply a final filter to ensure challengingness: only riddles with $\text{PassRate}(e^{(r)}, t) \geq \theta_p$ (Equation 4) are retained; others are rejected.

Riddles passing all five stages are *valid riddles*, forming well-verified reasoning tasks. We adopt GPT-5 as generator, reviewer, and reviser[4]. With regard to the solver set[5], we adopt o1-mini (OpenAI, 2024), Deepseek-R1 (Guo et al., 2025), GPT-oss-120b (OpenAI, 2025a), GPT-4o (Hurst et al., 2024), GPT-4o-mini (Hurst et al., 2024), and o3-mini (OpenAI, 2025c). Several parameters control riddle difficulty: (1) *Solver Set*: stronger solvers produce harder riddles; (2) *Pass-rate Threshold*: lower thresholds produce harder riddles. Detailed prompts and hyperparameter settings for each agent can refer to Appendix A.1.

4.4 Riddle Statistics

Riddle Answer Taxonomy Statistics. Figure 3 presents taxonomy statistics: 5,466 riddle answers demonstrating (1) *diversity*: 3 H_1, 11 H_2, 44 H_3 categories; (2) *balance*: balanced H_3 categories under each H_2.

Riddle Generation Statistics. Using six solvers, we generated 3,444 valid riddles satisfying all five

[4]Specifically, the version of the adopted models are medium version of GPT-5-2025-08-17.

[5]Specifically, the version of the adopted solvers are o1-mini-2024-09-12, GPT-4o-2024-11-20, GPT-4o-mini-2024-07-18

quality dimensions[6]. Figure 4 (left) shows the distribution of correct solver answer counts per riddle, indicating varying difficulty; median/mean suggest moderate overall difficulty. Figure 4 (right) shows solver performance: o1-mini perform the best while GPT-4o-mini perform the worst. Finally, Figure 5 shows iteration count distribution.

5 Training Recipe

We adopt the DAPO algorithm (Yu et al., 2025) for reinforcement learning with verifiable rewards (RLVR), which has demonstrated significant improvements in reasoning (Guo et al., 2025; Yu et al., 2025). The loss function is:

$$\mathcal{J}_{\text{DAPO}}(\theta) = \mathbb{E}_{(e,t)\sim\mathcal{D},\{o_i\}_{i=1}^{G}\sim\pi_{\theta_{\text{old}}}(\cdot|e)} \Bigg[\frac{1}{\sum_{i=1}^{G}|o_i|} \sum_{i=1}^{G}\sum_{j=1}^{|o_i|} \min\Big(r_{i,j}(\theta)\hat{A}_{i,j}, \quad (6)$$
$$\text{clip}\Big(r_{i,j}(\theta), 1-\varepsilon_{\text{low}}, 1+\varepsilon_{\text{high}}\Big)\hat{A}_{i,j}\Big)\Bigg]$$

where θ denotes model parameters, $\mathcal{D}$ is the training dataset, (e,t) is a riddle-answer pair, G is the group size, $\pi_{\theta_{\text{old}}}$ is the old policy for importance sampling, o_i is the i-th output sequence of length $|o_i|$, $r_{i,j}(\theta) = \frac{\pi_\theta(o_{i,j}|e,o_{i,<j})}{\pi_{\theta_{\text{old}}}(o_{i,j}|e,o_{i,<j})}$ is the importance sampling ratio at position j, $\hat{A}_{i,j}$ is the advantage function, and ε_{low} and $\varepsilon_{\text{high}}$ are clipping parameters. DAPO's *dynamic sampling* filters out groups with uniformly zero or maximum rewards, enabling effective and stable learning signals.

The advantage function uses group normalization: $\hat{A}_{i,j} = R_i - \bar{R}$, where $R_i \in \{0,1\}$ is the reward for output i, and $\bar{R} = \frac{1}{G}\sum_{i=1}^{G} R_i$ is the baseline. Since generated riddles are verifiable, rewards are computed via exact match (1 if extracted answer matches ground truth, 0 otherwise).

6 Experiments

6.1 Experiment Setup

Benchmark. To evaluate the impact of Metaphor Reasoning on reasoning abilities in other domains, we employ six categories of Out-of-Distribution (OOD) reasoning across 7 benchmarks: Logical Reasoning (Ma et al.; Chen et al., 2025); Commonsense Reasoning (Talmor et al., 2019); Natural Language Inference (Zellers et al., 2019); Math Reasoning[7]; Science, Technology, Engineering, and Mathematics (STEM) Reasoning (Rein et al.); Out-of-Distribution (OOD) Riddle Reasoning (Lin et al., 2021). Detailed descriptions of each benchmark are provided in Appendix A.2.1.

Backbones. We employ reasoning models as the backbones for reinforcement learning (RL) training, as these models have been pre-trained with explicit reasoning capabilities that provide a solid foundation for further enhancement through metaphor reasoning (Guo et al., 2025). To comprehensively evaluate the generalization of our approach, we investigate models across different *parameter scales* and *architectural generations*, including Qwen3-8B (Yang et al., 2025), Qwen3-14B (Yang et al., 2025), and QwQ-32B (Team, 2025), which is a reasoning model training on Qwen2.5-32B-instruct (Qwen et al., 2025).

The implementation details of training and evaluation are provided in Appendix A.2.2.

6.2 Experiment Results

Metaphor reasoning = Meta-reasoning?. According to Table 2, Metaphor reasoning can broadly enhance the reasoning abilities of models across diverse reasoning domains. Remarkably, training on only 3,444 metaphorical riddles yields substantial improvements across six out-of-distribution reasoning domains, demonstrating the efficiency and meta-reasoning nature of metaphor reasoning. This small-scale training dataset highlights that metaphor reasoning functions as a meta-ability that can be effectively transferred with minimal data. First, QwQ-32B exhibits remarkable improvements across all six reasoning categories relative to its backbone. Also, models of different scales and versions also demonstrate improvements, further validating the generalization of our approach. With regard to overall performance, QwQ-32B$_{\text{MetaR}}$ surpasses significantly larger models such as DeepSeek-R1 (671B parameters), as well as closed-source models such as GPT-4o. Notably, QwQ-32B$_{\text{MetaR}}$ outperforms GPT-oss-120B on four benchmarks across reasoning domains.

Model Scale. Generalization depends critically on model scale. QwQ-32B shows consistent improvements across all categories, while Qwen3-14B exhibits selective gains with slight declines in

[6]The reduction from 5,466 answers to 3,444 riddles occurs because some answers fail to generate qualifying riddles even after $R_{\max}$ refinement rounds and are thus discarded.

[7]https://artofproblemsolving.com/wiki/index.php/AIME_Problems_and_Solutions

Model	Math Reasoning	STEM Reasoning	Commonsense Reasoning	NL Inference	Logical Reasoning		OOD Riddle Reasoning	Overall
	AIME2024	GPQA	CommonQA	HellaSwag	Enigmata	KORBench	RiddleSense	
Closed-Source Models								
GPT-4o-mini	9.8	42.4	83.4	85.0	10.6	43.2	82.1	50.9
GPT-4o	11.9	46.5	84.3	90.8	21.8	51.6	88.7	56.5
o1-mini	87.7	75.6	85.3	88.5	89.0	61.7	94.4	83.2
o3	89.5	83.1	87.2	**93.7**	**92.7**	64.2	**95.3**	86.5
GPT-5	**93.7**	**89.4**	**88.0**	92.9	89.4	**67.2**	94.1	**87.8**
Open-Source Models								
Qwen3-8B	64.2	59.2	83.0	79.3	43.2	51.4	81.3	65.9
QwQ-32B	68.2	60.0	**85.0**	78.5	30.0	**66.0**	84.6	67.5
Qwen3-14B	69.0	63.3	84.2	**86.7**	44.5	56.2	85.2	69.9
DeepSeek-R1	77.4	69.2	69.9	85.1	49.8	62.8	79.2	70.3
GPT-oss-120B	**78.9**	**71.2**	84.4	83.1	**87.5**	63.6	**90.6**	**79.9**
Ours								
$\text{Qwen3-8B}_{\text{MetaR}}$	$64.0^{-0.2\%}$	$56.0^{-3.2\%}$	$84.2^{+1.2\%}$	$81.2^{+1.9\%}$	$41.3^{-1.9\%}$	$51.4^{+0.0\%}$	$84.3^{+3.0\%}$	$66.1^{+0.3\%}$
$\text{Qwen3-14B}_{\text{MetaR}}$	$68.9^{-0.1\%}$	$62.9^{-0.4\%}$	$84.6^{+0.4\%}$	$87.7^{+1.0\%}$	$\mathbf{51.8}^{+7.3\%}$	$57.6^{+1.4\%}$	$88.3^{+3.1\%}$	$71.7^{+2.6\%}$
$\text{QwQ-32B}_{\text{MetaR}}$	$\mathbf{74.4}^{+6.2\%}$	$\mathbf{63.6}^{+3.6\%}$	$\mathbf{86.1}^{+1.1\%}$	$\mathbf{88.5}^{+10.0\%}$	$40.3^{+10.3\%}$	$\mathbf{70.4}^{+4.4\%}$	$\mathbf{91.4}^{+6.8\%}$	$\mathbf{73.5}^{+8.9\%}$

Table 2: Main experimental results comparing baseline models and MetaR-enhanced models across different reasoning tasks. The superscript differences represent changes relative to the backbone model.

Model	AIME2024	GPQA	CommonQA	HellaSwag	Enigmata	KORBench	RiddleSense	Overall
$\text{Qwen3-14B}_{\text{MetaR}}$	68.9	62.9	84.6	87.7	51.8	57.6	88.3	71.7
$\text{Qwen3-14B}_{\text{MetaR\_DS}}$	$67.2^{-1.7\%}$	$63.0^{+0.1\%}$	$84.2^{-0.4\%}$	$86.2^{-1.5\%}$	$50.9^{-0.9\%}$	$55.6^{-2.0\%}$	$86.1^{-2.2\%}$	$69.8^{-1.9\%}$
$\text{Qwen3-8B}_{\text{MetaR}}$	64.0	56.0	84.2	81.2	41.3	51.4	84.3	66.1
$\text{Qwen3-8B}_{\text{MetaR\_DS}}$	$58.5^{-5.5\%}$	$53.2^{-2.8\%}$	$83.6^{-0.6\%}$	$80.4^{-0.8\%}$	$42.3^{+1.0\%}$	$52.8^{+1.4\%}$	$83.7^{-0.6\%}$	$63.7^{-2.4\%}$

Table 3: Dynamic ablation experimental results comparing MetaR models with and without dynamic reasoning.

Math and STEM. The 8B model shows limited generalization, improving only in OOD riddle reasoning, commonsense reasoning, and natural language inference. This scale-dependent pattern indicates that larger models better internalize and transfer metaphor reasoning strategies.

Reasoning Domain. Metaphor reasoning training exhibits differential effectiveness across reasoning domains. As shown in Table 2, OOD riddle reasoning shows the strongest improvements across all model scales, while commonsense reasoning and natural language inference also show consistent improvements. Math and STEM reasoning show gains only in larger models, suggesting greater reliance on domain-specific knowledge.

Training Recipe. Dynamic sampling is crucial for RL training. According to Table 3, removing dynamic sampling leads to performance degradation across different domains. We analyze the underlying reasons as follows. First, as shown in Figure 6, dynamic sampling yields smoother reward curves. Second, dynamic sampling increases entropy, ensuring greater sampling diversity (Cui et al., 2025). Moreover, we found that dynamic sampling produces shorter responses, improving token efficiency (Luo et al., 2025) (Figure 7).

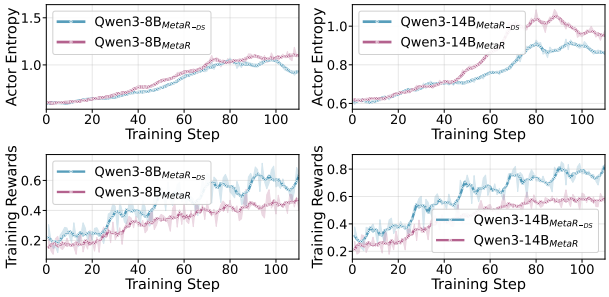

Figure 6: The training dynamics of models with and without dynamic sampling (indicated by "-DS").

6.3 Analysis

We analyze the underlying mechanisms of metaphor reasoning training on general reasoning capabilities from two perspectives.

Reasoning Domain Similarity. To analyze why generalization results vary across domains, we take Qwen3-8B as an example and examine the outputs

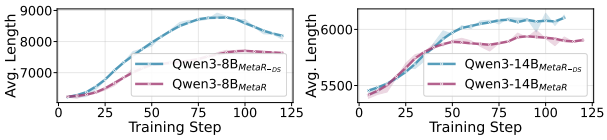

Figure 7: The average response length of models on the test set across different training steps.

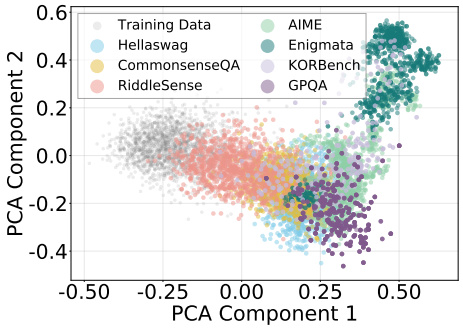

Figure 8: Response similarity analysis across different reasoning domains.

of the final model checkpoint on the training set and test sets from both embedding and vocabulary overlap perspectives. First, we perform semantic similarity analysis using Principal Component Analysis (PCA) visualization (Abdi and Williams, 2010)[8]. According to Figure 8, RiddleSense responses are closest to the training set in the semantic embedding space, while AIME and GPQA responses are farthest, indicating that RiddleSense shares more similar reasoning patterns with the metaphor reasoning training data. Also, we compute vocabulary overlap using Jaccard similarity (Jaccard, 1912) and Dice Coefficient (Dice, 1945), and according to Figure 9, the lexical patterns further validate this alignment. In conclusion, domains with aligned patterns benefit directly from metaphor reasoning training, while those requiring different approaches need larger model capacity to adapt.

Reasoning Pattern Evolution. We analyze the top-10 words with the largest frequency increases after metaphor reasoning training (Figure 10). These words fall into three key categories: *reflection* (e.g., "check", "not"), *perspective switching* (e.g., "let's", "maybe"), and *careful deliberation* (e.g., "think", "so"). This lexical shift demonstrates that metaphor reasoning training stimulates the model's deep thinking capabilities.

[8]Sentence-transformers embeddings (Reimers and Gurevych, 2020) are used for semantic similarity analysis. Component 1 and Component 2 capture the largest variance in response embeddings for dimensionality reduction.

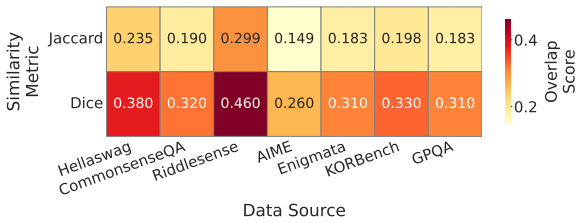

Figure 9: Vocabulary overlap analysis across different reasoning domains. The overlap score measures lexical similarity. Higher scores indicate greater shared vocabulary between domains.

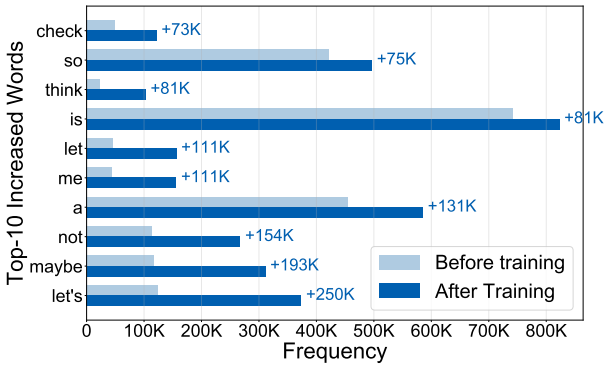

Figure 10: Top-10 increased words after metaphor reasoning training. K represents thousands. The counts are based on the number of words separated by spaces.

7 Conclusion

We present the first systematic study of metaphor reasoning's impact on broader reasoning capabilities in LLMs. We propose METAR, an automated framework synthesizing high-quality metaphorical riddles via answer taxonomy and multi-agent refinement. Training on 3,444 metaphorical riddles improves performance across six out-of-distribution reasoning domains.

Despite using only riddle solving with thousands of examples, our experiments demonstrate that metaphor reasoning functions as a meta-reasoning ability that transfers effectively with minimal data. Our work offers an important perspective: rather than developing task-specific solutions, we should identify core meta-abilities that generalize across diverse scenarios. Once properly trained, such meta-abilities serve as powerful transferable skills, reducing computational costs and data annotation requirements. Future work will explore additional task formulations beyond riddles to further investigate metaphor reasoning's transfer potential.

Limitations

While our work demonstrates the effectiveness of metaphor reasoning as a meta-reasoning ability, several limitations should be acknowledged. First, despite our multi-agent iterative refinement framework, guaranteeing completely unique answers for every riddle remains challenging due to natural language ambiguity. However, our reviewer agent employs independent assessment and cross-validation (§4.3), ensuring only riddles passing the *Verifiable* criterion are included, with experimental results demonstrating effectiveness (Table 2). Second, while our evaluation covers six reasoning categories across 7 benchmarks, important domains like causal or temporal reasoning are excluded. However, our evaluation breadth—from abstract logical to concrete mathematical reasoning—provides strong evidence for the meta-reasoning nature. Finally, experiments are conducted exclusively on the Qwen family (Qwen3-8B, Qwen3-14B, QwQ-32B), though other architectures like Llama and DeepSeek (Guo et al., 2025) exist. However, comprehensive validation across different versions, parameter scales (8B, 14B, 32B), and training paradigms provides robust evidence for generalization.

Acknowledgments

We acknowledge the use of Cursor (https://github.com/cursor/cursor) as an AI-assisted writing tool in the preparation of this manuscript. Specifically, Cursor assisted with writing polishment for the initial draft of this paper. Additionally, for the experiments section, Cursor helped generate code for data visualization and figure generation. During the development of the riddle answer taxonomy in §4 and the multi-agent framework, Cursor assisted with debugging and code development. The core ideas, framework design, and experimental design of this work were independently conceived and developed by the authors.

References

Hervé Abdi and Lynne J Williams. 2010. Principal component analysis. *Wiley interdisciplinary reviews: computational statistics*, 2(4):433–459.

Abhishek Arora, Emily Silcock, Melissa Dell, and Leander Heldring. 2024. Contrastive entity coreference and disambiguation for historical texts. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6174–6186.

Tuhin Chakrabarty, Xurui Zhang, Smaranda Muresan, and Nanyun Peng. 2021. MERMAID: Metaphor generation with symbolism and discriminative decoding. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4250–4261, Online. Association for Computational Linguistics.

Jiangjie Chen, Qianyu He, Siyu Yuan, Aili Chen, Zhicheng Cai, Weinan Dai, Hongli Yu, Qiying Yu, Xuefeng Li, Jiaze Chen, and 1 others. 2025. Enigmata: Scaling logical reasoning in large language models with synthetic verifiable puzzles. In *The Thirty-Ninth Annual Conference on Neural Information Processing Systems*.

Puli Chen, Cheng Yang, and Qingbao Huang. 2024. Merely judging metaphor is not enough: Research on reasonable metaphor detection. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 5850–5860, Miami, Florida, USA. Association for Computational Linguistics.

Ganqu Cui, Yuchen Zhang, Jiacheng Chen, Lifan Yuan, Zhi Wang, Yuxin Zuo, Haozhan Li, Yuchen Fan, Huayu Chen, Weize Chen, Zhiyuan Liu, Hao Peng, Lei Bai, Wanli Ouyang, Yu Cheng, Bowen Zhou, and Ning Ding. 2025. The entropy mechanism of reinforcement learning for reasoning language models. *CoRR*, abs/2505.22617.

Lee R Dice. 1945. Measures of the amount of ecologic association between species. *Ecology*, 26(3):297–302.

Giancarlo Guizzardi, Alessander Botti Benevides, Claudenir M. Fonseca, Daniele Porello, João Paulo A. Almeida, and Tiago Prince Sales. 2022. UFO: unified foundational ontology. *Appl. Ontology*, 17(1):167–210.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Qianyu He, Sijie Cheng, Zhixu Li, Rui Xie, and Yanghua Xiao. 2022. Can pre-trained language models interpret similes as smart as human? In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7875–7887.

Qianyu He, Xintao Wang, Jiaqing Liang, and Yanghua Xiao. 2023a. Maps-kb: A million-scale probabilistic simile knowledge base. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 6398–6406.

Qianyu He, Yikai Zhang, Jiaqing Liang, Yuncheng Huang, Yanghua Xiao, and Yunwen Chen. 2023b. HAUSER: Towards holistic and automatic evaluation

of simile generation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12557–12572, Toronto, Canada. Association for Computational Linguistics.

Jingcheng Hu, Yinmin Zhang, Qi Han, Daxin Jiang, Xiangyu Zhang, and Heung-Yeung Shum. 2025. Open-reasoner-zero: An open source approach to scaling up reinforcement learning on the base model. In *The Thirty-Ninth Annual Conference on Neural Information Processing Systems*.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Paul Jaccard. 1912. The distribution of the flora in the alpine zone. *New phytologist*, 11(2):37–50.

Yifan Jiang, Filip Ilievski, and Kaixin Ma. 2024. SemEval-2024 task 9: BRAINTEASER: A novel task defying common sense. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 1994–2008, Mexico City, Mexico. Association for Computational Linguistics.

Omid Khatin-Zadeh, Hassan Banaruee, and Babak Yazdani-Fazlabadi. 2022. A cognitive perspective on basic generic metaphors and their specific-level realizations. *Polish Psychological Bulletin*, pages 60–65.

Oliver Kramer. 2025. Conceptual metaphor theory as a prompting paradigm for large language models. *arXiv preprint arXiv:2502.01901*.

George Lakoff. 1993. The contemporary theory of metaphor.

George Lakoff and Mark Johnson. 2024. *Metaphors we live by*. University of Chicago press.

Duy Le, Kent Ziti, Evan Girard-Sun, Sean O'Brien, Vasu Sharma, and Kevin Zhu. 2025. Filtering for creativity: Adaptive prompting for multilingual riddle generation in llms. *arXiv e-prints*, pages arXiv–2508.

Yu Xi Li, Bo Peng, Yu-Yin Hsu, and Chu-Ren Huang. 2024. EmbodiedBERT: Cognitively informed metaphor detection incorporating sensorimotor information. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 16868–16876, Miami, Florida, USA. Association for Computational Linguistics.

Yucheng Li, Frank Guerin, and Chenghua Lin. 2022a. The secret of metaphor on expressing stronger emotion. In *Proceedings of the 3rd Workshop on Figurative Language Processing (FLP)*, pages 39–43, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Yucheng Li, Chenghua Lin, and Frank Guerin. 2022b. CM-gen: A neural framework for Chinese metaphor generation with explicit context modelling. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6468–6479, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Bill Yuchen Lin, Ziyi Wu, Yichi Yang, Dong-Ho Lee, and Xiang Ren. 2021. Riddlesense: Reasoning about riddle questions featuring linguistic creativity and commonsense knowledge. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1504–1515.

Haotian Luo, Li Shen, Haiying He, Yibo Wang, Shiwei Liu, Wei Li, Naiqiang Tan, Xiaochun Cao, and Dacheng Tao. 2025. O1-pruner: Length-harmonizing fine-tuning for o1-like reasoning pruning. *CoRR*, abs/2501.12570.

Kaijing Ma, Xeron Du, Yunran Wang, Haoran Zhang, Xingwei Qu, Jian Yang, Jiaheng Liu, Xiang Yue, Wenhao Huang, Ge Zhang, and 1 others. Kor-bench: Benchmarking language models on knowledge-orthogonal reasoning tasks. In *The Thirteenth International Conference on Learning Representations*.

OpenAI. 2024. Learning to reason with llms.

OpenAI. 2025a. gpt-oss-120b & gpt-oss-20b model card.

OpenAI. 2025b. Introducing gpt-5.

OpenAI. 2025c. Introducing o3 and o4 mini.

Ioannis Panagiotopoulos, George Filandrianos, Maria Lymperaiou, and Giorgos Stamou. 2025. Riscore: Enhancing in-context riddle solving in language models through context-reconstructed example augmentation. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 9431–9455.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, and 25 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R Bowman. Gpqa: A graduate-level google-proof q&a benchmark. In *First Conference on Language Modeling*.

Elisa Sanchez-Bayona and Rodrigo Agerri. 2025. Metaphor and large language models: When surface features matter more than deep understanding. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 17462–17477.

ByteDance Seed, Jiaze Chen, Tiantian Fan, Xin Liu, Lingjun Liu, Zhiqi Lin, Mingxuan Wang, Chengyi Wang, Xiangpeng Wei, Wenyuan Xu, and 1 others. 2025. Seed1. 5-thinking: Advancing superb reasoning models with reinforcement learning. *arXiv preprint arXiv:2504.13914*.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *CoRR*, abs/2402.03300.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. Commonsenseqa: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4149–4158.

Kimi Team, Yifan Bai, Yiping Bao, Guanduo Chen, Jiahao Chen, Ningxin Chen, Ruijue Chen, Yanru Chen, Yuankun Chen, Yutian Chen, and 1 others. 2025. Kimi k2: Open agentic intelligence. *arXiv preprint arXiv:2507.20534*.

Qwen Team. 2025. Qwq-32b: Embracing the power of reinforcement learning.

Paul H Thibodeau and Lera Boroditsky. 2011. Metaphors we think with: The role of metaphor in reasoning. *PloS one*, 6(2):e16782.

Yuan Tian, Nan Xu, and Wenji Mao. 2024. A theory guided scaffolding instruction framework for LLM-enabled metaphor reasoning. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7738–7755, Mexico City, Mexico. Association for Computational Linguistics.

Xiaoyu Tong, Rochelle Choenni, Martha Lewis, and Ekaterina Shutova. 2024. Metaphor understanding challenge dataset for LLMs. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3517–3536, Bangkok, Thailand. Association for Computational Linguistics.

volcengine. 2025. verl: Volcano engine reinforcement learning for llms.

Shun Wang, Ge Zhang, Han Wu, Tyler Loakman, Wenhao Huang, and Chenghua Lin. 2024. MMTE: Corpus and metrics for evaluating machine translation quality of metaphorical language. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 11343–11358, Miami, Florida, USA. Association for Computational Linguistics.

Shuhang Xu and Fangwei Zhong. 2025. CoMet: Metaphor-driven covert communication for multi-agent language games. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7892–7917, Vienna, Austria. Association for Computational Linguistics.

Yu Yan, Sheng Sun, Zenghao Duan, Teli Liu, Min Liu, Zhiyi Yin, LeiJingyu LeiJingyu, and Qi Li. 2025. from benign import toxic: Jailbreaking the language model via adversarial metaphors. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4785–4817.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, and 1 others. 2025. Dapo: An open-source llm reinforcement learning system at scale. In *The Thirty-Ninth Annual Conference on Neural Information Processing Systems*.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. Hellaswag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800.

Yunxiang Zhang and Xiaojun Wan. 2022. Birdqa: A bilingual dataset for question answering on tricky riddles. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11748–11756.

A Appendix

A.1 Riddle Generation

A.1.1 Riddle Generation Algorithm

We formalize our multi-agent iterative refinement framework for riddle generation as Algorithm 1.

A.1.2 Prompt Templates for Riddle Generation

This section presents the prompt templates used in our multi-agent iterative refinement framework for riddle generation. Each stage of the framework employs a specialized prompt template to guide the language models in their respective roles.

Table 4 shows the prompt template for Stage 1 (Riddle Draft Generation), which guides the generator to create metaphor-based riddles using the ground truth answer and its descriptions. The template variables are populated as follows (referring to Algorithm 1):

1. `{label}`: The ground truth answer $t \in \mathcal{T}$ from the Riddle Answer Taxonomy, provided as input to the algorithm (line 11 in Algorithm 1).

2. `{descriptions}`: The Wikidata descriptions d_t for the ground truth answer, which provide rich contextual information to enable the model to craft sophisticated metaphorical connections (line 11 in Algorithm 1).

Table 5 shows the prompt template for Stage 2 (Solver Answer Collection), which guides language models to solve riddles by analyzing metaphors and logical clues from multiple perspectives. The template variables are populated as follows:

1. `{riddle}`: The current riddle draft $e^{(r)}$ at revision round r, where $e^{(0)}$ is the initial draft generated in Stage 1 (line 17 in Algorithm 1), and $e^{(r)}$ for $r > 0$ is the refined riddle from previous revision rounds (line 65 in Algorithm 1).

Table 6 shows the prompt template for Stage 3 (Reviewer Assessment), which guides the reviewer to independently evaluate each solver's answer based solely on the riddle itself, without access to the ground truth, thus avoiding bias caused by exposing the ground truth. The template variables are populated as follows:

1. `{riddle}`: The current riddle draft $e^{(r)}$ being evaluated (line 31 in Algorithm 1).

2. `{solver_summary}`: A formatted summary of all solver answers from Stage 2, constructed from the answer set $\mathcal{A}^{(r)} = \{a_1, a_2, \ldots, a_k\}$ collected in lines 22-26 in Algorithm 1, where each a_i is generated by solver $M_s^{(i)} \in \mathcal{M}$.

Table 7 shows the prompt template for Stage 4 (Reviser Refinement), which guides the reviser to refine problematic riddles based on diagnostic outcomes from Stage 3, addressing ambiguity (`MULTI`), insufficient challenge (`EASY`), or unsolvability (`UNSOLV`). The template variables are populated as follows:

1. `{riddle}`: The previous riddle draft $e^{(r-1)}$ that needs refinement (line 65 in Algorithm 1).

2. `{groundtruth_answer}`: The ground truth answer t from the algorithm input (line 11 in Algorithm 1).

3. `{reviewer_feedback}`: The diagnostic outcome $d^{(r-1)}$ determined in Stage 3, which can be `MULTI` (line 42 in Algorithm 1), `EASY` (line 44 in Algorithm 1), or `UNSOLV` (line 46 in Algorithm 1), indicating the specific issue that needs to be addressed.

4. `{solver_feedback}`: Information derived from the solver answer set $\mathcal{A}^{(r-1)}$ and correctness judgments $\mathcal{C}^{(r-1)}$ collected in previous stages, providing additional context about how different solvers interpreted the riddle (line 48-52 in Algorithm 1).

A.1.3 Hyperparameters

We set the following hyperparameters in our riddle generation pipeline: (1) *Quality threshold* $\tau = 20$ for filtering entities with sitelinks below the minimum threshold; (2) *Taxonomic quota* $K = 1000$ allocated to each second-level category; (3) *Maximum revision rounds* $R_{\max} = 10$; (4) *Pass-rate threshold* $\theta_p = 0.9$.

A.1.4 Examples

Riddle Answer Examples. Table 8 presents examples of riddle answers organized by the hierarchical taxonomy structure (H1, H2, H3), demonstrating the diversity of concepts, objects, and phenomena covered in our dataset.

Riddle Examples. Tables 9 and 10 present example riddles from different taxonomic categories, demonstrating the diversity and quality of riddles generated by our framework.

A.2 Experimental Setup

A.2.1 Benchmark Details

To evaluate the impact of Metaphor Reasoning on reasoning abilities in other domains, we employ six categories of Out-of-Distribution (OOD) reasoning across 7 benchmarks: (1) *Logical Reasoning*: KORBench (Ma et al.), a knowledge-orthogonal benchmark, and Enigmata (Chen et al., 2025), containing 36 puzzle reasoning tasks across seven categories; (2) *Commonsense Reasoning*: CommonsenseQA (Talmor et al., 2019), which tests the ability to reason about everyday situations and world knowledge; (3) *Natural Language Inference*: HellaSwag (Zellers et al., 2019), which requires inferring the most plausible next situation based on context; (4) *Math Reasoning*: American Invitational Mathematics Examination (AIME) 2024[9], containing challenging problems from mathematical competitions; (5) *Science, Technology, Engineering, and Mathematics (STEM) Reasoning*: GPQA Diamond (Rein et al.), a graduate-level science test; (6) *Out-of-Distribution (OOD) Riddle Reasoning*: RiddleSense (Lin et al., 2021), a collection of manually curated riddles in multiple-choice format.

A.2.2 Implementation Details

Training Details. We implement DAPO training using the verl framework (volcengine, 2025), a scalable reinforcement learning system for large language models. DAPO (Yu et al., 2025) is a variant of GRPO (Shao et al., 2024) that employs group-based advantage normalization and asymmetric policy clipping for stable reinforcement learning with verifiable rewards. Unlike traditional PPO-based methods, DAPO does not require a critic network and computes advantages directly from group-normalized rewards within each group of generated outputs. All hyperparameters used in our experiments are summarized in Table 11.

Evaluation Details. To ensure sufficient training signal and stable learning, we employ dataset repetition during training. Each dataset is repeated multiple times to balance the representation of different reasoning domains. Specifically, the repetition counts for each dataset are as follows: AIME is repeated 32 times, GPQA_RuleVerifier is repeated 16 times, and all other datasets are repeated 3 times. All results reported in this paper are statistically significant with $p < 0.001$.

[9] https://artofproblemsolving.com/wiki/index.php/AIME_Problems_and_Solutions

Algorithm 1: Riddle Generation via Solvers-Reviewer-Reviser Workflow

Input: Ground truth answer $t \in \mathcal{T}$ from Riddle Answer Taxonomy $\mathcal{T}$; Solver set $\mathcal{M} = \{M_s^{(1)}, M_s^{(2)}, ..., M_s^{(k)}\}$; Reviewer M_r; Reviser M_v; Generator M_g; Example e_t; Descriptions d_t; Maximum revision rounds $R_{\max}$; Pass-rate threshold θ_p.

Output: Final riddle $e^{(r)}$ or `REJECTED`.

`// Stage 1: Riddle Draft Generation`

$e^{(0)} \leftarrow M_g(t, e_t, d_t)$

$r \leftarrow 0$

while $r < R_{\max}$ **do**

  `// Stage 2: Solver Answer Collection`

  $\mathcal{A}^{(r)} \leftarrow \{\}$

  for $M_s^{(i)} \in \mathcal{M}$ **do**

    $a_i \leftarrow M_s^{(i)}.\text{Solve}(e^{(r)})$

    $\mathcal{A}^{(r)}.\text{add}(a_i)$

  `// Stage 3: Reviewer Assessment`

  $\mathcal{C}^{(r)} \leftarrow \{\}$

  for $a_i \in \mathcal{A}^{(r)}$ **do**

    $c_i \leftarrow M_r(a_i, e^{(r)}) \in \{0, 1\}$

    $\mathcal{C}^{(r)}.\text{add}(c_i)$

  pass_rate $\leftarrow \frac{|\{i:(c_i=1)\wedge(a_i=t)\}|}{k}$

  solvable $\leftarrow \exists i : (c_i = 1) \wedge (a_i = t)$

  verifiable $\leftarrow$ solvable $\wedge \neg(\exists i : (c_i = 1) \wedge (a_i \neq t))$

  challenging $\leftarrow$ (pass_rate $\geq \theta_p$)

  `// Determine diagnostic outcome`

  if $\exists i : (c_i = 1) \wedge (a_i \neq t)$ **then**

    $d^{(r)} \leftarrow$ MULTI

  else if $\forall i : a_i = t$ **then**

    $d^{(r)} \leftarrow$ EASY

  else if $\neg$*solvable* **then**

    $d^{(r)} \leftarrow$ UNSOLV

  else if *verifiable* $\wedge$ *challenging* **then**

    $d^{(r)} \leftarrow$ PASS

  `// Stage 5: Pass Rate Filter`

  if $d^{(r)} =$ *PASS* **then**

    return $e^{(r)}$

  `// Stage 4: Reviser Refinement`

  if $d^{(r)} =$ *MULTI* **then**

    revision_type $\leftarrow$ "clarify ambiguity"

  else if $d^{(r)} =$ *EASY* **then**

    revision_type $\leftarrow$ "increase challenge"

  else if $d^{(r)} =$ *UNSOLV* **then**

    revision_type $\leftarrow$ "address unsolvability"

  $r \leftarrow r + 1$

  $e^{(r)} \leftarrow M_v(e^{(r-1)}, d^{(r-1)})$

return `REJECTED`

Prompt for Riddle Draft Generation

You are an expert riddle creator specializing in crafting challenging, metaphor-based puzzles that test advanced reasoning abilities.

Background
Riddles achieve their mystique by using metaphors to obscure the answer (A) through comparisons to seemingly unrelated objects (B, C, D), creating an atmosphere of mystery and intrigue.
Example:

- **Riddle:** "I am a fortress with no doors, a treasure chest with golden stores."
- **Answer:** An egg
- **Analysis:** The "fortress" and "treasure chest" serve as metaphors for the egg. The shell resembles a sturdy "fortress," while the yolk represents the golden "treasure."

Task
Please generate a riddle using the following label as the answer. I have provided extensive information that you can use to design your riddle. Ensure there is no obvious information leakage—the riddle should be as cryptic as possible through the use of metaphors.
The answer of the riddle:
{label}
Description:
{descriptions}

Requirements:

- Use metaphorical language to obscure the answer
- Avoid direct information leakage
- Create a challenging, multi-layered puzzle
- Consider nested metaphors for increased difficulty
- Ensure the riddle tests genuine reasoning abilities

Output Format
Please wrap your riddle in the following XML tags:
`<riddle>Your riddle here</riddle>`

Table 4: The prompt template used for riddle draft generation in Stage 1. The template guides the language model to create metaphor-based riddles using the ground truth answer and its descriptions.

Prompt for Solver Answer Collection

You are an expert riddle solver with deep knowledge and strong logical reasoning skills. You have the ability to analyze complex riddles from multiple perspectives and angles.

When solving the given riddle, please:

- Think broadly and explore various possible answers from different dimensions
- Identify and analyze key words, metaphors, and logical clues in the riddle
- Synthesize all possibilities and select the most fitting and elegant answer
- Present only your final answer as the conclusion

Please enclose your answer within `<answer></answer>` tags.

Riddle:
{riddle}

Table 5: The prompt template used for solver answer collection in Stage 2. The template guides language models to solve riddles by analyzing metaphors and logical clues from multiple perspectives.

Prompt for Reviewer Assessment

You are a senior puzzle editor evaluating whether proposed answers correctly solve a riddle. Judge each answer based solely on the riddle text, independent of any groundtruth answer.

Riddle:
{riddle}

Solver responses:
{solver_summary}

Instructions:
For each answer provided, analyze and judge whether it correctly solves the riddle:

1. Explain your reasoning: Evaluate how well the answer fits the riddle constraints, clues, and requirements.
2. Determine correctness: Based on your reasoning, decide if the answer is correct or incorrect.
3. Note that answers can be: all correct, all incorrect, or a mix of both.
4. If the riddle has ambiguities or multiple valid interpretations, explicitly note them.

Return your judgments using ONLY the per-solver tagged format below. Preserve the solver order from the summary and emit no extra narration.

```
<solver solver_id="solver_precise">
<answer>the solver's final answer</answer>
<analysis>concise single-line justification</analysis>
<correctness>true</correctness>
</solver>
```

Formatting requirements:

- Use lowercase true or false inside `<correctness>`.
- Keep each `<analysis>` on one line; replace newlines with spaces.
- Include exactly one `<solver>` block for every solver shown in the summary.
- Do not add any other tags or text outside the solver blocks.

Table 6: The prompt template used for reviewer assessment in Stage 3. The template guides the reviewer to independently evaluate each solver's answer based solely on the riddle itself, without access to the ground truth, thus avoiding bias caused by exposing the ground truth.

Prompt for Reviser Refinement

You are an expert riddle editor. Given an existing riddle and editorial feedback, produce a revised riddle that preserves the original mystery while ensuring only the groundtruth answer fits.

Old Riddle
{riddle}

Groundtruth Answer
{groundtruth_answer}

Reviewer Feedback
{reviewer_feedback}

Solver Feedback
{solver_feedback}

Task
Write a brand-new riddle that resolves the reviewer feedback, avoids leaking the groundtruth answer directly, and keeps a comparable difficulty level.

Return only the revised riddle wrapped exactly in `<riddle></riddle>` tags with no extra commentary.

Table 7: The prompt template used for reviser refinement in Stage 4. The template guides the reviser to refine problematic riddles based on diagnostic outcomes from Stage 3, addressing ambiguity (`MULTI`), insufficient challenge (`EASY`), or unsolvability (`UNSOLV`).

H_1	H_2	H_3	**Example of Riddle Answers**
Abstract Entities	Conceptual Systems	ideology	nationalism, conservatism, fascism, feminism
Abstract Entities	Conceptual Systems	mathematical concept	area, volume, length, height
Abstract Entities	Roles	social status	citizen, prisoner, slave, reputation
Abstract Entities	Traits	personality trait	curiosity, geek, cardinal virtues
Objects	Mixed Origin	drink	coffee, tea, milk, beer
Objects	Mixed Origin	food	bread, rice, apple, pizza
Objects	Natural Entities	natural geographic object	mountain, ocean, valley, volcano
Objects	Physical Objects	architectural structure	house, building, bridge, road
Objects	Physical Objects	clothing	dress, hat, trousers, jeans
Objects	Physical Objects	home appliance	refrigerator, washing machine, mobile phone, desktop computer
Objects	Physical Objects	musical instrument	guitar, violin, drum, flute
Objects	Physical Objects	weapon	sword, rifle, pistol, firearm
Objects	Regions	administrative territorial entity	country, state, province, county
Objects	Regions	city	metropolis, big city, capital city, global city
Objects	Regions	country	empire, sovereign state, nation state, developing country
Objects	Social Entities	educational institution	school, university, college, kindergarten
Objects	Social Entities	organization	company, government, army, religion
Objects	Social Entities	political party	communist party, green party, big tent
Phenomena	Natural Phenomena	astronomical phenomenon	eclipse, meteor, sunset, solar eclipse
Phenomena	Natural Phenomena	disease	cancer, diabetes, asthma, influenza
Phenomena	Natural Phenomena	meteorological phenomenon	rain, snow, storm, thunder
Phenomena	Natural Phenomena	natural disaster	earthquake, flood, tsunami, avalanche
Phenomena	Natural Phenomena	season	spring, summer, autumn, winter
Phenomena	Social Processes	historical event	battle, armistice, ceasefire
Phenomena	Social Processes	public election	general election, presidential election
Phenomena	Social Processes	social process	education, migration, revolution, reform

Table 8: Example of Riddle Answers organized by taxonomic hierarchy.

H1	H2	H3	Riddle Answer	Riddle
Objects	Mixed Origin	Food	Banana	I am a crescent wrought by a moonsmith, sheathed in a cloak with three long seams. My choir hangs like a chandelier near the ceiling of a giant that is grass yet mimics a tree. As days pass, night-dust freckles my garment; the bread in my heart turns slowly into honey. By the scholars' ledger I am called a berry, though my beads are ghosts and I'm born without a wedding. Some of my kin must learn the language of fire before their song becomes sweet. What am I?
Objects	Social Entities	Organization	College	A learning village rings its hours by a tower; paths braid toward a central green. Porches don Greek in the swift season of choosing; whispers and handclasps bind those lettered doors. Three wardens tend my thresholds: one keeps the purse, one the rolls and times, one the yes-or-no. At a bell's strike, blue books bloom; hush gathers while credits accrue by the hour. Some houses send you off in two winters with short initials; others keep you four for the longer pair. My kin answer to surnames like Community, Junior, and Liberal Arts; crisp Saturdays drum my name across the field. Between red margin and blue lines, the ruling bears my title; a numbered plan—five-two-nine—saves toward me by name. Here, they say they're going to me; across the sea, they go to "uni." When tassels turn, AA, AS, BA, or BS trail your name; taller hoods are seldom cut beneath my roof. And when grit must do, they urge the old try that carries my name. What am I?
Objects	Physical Objects	Clothing	Jeans	I am a night-skinned map that brightens where the world keeps touching me. Two hollow roads run my length—pilgrims stand inside me to meet the ground. My edges wear paired scars, and the mouths of my caves are pinned by tiny copper suns that do not set. Within one cavern, a smaller echo hides—fit for a spark, a coin, a whisper. My front can grin with tempered teeth, then hush with a single cold kiss. Five mute moons circle my crown to leash a tamed serpent. Born to wrestle grit and gravel, I learned the manners of parlors, yet I never forgot the language of dust. Name me.
Phenomena	Natural Phenomena	Disease	Vaccine	I was christened with a pastoral echo, though I graze nowhere. I knock at the cistern before the siege; I mend no wounds—I drill the watch. I come disguised: a ghost of the raider, a splinter, or a rumor written in letters. I travel with winter as escort; through a small hill of flesh I pin my notice. At times I hid in sweetness; at others I rode a breeze to the gate. I hire no mercenaries; I teach your forgemasters to cast their own steel and keep the plans. Rehearsals ring by appointment; the red river's numbers speak of my lesson. My stamped trail opens ports and halls; batch marks and dates betray my path. When true banners rise, the city answers at once, already sure.

Table 9: Example riddles from different taxonomic categories (Part 1 of 2).

H1	H2	H3	Riddle Answer	Riddle
Phenomena	Social Processes	Social Process	Middle Ages	My gate is the hinge between marble courts and untested mirrors. I woke when the western eagle's colors bled from its robe. Deep stone forests rose; praying boughs caught wind and brewed saints in light. Debt sworn on steel parceled fields; bread bowed to blade and altar. Lamplight tillers of parchment sowed letters; iron tongues weighed the hours. Empty fairs were counted by a pale collector; earth, silk, and incense ate together. Across seals and hands, towns bought voices; shells and crosses stitched roads to hungry shrines. Grain of thunder humbled glittering mail; a lodestone taught keels to speak straight. Etched metal whispered more words than throats; halls knotted three above four beneath contentious roofs. Speak me plain—ten letters and a gap; no herald walks before my name.
Abstract Entities	Roles	Social Status	Reputation	I am the foreword they read before your mouth writes a line. I am woven from footprints on days now dry; the weavers are not on your payroll. Parliament and notaries may declare me anew, yet their stamps pass through me. Scales quarrel with digits and calm at a sigh; light air edits my measure. I unlatch iron paths and make velvet doors grow thorns with no hand upon a bolt. I leave a colorless patina; baths and polish do nothing. Your mirror will not find me; their memory will; a shaped tale can bruise me while your truth stands. A crown hoards me in silence; a fool spends me with a grin and wakes to poverty. A syllable can split me; a decade can stitch me; I refuse pockets. I may be born of roofs you never raised and die for nights you never lived. Marks, brands, and charters pretend to leash me; merchants hire keepers to herd my smoke. No ledger numbers me, yet prices bow when I swell or sink. In rooms you enter, the air has already chosen its lean; I set the chairs.
Abstract Entities	Traits	Personal Trait	Geek	In crowded rooms I dodge the clink of shallow glasses, choosing one tick, one topic, to tune until it sings. I string stray facts like LEDs on a midnight cable, small sparks others swept from the floor. The name they stuck on me first pricked like a burr— I hammered it flat into a badge I wear at cons. Under sawdust lights a lifetime ago it meant a sideshow, a tent-oddity with a chicken and a crowd. Now it powers squads that come to mend your screens, and turns verb when joy makes me explain. Not Greek, though many write me so; my middle peers through twin lenses. What word am I?

Table 10: Example riddles from different taxonomic categories (Part 2 of 2).

Hyperparameter	**Value**
Training Framework	verl (volcengine, 2025)
Train Batch Size	256
Mini Batch Size	256
Group Size (G, num_bon)	16
Input Length	2K
Response Length	32K
Training Steps	
QwQ-32B	125
Qwen3-14B	130
Qwen3-8B	120
Actor Learning Rate	1×10^{-6}
Clip Ratio (ε_{low} / $\varepsilon_{\text{high}}$)	0.28 / 0.2
KL Coefficient	0.0
Number of GPUs	
QwQ-32B	256 A100
Qwen3-14B	64 A100
Qwen3-8B	64 A100

Table 11: DAPO training hyperparameters for different model scales. Note that K denotes thousands of tokens.