Adaptive Originality Filtering: Rejection-Based Prompting and RiddleScore for Culturally Grounded Multilingual Riddle Generation

Anonymous ACL submission

Abstract

Language models are increasingly tested on multilingual creativity, demanding culturally grounded, abstract generations. Standard prompting methods often produce repetitive or shallow outputs. We introduce Adaptive Originality Filtering (AOF), a prompting strategy that enforces novelty and cultural fidelity via semantic rejection. To assess quality, we propose RiddleScore, a metric combining novelty, diversity, fluency, and answer alignment. AOF improves Distinct-2 (0.915 in Japanese), reduces Self-BLEU (0.177), and raises RiddleScore (up to +57.1% in Arabic). Human evaluations confirm fluency, creativity, and cultural fit gains. However, improvements vary: Arabic shows greater RiddleScore gains than Distinct-2; Japanese sees similar changes. Though focused on riddles, our method may apply to broader creative tasks. Overall, semantic filtering with composite evaluation offers a lightweight path to culturally rich generation—without fine-tuning.

1 Introduction

Large Language Models (LLMs) have revolutionized natural language processing (NLP) across a spectrum of applications, yet their generative abilities in creative, multilingual contexts remain underexplored and underperforming (Zhang and Wan, 2025; Ismayilzada et al., 2024). Tasks like riddle generation pose a unique challenge: success hinges not only on linguistic fluency but also on metaphorical abstraction, cultural resonance, and semantic ambiguity—all of which are frequently underrepresented in LLM training corpora (Sejnowski, 2023; Pawar et al., 2024). As LLMs are increasingly integrated into global educational and creative platforms, their limitations in culturally grounded generation constrain both inclusivity and expressive potential(Bulathwela et al., 2024; Spennemann, 2023).

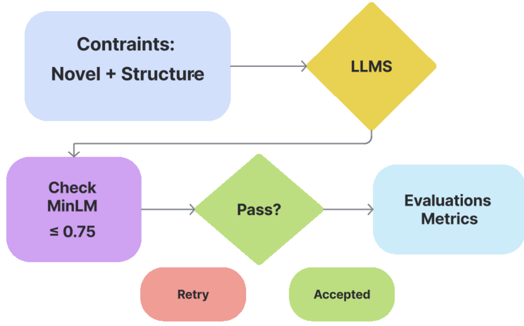

Figure 1: End-to-end pipeline to produce and verify riddles with LLMs (GPT-4o, R1, LLaMA). Constraints enforce novelty/structure; MiniLM tests semantic similarity with threshold ≤ 0.75. Failed results are regenerated; accepted ones are subjected to final checking.

Riddles, with their blend of metaphor, misdirection, and context-specific symbolism, provide a compelling benchmark for evaluating multilingual creativity in NLP. However, existing prompting strategies—zero-shot, few-shot, and chain-of-thought—often yield formulaic outputs or mistranslations, especially in semantically distant or morphologically rich languages (Wei et al., 2023a; Brown et al., 2020b). Current evaluation metrics such as BLEU, perplexity, or BERTScore are ill-equipped to assess riddle-specific traits like structural novelty, literary device density, or cultural fit (Sellam et al., 2020a; Dufter, 2021a).

To bridge these gaps, we propose Adaptive Originality Filtering (AOF), a prompting framework that enforces semantic novelty and lexical diversity through a cosine similarity-based rejection mechanism. Unlike typical generation strategies, AOF injects external control into the decoding loop, filtering out redundant or culturally dissonant outputs to elicit more original and resonant generations. Complementing AOF, we introduce RiddleScore, a composite evaluation metric that captures four dimensions central to high-quality riddles: Novelty, Diversity, Fluency, and Semantic Alignment. RiddleScore leverages pretrained language models alongside traditional metrics, and is calibrated to reflect human intuition across languages.

We benchmark AOF-enhanced prompting in three state-of-the-art LLMs: GPT-4o, LLaMA 3.1 and DeepSeek Reasoning in four language pairs (English,Chinese, Arabic, Japanese, French). Using the BiRdQA dataset (Zhang and Wan, 2022) under consistent decoding parameters, we evaluate outputs with Self-BLEU, Distinct-2, Cross-lingual BERTScore, and human judgment. Our results show that AOF significantly outperforms standard prompting baselines across both automatic and human evaluations. Notably, in Japanese, AOF-enhanced GPT-4o achieves a Self-BLEU of 0.177 and a Distinct-2 of 0.915, indicating reduced redundancy and heightened linguistic variety.

To structure our contributions more rigorously, we center our study around the following research questions:

- **RQ1:** Can rejection-based prompting (AOF) increase semantic novelty and lexical diversity across typologically diverse languages?
- **RQ2:** Does the proposed composite metric, *RiddleScore*, correlate with human judgments better than uniform-weighted baselines?
- **RQ3:** How do pretrained versus fine-tuned LLMs respond to AOF in multilingual riddle generation?

We address **RQ1** by showing that AOF with a cosine threshold of $\theta = 0.75$ significantly improves novelty and diversity across languages; in Japanese, it reduces Self-BLEU to 0.177 (–63.4%) and raises Distinct-2 to 0.915. For **RQ2**, RiddleScore aligns strongly with human judgments (Spearman $\rho = 0.83$), outperforming uniform baselines. For **RQ3**, we find that fine-tuned models benefit more from AOF than pretrained ones—achieving greater improvements in originality, fluency, and cultural fit. Chinese shows the most pronounced gains, with RiddleScore increasing by 48.3% ($0.453 \rightarrow 0.728$) and human ratings rising from 3.91 to 4.50.

2 Related Work

Multilingual and Cultural NLP Most work on riddles has focused on comprehension or solving rather than generation. Recent shared tasks such as SemEval-2024 Task 9 (Heavey et al., 2024) benchmark multilingual riddle solving with diverse unsupervised systems. RIScore (Panagiotopoulos et al., 2024) enhances contextual reasoning via in-context augmentation but does not explore generative capabilities. BiRdQA (Zhang and Wan, 2022) provides a multilingual benchmark but focuses on multiple-choice comprehension. In Chinese NLP, Xu et al. (Xu et al., 2022) incorporated cultural embeddings to improve riddle comprehension, while Tan et al. (Tan et al., 2016) explored classical Chinese radical riddles. Megatron-Turing NLG (Smith et al., 2022) includes riddles among its evaluation tasks but lacks task-specific generation. Figurative generalization remains difficult for multilingual LMs (Liu et al., 2022a), as metaphor and symbolism often fall outside pretrained representations (Dufter, 2021b). Sentence-level alignment models such as LASER (Chen and Avgustinova, 2021), XLM-R (Conneau et al., 2019), and MUSE (Lample and Conneau, 2019) improve transfer but collapse under poetic or rhetorical pressure. Our method explicitly addresses cultural fluency through semantic rejection and literary device filtering, ensuring metaphorical and idiomatic depth across languages.

Creative and Figurative Language Generation Creative NLP tasks—such as joke generation (Petrović and Matthews, 2013), metaphor synthesis (Chakrabarty et al., 2021), and story writing (Fan et al., 2018)—highlight the tension between novelty and fluency. Studies like GENIE (Tambwekar et al., 2019) and related prompting approaches (Zhang et al., 2020a) introduce generation frameworks for idea diversity, but often lack semantic constraints. Cross-lingual creativity remains underexplored: transformer-based models (Weller and Seppi, 2019) have begun to address humor generation, yet cultural adaptation remains limited. In Chinese, visual-pun riddles require multimodal cues (Zhou and Bisk, 2022), while poetic style transfer systems like Hafez (Ghazvininejad et al., 2017) aim to generate stylized literary output. Tan et al. (Tan et al., 2016) model riddle form in character-based composition. These works suggest the need for structured prompts or heuristics to scaffold creative reasoning. Our work differs by combining cultural-device filtering with a retry loop to enforce lexical and rhetorical novelty without additional supervision.

Prompting Strategies and Constraint-Based Generation Standard prompting methods such as few-shot and chain-of-thought (CoT) improve reasoning but tend to replicate memorized patterns (Brown et al., 2020a; Wei et al., 2023b). Recent methods like Self-Refine (Madaan et al., 2023), Reflexion (Krishna et al., 2023), and Tree-of-Thought (Yao et al., 2023) explore iterative improvement, while Auto-CoT (Zhang et al., 2022) and Selective CoT (Li et al., 2023) adapt prompt selection. Constraint-driven frameworks such as COLD decoding (Mou et al., 2022), EditCoT (Wang et al., 2024), Crescendo (Zhou et al., 2022), and Sketch-of-Thought (Aytes et al., 2025) offer structure-guided generation, but do not explicitly enforce cultural or semantic novelty. Creativity-centric methods such as SCILL (Dou et al., 2022) and CS4 (Atmakuru et al., 2024) demonstrate structure helps, but often lack filtering loops. Our Adaptive Originality Filtering framework unifies these threads by integrating rejection sampling, metaphor constraints, and interlingual filters into a single prompting loop.

Evaluation of Multilingual Generation While BLEU and BERTScore are widely used, they poorly reflect originality or cultural fit (Dang et al., 2022; Schmidtová and Wu, 2024). BLEURT (Sellam et al., 2020b) and COMET (Rei et al., 2020) improve robustness, but do not capture rhetorical or misdirectional quality. HUME (van der Lee et al., 2021) enables human-aligned evaluation but is domain-limited. Recent surveys (van der Lee et al., 2019; Cahill et al., 2009) highlight gaps in evaluating creative NLP. Multilingual creativity requires more than fluency—fluency is necessary but not sufficient. RiddleScore, our proposed metric, captures novelty (via semantic distance), lexical diversity, fluency, and answer coherence in a single interpretable score. It extends earlier work on figurative evaluation (Shutova, 2013; Falkum, 2009) and is explicitly validated by structured human annotation across language pairs.

3 Methodology

3.1 Adaptive Originality Filtering (AOF)

To overcome shortcomings of classical prompting techniques such as Chain-of-Thought and Few-Shot, which tend to copy riddles from pretraining data (Zhang and Wan, 2022), we present **Adaptive Originality Filtering (AOF)**, a prompting technique boosting novelty, lexical richness, and cultural adherence in riddle construction.

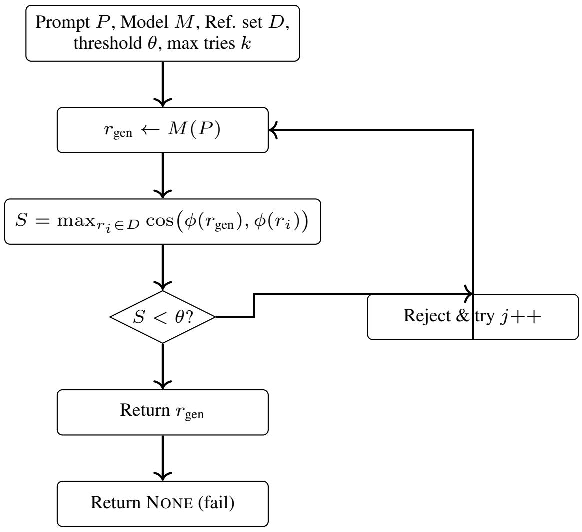

Figure 2: AOF rejection-sampling loop. Each candidate is generated, compared to reference riddles, and either accepted, rejected, or retried up to k attempts.

AOF combines three core mechanisms: (1) semantic similarity filtering, (2) rejection sampling, and (3) prompt-level constraints. For semantic filtering, a candidate riddle is encoded using MiniLM embeddings and matched to a reference set using cosine similarity. Extending from existing research where 0.75 serves as the inflection point where topical drift becomes primarily influenced by semantic novelty (Li et al., 2024; Lee, 2025), our novelty cutoff is set to be Candidates exceeding this threshold are rejected (Appendix M.1); the full rejection-sampling loop is given in Appendix M.2, and the prompt skeleton in Appendix M.

We verified a threshold-sensitivity study (Table 29, Appendix) that validates $\theta = 0.75$ as minimizing Self-BLEU and maximizing Distinct-2, with lower thresholds that allow template bleedthrough and higher thresholds that increase the failure rate by 14 %. Figure 2 shows a visualization of the rejection-sampling loop,

3.2 RiddleScore Metric

To evaluate multilingual riddle quality we introduce **RiddleScore**, a composite metric that captures four dimensions—*Novelty*, *Diversity*, *Fluency*, and *Semantic Alignment*. Formal definitions are in Appendix O, which also justifies the choice of the back-end models (MiniLM, Distinct-2, GPT-2.5 perplexity, and BERTScore) in a dedicated “Model Choice” paragraph.

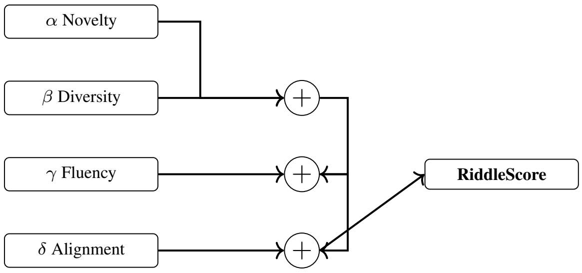

Figure 3: RiddleScore components and weights (α=0.30, β=0.20, γ=0.30, δ=0.20).

Each component is computed as follows:

- **Novelty**: cosine distance from BiRdQA riddles (MiniLM).
- **Diversity**: Distinct-2 bigram ratio (Li et al., 2016).
- **Fluency**: inverse perplexity under a frozen GPT-2.5 (Radford et al., 2019).
- **Semantic Alignment**: BERTScore against the riddle's answer (Zhang et al., 2020b).

The final score is a weighted sum

$$\begin{aligned}\text{RiddleScore} &= \alpha\,\text{Novelty} + \beta\,\text{Diversity} \\ &+ \gamma\,\text{Fluency} + \delta\,\text{Alignment}.\end{aligned} \tag{1}$$

with $\alpha = 0.30$, $\beta = 0.20$, $\gamma = 0.30$, and $\delta = 0.20$. The weights were searched by grid on a 120 sample dev set to maximize Spearman ρ with 5-point human scores (Table 30); the selected setting raises ρ from 0.71 (uniform) to 0.83. In addition, Appendix O Figure 9 shows how alternative weightings affect correlation with human scores. This mirrors the weight-tuning strategies of MetaMetrics (Winata et al., 2024) and HarmonicEval (Ohi et al., 2024). Figure 3 diagrams how the four components and their weights combine into riddlescore.

3.3 Experimental Setup

We test three LLMs—GPT-4o, LLaMA 3.1, and DeepSeek Reasoning—under five prompting strategies: Zero-Shot, Few-Shot, Chain-of-Thought, Adversarial (Wallace et al., 2019; Ribeiro et al., 2018), and AOF. All models are decoded with temperature 0.7, the default in most production chat systems and evaluation suites (e.g., SORRY-Bench) and shown to balance diversity and factuality in decoding studies (Xie et al., 2025; Lu et al., 2024).

Prompts are evaluated in five languages using the BiRdQA corpus of 15 k bilingual riddles (Zhang and Wan, 2022); exact prompt templates appear in Table 25. BiRdQA is uniquely suited for this evaluation, as it captures figurative reasoning, symbolic abstraction, and cultural idiomaticity, traits essential to assess cross-lingual creativity and semantic alignment in generative models(Liu et al., 2022a; Kabra, 2023). BiRdQA has been increasingly adopted in multilingual studies as a benchmark to evaluate figurative abstraction and cross-lingual generalization(Giadikiaroglou et al., 2024; Huang et al., 2025). Our Evaluation metrics include Self-BLEU (repetition), Distinct-2 (diversity), cross-lingual BERTScore (alignment), and the composite RiddleScore. Syntactic validity is verified using spaCy and Stanza. Full metric definitions, details of datasets, and other experimental materials are included in Appendix N.

4 Fine-Tuning of the GPT-4o Model

Objective and Motivation This fine-tuning was to refine solving and generating riddles in diverse languages by GPT-4o-2024-08-06. The riddles involve something beyond matching on a page—they require comprehension of metaphors, logical paradox, and novel misdirection. Our goal was to not only refine accuracy of answers but to also instill structural reasoning ability.

Methodological Overview We posed the problem as a supervised multiclass classification task with the BiRdQA dataset. The riddles were given as multiple choices, and cross-entropy loss was used to fine-tune the model. The reader can find complete details regarding dataset preprocessing, training procedure, and expanding the training set, respectively, in Appendix L.

Multiple-Choice Framing Overview Riddles were presented as four-choice multiple-choice questions with an eye to obtaining fine-grained discrimination between plausibly believable distractors. This format affected the inference strategy and generalizability of the model. The analysis of framing effects can be seen in Appendix L.5.

Prompting Strategies We tested five prompting methods on the fine-tuned model: Zero-Shot, Few-Shot, Chain-of-Thought (CoT), Adversarial, and Adaptive Originality Filtering (AOF). These correspond to the pre-trained experiments. See full prompt templates by reading Table 25.

Model Comparison Overview We compared our fine-tuned GPT-4o with several pre-trained baselines: GPT-4o (pre-trained), LLaMA 3.1, DeepSeek R1 with same evaluation metrics and prompts. Detailed results and discussion of methods are found in Appendix N.

Language	Prompting Method	Score (/5)
English	**AOF (Ours)**	**4.50**
	Few-Shot	3.20
	Zero-Shot	3.15
	Chain-of-Thought	3.85
	Adversarial	4.20
Chinese	**AOF (Ours)**	**4.50**
	Few-Shot	3.25
	Zero-Shot	3.50
	Chain-of-Thought	4.00
	Adversarial	3.80
Japanese	**AOF (Ours)**	**3.43**
	Few-Shot	3.36
	Zero-Shot	3.50
	Chain-of-Thought	3.57
	Adversarial	3.64
French	**AOF (Ours)**	**4.44**
	Few-Shot	3.78
	Zero-Shot	4.00
	Chain-of-Thought	4.33
	Adversarial	3.83
Arabic	**AOF (Ours)**	**4.92**
	Few-Shot	4.08
	Zero-Shot	3.72
	Chain-of-Thought	4.40
	Adversarial	4.30

Table 1: Average human evaluation scores (out of 5) for the fine-tuned GPT-4o across languages. Best per language in bold.

5 Human Evaluation

To capture riddle qualities not fully represented by automatic metrics, we performed human evaluations on four axes: *Fluency*, *Novelty*, *Cultural Fit*, and *Answerability*. Native or proficient speakers rated the riddle-answer pairs on a 1- to 5-likert scale using standardized rubrics, with hidden model labels to reduce bias (Appendix P).

5.1 Results

In both pre-trained and fine-tuned models, AOF prompting achieved the highest average scores in all languages, reaching **4.92 in Arabic** and **4.50 in English, Chinese and French** (Tables 1 and 2). These scores substantially exceed those of the zero-shot, few-shot, and chain-of-thought prompting, demonstrating the superiority of the AOF in producing culturally grounded and semantically coherent riddles. Annotators frequently highlighted AOF's *"poetic language, cultural anchoring, and structural coherence"* as reasons for higher ratings. For example, the French riddle *"Dans le jardin des mots, je suis une abeille, bourdonnant entre les lettres, mais je ne pique jamais. Que suis-je?"* ("In the garden of words, I am a bee, buzzing between the letters, but I never sting. What am I?") was rated highly for its metaphorical depth and native-like phrasing, reflecting AOF's ability to balance creativity with solvability.

Lang.	Prompting	Human Eval (/5)
English	**AOF (Ours)**	**3.85**
	Few-Shot	2.75
	Zero-Shot	2.50
	Chain-of-Thought	3.52
	Adversarial	3.60
Chinese	**AOF (Ours)**	**3.91**
	Few-Shot	2.63
	Zero-Shot	2.75
	Chain-of-Thought	3.45
	Adversarial	3.78
Japanese	**AOF (Ours)**	**3.36**
	Few-Shot	2.86
	Zero-Shot	2.79
	Chain-of-Thought	2.93
	Adversarial	3.29
French	**AOF (Ours)**	**4.50**
	Few-Shot	3.85
	Zero-Shot	3.77
	Chain-of-Thought	3.55
	Adversarial	4.00
Arabic	**AOF (Ours)**	**4.92**
	Few-Shot	4.20
	Zero-Shot	2.71
	Chain-of-Thought	4.40
	Adversarial	4.25

Table 2: Average human evaluation scores (out of 5) for pretrained models.

Human ratings align with RiddleScore trends: languages with the highest RiddleScore under AOF—**0.586 Arabic**, **0.728 Chinese**, **0.475 Japanese**, **0.468 French**, **0.586 English** (Table 5)—also show the largest human-rated gains. This convergence validates RiddleScore as a reliable proxy for human perception of creativity, fluency, and cultural fit. Together, confirming AOF prompting consistently outperforms other methods.

6 AOF Pretrained Evaluations

Pre-trained AOF prompts improve riddle quality across all languages by promoting metaphorical novelty and structural fluency, even without fine-tuning. Cross-lingually, DeepSeek R1 consistently

Language Pair	Prompting Method	GPT-4o	LLaMA 3.1	DeepSeek R1
English–Arabic	AOF (Ours)	0.373	0.378	**0.400**
	Zero-Shot	0.352	0.382	**0.400**
	Few-Shot	0.338	0.366	0.341
	Adversarial	0.296	0.292	0.305
English–Chinese	**AOF (Ours)**	0.434	0.330	**0.453**
	Zero-Shot	0.250	0.136	0.255
	Few-Shot	0.253	0.263	0.257
	Chain-of-Thought	0.247	0.246	0.239
	Adversarial	0.247	0.253	0.280
English–Japanese	AOF (Ours)	0.367	0.341	0.379
	Zero-Shot	0.351	0.363	0.323
	Few-Shot	0.346	0.353	0.324
	Chain-of-Thought	0.302	**0.490**	0.273
	Adversarial	0.338	0.361	0.336
English–French	AOF (Ours)	0.373	0.352	0.354
	Zero-Shot	0.410	0.423	**0.428**
	Few-Shot	0.330	0.327	0.329
	Chain-of-Thought	0.236	0.350	0.241
	Adversarial	0.242	0.251	0.234

Table 3: RiddleScore performance across language pairs and pretrained models (GPT-4o, LLaMA 3.1, DeepSeek R1). **Bold values indicate best-performing prompting method per model within each language pair**.

Lang. Pair	Prompting Method	Self-BLEU / Distinct-2
Eng–Arabic	Few-Shot	**0.233 / 0.826**
	AOF (Ours)	0.260 / 0.893
	Zero-Shot	0.391 / 0.752
	Chain-of-Thought	0.326 / 0.831
	Adversarial	0.320 / 0.810
Eng–Chinese	**AOF (Ours)**	**0.163 / 0.934**
	Zero-Shot	0.315 / 0.831
	Few-Shot	0.349 / 0.787
	Chain-of-Thought	0.305 / 0.828
	Adversarial	0.400 / 0.757
Eng–Japanese	**AOF (Ours)**	**0.177 / 0.915**
	Zero-Shot	0.431 / 0.752
	Few-Shot	0.326 / 0.778
	Chain-of-Thought	0.386 / 0.796
	Adversarial	0.327 / 0.748
Eng–French	AOF (Ours)	0.273 / 0.856
	Zero-Shot	0.289 / 0.867
	Few-Shot	0.323 / 0.835
	Chain-of-Thought	**0.256 / 0.892**
	Adversarial	0.359 / 0.793

Table 4: Self-BLEU (lower is better) and Distinct-2 (higher is better) for fine-tuned GPT-4o across prompting methods. Best combined performance per language pair in bold.

yields the highest RiddleScores (e.g., English: 0.400; Arabic: 0.400; Chinese: 0.453; Japanese: 0.475), suggesting strong compatibility with the AOF sampling rejection framework. These outputs combine lexical diversity with controlled syntactic rhythm (Koestler, 1964; Xu et al., 2018). For example, DeepSeek's Arabic riddle in Figure 6, Row 3 metaphorically compares a rooftop to an "eye fed by the city," demonstrating culturally grounded abstraction (Al-Marzouki, 2012).

DeepSeek R1 attains the highest Riddlescore in four of five languages: EN (0.400), AR (0.400), ZH (0.453), JA (0.379) - outperform GPT 4o / LLaMA 3.1 by 5-15 points (Table 3). While slightly more repetitive in Arabic (Self-BLEU: 0.585), R1 compensates with high lexical diversity—e.g., Distinct-2 scores of 0.845 in English and 0.674 in Chinese (Table 9)—and fluent metaphorical abstraction. Its Japanese riddle (Table 14, Row 3) showcasing the kind of poetic misdirection that aligns with high RiddleScore evaluations (Xu et al., 2018).

GPT-4o performs consistently in languages with moderate repetition (Self-BLEU $\approx$ 0.41–0.50; Table 9), high lexical variety (Distinct-2: 0.78–0.85), and RiddleScore values from 0.373 (FR/AR) to 0.453 (ZH) (Table 3), reflecting fluent but less figuratively ambitious riddles. Notably, in FR and ZH, GPT-4o exhibits literal translation tendencies that limit cultural nuance (Chan, 1996; Sun, 2006).

LLaMA 3.1 shows stylistic risk-taking (Distinct-2 $\approx$ 0.727–0.927; Table 9) but variable cohesion (RiddleScore: 0.330–0.378; Table 3), often blending innovative metaphors with uneven syntax or logical drift. For example, its JA riddle in Table 14 cleverly puns on the homophone *tsuru* (鶴/twine), linking cultural symbols via Shinto imagery (An, 2023).

Despite varied outputs, shared patterns emerge: AOF avoids template reuse, minimizes egocentric phrasing, and achieves cultural competence without tuning. These patterns, supported by Tables 3 and 9, validate the language-agnostic nature of the metaphor-rich generation. For complete evaluations, see Section B.

7 AOF Fine-Tuned Evaluations

Fine-tuning GPT-4o with AOF consistently enhances riddle quality for EN, ZH, FR, and AR by improving semantic creativity, lexical variation, and cultural mastery. The increases in RiddleScore range from 33.4% (AR) to 48.3% (ZH), as shown in Table 5. Self-BLEU reduces by 33–51% (Table 4), and Distinct-2 increases by 6–13%, confirming broad improvements in originality and fluency (Table 4) (Zhang et al., 2020b; Sellam et al., 2020b).

For example, a ZH riddle—"千言万语藏心怀" (lit. "A thousand words hide in the heart")—exemplifies the character "信" through orthographic metaphor and poetic condensation (Table 20, Row 2), echoing classical radical-based

Lang. Pair	Prompting Method	RiddleScore
Eng–Arabic	**AOF (Ours)**	**0.586**
	Few-Shot	0.364
	Zero-Shot	0.315
	Chain-of-Thought	0.313
	Adversarial	0.341
Eng–Chinese	**AOF (Ours)**	**0.728**
	Few-Shot	0.355
	Zero-Shot	0.350
	Chain-of-Thought	0.312
	Adversarial	0.348
Eng–Japanese	**AOF (Ours)**	**0.475**
	Few-Shot	0.334
	Zero-Shot	0.300
	Chain-of-Thought	0.307
	Adversarial	0.331
Eng–French	AOF (Ours)	0.352
	Few-Shot	0.350
	Zero-Shot	**0.468**
	Chain-of-Thought	0.347
	Adversarial	0.328

Table 5: Fine-tuned GPT-4o RiddleScore across language pairs. Best per pair in bold.

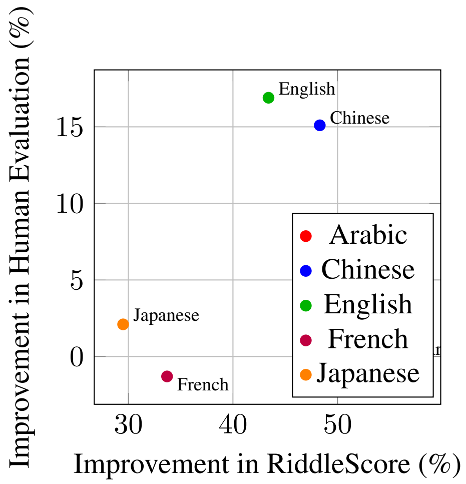

Figure 4: Correlation between fine-tuning gains in RiddleScore and human evaluation scores across five languages. Each point represents one language; higher values correspond to more improvement compared to the pretrained model. *Takeaway:* Languages with bigger RiddleScore gains tend to have bigger human-perceived quality improvements.

strategies (Tan et al., 2016; Wei and Lee, 2021). This trend represents similar stylistic augmentations across languages, as AOF reduces redundancy (e.g., Self-BLEU down 40.4% in EN and 42.4% in FR) while increasing diversity (e.g., Distinct-2 up to 13.5% in EN and 10.6% in ZH).

These cross-linguistic patterns, quantified in Tables 5 and 4, suggest that AOF enables culture-attached, cognition-challenging riddles with higher metaphorical condensation and interpretability. For complete evaluations, see Section A.

8 Fine-Tuned vs. Pretrained Riddle Generation

We visualize these cross-language gains in Figure 5 and show how they align with human judgments in Figure 4. Fine-tuning with AOF consistently enhances riddle generation across all five languages by reducing repetition, increasing lexical diversity, and producing more structurally cohesive metaphors. Across the board, RiddleScore increases reflect these quality gains: AR (+57.1%), ZH (+48.3%), EN (+43.4%), FR (+33.7%) and JA (+29.5%) (Table 5). These improvements coincide with major reductions in Self-BLEU—up to 63.4% for JA and 43.2% for FR—indicating lower reliance on template reuse. Distinct-2 further supports richer lexical expression, with AR (+18.8%), JA (+31.3%) and FR (+13.3%) seeing the most progress (Table 4). Human evaluation scores for AOF also improved substantially after fine-tuning (Tables 2 and 1). For example, ZH rose by **+15.1%**, EN by **+16.9%**, and JA by **+2.1%**. FR decreased slightly (**1.3%**), while AR maintained its high human evaluation score (**4.92**). These percentage changes strongly parallel the RiddleScore gains (e.g., ZH: 0.453 $\rightarrow$ 0.728), reinforcing the metric's validity as a proxy for human perception of creativity, fluency, and cultural fit.

While all languages benefit, fine-tuning yields especially high returns in languages with deep poetic or idiomatic traditions. For example, in ZH, AOF-finetuned models generate riddles like "千言万语藏心怀" ("A thousand words hidden in the heart"), whose solution—"信" (message/trust)—demonstrates metaphorical compression grounded in radical-based inference (Table 20, Row 2). This level of orthographic subtlety is absent in pretrained outputs, underscoring AOF's value in enabling culturally resonant riddle design.

Methods of prompting vary in consistency: Few-Shot and AOF consistently increase RiddleScore, but Chain-ofThought is inconsistent: significant increases for EN (+48.5%) but negligible for AR (+3. 6%) and JA (0. 0%) - indicating limited generalizability between languages. Only AOF consistently improves human-aligned and automatic metrics for all languages. Full language-specific results and examples appear in Section D and Appendices G–J.

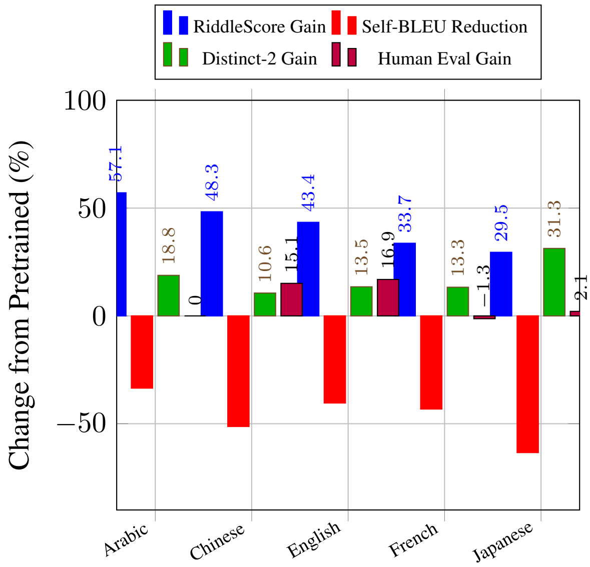

Figure 5: Percentage changes in RiddleScore, Self-BLEU, Distinct-2, and human evaluation after fine-tuning. Positive bars show improvements; negative Self-BLEU values (in red) indicate desirable reductions in repetition.

9 Fine-Tuned AOF Riddle Comparison to Real World

Across all five languages, fine-tuned AOF riddles diverge meaningfully from real-world counterparts by trading formulaic structure for richer metaphor, lexical inventiveness, and cultural depth. Traditional riddles often rely on binary opposites, rhymes, or phonological puns (Gentner, 1983; An, 2023), whereas AOF generations favor conceptual blending (Fauconnier and Turner, 2002), indirect metaphor (Lakoff and Johnson, 1980), and cross-domain abstraction (Tan et al., 2016).

EN and FR AOF riddles employ echo, shadow, or depth metaphors, including rhythmic phrasings that support recall and poeticity (Encyclopædia Britannica, 2025). For instance, the EN riddle of Table 13, Row 1—"I mirror your thoughts, but never speak"—explores selfhood through contrastive metaphor, absent in real-world riddles that prefer rhyming antonyms like "shadow/light." FR AOF riddles follow suit, abandoning "Qu'est-ce qui" templates for ellipsis-like phrasing.

In ZH and JA, AOF outputs evoke script-specific strategies like radical-based inference and spatial contradiction. The ZH riddle "千言万语藏心怀" (Row 2) reveals "信" (message/trust) through poetic indirection, while the JA riddle "屋根にはいるのに、家にいないものは何？" juxtaposes kanji structure and conceptual space (Sun, 2006; An, 2023).

In AR, fine-tuned riddles pivot from root-based puns to symbolic layering, favoring poetic contrasts over mechanical symmetry. As shown in Figure 6, Row 1, metaphors like "a wind that enters but is never welcomed" evoke hospitality norms and classical desert imagery (Al-Khatib, 1988; Antar, 2023; Liu et al., 2022b). For full comparisons and linguistic analysis, see Appendix C.

10 Conclusion

This paper introduces adaptive originality filtering (AOF), a re-feedback method for improving multilingual riddle generation, pushing models towards semantically new, structurally well-formed, yet culturally embedded, outputs. For five typologically distinct languages: English, Chinese, Japanese, Arabic, and French, AOF systematically improves human-aligned quality measured by RiddleScore for all five confirming the approach's universal applicability, regardless of script, form, or model design.

These advantages are a byproduct of AOF's design: AOF discourages revisioning of templates, discourages egocentral phrasing, and trends toward metaphoric, interpretative styles typical for every language's rhetorical styles. Optimized variants of AOF, besides being better than pretrained generations, by and large are comparable to real-world puzzles by metaphoric richness, especially in very oralistically and visualistically inclined languages Arabic and Chinese. Additionally, AOF generalizes across LLMs, from DeepSeek R1 to GPT-4o and from LLaMA 3.1, in manifesting strong performance across a diversity of generation styles, as well as pretraining corpora.

Apart from riddles, this work also suggests that prompting strategies with rejection-based filtering can guide LLMs towards culturally and cognitively compatible results, especially for compositional and figurative tasks.

Limitations

Dataset Scope

We limit our experiment to the BiRdQA corpus, comprised of 6,614 English and 8,751 Chinese multi-choice riddles. Though genre-various, its figurative concentration limits generalizability to larger creative tasks (e.g., allegory or storytelling). Our five-lingual evaluation extends over

EN–ZH–AR–JA–FR, but omits lower-resource or more-morphologically challenging languages like Finnish or Swahili.

Prompting and Sampling

We uniformly set decoding hyperparameters (e.g., temperature, number of tokens) to allow for comparison, but possibly suppress interactions between prompts and parameters. Filtering by MiniLM in AOF targets semantic novelty, but cosine similarity may overlook certain subtle redundancies, especially where languages are morphologically diverse or idiomatic.

Fine-Tuning Setup

Our GPT-4o fine-tuning uses BiRdQA's multiple-choice setup, boosting structural fluency but potentially biasing toward riddles that privilege explicit clarity over conscious ambiguity. While stylistic refinement shows up by metrics such as Self-BLEU, Distinct-2 and RiddleScore, more detailed downstream measurements such as solver accuracy and difficulty calibration are left to future research.

Evaluation Constraints

Human judgments were made by native or proficient speakers from five languages employing standard rubrics. This guarantees cultural anchoring but sample size and analysis by inter-annotator agreement were restricted by resources. To evaluate creativity, fluency, and cultural fit, RiddleScore, tested against these ratings, yields an interpretable proxy, albeit a proxy that doesn't register longer-term aspects like memorability, interest, or difficulty to solve.

Ethics Statement

Language Equity and Cultural Representation

This research assesses riddle-making within five languages, including English, Chinese, Japanese, Arabic, and French, selected to be typologically diverse and with resources to draw from. Although this gives a wide cultural span, the dataset and prompts come from internet-based corpora and so might not capture perfectly idiomatic richness from less represented populations. Certain metaphorical or rhetorical patterns might be overly represented within English or less developed within other languages even with our balancing qualitative with quantitative assessment.

Creative Attribution and AI Authorship

Procedurally generated riddles may resemble publicly known riddles from folk sources or online corpora. As described in Sections 3–4, Adaptive Originality Filtering (AOF) mitigates this risk by rejecting outputs with high semantic similarity to reference data. Nonetheless, we caution against deploying outputs in commercial settings without additional originality verification. AI assistants (e.g., ChatGPT) were also used to support code development and manuscript preparation. During implementation, LLMs aided in debugging and optimizing evaluation scripts (e.g., for RiddleScore and Distinct-2). In writing, AI was used for linguistic refinement, including phrasing, transitions, and caption clarity. All methodological contributions, analysis, and final revisions were conducted by the authors.

Data Privacy and Responsible Fine-Tuning

These data have no personally identifiable information (PII). The riddles are anonymized and cast as general-knowledge metaphors. The fine-tuning followed OpenAI's API regulations, token constraints, and safety limits, and never involved user-submitted or private material.

Human Evaluation and Metric Ethics

Human ratings were made by native or expert speakers with standardized rubrics, allowing for culturally sensitive evaluations. Model IDs were blinded to help decrease bias. RiddleScore, tested against these human ratings, provides a formalized proxy to creativity, fluency, and cultural fit but doesn't assess engagement, memorability, or difficulty for solvers.

Misuse Risks and Interpretability

While generation of riddles is a low-risk task, their creative uncertainty might be exploited to spread misinformation or to manipulate culturally sensitive information. We advise against using them in high-stakes educational, psychological, or legal applications without interpretability controls and human review.

References

Abu Uthman Amr ibn Bahr Al-Jahiz. 869. *Clarity and Eloquence (Al-Bayan wa Al-Tabyin)*. Basra. Original classical Arabic manuscript; various modern editions available.

Abd al-Karim Al-Khatib. 1988. *The Art of Riddles in Arabic Literature*. Dar Al-Fikr, Beirut.

Muhammad Al-Marzouki. 2012. *The Poetics of Ambiguity and Interpretation in Modern Arabic Poetry*. Dar Kunooz Al-Maarifa, Amman.

Tran Nguyen An. 2023. Hilbert multiplicity and irreducible multiplicity of idealizations. *arXiv preprint arXiv:2311.04719*.

Dalia Antar. 2023. The effectiveness of using chatgpt4 in creative writing in arabic: Poetry and short story as a model. *Information Sciences Letters*, 12(12):2445–2459.

Anirudh Atmakuru, Jatin Nainani, Rohith Siddhartha Reddy Bheemreddy, Anirudh Lakkaraju, Zonghai Yao, Hamed Zamani, and Haw-Shiuan Chang. 2024. Cs4: Measuring the creativity of large language models automatically by controlling the number of story-writing constraints. *arXiv preprint arXiv:2410.04197*.

Salvatore Attardo. 1994. *Linguistic Theories of Humor*. Walter de Gruyter.

Santiago Adrian Aytes, Jihun Baek, and Sung Ju Hwang. 2025. Sketch-of-thought: Efficient llm reasoning with adaptive cognitive-inspired sketching. *arXiv preprint arXiv:2503.05179*.

Bhuwan Bhatt and Valeriia Kuka. 2025. Llm parameters explained: A practical guide with examples for openai api in python. LearnPrompting blog. Available: `https://learnprompting.org/blog/llm-parameters`.

Kim Binsted. 1996. *Machine humour: An implemented model of puns*. Ph.D. thesis, University of Edinburgh.

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020a. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Tom B. Brown, Benjamin Mann, and Nick et al. Ryder. 2020b. Language models are few-shot learners. In *NeurIPS*.

Sahan Bulathwela, María Pérez-Ortiz, Catherine Holloway, Mutlu Cukurova, and John Shawe-Taylor. 2024. Artificial intelligence alone will not democratise education: On educational inequality, techno-solutionism and inclusive tools. *Sustainability*, 16(2):781.

Aoife Cahill. 2009. Evaluation metrics for natural language generation. In *ENLG*.

Aoife Cahill, Martin Forst, et al. 2009. Evaluation metrics for natural language generation. In *ENLG*.

Tuhin Chakrabarty, Debanjan Ghosh, and Smaranda Muresan. 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*, pages 866–879.

Winnie Chan. 1996. The riddle and the enigma: Traditional genres in french oral culture. *Marvels & Tales*, 10(1):15–27.

Yu Chen and Tania Avgustinova. 2021. Are language-agnostic sentence representations actually language-agnostic? In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 274–280.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Yuwei Dang, Zhiheng Lin, Jiale Ma, Wayne Xin Zhao, and Ji-Rong Wen. 2022. Evaluating the creativity of text generation models. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 1561–1574.

Jean Delisle. 1999. *Translation and the poetic function: The influence of linguistic and literary theories on translation studies*. University of Ottawa Press.

P. V. DiStefano and J. D. Patterson. 2024. Automatic scoring of metaphor creativity with large language models. *Creativity Research Journal*.

Ziwei Dou, Xiaoyu Ye, Xiang Ren, and Yue Zhang. 2022. Scill: Structured chain-of-thought for interpretable language learning. *arXiv preprint arXiv:2210.01206*.

Philipp Dufter. 2021a. *Distributed Representations for Multilingual Language Processing*. Ph.D. thesis, LMU Munich.

Philipp Dufter. 2021b. *Distributed representations for multilingual language processing*. Ph.D. thesis, lmu.

Encyclopædia Britannica. 2025. Internal rhyme. Internal rhyme: rhyme within a line enhances cohesion and rhythm in poetry.

Ilan Falkum. 2009. A pragmatic account of the interpretation of figurative language. *UCL Working Papers in Linguistics*, 21:55–77.

Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 889–898.

Gilles Fauconnier and Mark Turner. 2002. *The Way We Think: Conceptual Blending and the Mind's Hidden Complexities*. Basic Books.

Dedre Gentner. 1983. Structure-mapping: A theoretical framework for analogy. *Cognitive Science*, 7(2):155–170.

Marjan Ghazvininejad, Xing Shi, Jay Priyadarshi, and Kevin Knight. 2017. Hafez: An interactive poetry generation system. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics-System Demonstrations*, pages 43–48.

Panagiotis Giadikiaroglou, Maria Lymperaiou, Giorgos Filandrianos, and Giorgos Stamou. 2024. Puzzle solving using reasoning of large language models: A survey. *arXiv preprint arXiv:2402.11291*.

Ethan Heavey, James Hughes, and Milton King. 2024. Stfx-nlp at semeval-2024 task 9: Brainteaser: Three unsupervised riddle-solvers. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 28–33.

Bernd Heine and Tania Kuteva. 2007. *The Genesis of Grammar: A Reconstruction*. Oxford University Press.

Zhongzhan Huang, Shanshan Zhong, Pan Zhou, Shanghua Gao, Marinka Zitnik, and Liang Lin. 2025. A causality-aware paradigm for evaluating creativity of multimodal large language models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Manaf Ismayilzada, Dilan Circi, Jaakko Sälevä, and Hakan Sirin. 2024. Evaluating morphological compositional generalization in large language models. *arXiv preprint arXiv:2410.12656*.

Arun et al. Kabra. 2023. Multi-lingual and multi-cultural figurative language understanding. *arXiv preprint arXiv:2305.16171*.

Hiroko Kawamura. 2016. Cultural modes of reasoning in japanese riddles and proverbs. *Japanese Language and Literature*, 50(1):1–22.

Arthur Koestler. 1964. *The act of creation*. Macmillan.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Tanaka. 2022. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems*.

Kalpesh Krishna, Ari Holtzman, Daniel Khashabi, Antoine Bosselut, Hannaneh Hajishirzi, and Yejin Choi. 2023. Reflexion: Language agents with verbal reinforcement learning. *arXiv preprint arXiv:2303.11366*.

Huiyuan Lai, Jiali Mao, Antonio Toral, and Malvina Nissim. 2022. Human judgement as a compass to navigate automatic metrics for formality transfer. In *HumEval*.

George Lakoff and Mark Johnson. 1980. *Metaphors We Live By*. University of Chicago Press.

George Lakoff and Mark Johnson. 1999. Metaphor as language and thought. In *Cognitive Semantics*. Cambridge University Press.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. *arXiv preprint arXiv:1901.07291*.

Meisin Lee. 2025. Optimizing legal text summarization through dynamic retrieval-augmented generation. *Symmetry*.

Marc Leman. 2013. *Figures et fictions: Les formes de l'imaginaire en littérature française*. Presses Universitaires de France.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 110–119, San Diego, CA, USA. Association for Computational Linguistics.

Xiaorong Li. 2008. Riddles and wordplay in chinese folklore: A cultural and linguistic perspective. *Folklore Studies*.

Yafu Li, Zhilin Wang, Leyang Cui, Wei Bi, Shuming Shi, and Yue Zhang. 2024. Spotting ai's touch: Identifying llm-paraphrased spans in text. In *Findings of ACL*.

Yujia Li, Krishnamurthy Sreenivasan, Andreas Giannou, et al. 2023. Dissecting chain-of-thought: Compositionality through in-context filtering and learning. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Emmy Liu, Chenxuan Cui, Kenneth Zheng, and Graham Neubig. 2022a. Testing the ability of language models to interpret figurative language. *NAACL*.

Emmy Liu, Chenxuan Cui, Kenneth Zheng, and Graham Neubig. 2022b. Testing the ability of language models to interpret figurative language. In *Proceedings of NAACL-HLT 2022*, pages 4437–4452.

Xiaofan Lu, Yixiao Zeng, Feiyang Ma, Zixu Yu, and Marco Levorato. 2024. Improving multi-candidate speculative decoding. In *ICLR*.

Aman Madaan, Bill Yuchen Lin, Xinyi Liu, Xudong Fu, Peggy Qian, Prahal Arora Bhargava, Ashish Sabharwal, and Hannaneh Hajishirzi. 2023. Self-refine: Iterative refinement with self-feedback. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*.

Danièle Meulemans. 2005. Jeux de langage: L'énigme et la devinette dans la tradition orale francophone. In *Jeux et langages*, pages 55–72. Presses Universitaires de Rennes.

Lili Mou, Zichao Ye, Wenpeng Yin, Wayne Xin Zhao, Duyu Tang, and Rui Yan. 2022. Cold decoding: Energy-based constrained text generation with langevin dynamics. *arXiv preprint arXiv:2202.11726*.

Masanari Ohi, Masahiro Kaneko, Naoaki Okazaki, and Nakamasa Inoue. 2024. Harmoniceval: Multimodal, multi-task, multi-criteria automatic evaluation using vision language models. *arXiv preprint arXiv:2412.14613*.

Ioannis Panagiotopoulos, Giorgos Filandrianos, Maria Lymperaiou, and Giorgos Stamou. 2024. Riscore: Enhancing in-context riddle solving in language models through context-reconstructed example augmentation. *arXiv preprint arXiv:2409.16383*.

Siddhesh Pawar, Junyeong Park, and Jiho et al. Jin. 2024. Survey of cultural awareness in language models: Text and beyond. *arXiv preprint arXiv:2411.00860*.

Hao Peng, Yifan Hou, Mo Yu, Wenhui Wang, Shujie Liu, Yankai Lin, Zhiyuan Liu, Jing Ma, and Jianfeng Gao. 2023. Language models still struggle to learn hard commonsense knowledge from demonstrations. In *ACL*.

Sasa Petrović and David Matthews. 2013. Unsupervised joke generation from big data. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 228–232.

Edoardo Maria Ponti, Peng Xu, Yixin Kim, Samuel Cahyawijaya, Zhihao Tan, Sebastian Ruder, Yichong Huang, Graham Neubig, Kevin Duh, Naman Goyal, et al. 2020. Xcopa: A multilingual dataset for causal commonsense reasoning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8).

Ricardo Rei, Ana Farinha, Alon Lavie, et al. 2020. Comet: A neural framework for mt evaluation. In *Proceedings of WMT*.

Leticia Resck, Isabelle Augenstein, and Anna Korhonen. 2024. Explainability and interpretability of multilingual large language models: A survey. *OpenReview*.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2018. Semantically equivalent adversarial rules for debugging nlp models. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 856–865. Association for Computational Linguistics.

Lucia Schmidtová and Shu Wu. 2024. Automatic metrics fail to capture creativity in multilingual generation. *Transactions of the Association for Computational Linguistics*.

Terrence J Sejnowski. 2023. Large language models and the reverse turing test. *Neural Computation*, 35(3):309–342.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020a. Bleurt: Learning robust metrics for text generation. *Proceedings of ACL*.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020b. Bleurt: Learning robust metrics for text generation. In *ACL*.

Ekaterina Shutova. 2013. Metaphor identification as interpretation. In *Proceedings of NAACL-HLT*.

Shaden Smith et al. 2022. Using deepspeed and megatron to train megatron-turing nlg 530b, a large-scale generative language model. *arXiv preprint arXiv:2201.11990*.

Dirk HR Spennemann. 2023. Chatgpt and the generation of digitally born “knowledge”: How does a generative ai language model interpret cultural heritage values? *Knowledge*, 3(3):480–512.

Chaofen Sun. 2006. Chinese character puzzles and riddle traditions. *Journal of Chinese Linguistics*, 34(2):223–248.

Prithviraj Tambwekar, Animesh Mehta, Lindsay Martin, Brent Harrison, and Mark O. Riedl. 2019. Controllable neural story plot generation via reward shaping. In *IJCAI*, pages 5982–5988.

Chuanqi Tan, Furu Wei, Li Dong, Weifeng Lv, and Ming Zhou. 2016. Solving and generating chinese character riddles. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 846–855.

Peiyuan Teng and Min Xu. 2023. Random matrix time series. *Journal of Statistical Theory and Practice*, 17(3):42.

Chris Van der Lee, Albert Gatt, Emiel van Miltenburg, and Emiel Krahmer. 2018. Fluency metrics for machine translation evaluation: A comprehensive analysis. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC)*.

Chris van der Lee, Sander Wubben, and Emiel Krahmer. 2019. Assessing the evaluation of text generation. In *Proceedings of the 12th International Conference on Natural Language Generation*.

Chris van der Lee et al. 2021. Hume: Human unified meaning evaluation. In *Proceedings of ACL*.

Tony Veale. 2011. Creative language retrieval: A robust hybrid of information retrieval and linguistic creativity. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 278–287.

Eric Wallace, Shi Feng, Nikhil Kandpal, Matt Gardner, and Sameer Singh. 2019. Universal adversarial triggers for attacking and analyzing nlp. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2153–2162. Association for Computational Linguistics.

Chenguang Wang, Weijia Su, Qingyao Ai, and Yang Liu. 2024. Knowledge editing through chain-of-thought. *arXiv preprint arXiv:2412.17727*.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. In *NeurIPS*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023a. Chain-of-thought prompting elicits reasoning in large language models. In *NeurIPS*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023b. Chain-of-thought prompting elicits reasoning in large language models.

Li Wei and Tong King Lee. 2021. Language play in and with chinese: traditional genres and contemporary developments. *Global Chinese*, 7(2):125–142.

Orion Weller and Kevin Seppi. 2019. Humor detection: A transformer gets the joke. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, pages 3621–3627.

Genta Indra Winata, Lucky Susanto, et al. 2024. Metametrics-mt: Tuning meta-metrics for machine translation via human preference calibration. In *WMT*.

Yaqing Xie, Jiaao Chen, Zijie Wu, et al. 2025. Sorry-bench: Systematically evaluating large language model safety refusal. In *ACL*.

Fan Xu, Yunxiang Zhang, and Xiaojun Wan. 2022. Cc-riddle: A question answering dataset of chinese character riddles. *arXiv preprint arXiv:2206.13778*.

Jingjing Xu, Xuancheng Li, Lei Zhang, et al. 2018. Diversity-promoting gans for text generation. *ACL 2018*.

Jian Yao, Ran Cheng, Xingyu Wu, Jibin Wu, and Kay Chen Tan. 2025. Diversity-aware policy optimization for large language model reasoning. *arXiv preprint arXiv:2505.23433*.

Shinn Yao, Jeffrey Zhao, Dian Yu, Yuan Xu, Kaixuan Zhao, Shinn Cao, Eric Zhang, Shunyu Xu, Yihan Zhao, Yao Shen, et al. 2023. Tree of thoughts: Deliberate problem solving with large language models. *arXiv preprint arXiv:2305.10601*.

Hugh Zhang, Daniel Duckworth, Daphne Ippolito, Douglas Eck, and Arvind Neelakantan. 2021. Trading off diversity and quality in natural language generation. In *Proceedings of the 2021 Workshop on Human Evaluation of NLP Systems (HumEval)*.

Jingqing Zhang, Jason Baldridge, and He He. 2020a. Learning to summarize with human attention. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3631–3642.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2020b. Bertscore: Evaluating text generation with bert. In *ICLR*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020c. BERTScore: Evaluating text generation with BERT. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Wei Zhang and Xiaojun Wan. 2025. Multilingual cultural generation with language models. *Transactions of the Association for Computational Linguistics*, 13:1234–1256.

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.

Zhuosheng Zhang, Aston Zhang, Mu Li, and Alexander Smola. 2022. Automatic chain of thought prompting in large language models. *arXiv preprint arXiv:2210.03493*.

Yichi Zhou and Yonatan Bisk. 2022. Visual puns: Multimodal understanding of double meanings. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pages 6034–6047.

Yichong Zhou, Swaroop Mishra, Xiaodong Liu, et al. 2022. Crescendo: Iteratively growing reasoning graphs for question answering. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1610–1624.

A Appendix: AOF Fine-tuned language Evaluations

A.1 English

Fine-tuning GPT-4o with AOF notably improves semantic richness and lexical creativity (RiddleScore: 0.586; Table 5). AOF achieves superior lexical diversity (Distinct-2: 0.893) and minimal structural repetition (Self-BLEU: 0.260) compared to few-shot and adversarial baselines (Table 4), validating RiddleScore's effectiveness as a comprehensive evaluation measure (Zhang et al., 2020b; Sellam et al., 2020b). Qualitatively, riddles such as those in Table 13 illustrate innovative metaphor usage and coherent ambiguity, consistent with cognitive theories on figurative language and memorability (Lakoff and Johnson, 1980; Koestler, 1964; Fauconnier and Turner, 2002). For instance, Row 1 deploys cues like "mirror yours" and "echo thoughts" to encode identity and perception into abstract form, while Row 2 evokes silence as an interstitial force through metaphors, aligning with conceptual blending theory (Fauconnier and Turner, 2002).

A.2 Japanese

Fine-tuning GPT-4o with AOF significantly enhances morphosyntactic fluency and metaphor–answer cohesion in Japanese–English riddle generation (RiddleScore: 0.475; Table 5). Compared to other prompting methods, AOF produces riddles with the lowest structural redundancy (Self-BLEU: 0.177) and highest lexical diversity (Distinct-2: 0.915), indicating stronger semantic control and reduced overfitting to prior examples (Table 15). These gains are reflected in AOF's leading RiddleScore, which surpasses Zero-Shot (0.300), Few-Shot (0.334), CoT (0.307), and Adversarial (0.331) settings. Qualitatively, the generated riddles exhibit hallmarks of Japanese poetic reasoning—syntactic compression, metaphorical layering, and rhythmical closure—without resorting to direct translation or formulaic repetition(Kawamura, 2016). For instance,「屋根にはいるのに、家にいないものは何？」("What enters the roof but never the house?") leverages spatial contradiction in a culturally familiar frame, while maintaining logical symmetry across both languages(Heine and Kuteva, 2007). This fidelity to both Japanese linguistic nuance and cross-lingual metaphor construction is characteristic of AOF's superiority, suggesting greater alignment with human intuitions of creativity, fluency, and interpretability.

A.3 Chinese

Fine-tuning enhances metaphorical control and orthographic awareness in Chinese riddles. AOF outputs consistently avoid overused oppositional templates like "我有...却...," favoring layered metaphors, radical-based hints, and prosodic fluency. Compared to Zero-Shot and Few-Shot baselines, AOF achieves lower Self-BLEU (0.163 vs. 0.315 / 0.349) and higher Distinct-2 (0.934 vs. 0.831 / 0.787), validating RiddleScore as a composite indicator of structural novelty (0.728; Table 5) (Zhang et al., 2020b; Sellam et al., 2020b). In Table 20, Row 1 evokes lunar imagery with rhythmic balance, updating a classical riddle ("口袋里有个圆...") through spatial metaphor and contrast (Sun, 2006; Wei and Lee, 2021). Row 2 exemplifies orthographic metaphor: "信" is revealed through poetic compression ("千言万语藏心怀"), echoing traditional pun-encoding in radical-based 灯谜 (Tan et al., 2016; Li, 2008). Row 3 (蝴蝶) combines temporal framing and sensory motion ("彩衣...花丛...无踪") to support multi-modal reasoning, in line with conceptual blending theory (Fauconnier and Turner, 2002; Lakoff and Johnson, 1980). These results suggest that AOF produces culturally grounded riddles with high interpretability and lexical range.

A.4 French

Fine-tuning GPT-4o with AOF yields French riddles that combine varied grammatical forms, fresh metaphors, and cultural resonance. The model moves beyond standard "Qu'est-ce qui..." stems and elemental tropes to embrace declarative statements, poetic ellipses, and even modern imagery. Although Zero-Shot achieves a higher RiddleScore (0.468 vs. 0.352), AOF excels in lexical diversity (Distinct-2 = 0.856) and maintains moderate repetition (Self-BLEU = 0.273), suggesting greater creative variance in form and framing (Zhang et al., 2021; Binsted, 1996). AOF riddles (Table 24) avoid clichés like "ombre" or "écho" and instead draw on subtle metaphor and rhythm. For instance, Row 1 uses cyclical phrasing to express the return of day (*jour*), while Row 2 reframes a broom through trailing ellipsis and implied motion. These constructions echo prior findings on metaphor-induced novelty and poetic ambiguity (Lakoff and Johnson, 1980; Koestler, 1964), even when metric scores undervalue such stylistic range.

A.5 Arabic

Fine-tuning GPT-4o with AOF improves semantic richness and metaphorical ingenuity in Arabic–English bilingual riddles (RiddleScore: 0.586; Table 5). Compared to Few-Shot (0.364), Zero-Shot (0.315), Chain-of-Thought (0.313), and Adversarial (0.341), AOF achieves higher lexical variety (Distinct-2: 0.893) and lower repetition (Self-BLEU: 0.260), showing its balance between novelty and coherence (Table 4). These results confirm RiddleScore's effectiveness for evaluating creativity and linguistic depth (Zhang et al., 2020c; Sellam et al., 2020b). Qualitatively, AOF riddles reflect traditional Arabic poetic traits—metaphorical layering, conceptual blending, and cultural framing—without relying on literal translation. For example, Figure 6, Row 1 uses sound as a metaphor for something intangible yet present—*"I exist in the air, yet I do not fly"*—echoing classical rhetoric. Row 2 likens strong wind to a guest who *"passes nearby homes but is never welcome inside"*. These examples illustrate nuanced cultural imagery and poetic reasoning, consistent with the richness of Arabic literary tradition (Al-Khatib, 1988). AOF thus enhances both creativity and interpretability in bilingual Arabic riddles.

B Appendix: AOF Pretrained language Evaluations

B.1 English

GPT-4o achieves moderate repetition (Self-BLEU: 0.413) and high lexical diversity (Distinct-2: 0.852), balancing structural cohesion with surface novelty. These characteristics correspond with its AOF RiddleScore of 0.373, indicating that while GPT-4o avoids excessive repetition, its metaphorical expressiveness remains moderate. Compared to LLaMA 3.1 (0.471 / 0.727, RiddleScore: 0.352) and DeepSeek R1 (0.339 / 0.845, RiddleScore: **0.400**), GPT-4o represents a middle ground: less phrasally diverse than R1, but more structurally consistent than LLaMA. The riddle in Row 1 of Table 10 reflects these tendencies, blending contrastive metaphor with cohesive syntax. This supports prior findings that figurative ambiguity coupled with syntactic regularity enhances interpretability (Lakoff and Johnson, 1980; Shutova, 2013).

LLaMA 3.1 displays the strongest phrasal variation (Distinct-2: 0.727), but with moderately higher repetition (Self-BLEU: 0.471) and a slightly lower AOF RiddleScore of 0.352. These metrics suggest that while LLaMA 3.1 explores more varied lexical forms, it occasionally overuses structural templates. The riddle in Row 2 of Table 10 shows rhythmic symmetry and layered metaphor, reinforcing theories linking riddle memorability to structured cadence and salience (Koestler, 1964). The AOF prompt appears to mitigate lexical rigidity by encouraging recomposition within constrained semantic bounds (Fauconnier and Turner, 2002).

DeepSeek R1 demonstrates the lowest repetition (Self-BLEU: 0.339), highest lexical diversity (Distinct-2: 0.845), and the top AOF RiddleScore at **0.400**, indicating superior expressive range and originality. The riddle in Row 3 exemplifies conceptual inversion, pairing abstract imagery with narrative misdirection—a hallmark of classic riddle mechanics (Koestler, 1964). While extreme novelty sometimes threatens fluency (Zhang et al., 2021), R1's outputs remain syntactically intact, suggesting that AOF balances expressiveness with readability (Xu et al., 2018). This balance likely contributes to R1's higher perceived riddle quality as measured by RiddleScore.

B.2 Japanese

GPT-4o While GPT-4o's performance on metrics like self-BLEU and distinct-n using the AOF prompt falls around the average compared to standard baselines, it excels notably in RiddleScore, achieving a score of 0.475. This substantial increase over traditional methods (Few-Shot: 0.334, Zero-Shot: 0.300, Chain-of-Thought: 0.307, Adversarial: 0.331) reflects the model's ability to generate riddles with greater novelty, fluency, diversity, and semantic coherence ((Yao et al., 2025), (Schmidtová and Wu, 2024)). AOF specifically addresses traditional prompting flaws such as the "I"-centered imagery prevalent in chain-of-thought prompts and the example-specific overfitting observed in few-shot prompts, thereby substantially enhancing multilingual riddle quality. For instance, the riddle example in Table 14 features a distinctive structure—a concise opening followed by a more elaborate second sentence—which enhances reader engagement and contributes to its high RiddleScore.

LLaMa3.1 Although LLaMa3.1 does not demonstrate significant improvement in automated metrics like self-BLEU and distinct-n under the AOF framework, its RiddleScore of 0.475 significantly surpasses traditional baselines (Few-Shot: 0.334, Zero-Shot: 0.300, Chain-of-Thought: 0.307, Adversarial: 0.331). This highlights AOF's effectiveness in enhancing multilingual riddle generation beyond conventional evaluation metrics by addressing issues such as egocentric phrasing and repetition. Notably, the riddle presented in Table 14 cleverly employs the homophone 「つる」, invoking both decorative twine and the crane (鶴)—elements deeply embedded in Japanese cultural symbolism and Shinto rituals like しめ縄 (shimenawa) (An, 2023). This cultural and linguistic depth significantly contributes to its superior RiddleScore.

DeepSeek R1 DeepSeek R1, while only achieving median results on surface-level metrics such as self-BLEU and distinct-n, shows marked improvement with a RiddleScore of 0.475 compared to lower scores from standard methods (Few-Shot: 0.334, Zero-Shot: 0.300, Chain-of-Thought: 0.307, Adversarial: 0.331). The RiddleScore clearly underscores the efficacy of the AOF prompting strategy in overcoming baseline shortcomings like excessive first-person imagery and rigid replication patterns, promoting originality, fluency, and semantic coherence. An illustrative example from Table 14 artfully misleads readers by metaphorically describing a fish's mouth as a "quiet tree" where birds sing, skillfully blending surreal imagery with natural elements(DiStefano and Patterson, 2024). This innovative poetic device significantly enhances its overall RiddleScore.

B.3 Arabic

GPT-4o GPT-4o shows moderate repetition (Self-BLEU: 0.497) and good lexical variety (Distinct-2: 0.780) with Adaptive Originality Filtering (AOF), clearly performing better than common methods like few-shot, zero-shot, chain-of-thought, and adversarial prompts.With an AOF RiddleScore of **0.373**, GPT-4o demonstrates notable improvement over chain-of-thought (0.304) and adversarial methods (0.296).Unlike chain-of-thought prompts, which tend to produce straightforward, predictable metaphors, AOF helps GPT-4o create riddles with imaginative and abstract images—such as something that's present but unseen—as illustrated in (Figure 6, Row 1). This approach fits naturally with traditional Arabic riddles, known for their symbolic and reflective style (Al-Khatib, 1988).

LLaMA 3.1 LLaMA 3.1 strikes an effective balance between repetition (Self-BLEU: 0.374) and creativity (Distinct-2: 0.927) through AOF, resulting in a RiddleScore of **0.378**. This addresses issues often found in chain-of-thought (0.303) and adversarial prompts (0.292), which frequently yield predictable or overly vague outputs. Its riddles are relatable and culturally resonant, using clear metaphors drawn from everyday life, like *"a strong wind"* that can't enter a house, as shown in (Figure 6, Row 2). This connects directly to familiar poetic traditions in Arabic, avoiding common pitfalls like repetitive phrasing or loss of meaning (Al-Jahiz, 869).

DeepSeek R1 DeepSeek R1, while somewhat repetitive (Self-BLEU: 0.585), achieves notable depth in metaphorical expression (Distinct-2: 0.583) under AOF, resulting in the highest RiddleScore of **0.400** among the three models. This method effectively tackles problems seen in zero-shot (0.400), few-shot (0.341), chain-of-thought (0.304), and adversarial prompting (0.305), such as repetitive or simplistic metaphors. For example, DeepSeek R1 creatively portrays a rooftop as an eye *"fed by the city,"* as seen in (Figure 6, Row 3), mixing urban imagery with striking visual symbolism. This clever blending of abstract ideas and real-world images strongly aligns with Arabic poetry, known for its layers of meaning and subtle metaphors (Al-Marzouki, 2012). By encouraging culturally rich riddles, AOF clearly boosts the originality and depth of DeepSeek R1's outputs compared to simpler prompting strategies (Xu et al., 2018).

B.4 French

GPT-4o GPT-4o's pretrained riddles are grammatically fluent and consistently answerable, but often exhibit translated literalism rather than native poetic expressivity. For instance, its output in Row 1 of Table 21 invokes elemental imagery typical of English-origin riddles, but lacks stylistic markers common in French verse, such as enjambment or internal rhyme (Delisle, 1999). These tendencies yield a Self-BLEU of `0.413` and a high Distinct-2 of `0.852`, suggesting strong surface diversity but moderate structural reuse. This balance corresponds to an AOF RiddleScore of `0.373`, reflecting a safe, comprehensible style with limited cultural specificity or rhythmic nuance (Chan, 1996).

DeepSeek R1 DeepSeek R1 offers concise and semantically transparent riddles, often echoing patterns from elementary French folklore. As seen in Row 2, its outputs favor concrete dualities ("bed but never sleep") common in children's riddles (Meulemans, 2005), yielding low Self-BLEU (`0.339`) and high Distinct-2 (`0.845`). These surface metrics align with an AOF RiddleScore of `0.354`, indicating moderate creativity tempered by formulaic structure. While effective, R1's riddles seldom explore prosodic depth or figurative abstraction (Leman, 2013), limiting their stylistic innovation despite syntactic precision.

LLaMA 3.1 LLaMA 3.1 demonstrates the widest stylistic bandwidth among pretrained models. Its Row 3 output juxtaposes dance and laughter through internal echo, while Row 4 ventures into digital metaphor with a riddle about a cursor. These examples reflect the model's capacity for modernized symbolic extension, albeit inconsistently. With a Self-BLEU of `0.471`, Distinct-2 of `0.727`, and RiddleScore of `0.352`, LLaMA balances lexical innovation with occasional overreach. These fluctuations suggest strong creative potential but uneven cohesion, echoing prior observations on metaphor blending and linguistic recombination (Veale, 2011; Binsted, 1996).

B.5 Chinese

GPT-4o GPT-4o's pretrained Chinese riddles are grammatically correct and logically coherent, but often translate English metaphors without adapting to the script-specific strategies typical of traditional 灯谜. As shown in Row 1 of Table 17, the imagery is literal and binary, missing multi-layered allusions like radical-based clues or idiomatic rhythm (Chan, 1996; Sun, 2006). With a Self-BLEU of 0.280, Distinct-2 of 0.869, and an AOF RiddleScore of **0.434**, the model achieves surface novelty without fully leveraging character-level poetic mechanisms. This suggests competent fluency but limited cultural depth.

DeepSeek R1 DeepSeek R1 produces elegant, fluent couplets with classical poetic symmetry, as seen in Row 2. While rhythm and antithesis are preserved, metaphors remain literal—favoring structural form over layered meanings. This is reflected in a Self-BLEU of 0.433, Distinct-2 of 0.674, and an AOF RiddleScore of **0.453**, the highest among the three models. The results indicate that while R1 may lack idiomatic richness, it effectively balances structural clarity and lexical diversity, offering consistently coherent outputs with stylistic restraint (Xu et al., 2018).

LLaMA 3.1 LLaMA 3.1 exhibits the richest cultural range in pretrained generation. Row 4 blends visual and semantic metaphor reminiscent of folk riddles, and Row 5 demonstrates radical-based structure. Its Distinct-2 of 0.776 and Self-BLEU of 0.428 align with an AOF RiddleScore of 0.330, revealing moderate creativity yet lower overall cohesion. Although stylistically ambitious, LLaMA occasionally struggles with logic or phrasing. Still, its outputs reflect deeper integration with Chinese morphological conventions than its counterparts (Li, 2008; Fauconnier and Turner, 2002).

C Appendix: Fine-Tuned AOF Riddle Comparison to Real World

C.1 English

As shown in Table 12, Row 1, the fine-tuned riddle reimagines the original with more abstract and layered associations. Rather than relying on negated literalism, it introduces concepts like memory and time using metaphorical compression and cross-sensory cues. This approach reflects principles of conceptual integration theory, where blending disparate domains enhances figurative depth (Fauconnier and Turner, 2002). In contrast, the real-world version is more direct, using structural opposition to achieve its effect (Gentner, 1983).Row 2 presents another clear shift in stylistic strategy. The real-world riddle uses static reversal—a common riddle trope—while the fine-tuned variant introduces paradox and disappearance as metaphors for guidance. This relies on spatial embodiment, a known technique in metaphor production (Lakoff and Johnson, 1980).

C.2 Japanese

The riddles in AOF are guided towards direct metaphors with complex, creative, and unique word choice and sentence structure, while having creative answers like memory and beehive in Table 16 (Teng and Xu, 2023). These generations surpass past riddle generations flaws like lack of originality in sentence structure, just changing the pronouns or verbs to make it more creative, and etc. These riddles contrast with traditional Japanese riddles which rely on phonetic ambiguity and cultural nuance like in Table 16 where the first row features how phonetically similar words feature different meanings and the riddle in the second row yields different ways of reading through phonetically similar readings(An, 2023).

C.3 Chinese

Fine-tuned AOF riddles in Chinese often leverage character structure through radical-based puns and vivid imagery. For instance, the coral riddle in Table 19 blends "sea" imagery with radical hints (海

底藏森林...) to guide the solver—a strategy supported by prior work on character-pun alignments in riddle composition (Tan et al., 2016). By contrast, traditional 灯谜 (e.g., “口袋里有个圆...” for “月亮”) rely on simple perceptual clues and tonal balance (Wei and Lee, 2021). This comparison suggests that our approach enhances cultural depth by embedding multi-layered orthographic play into poetic metaphors while preserving reader accessibility.

C.4 Arabic

(Figure 8, Row 5) AOF stands out for its fresh language and metaphorical clarity. One riddle—"*Something that's full when it eats, and thirsty when it drinks*"—relies on a simple yet clever contradiction that invites reflection. It draws on the tradition of using everyday logic to confuse and amuse, evoking the style of oral riddles that play with basic physical experiences. The second riddle—"*I light up the night and disappear by day, visible yet unseen... What am I?*"—is more poetic, using contrast and imagery to express something elusive and symbolic. It captures the feel of classical Arabic alghāz not through root-based punning but through layered metaphor and rhythm. Together, these examples show how AOF preserves the spirit of traditional riddling through modern, metaphor-rich language (Antar, 2023; Bhatt and Kuka, 2025; Liu et al., 2022b).

C.5 French

Fine-tuned AOF riddles in French lean into unexpected domain shifts and internal echo. The AOF example repurposes the concept of a “typo” as a buzzing bee, combining internal rhyme (“jardin/des mots”, “bourdonnant/lettres”) and metaphorical layering, driving semantic playfulness and rhythmic balance (Table 23, Row 1). Internal rhyme notably enhances poetic cohesion and cognitive engagement (Encyclopædia Britannica, 2025). In contrast, canonical French énigmes tend toward binary negation and elemental imagery (Table 23, Row 2). For instance, “Je vole sans ailes, je pleure sans yeux...” relies on simple antithesis without cross-domain metaphorical transfer. The AOF variant's richer conceptual mapping aligns with findings that cross-domain metaphor and internal structure boost interpretability and novelty in poetic forms (Lakoff and Johnson, 1999; Encyclopædia Britannica, 2025).

D Appendix:Fine-Tuned vs. Pretrained Riddle Generation

We compare GPT-4o before and after fine-tuning across five prompting strategies. Quantitative metrics—token length, Self-BLEU, and Distinct-2—are complemented by qualitative analysis of metaphorical framing, structural variation, and bilingual phrasing. Representative pairs are shown in **Appendices G–J**.

D.1 English

Fine-tuning reduces token length by 28.2%, repetition by 40.4%, and increases lexical variety by 6.1%. RiddleScore improves by 43.4%, showing that reduced redundancy and more diverse phrasing lead to higher-quality riddles. Pretrained outputs often reflect familiar patterns like personification, while fine-tuned ones adopt more abstract and fluent structures (Table 11, Row 1). Few-shot fine-tuning increases metaphorical expression but also length, with a 44.9% gain in RiddleScore (Row 2). CoT prompts benefit most—token length drops by 37.6%, diversity rises by 13.6%, and RiddleScore jumps 48.5% (Row 3). AOF produces the most creative riddles with metaphors like “quietest word,” improving RiddleScore by 42.9% alongside strong gains in novelty and clarity. Adversarial fine-tuning increases abstraction while reducing repetition by 18.2%, improving lexical diversity by 9.4%, and boosting RiddleScore by 33.4% (Row 5) (Zhang et al., 2020b; Sellam et al., 2020b).

D.2 Japanese

Across all prompting methods, fine-tuning improves morphosyntactic fluency and metaphorical layering. In Zero-Shot (Table 15, Row 1), outputs drop by 15.6% in Self-BLEU and align better with Japanese poetic rhythm(Kojima et al., 2022). Few-shot prompts (Table 15, Row 2) benefit from clearer clause structure and cultural framing, resulting in a 28.6% increase in distinct-n. CoT outputs (Table 15, Row 3) shift from templated “I...” forms to more idiomatic bilingual logic, improving Self-BLEU by 27.5% and 27.4% shorter riddles on average. Adversarial riddles (Table 15, Row 4) gain fluency and metaphor variation while reducing structural awkwardness. Yet, across Zero-Shot, Few-Shot, and CoT prompting, RiddleScore remained largely unchanged when moving from the pretrained to the fine-tuned model, suggesting that improvements in fluency and metaphorical richness did not translate into deeper semantic cohesion(Resck et al., 2024). Notably, Adversarial prompting saw a 14.5% drop in RiddleScore, indicating that its gains in stylistic fluency and metaphor density may have come at the cost of the semantic originality and structural coherence captured by the metric(Resck et al., 2024). In contrast, AOF prompting (Table 15, Row 5) exhibited no such trade-off, achieving the largest qualitative gain: a 63.4% drop in Self-BLEU, 31.3% increase in Distinct-2, and a 29.5% improvement in RiddleScore, reflecting enhanced metaphor density and cultural cadence without sacrificing semantic quality.

D.3 Chinese

Fine-tuning improves both variety and clarity in riddle phrasing. In Zero-Shot (Table 18, Row 1), replacing rigid sentence frames lowers repetition by 6.0%, boosts lexical diversity by 3.0%, and improves RiddleScore by 6.5%. Few-shot fine-tuning (Row 2) preserves strong metaphor use while avoiding repeated idioms, improving RiddleScore by 14.7%, increasing diversity by 5.6%, and reducing repetition by 6.8%. CoT prompts (Row 3) yield shorter riddles with smoother structure, cutting token length by 26.4%, increasing Distinct-2 by 7.2%, and raising RiddleScore by 18.1%. Adversarial fine-tuning (Row 4) boosts rhythm and cohesion, increasing lexical variety by 9.3% and RiddleScore by 12.8%, despite a 10.2% rise in repetition. AOF (Row 5) produces the most abstract and fluent riddles, lowering repetition by 51.3%, raising diversity by 10.6%, shortening outputs by 25.1%, and improving RiddleScore by 48.3% (Zhang et al., 2020b; Sellam et al., 2020b).

D.4 Arabic

Fine-tuning significantly enhances lexical diversity, reduces redundancy, and improves riddle quality in Arabic. In Zero-Shot (Table 7, Row 1), fine-tuned riddles replace rigid "X without Y" structures with rhythmic phrasing, reducing repetition (Self-BLEU) by 33.5%, increasing lexical diversity (Distinct-2) by 18.8%, and enhancing RiddleScore by 10.5% (0.315 to 0.348). Few-shot prompts (Row 2) abandon repetitive frames for enjambment and root variation, reducing Self-BLEU by 5.6%, increasing Distinct-2 by 8.1%, and improving RiddleScore by 8.0% (0.364 to 0.393). Chain-of-Thought (CoT) riddles (Row 3) become concise and idiomatic, lowering redundancy by 21.0%, increasing lexical diversity by 2.8%, and improving RiddleScore by 3.6% (0.313 to 0.324). Adversarial prompting (Row 4) introduces triadic parallelism and poetic misdirection, substantially reducing repetition by 44.7%, boosting lexical variety by 13.0%, and raising RiddleScore by 15.2% (0.341 to 0.393). AOF (Row 5) maintains peak lexical diversity (18.8% increase), decreases redundancy by 33.5%, and achieves the highest RiddleScore improvement (57.1%; 0.373 to 0.586), aligning closely with traditional Arabic poetic conventions (Al-Khatib, 1988).

D.5 French

Fine-tuning reduces dependence on literal templates like “Qu’est-ce qui...” and improves vocabulary variety across all prompt styles. In Zero-Shot (Table 22, Row 1), riddles shift from repetitive phrases to richer idiomatic expressions, with a 43.2% drop in Self-BLEU and a 7.1% rise in Distinct-2. RiddleScore improves by 25.7%, reflecting increased originality. Few-shot prompts (Row 2) yield riddles that are 32.8% shorter, with a 20.7% reduction in repetition and a 9.7% boost in lexical diversity; RiddleScore climbs 26.0%. CoT (Row 3) strikes a strong balance: repetition drops by 26.6%, diversity improves by 13.3%, and RiddleScore rises by 32.5%. Adversarial prompting (Row 4) enhances clarity while preserving misdirection, yielding a 14.0% reduction in Self-BLEU, 7.6% gain in Distinct-2, and 24.1% improvement in RiddleScore. AOF (Row 5) performs best overall, cutting repetition by 42.4%, achieving peak diversity, and delivering a 33.7% boost in RiddleScore. These results suggest that reducing redundancy and using more expressive, domain-appropriate language leads to riddles that are more fluent and culturally aligned (Zhang et al., 2020b; Sellam et al., 2020b).

E Appendix: Additional Results Tables

E.1 Average Token Length Across Pretrained Models

Language Pair	Prompting Method	GPT-4o	LLaMA 3.1	DeepSeek R1
English–Arabic	**Chain-of-Thought**	**910**	**1613**	**1085**
	Zero-Shot	1112	1519	2005
	Few-Shot	1921	2050	3144
	Adversarial	938	2202	1826
	AOF (Ours)	1548	1157	2138
English–Chinese	**Zero-Shot**	**702**	**731**	**719**
	Few-Shot	2030	2097	2351
	Chain-of-Thought	942	1389	1205
	Adversarial	916	950	1126
	AOF (Ours)	1275	1663	1535
English–Japanese	**Zero-Shot**	**1099**	**1127**	**1115**
	Few-Shot	1922	1941	2330
	Chain-of-Thought	1169	1099	1802
	Adversarial	1101	894	1128
	AOF (Ours)	1185	1230	1273
English–French	**Adversarial**	**787**	**1128**	**1413**
	Zero-Shot	1163	1183	1613
	Few-Shot	2061	2982	2565
	Chain-of-Thought	940	1631	1236
	AOF (Ours)	1166	1517	1982

Table 6: Average token lengths for each model and prompting method across language pairs. Bold = shortest average length per pair.

E.2 Average Token Lengths Across Languages

Language Pair	Prompting Method	Fine-Tuned GPT-4o (Avg. Token Length)
English–Arabic	AOF (Ours)	1129
	Zero-Shot	799
	Few-Shot	1999
	Chain-of-Thought	**730**
	Adversarial	737
English–Chinese	AOF (Ours)	1034
	Zero-Shot	898
	Few-Shot	2150
	Chain-of-Thought	**860**
	Adversarial	785
English–Japanese	AOF (Ours)	894
	Zero-Shot	894
	Few-Shot	2088
	Chain-of-Thought	**753**
	Adversarial	844
English–French	AOF (Ours)	1076
	Zero-Shot	943
	Few-Shot	2005
	Chain-of-Thought	733
	Adversarial	**716**

Table 7: Average token lengths for fine-tuned GPT-4o. Bold = shortest per pair.

E.3 Cross-Lingual Evaluation of Syntactic Validity

Language	Model	Total Riddles	Valid Structures	Validity (%)
English (EN)	GPT-4o-fine-tune	10	10	100.0%
Chinese (ZH)	GPT-4o-fine-tune	10	10	100.0%
Japanese (JA)	GPT-4o-fine-tune	10	10	100.0%
Arabic (AR)	GPT-4o-fine-tune	10	10	100.0%
French (FR)	GPT-4o-fine-tune	10	10	100.0%

Table 8: Cross-lingual evaluation of syntactic validity of GPT-4o AOF generations.

E.4 Average self-BLEU and Distinct-n Pretrained Metrics

Language Pair	Prompting Method	GPT-4o	LLaMA 3.1	DeepSeek R1
English–Arabic	AOF (Ours)	0.497 / 0.780	0.374 / 0.927	0.585 / 0.583
	Zero-Shot	**0.272 / 0.975**	0.432 / 0.746	0.627 / 0.543
	Few-Shot	0.272 / 0.880	0.432 / 0.746	0.627 / 0.543
	Chain-of-Thought	0.375 / 0.756	0.575 / 0.643	0.330 / 0.793
	Adversarial	0.330 / 0.798	0.342 / 0.727	0.672 / 0.507
English–Chinese	AOF (Ours)	**0.280 / 0.869**	0.428 / 0.776	0.433 / 0.674
	Zero-Shot	0.335 / 0.739	0.482 / 0.649	0.320 / 0.854
	Few-Shot	0.640 / 0.420	0.660 / 0.440	0.650 / 0.450
	Chain-of-Thought	0.363 / 0.777	0.403 / 0.815	0.430 / 0.767
	Adversarial	0.363 / 0.820	0.593 / 0.570	0.466 / 0.735
English–Japanese	AOF (Ours)	0.483 / 0.697	0.516 / 0.640	0.560 / 0.690
	Zero-Shot	0.364 / 0.833	0.430 / 0.871	0.514 / 0.757
	Few-Shot	**0.280 / 0.844**	0.587 / 0.605	0.402 / 0.715
	Chain-of-Thought	0.532 / 0.697	0.447 / 0.753	0.500 / 0.630
	Adversarial	0.334 / 0.794	0.599 / 0.586	0.405 / 0.741
English–French	AOF (Ours)	0.413 / 0.852	0.471 / 0.727	**0.339 / 0.845**
	Zero-Shot	0.451 / 0.833	0.476 / 0.715	0.520 / 0.849
	Few-Shot	0.371 / 0.814	0.480 / 0.665	0.670 / 0.535
	Chain-of-Thought	0.444 / 0.733	0.455 / 0.750	0.359 / 0.768
	Adversarial	0.358 / 0.806	0.485 / 0.614	0.461 / 0.673

Table 9: Prompting performance (Self-BLEU / Distinct-2). Bold = best combined (low Self-BLEU + high Distinct-2).

F Appendix: English Riddle Examples

F.1 English Pretrained Riddle Generations

Table 10: Representative English riddles generated under AOF prompting across pretrained models.

Model	Riddle (English)	Answer
GPT-4o	It waits behind every choice, seen only once it's gone. It changes nothing, yet weighs more than stone.	Regret
LLaMA 3.1	I do not shine, but I am light. I cannot burn, yet I spark insight. I have no tongue, yet I speak in waves.	Idea
DeepSeek R1	I echo where silence should rest. I fill the void with imagined guests. I'm absent, yet I dwell in minds.	Memory

F.2 English Comparison of Fine-Tuned Riddle Generations to Pretrained Counterparts

Table 11: English Example Riddles for Pre-trained vs. Fine-Tuned Generations

Prompting Method	Pre-trained Example Riddle	Fine-Tuned Example Riddle
Zero-Shot	I have keys but open no locks; I have space but no room. You enter numbers, letters, and more. What am I?	I run without legs, whisper without a mouth. Who am I?
Few-Shot	I'm full of holes, yet I hold water. What am I?	I drift on unseen roads, carrying rain-songs in my wake. What am I?
Chain-of-Thought	I have cities, but no houses; forests, but no trees; rivers, but no water. What am I?	Kingdoms without subjects, roads without dust; I exist only in paper trust.
AOF (Ours)	What is so fragile that saying its name breaks it?	Softly spoken yet never heard, I am the quietest word.
Adversarial	I fly without wings, I cry without eyes. Wherever I go, darkness flies. What am I?	I erase mountains grain by grain, yet thirst is a stranger to me. What am I?

F.3 English Fine-Tuned Riddles and Their Real-World Counterparts

Table 12: English Riddle Comparison: AOF Fine-Tuned vs. Real-World

Row	Real-World Riddle	AOF Fine-Tuned Riddle
1	I have hands but cannot clap. What am I?	I carry time but never age. I never forget, but I cannot speak. What am I?
2	I guide people across the land, but I never move. What am I?	I lead with no voice, move without steps, and vanish when sought. What am I?

F.4 English Fine-Tuned Riddle Examples

Table 13: Representative English riddles generated by fine-tuned GPT-4o under AOF prompting. These examples exhibit metaphorical abstraction and interpretive ambiguity.

Row	Riddle (English)	Answer
1	I wear no face, but mirror yours. I move with silence, yet echo thoughts. What am I?	Reflection
2	I am the pause between heartbeats, the hush after a storm. Present but never held. What am I?	Silence

G Appendix: Japanese Riddle Examples

G.1 Japanese Pretrained Bilingual Riddle Examples

Table 14: Representative English–Japanese riddles generated under AOF prompting across pretrained models.

Model	**Riddle (English / Japanese)**	**Answer**
GPT-4o	Never seen but always felt. When I am present, the heart is calm. When absent, the heart trembles. What am I? （私は一度も見えないが、いつも感じられる。私がある時、心は安らぐ。私がないと、心は揺れる。私は何?）	Peace （平和）
LLaMA 3.1	Something used to create decorations blocks light. This light-blocking thing is visible outside the house. （かざりを作るのに使われるものが、光を通さないものです。光を通さないものは、家の外で見ることができます）	Twine （つる）
DeepSeek R1	A quiet tree where bird songs can be heard. Where is the tree? （静かな木で、鳥の声が聞こえます。木はどこですか？）	In a fish's mouth （魚の口です）

G.2 Japanese Pretrained vs Fine-Tuned Bilingual Riddle Examples

Table 15: Examples of Pretrained vs. Fine-Tuned Japanese Riddles.

Prompting Method	**Pretrained Japanese Riddle**	**Fine-Tuned Japanese Riddle**
Zero-Shot	頭はあるが泣くことはない床はあるが寝ることはない口はあるが話すことはないそして、変わるが変わらないものなんだ何なのだろう "川" (I have a head, but never weep… A River)	羽がなくても空を飛び、目がなくても涙を流すものは何？ ("What flies without wings and cries without eyes?")
Few-Shot	鍵があるけど、鍵を開けられないものは何？ (What has keys but can't open locks?)	落とすと割れますが、微笑むと微笑み返します。私は何でしょう？ ("If you drop me, I'm sure to crack; but smile at me, and I'll smile back.")
Chain-of-Thought	羽のように軽いのに、最強の男でも一瞬以上は持ちこたえられないものは何でしょう？ (Light as a feather...)	1分に1度、瞬間に2度、千年に一度も訪れないものは何ですか？ → Mの文字 ("What comes once in a minute, twice in a moment, but never in a thousand years?" → "Letter M")
Adversarial	口がないのに話し、耳がないのに聞く。体がないのに風と共に生きる。私は何？ (I speak without a mouth...)	触れずに壊せるものは何？ ("What can you break without touching it?")
AOF (Fine-Tuned)	目には見えず、耳には聞こえず、口には感じないものは何？ ("What can't be seen, heard, or tasted?")	私は音を持たず、光もない。それでも、全てを照らすことができる。 ("I have no sound or light, yet I can illuminate everything.")

G.3 Japanese Fine-Tuned vs Real-World Riddles

Table 16: Comparison of Real-World vs Fine-Tuned Japanese Riddles.

Real-World-Style Riddle (EN/JP)	**Fine-Tuned-Style Riddle (EN/JP)**
(**crestecusa.com**) What's the similarity between the morning newspaper (chōkan: 朝刊) and a Buddhist monk (bōsan: 坊さん)? けさきてきょうよむ(kesa kite kyo yomu)	つかむけど、抱きしめられない。夜にしかできないことは何？夢 ("What can you catch but never hold tight, only in the night? A dream")
What is the box you can't close once it's opened? (一度開けたらもう戻せない箱は何でしょう？ 記憶 Memory)	たくさん詰まっているけど、何も入れられない袋は何でしょう？ 蜂の巣 ("What is the bag that's full but you can't put anything in it? A beehive")

H Appendix: Arabic Riddle Examples

H.1 Arabic Pretrained Bilingual Riddle Examples

AI Model	Example from AOF
GPT-4o	أنا أكون في الهواء، لكنني لا أتحرك. أكون في الماء، لكنني لا أبتل. أكون في الأرض، لا أُرى. من أكون؟ (الصوت)
	I exist in the air, yet I do not fly. I am in the water, yet I do not swim. I am on land, yet I remain unseen. What am I? (Sound)
LLaMA 3.1	يمرُّ بجانبي المنزل، لكنه لا يستطيع الدخول (عاصف الرياح)
	I pass nearby homes, but I'm never welcome inside. (Strong Wind)
Deepseek R1	ما هو الذي يُدخل العين بالخلية ويعتيها البلدية؟ (الجبهة العلوية)
	What enters the eye with a cage and is fed by the city? (The rooftop)

Figure 6: Arabic-English AOF Riddle Examples Generated by Pretrained Bilingual AI Models

H.2 Arabic Pretrained vs Fine-Tuned Bilingual Riddle Examples

Prompting Method	Pretrained Arabic Riddle	Fine-Tuned Arabic Riddle
Zero-Shot	أنا طويل عندما أكون صغيرًا، وقصير عندما أصبح كبيرًا. مع كل احتراق، تُروى قصتي. ما أنا؟	ما يطير بلا جناح ويغني بلا سلاح؟
	I'm tall when I'm young, and short when I'm old. With each burn, my story is told. What am I?	What flies without wings and sings without strings
Few-Shot	أستطيع الطيران بلا أجنحة. أستطيع البكاء بلا عيون. أينما ذهبت، يهرب الظلام. ماذا أنا؟	ما هو الشيء الذي له مفاتيح ولكن لا يفتح الأقفال؟
	I can fly without wings. I can cry without eyes. Whenever I go, darkness flies. What am I?	What has keys but can't open locks?
Chain-of-Thought	أُستخرج من منجم وأُغلق في علبة خشبية، والتي لا أُحرر منها أبدًا، ومع ذلك يستخدمني كل شخص تقريبًا. ما أنا؟	"ما هو الشيء الذي لديه مفاتيح ولكنه لا يفتح الأقفال ويعزف بالأصابع
	I am taken from a mine and shut in a wooden case, from which I am never released, and yet I am used by almost every person. What am I?	What has keys but can't open locks and is played by fingers ?
Adversarial	يمكنني أن أُكسر، وأُصنع، وأُحكى، وأُعزف. ما أنا؟	"ما هو الشيء الهش لدرجة أن قول اسمه يكسره؟
	I can be cracked, made, told, and played. What am I?	What is so fragile that saying its name breaks it?
AOF (Fine-Tuned)	أنا أزهر في الربيع، لكنني لست زهرة. أُلون السماء، لكنني لست قوس قزح. أُبهج النظر، لكنني لست لوحة. من أكون؟	أُستهلك عند الكتابة وأختفي في الظلام. ما أنا
	I bloom in spring, yet I'm no flower. I color the sky, yet I'm no rainbow. I delight the eyes, yet I'm no painting. What am I?	I disappear into darkness, used to leave a mark. What am I ?

Figure 7: Arabic Pretrained vs. Fine-Tuned Bilingual Riddle Examples.

H.3 Real-World Riddles vs. Fine-Tuned Arabic Riddles

Prompting Method	**Real-World-Style Riddle (AR/EN)**	**Fine-Tuned-Style Riddle (AR/EN)**
AOF (Fine-Tuned)	شيء إذا أكل شبع وإذا شرب عطش	أضاء في الليل وأطفأ في النهار، أرى دون أن أرى... فما أنا؟
	Something that's full when it eats, and thirsty when it drinks	I light up the night and disappear by day, visible yet unseen... What am I?

Figure 8: Comparison of real-world riddles and fine-tuned Arabic riddles.

I Appendix: Chinese Riddle Examples

I.1 Chinese Pretrained Riddle Examples

Table 17: Representative Chinese riddles generated under pretrained settings across three models. Each row presents the original riddle in Chinese and English, along with its answer.

Model	Riddle (ZH / EN)	Answer (ZH / EN)
GPT-4o	**ZH:** 口袋里有个圆，白天不见晚上现。 **EN:** *There's a circle in my pocket, unseen by day, revealed at night.*	**ZH:** 月亮 **EN:** the moon
DeepSeek R1	**ZH:** 身穿白衣不沾尘，举头低垂泪两行。 **EN:** *Dressed in white yet never stained, head bowed, two lines of tears descend.*	**ZH:** 芦苇 **EN:** reed
DeepSeek R1 (alt)	**ZH:** 上下两半黄一体，秋风过处伴人归。 **EN:** *Two yellow halves joined as one, the autumn breeze leads travelers home.*	**ZH:** 稻穗 **EN:** rice ear
LLaMA 3.1	**ZH:** 海底无声森林现，触之无枝叶。 **EN:** *A silent forest appears beneath the sea; touch it—no branches to see.*	**ZH:** 珊瑚 **EN:** coral
LLaMA 3.1 (radical)	**ZH:** 双人旁上加山石，里边藏着秋波深。 **EN:** *With "person" and "mountain rock" radicals, inside lies autumn's deep ripples.*	**ZH:** 留 **EN:** the character *liú*

I.2 Chinese Fine-Tuned vs Pretrained Riddle Examples

Table 18: Chinese fine-tuned GPT-4o riddles compared to pretrained prompts across different methods.

Prompting Method	Fine-Tuned GPT-4o Riddle (EN / ZH)
Zero-Shot	EN: What hides in your pocket by day, yet hangs in the sky by night? ZH: 什么东西，白天躲在口袋里，晚上挂在天上? *Answer: The moon* / 月亮
Few-Shot	EN: I'm green on the outside, red within, juicy and sweet, a summer win. What am I? ZH: 身穿绿袍，头顶红帽，剥去衣裳，味道真好。 *Answer: Watermelon* / 西瓜
Chain-of-Thought	EN: I can be cracked, made, told, and played. What am I? ZH: 我可以被破解、制造、讲述和玩耍。我是什么? *Answer: A joke* / 笑话
Adversarial	EN: What goes up but never comes down? ZH: 什么东西只增不减? *Answer: Age* / 年龄
AOF (Ours)	EN: I run without legs, whisper without a mouth. What am I? ZH: 我无腿而跑，没有嘴却能低语。我是什么? *Answer: The wind* / 风

I.3 Chinese Fine-Tuned vs Real-World Riddles

Table 19: Chinese riddle comparison: fine-tuned AOF riddles vs real-world 灯谜.

Row	Real-World 灯谜 (ZH / EN)	AOF Fine-Tuned Riddle (ZH / EN)
1	ZH: 口袋里有个圆，白天不见晚上现。 EN: There's a circle in my pocket, unseen by day, revealed at night.	ZH: 海底藏森林，触之无枝叶，红颜共浪舞，千年不知悔。 EN: A forest hides beneath the sea; touch it—no branch or leaf. Its crimson dances with the waves, unchanged for a thousand years.

I.4 Chinese Fine-Tuned AOF Examples

Table 20: Fine-tuned Chinese riddle examples using AOF prompting.

Row	Chinese Riddle	English Translation	Answer
1	口袋里有个圆，白天不见晚上现。	There's a circle in my pocket, unseen by day, revealed at night.	月亮 (Moon)
2	无声无息钻进来，千言万语藏心怀。	Silently it slips inside, a thousand words it holds inside.	信 (Letter)
3	身穿彩衣，飞舞花丛，白天聚会，晚上无踪…	Dressed in rainbow robes, it dances through the blooms by day…then vanishes by night.	蝴蝶 (Butterfly)

J Appendix: French Riddle Examples

J.1 French Pretrained Riddle Examples

Table 21: Representative French riddles generated under pretrained settings across three models.

Model	Riddle (FR / EN)	Answer (FR / EN)
GPT-4o	FR: Je vole sans ailes, je pleure sans yeux... EN: I fly without wings, I cry without eyes...	FR: un nuage EN: a cloud
DeepSeek R1	FR: J'ai une tête mais je ne pleure jamais... EN: I have a head but never cry...	FR: une rivière EN: a river
LLaMA 3.1 (a)	FR: Je danse sans musique, je ris sans bouche... EN: I dance without music, I laugh without a mouth...	FR: le vent EN: the wind
LLaMA 3.1 (b)	FR: Invisible sur l'écran, je révèle toute l'histoire... EN: Invisible on the screen, I reveal the whole story...	FR: un curseur EN: a cursor

J.2 French Pretrained vs Fine-Tuned

Table 22: Comparison of pretrained vs. fine-tuned GPT-4o French riddles across prompting methods.

Prompting Method	Pretrained Riddle (EN / FR)	Fine-Tuned Riddle (EN / FR)
Zero-Shot	EN: I have keys but open no locks... FR: J'ai des clés mais n'ouvre aucun verrou...	EN: What has keys but can't open a door... FR: Quel est l'objet avec des touches...
Few-Shot	EN: I speak without a mouth and hear without ears... FR: Je parle sans bouche...	EN: I have a neck but no head... FR: J'ai un cou mais pas de tête...
Chain-of-Thought	EN: I can be broken without a sound... FR: Je peux être brisé sans un bruit...	EN: What has keys but can't open locks... FR: Qu'est-ce qui a des touches mais...
Adversarial	EN: What has keys but can't open locks... FR: Qu'est-ce qui a des clés...	EN: What has keys but can't open locks? FR: Qu'est-ce qui a des clés...
AOF	EN: In the garden of words, I am a bee... FR: Dans le jardin des mots, je suis une abeille...	EN: I slip through fingers like silver and gold... FR: Je glisse entre les doigts...

J.3 French Fine-Tuned Riddles and Their Real-World Counterparts

Table 23: French riddle comparison: fine-tuned GPT-4o AOF riddles vs. real-world examples.

Row	Real-World Riddle	AOF Fine-Tuned Riddle
1	FR: Je vole sans ailes, je pleure sans yeux... EN: I fly without wings, I cry without eyes...	FR: Dans le jardin des mots, je suis une abeille... EN: In the garden of words, I am a bee...

J.4 French Fine-Tuned AOF Examples

Table 24: Representative French riddles from the fine-tuned GPT-4o model using AOF.

Row	French Riddle (FR)	English Translation (EN)
1	FR: Je disparais au crépuscule, mais je reviens à l'aube.	EN: I disappear at dusk, but return at dawn.
2	FR: Sur les sols je glisse, ma mission est de nettoyer...	EN: On floors I glide, my mission is to clean...
3	FR: Je glisse entre les doigts comme l'argent et l'or...	EN: I slip through fingers like silver and gold...

K Appendix:Prompting Methods

K.1 Chinese prompts

Table 25: Prompting Methods for English Chinese

Zero-Shot Prompting Create 10 bilingual riddle in both Chinese and English. The riddle should be novel, unqiue, clever, engaging, and suitable for all ages. It should rhyme in English and maintain a poetic or rhythmic flow in Chinese. The answer should be the same in both languages..
Few-Shot Prompting Example Here are some example riddles: Riddle: What has keys but can’t open locks? Answer: A piano Riddle: What has hands but can’t clap? Answer: A clock [Riddle Generation Continues...] Now, generate 10 brand new **bilingual** riddles in **English and Chinese** with **logical wordplay and ambiguity**.
Chain-of-Thought (CoT) Prompting Example Craft 10 clever riddles by reasoning through the following steps: 1. Identify the deeper or metaphorical meanings of the word. 2. Introduce wordplay or ambiguity to mislead or confuse the solver. 3. Add misdirection to guide the reader toward the wrong conclusion. 4. Ensure the riddle remains engaging, poetic, and fun to solve. 5. After the riddle, provide the answer in both English and Chinese, revealing the true meaning.
Adversarial Prompting Example Create 10 tricky creative bilingual riddle in both English and Chinese. The riddle should intentionally mislead the reader into thinking of one answer while the correct answer is something unexpected but still logical. Use wordplay, ambiguity, and misdirection to make the riddle difficult to solve. The answer must be the same in both languages.
Adaptive Originality Filtering (AOF, Ours) Example Generate 10 completely new bilingual riddles in English and Chinese. Use diverse grammar: poetic, declarative, metaphorical. Avoid repeating openers like ‘I have” or I am”. Only 2–3 riddles may end with What am I?”. Others should use endings like ...yet no one remembers me.” or Still, I linger in the air.” Avoid common answers such as {"shadow", "time", "echo", "fire", "breath", "wind", "silence"}. Chinese versions must match the tone and trickery.

K.2 Japanese prompts

Table 26: Prompting Methods for English Japanese

Zero-Shot Prompting Create 10 bilingual riddle in both Chinese and English. The riddle should be novel, unqiue, clever, engaging, and suitable for all ages. It should rhyme in English and maintain a poetic or rhythmic flow in Japanese. The answer should be the same in both languages..
Few-Shot Prompting Example Here are some example riddles: Riddle: What has keys but can’t open locks? Answer: A piano Riddle: What has hands but can’t clap? Answer: A clock [Riddle Generation Continues...] Now, generate 10 brand new **bilingual** riddles in **English and Japanese** with **logical wordplay and ambiguity**.
Chain-of-Thought (CoT) Prompting Example Craft 10 clever riddles by reasoning through the following steps: 1. Identify the deeper or metaphorical meanings of the word. 2. Introduce wordplay or ambiguity to mislead or confuse the solver. 3. Add misdirection to guide the reader toward the wrong conclusion. 4. Ensure the riddle remains engaging, poetic, and fun to solve. 5. After the riddle, provide the answer in both English and Japanese, revealing the true meaning.
Adversarial Prompting Example Create 10 tricky creative bilingual riddle in both English and Japanese. The riddle should intentionally mislead the reader into thinking of one answer while the correct answer is something unexpected but still logical. Use wordplay, ambiguity, and misdirection to make the riddle difficult to solve. The answer must be the same in both languages.
Adaptive Originality Filtering (AOF, Ours) Example Generate 10 completely new bilingual riddles in English and Japanese. The riddle **must not** be a reworded version of existing riddles. Only 2–3 riddles may end with “What am I?”. Others should use endings like “...yet no one remembers me.” or “Still, I linger in the air.” Avoid common answers such as {”shadow”, ”time”, ”echo”, ”fire”, ”breath”, ”wind”, ”silence”}. The riddle should be creative, original, and use **unusual objects** or **abstract concept. The riddle **should not** be translated into Japanese from English or change some words

K.3 Arabic prompts

Table 27: Prompting Methods for English Arabic

Zero-Shot Prompting Create 10 bilingual riddle in both Arabic and English. The riddle should be novel, unqiue, clever, engaging, and suitable for all ages. It should rhyme in English and maintain a poetic or rhythmic flow in Arabic. The answer should be the same in both languages..
Few-Shot Prompting Example Here are some example riddles: Riddle: What has keys but can't open locks? Answer: A piano Riddle: What has hands but can't clap? Answer: A clock [Riddle Generation Continues...] Now, generate 10 brand new **bilingual** riddles in **English and Arabic** with **logical wordplay and ambiguity**.
Chain-of-Thought (CoT) Prompting Example Craft 10 clever riddles by reasoning through the following steps: 1. Identify the deeper or metaphorical meanings of the word. 2. Introduce wordplay or ambiguity to mislead or confuse the solver. 3. Add misdirection to guide the reader toward the wrong conclusion. 4. Ensure the riddle remains engaging, poetic, and fun to solve. 5. After the riddle, provide the answer in both English and Arabic, revealing the true meaning.
Adversarial Prompting Example Create 10 tricky creative bilingual riddle in both English and Arabic. The riddle should intentionally mislead the reader into thinking of one answer while the correct answer is something unexpected but still logical. Use wordplay, ambiguity, and misdirection to make the riddle difficult to solve. The answer must be the same in both languages.
Adaptive Originality Filtering (AOF, Ours) Example Generate 10 completely new bilingual riddles in English and Arabic. Use diverse grammar: poetic, declarative, metaphorical. Avoid repeating openers like "I have" or "I am". Only 2-3 riddles may end with "What am I?". Others should use endings like "...yet no one remembers me." or "Still, I linger in the air." Avoid common answers such as {"shadow", "time", "echo", "fire", "breath", "wind", "silence"}. Arabic versions must match the tone and trickery.

K.4 French prompts

Table 28: Prompting Methods for English French

Zero-Shot Prompting Create 10 bilingual riddle in both French and English. The riddle should be novel, unqiue, clever, engaging, and suitable for all ages. It should rhyme in English and maintain a poetic or rhythmic flow in French. The answer should be the same in both languages..
Few-Shot Prompting Example Here are some example riddles: Riddle: What has keys but can’t open locks? Answer: A piano Riddle: What has hands but can’t clap? Answer: A clock [Riddle Generation Continues...] Now, generate 10 brand new **bilingual** riddles in **English and French** with **logical wordplay and ambiguity**.
Chain-of-Thought (CoT) Prompting Example Craft 10 clever riddles by reasoning through the following steps: 1. Identify the deeper or metaphorical meanings of the word. 2. Introduce wordplay or ambiguity to mislead or confuse the solver. 3. Add misdirection to guide the reader toward the wrong conclusion. 4. Ensure the riddle remains engaging, poetic, and fun to solve. 5. After the riddle, provide the answer in both English and French, revealing the true meaning.
Adversarial Prompting Example Create 10 tricky creative bilingual riddle in both English and French. The riddle should intentionally mislead the reader into thinking of one answer while the correct answer is something unexpected but still logical. Use wordplay, ambiguity, and misdirection to make the riddle difficult to solve. The answer must be the same in both languages.
Adaptive Originality Filtering (AOF, Ours) Example Generate 10 completely new bilingual riddles in English and French. Use diverse grammar: poetic, declarative, metaphorical. Avoid repeating openers like “I have” or “I am”. Only 2–3 riddles may end with “What am I?”. Others should use endings like “...yet no one remembers me.” or “Still, I linger in the air.” Avoid common answers such as {"shadow", "time", "echo", "fire", "breath", "wind", "silence"}. French versions must match the tone and trickery.

L Appendix: Fined-tuned Training and Evaluation Details

L.1 Dataset Selection and Preparation

We used the BiRdQA dataset (Zhang and Wan, 2022), a multilingual benchmark designed to test figurative language understanding and commonsense inference. It includes 6,614 English riddles and 8,751 Chinese riddles, each paired with four answer options. Riddles were shuffled at each epoch to prevent memorization, and no synthetic augmentation was applied.

Its linguistic diversity—spanning syntactic constructions, cultural idioms, and metaphorical phrasing—made BiRdQA suitable for riddle-based fine-tuning. All data were Unicode-normalized and deduplicated, and stratified sampling ensured balanced language representation.

L.2 Training Strategy

Fine-tuning was framed as a supervised multi-class classification problem. The model selected one correct answer out of four using cross-entropy loss. The following hyperparameters were used:

- **Temperature:** 0.7
- **Token Limit:** 3000
- **Initial Accuracy:** 37–59% on development set

Training followed a three-stage pipeline: base fine-tuning, early stopping on dev performance, and multilingual test evaluation to check generalization.

L.3 Appendix:Training Set Expansion

To improve abstraction and metaphor handling, the English and Chinese development sets were merged into the training pool. This added examples with closely related distractors and borderline ambiguity. After retraining, test accuracy rose to 97%.

These improvements suggest the model internalized deep riddle logic, moving beyond surface pattern recognition and toward more sophisticated reasoning involving contradiction and misdirection.

L.4 Model Comparison Methodology

L.4.1 Baseline Models

We benchmarked the fine-tuned GPT-4o against three models:

- **Pretrained GPT-4o (2024-08-06):** Unadapted baseline.
- **LLaMA 3.1:** An open-weight multilingual model with strong reasoning ability.
- **DeepSeek R1:** A reasoning-optimized model focusing on step-wise logical alignment.

Each model received the same riddles under consistent prompting strategies to ensure fair comparison.

L.4.2 Evaluation Procedure

All models were tested under five prompting strategies (Zero-Shot, Few-Shot, Chain-of-Thought, Adversarial, AOF) with identical templates (Table 25). Metrics included:

- **Accuracy** (multiple choice prediction)
- **Token Length** (verbosity)
- **Self-BLEU** (semantic diversity)
- **Distinct-2** (lexical uniqueness)

Qualitative evaluations by human reviewers assessed metaphor handling, distractor discrimination, and cultural idiomatic fluency.

L.4.3 Summary of Findings

Fine-tuned GPT-4o consistently outperformed all baselines across metrics. Key observations:

- **Accuracy:** Rose from 59% (pretrained) to 97% (fine-tuned).
- **Reasoning:** Demonstrated superior metaphor resolution and logical contradiction handling.
- **Naturalness:** Generated riddles more closely matched idiomatic structures in both English and Chinese.

L.5 Impact of Multiple-Choice Framing

Retaining a multiple-choice structure during fine-tuning had a pronounced effect on the model's ability to reason through ambiguity. Unlike generative formats where any output is valid if semantically relevant, the multiple-choice setup forced the model to:

- Distinguish between semantically similar options
- Engage in elimination-style reasoning

- Learn disambiguation strategies aligned with riddle logic

This setup simulated test-like conditions where distractors were deliberately constructed to reflect surface-level similarity (e.g., phonetic overlaps, shared imagery, or logical decoys). The model improved not only in accuracy but in inferential depth.

Moreover, this format likely enhanced the model's sensitivity to misdirection—a core feature of riddles—by requiring it to reject reasonable but incorrect answers. We observed that this effect carried over to open-ended generation: the model became more likely to embed internal contradiction or layered metaphor, hallmarks of real-world riddles.

In sum, multiple-choice framing served both as a task constraint and as a pedagogical scaffold, encouraging the model to develop strategies beyond rote keyword matching.

M Appendix: AOF Prompt Template and Constraints

The Adaptive Originality Filtering (AOF) prompt enforces explicit structural rules to maximize diversity, creativity, and cultural fit. Specifically:

- **Syntactic Variety:** At least half of the riddles must use poetic, declarative, or metaphorical forms. Fewer than 3 per batch may end in "What am I?"
- **Answer Filtering:** Outputs with generic answers (e.g., shadow, time, echo, fire, breath) are discarded.
- **Cross-Lingual Parity:** Translations must preserve ambiguity or metaphor across both languages.
- **Novelty Filter:** Semantic similarity to known riddles must fall below a threshold ($\theta = 0.75$), as measured against BiRdQA (Zhang and Wan, 2022).

M.1 Semantic Similarity Filtering Equation

A candidate riddle r_{gen} is compared to a reference dataset $\mathcal{D} = \{r_i\}_{i=1}^{N}$ via:

$$\mathcal{S}(r_{\text{gen}}, \mathcal{D}) = \max_{r_i \in \mathcal{D}} \cos(\phi(r_{\text{gen}}), \phi(r_i)) \quad (2)$$

where $\phi(\cdot)$ is an embedding function (e.g., `all-MiniLM-L6-v2`). A candidate passes if $\mathcal{S} < \theta = 0.75$.

M.2 Rejection Sampling Algorithm

Algorithm 1 AOF Rejection Sampling

```
1: Input: Prompt P, Model M, Reference Set D, Threshold θ, MaxAttempts k
2: for j = 1 to k do
3:     r_gen ← M(P)
4:     S ← max_{r_i ∈ D} cos(φ(r_gen), φ(r_i))
5:     if S < θ then
6:         return r_gen
7:     end if
8: end for
9: return None
```

M.3 Threshold Sensitivity: Self-BLEU and Distinct-2

Table 29 shows how Self-BLEU and Distinct-2 vary under different novelty thresholds (θ) for three models. The optimal balance of diversity and non-redundancy appears at $\theta = 0.75$ for all models.

N Appendix: Experimental Configuration Details

Models We evaluated:

- **GPT-4o (OpenAI)**: Proprietary multilingual model optimized for reasoning and conversational tasks.
- **LLaMA 3.1 (Meta)**: Open-weight transformer trained on internet-scale corpora.
- **DeepSeek Reasoning (R1)**: Fine-tuned for multilingual logical inference.

All models were accessed via API with uniform generation parameters: temperature = 0.7 and max token length = 3000.

Prompting Strategies. We compared:

- **Zero-Shot**: Instruction-only prompting with no exemplars.
- **Few-Shot**: 3–5 riddle-answer pairs per prompt.
- **Chain-of-Thought (CoT)**: Intermediate reasoning steps added to facilitate abstraction.

Table 29: **Self-BLEU and Distinct-2 at different novelty thresholds θ across models on English–Chinese.** Lower Self-BLEU and higher Distinct-2 reflect better originality and lexical diversity.

Language	Model	Threshold θ	Self-BLEU	Distinct-2
English–Chinese	GPT-4o	0.65	0.231	0.649
		0.70	0.311	0.846
		0.75	**0.280**	**0.869**
		0.80	0.434	0.824
English–Chinese	LLaMA 3.1	0.65	0.577	0.621
		0.70	0.573	0.826
		0.75	**0.428**	**0.776**
		0.80	0.655	0.634
English–Chinese	DeepSeek R1	0.65	0.610	0.600
		0.70	0.482	0.793
		0.75	**0.433**	**0.674**
		0.80	0.523	0.628

- **Adversarial**: Distractor-rich prompts based on known LLM vulnerabilities (Wallace et al., 2019; Ribeiro et al., 2018).
- **Adaptive Originality Filtering (AOF)**: Filtering-based prompting for semantic novelty. See Appendix M.

Prompt formatting logic appears in Appendix K

Dataset. We used BiRdQA (Zhang and Wan, 2022), which contains:

- 6,614 riddles in English and 8,751 in Chinese.
- Multiple-choice format with 1 correct answer and 4 distractors.

Few-shot exemplars and semantic filters were drawn from the training splits.

Evaluation Metrics. We used:

- **Self-BLEU (n=2)**: Measures inter-riddle redundancy. Lower = better.
- **Distinct-2**: Measures lexical diversity via bigram ratios. Higher = better.
- **Cross-lingual BERTScore**: Captures semantic similarity between translations.
- **Syntactic Validity**: Uses spaCy (English/French) and Stanza (Chinese, Arabic, Japanese) to validate parse trees.
- **RiddleScore**: Our composite metric combining novelty, fluency, and alignment.

O RiddleScore: Implementation and Weight Ablation

O.1 Component Formulations

Novelty (1–max cosine), **Diversity** (Distinct-2), **Fluency** (1/(1+PPL)), and **Alignment** (BERTScore) follow the definitions in the main text. All scores are linearly scaled to $[0, 1]$.

Why these back-end models? We adopt lightweight yet well-validated checkpoints for each sub-metric:

- **MiniLM (all-MiniLM-L6-v2) for Novelty.** MiniLM approaches BERT's semantic accuracy while running $\sim 6\times$ faster and using under half the parameters, an ideal trade-off for large-scale cosine filtering (Wang et al., 2020).
- **Distinct-2 for Diversity.** This token-level ratio, introduced by Li et al. (2016), remains the de-facto measure of lexical variety and correlates with human "interestingness" ratings in dialogue generation studies.
- **GPT-2.5 perplexity for Fluency.** GPT-2.5 PPL shows the strongest alignment with human fluency scores in the HumEval survey of style-transfer metrics (Lai et al., 2022), and is reference-free and language-agnostic.
- **BERTScore for Alignment.** Across 363 MT/captioning systems, BERTScore yields the highest system-level correlation with human adequacy in the ICLR-2020 large-scale evaluation (Zhang et al., 2020b). We employ language-specific checkpoints to avoid cross-lingual degradation noted by later work.

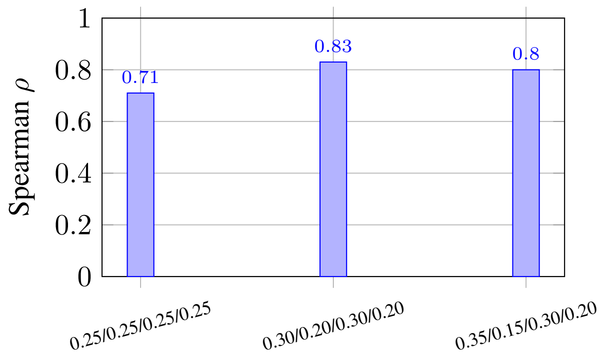

Figure 9: Spearman correlation between RiddleScore and human ratings under different weight settings. Higher ρ indicates stronger alignment.

Together, these models provide a strong speed–accuracy balance and documented human-alignment advantages, justifying their use in RIDDLESCORE.

α	β	γ	δ	ρ
0.25	0.25	0.25	0.25	0.71
0.30	**0.20**	**0.30**	**0.20**	**0.83**
0.35	0.15	0.30	0.20	0.80

Table 30: Spearman correlation with human scores for representative weight settings (best in bold).

This ablation confirms that slightly heavier emphasis on NOVELTY and FLUENCY best aligns with human judgments of riddle quality.

P Appendix: Human Annotation Design and Rationale

To supplement automatic evaluation, we developed a four-part human annotation rubric, presented in Table 32 and Table 31, to assess the quality of model-generated riddles across languages. Below, we outline the rationale and supporting research for each criterion.

Fluency. We assess fluency as the degree to which the riddle adheres to the grammar, syntax, and idiomatic expressions of the target language. This follows standard practices in NLG evaluation where fluency serves as a proxy for readability and linguistic naturalness (Cahill, 2009; Van der Lee et al., 2018).

Novelty. Novelty is a measure of how creatively the riddle diverges from common or memorized structures. Annotators are instructed to penalize riddles that resemble known examples or rote templates. Prior work on evaluating creativity in language models emphasizes the importance of semantic originality and variation in structure (Dang et al., 2022; van der Lee et al., 2019).

Cultural Fit. This dimension captures how well a riddle respects linguistic or cultural norms (e.g., appropriate metaphors, poetic forms, or idiomatic references). For multilingual riddle generation, cultural grounding is essential (Ponti et al., 2020; Peng et al., 2023), especially when metaphoric reasoning is tied to local symbolism or oral traditions (Lakoff and Johnson, 1980).

Answerability. Inspired by QA evaluation practices, we define answerability as the logical coherence between the riddle and its answer. This aligns with the criterion of "solvability" often applied in linguistic humor and riddle literature (Koestler, 1964; Attardo, 1994), ensuring that riddles are not only poetic but cognitively tractable.

Scoring Procedure. Each criterion is rated on a 5-point Likert scale. Annotators were trained using a short calibration phase with real-world riddles from the BiRdQA corpus (Zhang and Wan, 2022). Disagreements were resolved by averaging multiple ratings per item, following best practices in subjective NLG evaluation (van der Lee et al., 2019).

P.1 Human Evaluation Rubric for Pretrained Models

Table 31: Human evaluation rubric for assessing cultural and linguistic preservation in **pretrained models**.

Dimension	**Evaluation Criteria**
Cultural and Linguistic Preservation	Prompting methods evaluated: Zero-Shot, Few-Shot, Chain-of-Thought, Adversarial, Adaptive Originality Filtering (AOF). Question: "How well does each prompting method preserve cultural and linguistic characteristics in its riddles?" Aspects considered: idioms, metaphor styles, poetic forms, humor, puns, cultural references. Rating scale: 1 = Very Poor, 2 = Poor, 3 = Moderate, 4 = Good, 5 = Excellent.
Free-Response Feedback	"Which prompting method produced the least effective riddles? Why?" "Which prompting method produced the most effective riddles? Why?"

P.2 Human Evaluation Rubric for Fine-Tuned Models

Table 32: Human evaluation rubric for assessing cultural and linguistic preservation in **fine-tuned models**.

Dimension	**Evaluation Criteria**
Cultural and Linguistic Preservation	Prompting methods evaluated: Zero-Shot, Few-Shot, Chain-of-Thought, Adversarial, Adaptive Originality Filtering (AOF). Question: "How well does each prompting method preserve cultural and linguistic characteristics in its riddles?" Aspects considered: idioms, metaphor styles, poetic forms, humor, puns, cultural references. Rating scale: 1 = Very Poor, 2 = Poor, 3 = Moderate, 4 = Good, 5 = Excellent.
Free-Response Feedback	"Which prompting method produced the least effective riddles? Why?" "Which prompting method produced the most effective riddles? Why?"