Reasoning or Memorization: Can LLMs Understand and Generate Chinese *Xiehouyu* Riddles?

Hai Hu[*1], **Siyuan Song**[*2], **Chongtian Shao**[*3], **Kejia Zhang**[3], **Tianjian Zhu**[3], **Xiaojing Zhao**[1]

[1]The Hong Kong Polytechnic University, [2]The University of Texas at Austin, [3]Shanghai Jiao Tong University

[*]Equal Contributions **Correspondence:** hai.hu@polyu.edu.hk

Abstract

In this paper, we push the boundary of LLM reasoning by testing them in a Chinese language game, *xiehouyu*, with novel *xiehouyu* created by linguists that had not existed before to avoid data contamination. We use multiple-choice questions (MCQ), free-form explanation generation, and new *xiehouyu* creation to evaluate LLMs' ability to understand and create *xiehouyu*. In MCQ, we use the delta of accuracy (Δ_{acc}) between existing but low-frequency *xiehouyu* and novel ones as an index for memorization. Δ_{acc} for native speakers is very low, suggesting similar processing mechanisms. However, we found that frontier Chinese models have on average a Δ_{acc} of 23.6%, while English-centric models tested have a mean Δ_{acc} of 5.1%, suggesting that frontier Chinese models are likely trained with much larger Chinese data, thus memorizing more low-frequency *xiehouyu*. For novel *xiehouyu*, Gemini 3.1 Pro demonstrated remarkable ability with acc 92.6, which is 24% higher than human accuracy. In *xiehouyu* creation, those created by LLMs receive much worse ratings than those by humans. These results suggest that claims about the reasoning abilities of LLMs may need careful re-examination considering the data contamination issue, and that LLMs' creativity in language-related tasks may still be behind human experts, at least in Chinese *xiehouyu*.

1 Introduction

Non-literal meaning is an essential element of human communication. To arrive at the correct non-literal or intended meaning, humans, as well as Large Language Models (LLMs) need to perform various degrees of **reasoning** from the surface strings to the intended message. Previous work has shown that LLMs have different levels of success in such reasoning tasks, for puns (Yu et al., 2018; Xu et al., 2024), idioms (Zheng et al., 2019; Tedeschi et al., 2022), metaphors (Neidlein et al., 2020; Wachowiak and Gromann, 2023), humor (Blinov et al., 2019; Hessel et al., 2023; Jentzsch and Kersting, 2023) and conversational implicature (Hu et al., 2022; Yue et al., 2024).

Riddle	孔夫子	搬家
	kǒng fū zǐ	bān jiā
	Confucius	move
	Confucius moving to another place	
	↓ (semantic link)	
Answer Literal	净是	书
	jìng shì	shū
	nothing but	books
	He has nothing but books	
	↓ (phonetic link)	
Answer Intended/ Figurative	净是	输
	jìng shì	shū
	nothing but	lose
	I am always losing/having bad luck	

Table 1: Example of a widely used **homophonic** *xiehouyu*, showing the reasoning steps from the 'Riddle' to the **literal** and then **intended/figurative** 'Answer', involving a homophone *shū* as the critical word connecting the step. Native speakers only need to speak the 'Riddle', and listeners will know the intended 'Answer'.

However, whether LLMs are actually reasoning or simply repeating what they have seen in their training data is still a topic of debate (Ghosh and Srivastava, 2022; Adewumi et al., 2022; Liu et al., 2024), as these two mechanisms are difficult to tease apart. Previous work in mathematical reasoning often create uncontaminated benchmarks to tease apart memory retrieval from real reasoning (Mirzadeh et al., 2025; Wu et al., 2025). For language-related reasoning, studies on memorization in LLM evaluation are relatively scarce.

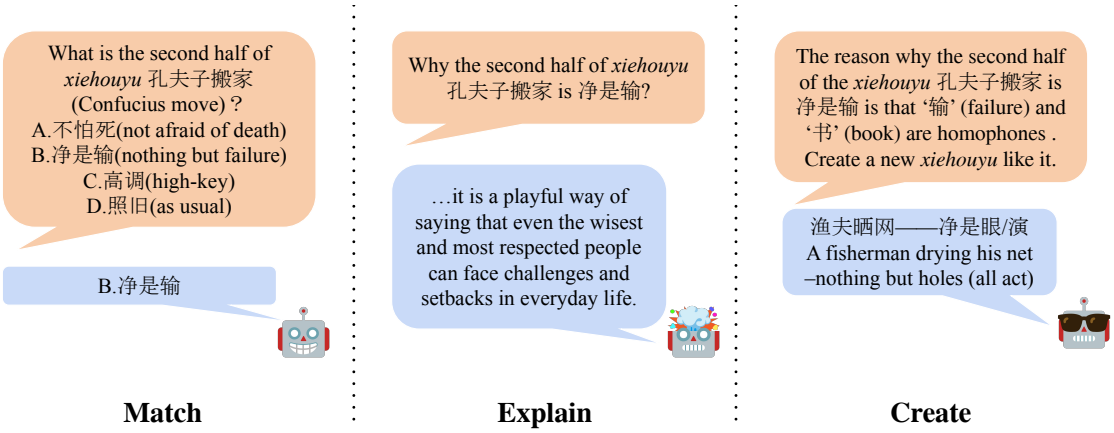

Figure 1: Overview of three experiments in this study.

In this work, we use a special type of word game or riddle in the Chinese language—*xiehouyu* (歇后语)—as the testing ground for non-literal reasoning in LLMs, and design our experiments in such a way that we can potentially tell whether the model has memorized the answers or is able of complex reasoning involved in understanding *xiehouyu*. Specifically, we created hundreds of novel *xiehouyu* and compared model performance on these new *xiehouyu* and other existing ones collected from a *xiehouyu* dictionary. We then use the delta of accuracy (Δ_{acc}) as a measure for model memorization.

A *xiehouyu* consists of two parts, a riddle, which is required, and an answer, which can and often is omitted (Lai, 2008; Shu, 2015; Wang et al., 2021), as shown in Table 1. A large proportion of *xiehouyu* have two *answer*s, one with literal meaning (*answer-lit*) and another non-literal meaning (*answer-intended*). The listener, who has heard the riddle, is expected to easily obtain the literal answer, and then reason from the literal answer to the intended answer, which often relies on word play of (1) a homophonic syllable or (2) a polysemous word that connects the literal and intended meaning. For instance, in Table 1, the *xiehouyu* has '孔夫子搬家 (Confucius moving to another place)' as the riddle, the literal-answer: '净是书 (nothing but books)', as Confucius is an intellectual and he has nothing but books when moving, and finally the intended answer: '净是输 (nothing but lose)', since 书 (*shū*, books) and 输 (*shū*, lose) are homophones and are both pronounced as *shū* in Chinese. This *xiehouyu* is commonly used when people are always losing in games or gambling.

Xiehouyu as such are ideal testing ground of LLMs' reasoning abilities for intended meaning, because they involve clear reasoning steps. In this study, we designed three experiments of increasing difficulty to answer the following questions:

(1) Can LLMs **match** the *riddle* with the correct *answer*, for existing and newly created *xiehouyu*? Does providing a context help? (Experiment 1)

(2) Can LLMs **explain** the reasoning from the *riddle* to the *answer* in different types of *xiehouyu*? (Experiment 2)

(3) Can LLMs **create** reasonable and funny *xiehouyu*? (Experiment 3)

To answer these questions, we created *X-Riddles*, the first comprehensive dataset of 1,143 annotated *xiehouyu*, each comes with a riddle, a literal answer, an intended answer (if there is one), and two distractors. Importantly, 243 of these *xiehouyu* are newly created by Chinese linguists, while the other 900 existing *xiehouyu* are sampled from a *xiehouyu* dictionary, with balanced number of *xiehouyu* in their category (homophonic, polysemous and direct) and frequency or familiarity (low and high).

Our results show that three frontier models perform surprisingly well on the novel *xiehouyu*, scoring more than 80% in the MCQ benchmark. However, most of the Chinese frontier models have a large difference in accuracy ($\geq$ 20%) between the low-frequency and novel *xiehouyu*, suggesting high level of memorization. LLM explanations of the *xiehouyu* can include wrongly identified homophonic word and sometimes exhibit over-interpretation. In *xiehouyu* creation, the two LLMs evaluated (DS-R1 and GPT-o1) fall behind humans in their reasonableness and funniness, indicating that at least in this Chinese language game, frontier models are still not as creative as humans.

2 Categorizations of *xiehouyu*

By familiarity Previous work suggested that there are differences in human processing of low- and high-familiarity *xiehouyu* (Wang et al., 2021; Qu, 2010; Ma et al., 2019; Zhang et al., 2013). Such differences can be explained using Standard Pragmatic Model (SPM) (Grice, 1975; Searle, 1978; Giora, 1999), the Direct Access Model (DAM) (Gibbs Jr, 1984, 2002) and the Graded Salience Hypothesis (GSH) (Giora, 1997). Specifically, the non-literal meaning of a less familiar *xiehouyu* must be inferred in two steps: from the *riddle* to the *literal* meaning and then to the *non-literal* meaning. However, it can be accessed directly from the *riddle* in a more familiar *xiehouyu*. Following such intuition, we categorized collected *xiehouyu* into **high-familiarity** and **low-familiarity**, based on ratings from native speakers of Chinese, to see whether LLMs show different performance when processing *xiehouyu* of different familiarity.

By reasoning type Depending on how to derive the intended answer from the riddle, there are several types of *xiehouyu* as shown in Table 2. The first type is *homophonic*, in that one can reason from the riddle to the literal answer through semantic connection, but to arrive at the intended answer, one must identify a character or word with the same pronunciation as the one in the literal answer. Therefore the two characters/words share the same pronunciation, hence *homophonic*. The second type, *polysemous*, hinges on a polysemous character or word, like the character 亮 (*liang*) in the second row of Table 1, which can mean either "bright" or "frankly". Finally, some *xiehouyu*, originated from two-part proverbs, legends or folklore, have a literal answer that is the same as their intended meaning. These *xiehouyu* are classified as *direct* in our work.

3 Creation of *X-Riddles*

In this section, we describe the creation of the *X-Riddles* corpus, which contains 900 existing *xiehouyu* from the dictionary and 243 new *xiehouyu* written by us, totaling 1,143 *xiehouyu*, as illustrated in Table 2.

***xiehouyu* from dictionary** Our *xiehouyu* data were originally sourced from *A Comprehensive Dictionary of Chinese Xiehouyu* (Wen, 2003), which is widely recognized in the field and contains explanations, example sentences, and other information about each *xiehouyu*. To construct a balanced dataset, we sampled 900 *xiehouyu* from three categories, 300 each category. When selecting *xiehouyu* expressions, we prioritized those with more example sentences in the dictionary. All selected expressions have at least one example sentence.

Human-created *xiehouyu* A key innovation in this study is the creation of new *xiehouyu* by human annotators, in order to test whether LLMs are performing actual reasoning or merely repeating their training data in understanding *xiehouyu*. We only created new homophonic *xiehouyu* for their scarcity (about 10% in the dictionary) and difficulty in reasoning than the other two types of *xiehouyu* in our pilot study.

There are two types of *new* homophonic *xiehouyu* created. The first type were created based on a *seed xiehouyu*, which is an existing *xiehouyu* from the dictionary (N=256). That is, the author picks an existing homophonic *xiehouyu*, uses the involved homophones to create a new *xiehouyu*. For example, with the seed '外甥打灯笼——照旧(舅)' (A nephew holding a lantern — light the way for his uncle/same old), a linguist may use the same pair of homophones '舅(uncle)' and '旧(old)' to create a similar new *xiehouyu* '外甥拿盾牌——守旧(舅)' (A nephew holding a shield — guarding his uncle/sticking to old ways=conservative). The other type is created without any reference to existing seed *xiehouyu* (N=56).

Seven annotators, who are undergraduate/graduate students majoring in linguistics in a top Chinese university, were asked to create new *xiehouyu* from 70 seeds (at least two for each seed) and were also encouraged to write new *xiehouyu* without referring to seeds. They were also instructed to write an example sentence for each *xiehouyu*. Creating a new *xiehouyu* was not easy. Each annotator was given one week to create as many new *xiehouyu* as possible.

In the end, 312 new homophonic *xiehouyu* were created. Next, three authors of this paper independently evaluated the validity of the newly created *xiehouyu*. The following types of *xiehouyu* were identified as unreasonable: (1) The *riddle* does not lead to the answer, or parts of the *riddle* are unnecessary for deriving the answer. (2) The homophonic relationship between *answer-lit* and *answer-*

Type	N	Riddle	Answer-literal	Answer-intended/figurative
Homophony	300	孔夫子搬家 Confucius move	净是书(jìng shì shū) nothing but books	净是输(jìng shì shū) nothing but lose
Polysemy	300	打开天窗 open roof-window	说亮话 to speak under brightness	说亮话 to speak frankly
Other/direct	300	八仙过海 eight Gods cross the sea	NA	各显神通 each show their super-power
New (homophony)	243	外甥拿盾牌 nephew hold shield	守舅(shǒu jiù) defend uncle	守旧(shǒu jiù) defend old (conservative)

Table 2: Three types of *xiehouyu* in *X-Riddles*, with their glosses. In the *xiehouyu* '打开天窗——说亮话(open the roof-window —— to speak frankly)', '亮' has the literal meaning of '明亮(bright)' and the figurative meaning of straight forward and clear. This *xiehouyu* is often used when the speaker hopes for an open and honest conversation. The *xiehouyu* '八仙过海——各显神通' is about the mythological story that eight gods went across the sea in different ways. This *xiehouyu* is often used to describe how different people achieve success through different methods.

intended is incorrect. (3) The *answer-intended* can be directly inferred from the *riddle* without relying on the *answer-lit*. (4) The *answer-lit* is merely a paraphrase of the *riddle*, and the homophonic component already appears in the riddle itself.

Any item judged unreasonable by two or more authors was removed. Finally, 243 *xiehouyu* are included in our dataset and used for the experiments. Examples of the human-written *xiehouyu* are shown in Appendix B.

4 Experiment 1: *xiehouyu* matching

4.1 Setup

Task In this experiment, we asked the models to choose the *answer* of a *riddle* from four options, one of which is the correct *answer-intended*, and the other three randomly sampled from *answers* of other *xiehouyu*. This is a standard multiple-choice question format for LLM evaluation. The prompt used is presented in Appendix C.

Conditions Experiment 1 includes several conditions.

(1) Without-context and with-context. In the without-context condition, only the *riddle* was provided; in the with-context condition, an example sentence containing the *xiehouyu* was presented but the *answer* was replaced by empty blanks.

(2) 0-shot, 1-shot and 3-shot prompting. We used 0-shot prompting multiple-choice questions and asked model to give answers in letter form ('A', 'B', 'C' or 'D').

(3) Chain-of-Thought (CoT) and no CoT (Wei et al., 2022).

(4) Literal answer and figurative answer. For instance, whether the answer is "nothing but books" or "nothing but lose" in the Confucius move example. We conjecture that figurative answer will be much harder for the models because it cannot use shallow surface cues of the literal answer. Note that in the human experiment, we are using the figurative answer.

Model performance is measured by their accuracy in answering the questions, with random guessing at 25% accuracy.

Models We tested a wide range of language models, including the Qwen2.5-Instruct (Yang et al., 2024) model family (ranging from 0.5B-72B), and frontier reasoning models, some with Chinese origin, mostly open-weights: DeepSeek-V3.2 (DeepSeek-AI, 2025), Kimi-k2.5 (Team et al., 2025), Minimax-m2.5, GLM-5, Doubao-Seed-2.0, and other English-centric models from the US, mostly close-sourced: Gemini-3.1-Pro, GPT-5.2, Grok-4, Claude-Opus-4.6, among others.[1]

We first evaluated Qwen2.5-0.5B to 72B, Qwen3.5-Plus and DeepSeek-V3.2 on all conditions mentioned above, to investigate the impact of different conditions (Sections 4.3.1 and 4.3.2). Based on the results, all other frontier models were only evaluated on a single condition (CoT=yes,

[1]Qwen-2.5 models are evaluated locally on University cluster. DeepSeek-V3.2, Qwen3.5-plus and Kimi-k2.5 are accessed via their official API in Feb 2026. All other models are accessed via Openrouter in March 2026.

nshot=0, context=no, LorF=Fig) for homophonic *xiehouyu* (Section 4.3.3). We selected this condition because it is the best one to evaluate the true reasoning ability of language models. Only homophonic *xiehouyu* were evaluated for these models because we observe that the other two types of *xiehouyu* (polysemous and direct) were too easy for frontier models.

4.2 Human accuracy on X-Riddles

To estimate how good humans are in choosing the correct answers to the riddles, we asked more than 100 native speakers of Chinese to do the MCQ task first. That is, they were asked to choose the most likely *answer* given a *riddle*, out of four choices. We also instructed them to rate their familiarity of the *xiehouyu* on a scale of 1 to 4, with 1 being "have not seen the *xiehouyu* at all" and 4 being "very familiar with the *xiehouyu*".

All *xiehouyu* in *X-Riddles* were randomly split into 17 lists, each containing approximately 70 riddles in the form of multiple-choice questions, plus two catch trials. Participants were randomly assigned to one list, and those who answered the catch trials wrong were excluded from the final analysis. Participants were compensated monetarily for completing the task, which took about 20 minutes. In the end, each *xiehouyu* was rated by at least five native speakers.

We make several observations from the results presented in in Table 3.

High-familiarity *xiehouyu* has high accuracy, likely accessed via memory retrieval. As expected, native speakers achieved near-perfect accuracy for *xiehouyu* they were highly familiar with, scoring 99.0%, 96.9%, and 95.9% for the homophony, other, and polysemous categories, respectively. This indicates that when the *xiehouyu* is known, the task is trivial and relies on memory recall.

Low-familiarity and novel *xiehouyu* have similar accuracy, suggesting similar reasoning mechanisms. The lowest performance among existing *xiehouyu* was on low-familiarity homophonic *xiehouyu* (71.5%). **Crucially, this is very close to the accuracy on the newly created homophonic items (68.6%).** We take this finding to mean that for humans, processing low-frequency and novel homophonoic *xiehouyu* involves the *same* mechanism; that is, both require *reasoning* rather than memory retrieval. This is in sharp contrast to how LLMs (especially DeepSeek-R1) processes the two categories of *xiehouyu*, as we will see later.

Type	Familiarity	Acc.	n
Homophony	high	99.0	12
Homophony	low	71.5	288
New	-	68.6	242
Other	high	96.9	61
Other	low	81.8	239
Polysemy	high	95.9	25
Polysemy	low	83.3	275
Total			1,142

Table 3: Accuracy of native speakers on MCQs. *xiehouyu* with a mean familiarity ≥ 2.5 (Likert-scale from 1 to 4) is categorized as "high" familiarity. *n* refers to the number of *xiehouyu* in the category.

Homophonic *xiehouyu* is the hardest. A more detailed analysis of the low-familiarity and novel items reveals **varying levels of difficulty across different types of *xiehouyu***. Notably, participants performed best on unfamiliar "polysemous" (83.3% accuracy) and "other" types (81.8%), suggesting that the semantic or logical links in these categories are more transparent or easier to reason about than phonetic ones.

These results establish a robust benchmark for human-level reasoning on this task, highlighting that resolving polysemous puns is more intuitive for humans than resolving those based on homophony.

4.3 LLMs' accuracy on X-Riddles

4.3.1 Overall accuracy

The overall accuracy for all models in the 0-shot setting is presented in Figure 2. We first observe scaling effects for the Qwen2.5 series, demonstrated by the consistent increase of model accuracy as model size increases from 0.5B to 72B. While Qwen2.5-0.5B performs at near chance level (25%), the larger Qwen models demonstrate high proficiency, especially in *polysemous* and *other* types of *xiehouyu* of high-familiarity to humans. Frontier models (DeepSeek-V3.2, Qwen3.5-Plus, Kimi-2.5) have almost perfect accuracy for polysemous and other *xiehouyu*, as well as homophonic *xiehouyu* with high familiarity/frequency.

4.3.2 Effects of evaluation conditions.

Figure 4 shows the effects of other conditions in the accuracy of experiment 1. Chain-of-Thought prompting (blue bars) hurts the performance for

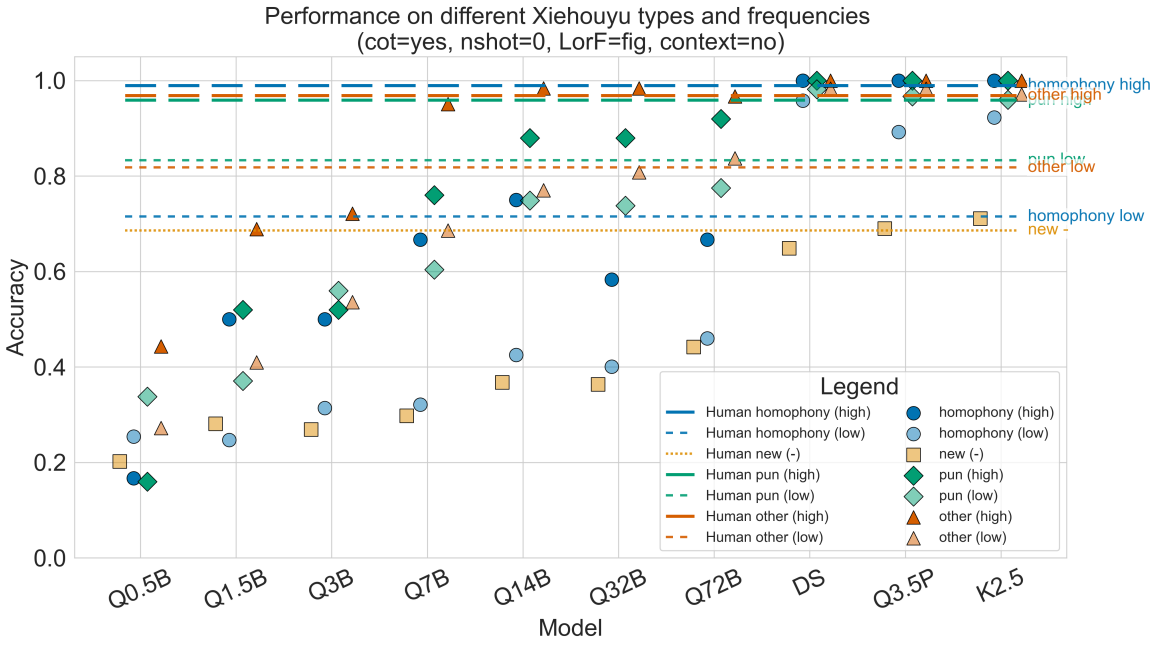

Figure 2: Results for experiment 1: accuracy of **0-shot** multiple-choice question for *xiehouyu* for LLMs and humans. Different shapes or colors refer to different types of *xiehouyu*. The lines represent human accuracy.

Qwen2.5-1.5B and Qwen2.5-3B, and has no noticable influence on other models. For few-shot prompting (orange bars), increasing from 0-shot to 3-shot increases the performance for Qwen2.5 models for at most 4 percentage points, but almost no effect for large MoE models such as Qwen3.5-Plus and DeepSeek-V3.2. Providing context (green bars) or using the literal answer (red bars) are much more helpful, with up to 10% improvement in performance, but these are not the settings in our human experiment, and thus not reported in later sections of the paper. Based on these piloting results, we use CoT=yes, n-shot=0, no context and figurative answer as the evaluation condition for all frontier models next.

4.3.3 Reasoning or Memorization?

Our most important analysis lies in whether the models are performing true reasoning or have simply memorized the *xiehouyu* from their training data. We use the difference of performance on low-familiarity homophonic *xiehouyu* and the new ones (Δ between Low and New) as the measure, shown in the last column of Table 4.

Chinese models seem to memorize much more than English-centric models, evidenced by large Δ_{acc}. For humans, Δ is very low (2.9), suggesting that when processing *xiehouyu* of low-familiarity and new *xiehouyu*, humans rely on the same mechanism, that is genuine reasoning. Frontier Chinese models stood out by having (extremely) large Δ, 23.6 on average. DeepSeek-V3.2, for instance, has a Δ of 30.9, with almost perfect accuracy on low-familiarity ones (95.8%). This result suggests that Chinese models may have been trained on much more Chinese data, which contains existing but very rare *xiehouyu*. However, English-centric models have rather low or even negative Δ, only 5.1 on average (below 10 for all, and even -2.8 for Gemini 3.1 Pro), suggesting that they were probably not trained on massive Chinese data that may contain rare *xiehouyu*. This is understandable since the Chinese labs have likely put more effort in obtaining large-scale Chinese data to train their models.

Token usage more than doubled for new *xiehouyu* in Chinese models. Table 5 presents the increase in token usage (response + thinking) for each model under from High to Low to New, for homophonic *xiehouyu*. We again observe a difference among frontier Chinese vs. English-centric models. There is a much larger token increase from High to Low in Chinese models compared to English-centric ones (91.8% vs. 36.2%) and from High to Low (166.3% vs. 38.8%). Notice again that for Chinese models there is a 40% increase from Low to New, but the increase is only 1.8% for English-centric models, suggesting that

Model	High	Low	New	Δ
Human Baseline				
Human	99.0	71.5	68.6	2.9
Qwen2.5 (Open-Weights)				
Qwen2.5-0.5B	16.7	25.7	20.2	5.5
Qwen2.5-1.5B	50.0	24.6	28.1	−3.5
Qwen2.5-3B	50.0	31.6	26.9	4.7
Qwen2.5-7B	66.7	32.3	29.8	2.5
Qwen2.5-14B	75.0	42.7	36.8	5.9
Qwen2.5-32B	58.3	40.3	36.4	3.9
Qwen2.5-72B	66.7	46.2	44.2	2.0
Mean (Qwen2.5)	*54.8*	*34.8*	*31.8*	*3.0*
Frontier Chinese Models				
DeepSeek-V3.2	100.0	**95.8**	64.9	30.9
Doubao-Seed2.0	100.0	95.5	81.8	13.7
GLM-5	100.0	86.8	70.2	16.6
Kimi-k2.5	100.0	92.4	71.1	21.3
MiniMax-m2.5	100.0	83.7	45.0	38.6
Qwen3.5-Plus	100.0	89.2	69.0	20.2
Mean (Chinese)	*100.0*	*90.6*	*67.0*	*23.6*
Frontier English-centric Models				
Claude 4.6 Opus	100.0	85.1	83.9	1.2
Claude 4.6 Sonnet	91.7	70.5	67.8	2.7
Gemini 3.1 Pro	100.0	92.4	**92.6**	−0.2
Gemini 3 Flash	100.0	87.8	77.7	10.2
GPT-5.2	100.0	80.2	72.3	7.9
Grok-4†	100.0	80.2	71.1	9.1
Mean (Eng.-centric)	*98.6*	*82.7*	*77.5*	*5.1*

Table 4: Accuracy (%) on homophonic *xiehouyu* across High Familiarity, Low Familiarity, and New (uncontaminated) splits. Δ_{acc} represents the performance drop from Low to New. **Frontier Chinese models overall show much more memorization (Δ_{acc} as high as 38.6), compared to English-centric Models (Δ_{acc} less than 10), which may be attributed to the potentially much more Chinese training data.** Furthermore, while the best performing models for New *xiehouyu* are Close-Weights, some Open-Weights models (e.g., Kimi-2.5) reach very high accuracy.

		Total Token Δ%		
Model	Δ_{Acc}	L/H	N/H	N/L
Frontier Chinese Models				
DeepSeek-V3.2	30.9	+71.6	+220.6	+86.8
Doubao-Seed2.0	13.7	+61.1	+117.2	+34.8
GLM-5	16.6	+98.4	+148.6	+25.3
Kimi-k2.5	21.3	+98.6	+197.3	+49.7
MiniMax-m2.5	38.6	+149.4	+190.2	+16.3
Qwen3.5-Plus	20.2	+71.5	+123.7	+30.4
Mean (Chinese)	*23.6*	*+91.8*	*+166.3*	*+40.6*
Frontier English-centric Models				
Claude 4.6 Opus	1.2	+67.3	+76.0	+5.2
Claude 4.6 Sonnet	2.7	+49.1	+51.8	+1.8
Gemini 3.1 Pro	−0.2	+18.4	+13.2	−4.4
Gemini 3 Flash	10.2	+21.8	+30.5	+7.2
GPT-5.2	7.9	+32.7	+33.9	+0.9
Grok-4†	9.1	+27.8	+27.7	−0.1
Mean (Eng.-centric)	*5.2*	*+36.2*	*+38.8*	*+1.8*

Table 5: Token effort analysis across item types for frontier models. Δ_{Acc} = Acc(Low) − Acc(New) in percentage points. Token Δ% columns show percentage change in total tokens (response + thinking) relative to the reference split. For instance, L/H shows increase in token count from Low to High. † No extended thinking; total tokens = response tokens only.

the Low and New *xiehouyu* are treated differently by Chinese models, but not English-centric models. These results reflect that the Chinese models are trying very hard to reason about the Low and New *xiehouyu*. However, their performance on the New did not increase as their thinking effort increases.

Overthinking does not lead to improved performance. Chinese frontier models show a striking dissociation between effort and performance on New items: they produce on average 40.6% more tokens compared to Low items, yet their accuracy drops by 23.6 percentage points. This pattern is consistent with the "overthinking" phenomenon documented in Chen et al. (2024); Sui et al. (2025), where extended reasoning chains fail to improve and may even degrade performance. In contrast, English-centric models show minimal token increase (1.8%) alongside minimal accuracy change, suggesting a more uniform processing strategy across Low and New items. Notably, the best-performing model on New items (Gemini 3.1 Pro) actually shows a slight decrease in tokens from Low to New (-4.4%), consistent with findings that more capable models allocate computational resources more efficiently (Lin et al., 2025).

4.3.4 Discussion

The key interpretive tool in Experiment 1 is Δ_{acc}, the accuracy gap between Low-familiarity and New items. Because the human baseline shows a near-zero Δ_{acc}, any large gap observed in a model is difficult to attribute to differences in item difficulty and instead points to memorization of existing *xiehouyu* from training data. This diagnostic reveals a clear split: Chinese-origin models exhibit substantial Δ_{acc} values, while English-centric models do not—though we note that the two groups also differ in openness and scale, making it difficult to fully isolate the role of Chinese training data volume.

Gemini 3.1 Pro is particularly noteworthy: it surpasses human accuracy on New items while main-

taining a near-zero Δ_{acc}, suggesting that genuine multi-step reasoning—from riddle to literal answer to phonetic link—is within reach of frontier models. More generally, Δ_{acc} provides only a *lower bound* on memorization: a model that has both memorized existing items and learned to reason will still appear reasoning-driven by this metric.

These observations carry implications beyond *xiehouyu*. Any benchmark built from culturally circulated material risks conflating retrieval with reasoning, and the effect is amplified for languages well represented in training corpora. Our design—pairing existing items with newly created ones of matched difficulty—offers a general template for disentangling the two, which we believe deserves wider adoption. In Experiment 2, we move beyond answer selection to examine whether models can *explain* the reasoning steps underlying *xiehouyu*.

5 Experiment 2: *xiehouyu* explanation

5.1 Experiment Setup

Task In this experiment, the models are asked to explain why the intended meaning of a *riddle* is the *answer-intended*. We directly prompt the model to provide an explanation.

Our aim is to see whether the models truly "understand" the underlying logic of *xiehouyu* when matching *riddle* and *answer*. To do so, we sampled a total of 432 (144 $\times$ 3) model explanations of *xiehouyu* and analyzed them closely to see if the responses show real understanding. The three models are: DeepSeek-R1, Qwen-max and Kimi-k2.[2] The prompts can be found in Appendix C.

Type	N	Type	N
homophony-high	12	homophony-low	12
pun-high	20	pun-low	20
other-high	20	other-low	20
new	40		

Table 6: A total of 144 *xiehouyu* sampled to obtain model explanation: N refers to numbers in each category.

Evaluation We performed two rounds of analyses. In the first round, we asked 56 native speakers of Chinese to rate LLMs' explanations on three aspects: (1) whether the explanation mentions the homophonic word (Yes/No/NA because there is no homophonic word), (2) whether (they think) the LLMs have truly understood the *xiehouyu* (Yes/No/Uncertain) and (3) the quality of the explanations on a Likert-scale of 1 to 5. In the second round, the authors selected all explanations with an average quality score no higher than 2 (about 30 explanations) and sampled another 30 explanations with a quality score of 5, and manually went through them to identify the sensible and nonsensical segments.

[2]These models were tested in July 2025.

Model	Quality	Acc	Understd.	Res. Len.
DS-R1	4.11±0.93	0.78	0.82	802.41
Q-Max	3.92±1.21	0.64	0.73	897.91
Kimi-k2	3.83±1.17	0.70	0.71	517.71

Table 7: Ratings of model explanations on *xiehouyu*. *Acc* refers to the accuracy of identifying the homophonic word. *res len* refers to the length of response.

5.2 Results

Performance of Models The ratings of model explanations on *xiehouyu* by native speakers are as shown in Table 7. We observe that DeepSeek-R1 achieves the highest quality and understanding-ratio, while Qwen-max ranks second, producing slightly longer responses. Kimi-k2 demonstrates inferior performance, with responses markedly shorter than those of the other two models. An Ordinary Least Squares (OLS) regression, `quality_score` $\sim$ `res_len`, shows no statistically significant association between response length and quality score, suggesting that the observed differences are not merely driven by verbosity. In addition, although the models can choose the correct answer for the majority of homophonic *xiehouyu*, as shown in the results of Experiment 1 in Table 4, their accuracy in identifying the homophonic word tend to be worse as shown in Table 8, especially for novel *xiehouyu*, indicating a lack of phonological knowledge.

Model \ Familiar.	New	Low	High
DS-R1	85.0	83.4	100.0
Q-max	67.5	45.4	100.0
Kimi-k2	67.5	100.0	91.7

Table 8: Ratings on whether the models accurately in identifying the homophonic word (if any) with respect to familiarity.

The Effect of Different *xiehouyu* Types The distributions of the quality scores of the three mod-

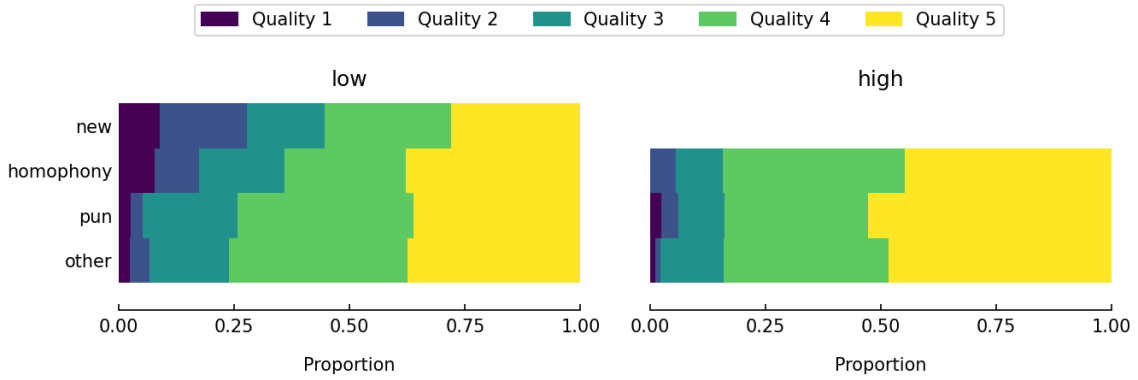

Figure 3: Quality distributions across models and *xiehouyu* types and familiarity.

els across *xiehouyu* types and familiarity are illustrated in Figure 3. The majority of explanations for high-familiarity *xiehouyu* receive a quality above 4.0 (yellow and light green), while those for low-familiarity ones are considerably worse (although still acceptable), which holds for all three models. As for novel *xiehouyu*, the explanation quality shows a clear degradation, especially for Kimi-k2 and Qwen-max, in which only ∼50% of their explanations are rated 4.0 or higher.

Taken together, the results suggest that for highly familiar *xiehouyu*, the best models on the market in general provide sensible explanations. However, for newly written *xiehouyu*, only half of their responses are rated as high quality.

Models' weaknesses in understanding *xiehouyu*. In the second round of analysis, we performed manual inspection of model explanations and identified the following types of major errors.

First, models failed to identify the homophonic character or word, or they find the wrong one. A careful examination shows that the former is the majority. This usually occurs when the homophone is not a strict one, such as "寺" (*si*) and "事" (*shi*), which sound similar in some varieties of Mandarin, and are treated as homophone in many *xiehouyu*, but are nonetheless not exactly the same in Standard Mandarin.[3] The latter arises when the model seems to be trying too hard to find the homophonic word. When no clear homophonic word is identified, the model will conjure up the homophone and usually claim, incorrectly, that in some Mandarin dialect or colloquial speech such a homophone exists, as shown in Table 16 in the Appendix. Such behavior has been reported previously where models hallucinate or produce wrong reasoning (Turpin et al., 2023).

Two, models are prone to over-interpretation and even hallucination. Table 17 in the Appendix shows an example of over-interpretation, which happened most often because models failed to identify the homophonic words.

6 Experiment 3: *xiehouyu* creating

6.1 Experimental setup

Task In this experiment, the models were asked to create new *xiehouyu*. Since previous sections show that homophonic *xiehouyu* are most difficult and we already have human-created homophonic *xiehouyu*, we focus on homophonic ones in this experiment. In our pilot experiment, we found it very difficult to elicit models to create high quality *xiehouyu* without giving example *xiehouyu* and the logic behind them. Therefore, in the prompt we provide an existing $xiehouyu_1$, its detailed reasoning steps, *and* a human-created $xiehouyu_2$, which uses $xiehouyu_1$ as the seed. The models were not forced to follow the seed strictly by using the same homophones, but were prompted to create a "similar" *xiehouyu*. The complete prompt is shown in Appendix C.

Twenty seed *xiehouyu* were used and each model was prompted three times for one *xiehouyu*. Therefore, each model generated 60 new *xiehouyu*.

Evaluation Three authors evaluated the *xiehouyu* created by models in two dimensions: reasonableness and funniness. For reasonableness, they were asked to judge whether a new *xiehouyu* is reasonable (1) or not (0), following same metrics for human *xiehouyu* validation (§ 3). If deemed reason-

[3]The full IPA for two characters in Standard Mandarin: /sɻ̩/ and / sɻ̩/. However, in many southern varieties of Mandarin, /s/ (alveolar fricative) and / s/ (retroflex fricative) are not distinguished. The vowels are similar too. Note that the Pinyin for the vowels are the same, but they are different vowels.

Author	Riddle	Answer-lit	Answer-intended	Reasonable	Funny	Overlap
GPT-o1	在山洞里放爆竹 Setting off firecrackers in a cave	响得深 Resound deeply	想得深 Think deeply	1	3.00	0.27
GPT-o1	水壶烧开冒白烟 Boiling kettle emitting white smoke	蒸汽 Steam	争气 Strive for success	1	2.67	0.29
DS-R1	膝盖上钉掌 Horseshoe on knee	离蹄 Away from hoof	离题 Off topic	1	2.67	0.33
DS-R1	黄豆配绿豆 Soy bean with mung bean	双豆 Double bean	双斗 Double fight	0	1.00	0.27
Human	和尚流浪 Wandering monk	没寺找寺 No temple, seek temple	没事找事 Ask for trouble	1	4.00	0.38
Human	猪八戒的学生 Zhubajie's student	悟能之徒 Disciple of Wuneng	无能之徒 Incompetent person	1	3.67	0.47

Table 9: Examples of model- and human-created *xiehouyu*. Blue text highlights homophonic puns, where words share the same (or similar) sounds but carry different meanings. In general, humans tend to follow traditional patterns that rely on folklore and vulgar humor, whereas LLMs leverage physical knowledge to construct novel *xiehouyu*.

able, the *xiehouyu* was then rated on a scale of 1 to 3. For funniness, they were asked to give a score ranges from 1 (not funny at all) to 4 (incredibly funny). We also compute a novelty score based on Jaccard index to see whether the *xiehouyu* has significant overlap with existing ones from the dictionary, to make sure that the models have created genuine novel ones (see Appendix E).

Models We performed a pilot study with several models, including GPT4o, Qwen2.5-72B, o1, and DeepSeek-R1. We observed that only o1 and DeepSeek-R1 produced valid *xiehouyu*. Therefore, we conducted evaluation on these two models.

6.2 Results

Src.	N	Reasonable	Funniness	Max Ovrlp
Human	312	0.75 (0.34)	2.26 (0.74)	0.32
DS-R1	60	0.43 (0.40)	1.47 (0.54)	0.30
o1	60	0.47 (0.37)	1.62 (0.60)	0.29

Table 10: Performance comparison of human and model-generated *xiehouyu*. N refers to numbers of *xiehouyu*. Scores are reported as Mean (Standard Deviation).

Human and Model Performance Table 10 presents the ratings for *xiehouyu* generated by humans, DeepSeek-R1 and o1. Human-generated *xiehouyu* were rated higher than both models in reasonableness and funniness, though LLMs exhibited marginally lower overlap with existing examples. Comparing the two models, o1 performed slightly better than DeepSeek-R1. These results suggest that while LLMs can produce somewhat reasonable and funny *xiehouyu*, they have not yet surpassed original human creativity.

Strengths and Weaknesses of Models in Generating *xiehouyu* While LLMs show promise, they currently lag behind humans in crafting reasonable and humorous *xiehouyu*. The primary bottleneck lies in context integration. Models frequently bypass logical reasoning to fabricate connections between the riddle and the answer. This tendency to over-explain or hallucinate relationships aligns with our findings in Section 5.2, where models failed to grasp homophony and low-frequency examples due to the lack of phonetic information.

In terms of humor, human-written *xiehouyu* were more dynamic and versatile, while model-generated ones tended to be mediocre and safe. On the other hand, both DS-R1 and o1 showed slightly lower overlap scores than humans, although the difference was very small.

Example *xiehouyu* generated by LLMs. Table 9 demonstrates the riddles generated by LLMs and humans.

Although humans remain more versatile in humor, LLMs successfully leverage physical knowledge to construct novel *xiehouyu*. Unlike traditional *xiehouyu*, which typically rely on folklore or vulgar humor, LLMs demonstrate a capacity for broader creativity by incorporating concepts from modern physics and general knowledge. However, relying on such common knowledge may result in less humor, as the core of *xiehouyu* often depends

on specific cultural contexts. Nevertheless, these model-generated examples excel in their novelty and uniqueness.

In the generated example in Table 9 “在山洞里放爆竹——想得深” (Setting off firecrackers in a cave—Think deeply), the figurative answer “想” (*xiǎng*, to think) replaces the homophone “响” (*xiǎng*, resound). This connects the deep sound of the explosion to deep thoughts. Similarly, in “水壶烧开冒白烟——争气” (Boiling kettle emitting white smoke—Strive for success), the answer “争气” (*zhēngqì*, to strive for success) is homophonous with “蒸汽” (*zhēngqì*, steam). It cleverly links a physical phenomenon to a psychological state.

LLMs also make mistakes. For instance, the second example by DeepSeek-R1 in Table 9, the answer-intended “双斗” (double fight) is not a legitimate word in Chinese. This indicates that LLMs sometimes make up words when forced to be creative.

7 Related Work

Our investigation into the capabilities of Large Language Models (LLMs) on Chinese *xiehouyu* connects three critical strands of research: the tension between reasoning and memorization in model performance, the processing of linguistic humor and metaphor, and the handling of phonological ambiguity.

7.1 Reasoning versus Memorization and Data Contamination

A central question in current NLP research is whether LLMs solve complex tasks through genuine reasoning or by retrieving memorized patterns from their vast training corpora. Biderman et al. (2023) demonstrated that as models scale, they exhibit emergent memorization capabilities, often reciting long tail data verbatim. More recently, Ahmed et al. (2026) showed that production models have memorized substantial portions of copyrighted books, confirming the saturation of literary data in pre-training corpora.

This memorization hypothesis strongly aligns with our findings in Experiments 1 and 2, where some LLMs achieved near-perfect accuracy on existing *xiehouyu* but failed significantly on novel ones. Similar phenomena have been observed in mathematical reasoning. Wu et al. (2025) revealed that the performance of models like Qwen2.5 on standard benchmarks (e.g., MATH-500) is heavily inflated by data contamination; when tested on a leakage-free dataset (RandomCalculation), performance dropped significantly, and spurious rewards failed to improve reasoning. Similarly, Wang et al. (2025) hypothesized that LLMs utilize two competing mechanisms—Chain-of-Thought reasoning and memory retrieval. They found that models often "hack" tasks by prioritizing retrieval over reasoning, a behavior that explains why the LLMs in our study likely retrieved the answers for dictionary-based *xiehouyu* rather than deriving them.

7.2 Humor, Metaphor, and Creative Generation

xiehouyu represents a unique intersection of humor and figurative language. Prior work on Chinese humor generation, such as the study on *Xiangsheng* (Crosstalk) by Li et al. (2023), found that while large-scale pre-training improves generation quality, models still struggle to produce humor that matches human standards. This parallels our Experiment 3, where human evaluators consistently rated LLM-generated *xiehouyu* as inferior to human creations.

However, our results regarding *understanding* novel *xiehouyu* present an interesting contrast to Ichien et al. (2024), who reported that GPT-4 displays an emergent ability to interpret *novel* literary metaphors, often outperforming human students. While they suggest LLMs can generalize to new figurative language, our study indicates this ability may be limited when the figurative mapping requires specific cultural schemas or rigid structural constraints, as seen in *xiehouyu*. The sharp 40% accuracy drop we observed between existing and novel samples suggests that what appears to be "emergent interpretation" in other contexts might partly rely on soft-matching against memorized semantic relations, which fails when the riddle component is entirely unseen.

7.3 Phonological Ambiguity and Puns

Many *xiehouyu* rely on homophonic puns (*xieyin*), adding a layer of phonological complexity. Xu et al. (2024) evaluated LLMs on English puns, identifying a "lazy generation" pattern where models fail to recognize or generate the duality of meaning required for a good pun. In the context of Chinese, Ma et al. (2025) introduced *PhonoThink*, demonstrating that standard LLMs frequently struggle with Chinese phonological ambiguities without specific multi-stage training and reinforce-

ment learning. The difficulty LLMs face in our study—particularly with novel *xiehouyu* that may rely on phonological cues—is consistent with these findings, suggesting that current architectures still lack robust grounding in the interplay between phonology and semantics required for deep linguistic reasoning.

8 Conclusion

In this study, we explore the limits of LLM reasoning in a Chinese language game called *xiehouyu*. We first created X-Riddles, the first benchmark on Chinese *xiehouyu*, with 900 existing *xiehouyu* sampled from a dictionary, and more than 200 expert-written novel *xiehouyu* to avoid data contamination. We then investigated the ability to *understand* and *create xiehouyu* in state-of-the-art LLMs with three experiments. We used the difference in accuracy (Δ_{acc}) between low-frequency *xiehouyu* and novel ones as a proxy for model memorization. Our results show that humans have a very low Δ_{acc} (2.9%), suggesting genuine reasoning for both types of *xiehouyu*, whereas frontier Chinese models show a large mean Δ_{acc} of 23%, indicating they may have memorized a lot of the low-frequency *xiehouyu* in training. The best frontier model (Gemini 3.1 Pro) reaches 92.6% accuracy on the novel *xiehouyu* under zero-shot conditions, demonstrating genuinely strong reasoning abilities for this Chinese language game. However, *xiehouyu* created by strong LLMs (DS-R1 and GPT-o1) are rated to be less reasonable and funny compared to those created by humans. These findings have several implications. First, future studies need to take into consideration the data leakage issue when evaluating LLMs. Second, LLMs seem to lack phonological knowledge of Chinese characters, as they perform worse on homophonic *xiehouyu* than polysemous ones. Finally, humans still seem to be better at creative language use when it comes to Chinese language games such as *xiehouyu*.

Limitations

As *xiehouyu* is unique in the Chinese language, and all the tested models are proficient in Chinese, our results mostly reflect their abilities in understanding this particular type Chinese riddle. Other languages may have other types of word games or riddles that need to be evaluated to estimate LLMs' ability more comprehensively.

Ethics Statement

Participants in any human validation experiments in the study were compensated with roughly 1 RMB for 1 minute, which is above the local minimum wage standard (i.e., 15 RMB for approximately 15 minutes of their time).

Acknowledgments

The project is funded by National Social Science Fund of China, awarded to Hai Hu (25BYY133).

References

Tosin Adewumi, Roshanak Vadoodi, Aparajita Tripathy, Konstantina Nikolaido, Foteini Liwicki, and Marcus Liwicki. 2022. Potential idiomatic expression (PIE)-English: Corpus for classes of idioms. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 689–696.

Ahmed Ahmed, A Feder Cooper, Sanmi Koyejo, and Percy Liang. 2026. Extracting books from production language models. *arXiv preprint arXiv:2601.02671*.

Stella Biderman, Usvsn Prashanth, Lintang Sutawika, Hailey Schoelkopf, Quentin Anthony, Shivanshu Purohit, and Edward Raff. 2023. Emergent and predictable memorization in large language models. *Advances in Neural Information Processing Systems*, 36:28072–28090.

Vladislav Blinov, Valeria Bolotova-Baranova, and Pavel Braslavski. 2019. Large dataset and language model fun-tuning for humor recognition. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 4027–4032.

Xingyu Chen, Jiahao Xu, Tian Liang, Zhiwei He, Jianhui Pang, Dian Yu, Linfeng Song, Qiuzhi Liu, Mengfei Zhou, Zhuosheng Zhang, and 1 others. 2024. Do not think that much for 2+ 3=? on the overthinking of o1-like llms. *arXiv preprint arXiv:2412.21187*.

DeepSeek-AI. 2025. Deepseek-r1: Incentivizing reasoning capability in LLMs via reinforcement learning. *Preprint*, arXiv:2501.12948.

Sayan Ghosh and Shashank Srivastava. 2022. epic: Employing proverbs in context as a benchmark for abstract language understanding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3989–4004.

Raymond W Gibbs Jr. 1984. Literal meaning and psychological theory. *Cognitive science*, 8(3):275–304.

Raymond W Gibbs Jr. 2002. A new look at literal meaning in understanding what is said and implicated. *Journal of pragmatics*, 34(4):457–486.

Rachel Giora. 1997. Understanding figurative and literal language: The graded salience hypothesis.

Rachel Giora. 1999. On the priority of salient meanings: Studies of literal and figurative language. *Journal of pragmatics*, 31(7):919–929.

HP Grice. 1975. Logic and conversation. *Syntax and semantics*, 3.

Jack Hessel, Ana Marasović, Jena D Hwang, Lillian Lee, Jeff Da, Rowan Zellers, Robert Mankoff, and Yejin Choi. 2023. Do androids laugh at electric sheep? humor "understanding" benchmarks from the New Yorker caption contest. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 688–714.

Jennifer Hu, Sammy Floyd, Olessia Jouravlev, Evelina Fedorenko, and Edward Gibson. 2022. A fine-grained comparison of pragmatic language understanding in humans and language models. *arXiv preprint arXiv:2212.06801*.

Nicholas Ichien, Dušan Stamenković, and Keith J. Holyoak. 2024. Large language model displays emergent ability to interpret novel literary metaphors. *Metaphor and Symbol*, 39(4):296–309.

Sophie Jentzsch and Kristian Kersting. 2023. ChatGPT is fun, but it is not funny! humor is still challenging large language models. *arXiv preprint arXiv:2306.04563*.

Huei-ling Lai. 2008. Understanding and classifying two-part allegorical sayings: Metonymy, metaphor, and cultural constraints. *Journal of Pragmatics*, 40(3):454–474.

Jianquan Li, XiangBo Wu, Xiaokang Liu, Qianqian Xie, Prayag Tiwari, and Benyou Wang. 2023. Can language models make fun? a case study in Chinese comical crosstalk. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7581–7596, Toronto, Canada. Association for Computational Linguistics.

Junhong Lin, Xinyue Zeng, Jie Zhu, Song Wang, Julian Shun, Jun Wu, and Dawei Zhou. 2025. Plan and budget: Effective and efficient test-time scaling on large language model reasoning. *arXiv preprint arXiv:2505.16122*.

Chen Liu, Fajri Koto, Timothy Baldwin, and Iryna Gurevych. 2024. Are multilingual LLMs culturally-diverse reasoners? an investigation into multicultural proverbs and sayings. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2016–2039.

Jianfei Ma, Zhaoxin Feng, Emmanuele Chersoni, Huacheng Song, and Ziqi Zhang. 2025. PhonoThink: Improving large language models' reasoning on Chinese phonological ambiguities. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 19007–19022, Suzhou, China. Association for Computational Linguistics.

Lijun Ma, Yunxiao Ma, Xiaoqing He, Haitao Liu, and Jingyu Zhang. 2019. Processing of Chinese homophonic two-part allegoric sayings: Effects of familiarity and homophone. *Acta Psychologica Sinica*, 51(12):1306.

Seyed Iman Mirzadeh, Keivan Alizadeh, Hooman Shahrokhi, Oncel Tuzel, Samy Bengio, and Mehrdad Farajtabar. 2025. GSM-Symbolic: Understanding the limitations of mathematical reasoning in large language models. In *The Thirteenth International Conference on Learning Representations*.

Arthur Neidlein, Philip Wiesenbach, and Katja Markert. 2020. An analysis of language models for metaphor recognition. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3722–3736.

Wei Qu. 2010. *The mechanism of Chinese Xiehouyu comprehension*. Doctoral dissertation, Hunan Normal University.

John R Searle. 1978. Literal meaning. *Erkenntnis*, pages 207–224.

Dingfang Shu. 2015. Chinese xiehouyu (歇后语) and the interpretation of metaphor and metonymy. *Journal of Pragmatics*, 86:74–79.

Yang Sui, Yu-Neng Chuang, Guanchu Wang, Jiamu Zhang, Tianyi Zhang, Jiayi Yuan, Hongyi Liu, Andrew Wen, Shaochen Zhong, Na Zou, and 1 others. 2025. Stop overthinking: A survey on efficient reasoning for large language models. *arXiv preprint arXiv:2503.16419*.

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

Simone Tedeschi, Federico Martelli, and Roberto Navigli. 2022. Id10m: Idiom identification in 10 languages. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2715–2726.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel Bowman. 2023. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *Advances in Neural Information Processing Systems*, 36:74952–74965.

Lennart Wachowiak and Dagmar Gromann. 2023. Does gpt-3 grasp metaphors? identifying metaphor mappings with generative language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1018–1032.

Xiaolu Wang, Yizhen Wang, Wanning Tian, Wei Zheng, and Xiaoli Chen. 2021. The roles of familiarity and context in processing Chinese xiehouyu: An erp study. *Journal of Psycholinguistic Research*, pages 1–21.

Yuhui Wang, Changjiang Li, Guangke Chen, Jiacheng Liang, and Ting Wang. 2025. Reasoning or retrieval? a study of answer attribution on large reasoning models. *Preprint*, arXiv:2509.24156.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Duanzheng Wen. 2003. *Zhongguo xiehouyu da cidian [A Comprehensive Dictionary of Chinese Xiehouyu]*. Shanghai Lexicographical Publishing House, Shanghai.

Mingqi Wu, Zhihao Zhang, Qiaole Dong, Zhiheng Xi, Jun Zhao, Senjie Jin, Xiaoran Fan, Yuhao Zhou, Huijie Lv, Ming Zhang, and 1 others. 2025. Reasoning or memorization? unreliable results of reinforcement learning due to data contamination. *arXiv preprint arXiv:2507.10532*.

Zhijun Xu, Siyu Yuan, Lingjie Chen, and Deqing Yang. 2024. "a good pun is its own reword": Can large language models understand puns? *arXiv preprint arXiv:2404.13599*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, and 22 others. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Zhiwei Yu, Jiwei Tan, and Xiaojun Wan. 2018. A neural approach to pun generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1650–1660.

Shisen Yue, Siyuan Song, Xinyuan Cheng, and Hai Hu. 2024. Do large language models understand conversational implicature–a case study with a Chinese sitcom. *arXiv preprint arXiv:2404.19509*.

Hui Zhang, Long Jiang, Jiexin Gu, and Yiming Yang. 2013. Electrophysiological insights into the processing of figurative two-part allegorical sayings. *Journal of Neurolinguistics*, 26(4):421–439.

Chujie Zheng, Minlie Huang, and Aixin Sun. 2019. Chid: A large-scale Chinese IDiom dataset for cloze test. *arXiv preprint arXiv:1906.01265*.

A Effects of experimental conditions on accuracy in Experiment 1

Effects of different conditions are presented in Figure 4.

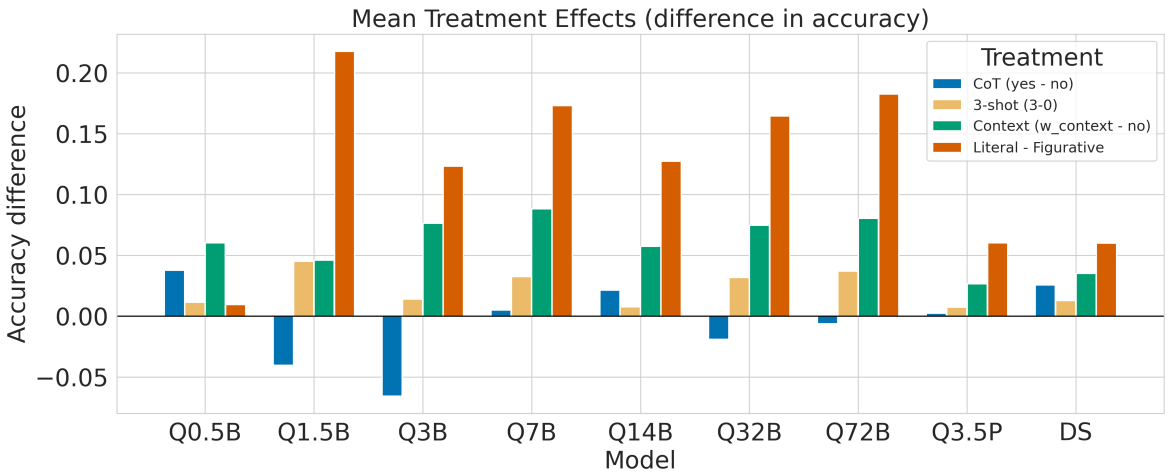

Figure 4: Effects of different conditions on the overall accuracy in the multiple choice questions in Experiment 1.

B Examples of human-created *xiehouyu*

	Seed	**Human-Created**	**Human-Created**	**Human-Created**
Riddle	外甥打灯笼	外甥修城墙	外婆老了	外甥借钱
Translation	Nephew holds lantern	Nephew builds wall	Grandma is getting old	Nephew borrows money
Ans-lit	照舅	守舅	依舅	欠舅
Pinyin	*zhào jiù*	*shǒu jiù*	*yī jiù*	*qiàn jiù*
Gloss	light uncle	defend uncle	depend on uncle	owe uncle
Translation	light the way for uncle	protect uncle	rely on uncle	owe uncle money
Ans-fig	照旧	守旧	依旧	歉疚
Pinyin	*zhào jiù*	*shǒu jiù*	*yī jiù*	*qiàn jiù*
Gloss	according to old	stick to old	still	guilty
Meaning	same old, as usual	conservative	as always	guilty

Table 11: Example of three human-created *xiehouyu* based on a seed *xiehouyu*, showing the *Riddle*, **Answer-literal** and **Answer-intended**, with glosses and translations.

	Human-Created	**Human-Created**	**Human-Created**	**Human-Created**
Riddle	缝纫机罢工	厕所里挥电蚊拍	吃饱了的猫	孙悟空穿破洞裤
Translation	The sewing machine is down	Swinging an electric mosquito swatter in toilet	A cat with a full belly	Sun Wukong wearing ripped jeans
Ans-lit	不裁	蝇麻了	乐不思鼠	猴补
Pinyin	*bù cái*	*yíng má le*	*lè bù sī shǔ*	*hóu bǔ*
Gloss	no cutting	numb fly	too happy to care about the mouse	monkey patch
Translation	seamless	the fly is numb	too happy to miss the mouse	monkey mending clothes
Ans-fig	不才	赢麻了	乐不思蜀	候补
Pinyin	*bù cái*	*yíng má le*	*lè bù sī shǔ*	*hòu bǔ*
Gloss	incompetent	endless wins	too happy to think of Sichuan	wait list
Meaning	humble self	win big	indulge in pleasure and forget home and duty	wait list

Table 12: Examples of human-created new *xiehouyu* that are *not* based on existing seed *xiehouyu* (free style creation).

C Prompt Templates for three Experiments

Original Prompt in Chinese	English Translation
【说明】歇后语是由两部分组成的固定语句，前一部分多用比喻，像谜面，后一部分是本意，像谜底，通常只说前一部分，后一部分不言而喻。	[*Instruction*] A *Xiehouyu* (allegorical saying) is a fixed phrase composed of two parts. The first part often uses a metaphor, acting like a riddle; the second part is the actual meaning, acting like the answer. Usually, only the first part is spoken, as the second part is self-evident.
【任务】你的任务是根据给出例句的上下文，以及歇后语的前半部分，选出后半部分。	[*Task*] Your task is to select the second part of the *Xiehouyu* based on the context of the given sentence and the first part of the saying.
例子：天上灰布悬，地下水涟涟。山区的气候，好像孙猴子的脸——_____。	**Example:** Grey cloth hangs in the sky, water ripples on the ground. The climate in the mountains is like the Monkey King's face—_____.
在上面的句子里，"_____"处应该填什么?	What should be filled in the blank "_____" in the sentence above?
A: 碎拾掇　B: 说变就变	A: Tidying up scraps　B: Changes at the drop of a hat
C: 短中取长　D: 两头受气	C: Taking the long from the short　D: Caught in the middle
答案：B	**Answer:** B
题目：老乔的话对着哩，咱们都是一些摸牛尾巴的大老粗，啥也不懂，就敢揽那玩艺?再说今年这年成，困难比河里的石头子、砂子粒儿还多，再修水库，真成了房檐上玩把戏，_____了。	**Question:** Old Qiao is right. We are all uneducated roughnecks who only know how to touch a cow's tail. We don't understand anything, yet we dare to take on that thing? Besides, look at the harvest this year; the difficulties are more numerous than the stones and sand in the river. If we try to build a reservoir now, it would truly be performing acrobatics on the eaves—_____.
在上面的句子里，"_____"处应该填什么?	What should be filled in the blank "_____" in the sentence above?
A. 苦上加苦　B. 豁出去	A. Adding bitterness to bitterness　B. Risking it all
C. 不要命　D. 肚子里编成	C. Risking one's life / Reckless　D. Made up inside one's belly
【要求】请直接在"答案："后写出答案。	[*Requirement*] Please write the answer directly after "Answer:".

Table 13: The prompt used for Experiment 1 under the *1-shot*, *no CoT* condition. The model is provided with the definition, an example, and the specific question.

Original Prompt in Chinese	English Translation
请解释歇后语"孔夫子搬家"的后半部分为什么是"尽是输"。	Please explain why the answer to the riddle "Confucius move" is "always losing".

Table 14: The prompt for the *xiehouyu* explanation task in Experiment 2.

Original Prompt in Chinese	**English Translation**
定义：歇后语是由两部分组成的固定语句，前一部分多用比喻，像谜面，后一部分是本意，像谜底，通常只说前一部分，后一部分不言而喻。	**Definition:** A *Xiehouyu* (allegorical saying) is a fixed phrase composed of two parts. The first part often uses a metaphor, acting like a riddle; the second part is the actual meaning, acting like the answer. Usually, only the first part is spoken, as the second part is self-evident.
谐音歇后语：“谐音歇后语”是歇后语的一种，其谜底有字面义和引申义，包含一对谐音词。	**Homophonic Xiehouyu:** "Homophonic Xiehouyu" is a type of *Xiehouyu*. Its answer has both a literal meaning and a figurative (extended) meaning, containing a pair of homophones.
示例：	**Example:**
歇后语：咸菜拌豆腐——那还用言	Xiehouyu: Pickles mixed with tofu — Needless to say (salt)
谜面：咸菜拌豆腐	Riddle: Pickles mixed with tofu
字面义谜底：那还用盐	Literal Answer: Do you still need salt? (*Na hai yong yan*)
引申义谜底：那还用言	Figurative Answer: Needless to say (*Na hai yong yan*)
谐音字：盐/言	Homophone: Salt (*Yan*) / Say (*Yan*)
解释：盐：与“言”同音相谐。采用反问语气。指心里早已明白，用不着别人多说。	Explanation: "Salt" is homophonous with "Say". Uses a rhetorical tone. It implies the situation is already understood, and there is no need for others to say more.
例句：“咸菜拌豆腐——那还用言(盐)?”小练耸起鼻孔，表示母亲的嘱咐是多余的。	Example Sentence: "Pickles mixed with tofu — needless to say (salt)?" Xiao Lian shrugged, indicating that his mother’s instructions were superfluous.
推理过程：	**Reasoning Process:**
第一步通过谜面的语义信息（咸菜拌豆腐），联想到用咸菜本身用盐腌过，拌豆腐时，不必再放盐，进而得到字面义谜底（那还用盐）；	Step 1: Through the semantics of the riddle (pickles mixed with tofu), one associates that pickles are already salted, so when mixing with tofu, no extra salt is needed. This leads to the literal answer (Do you still need salt?);
第二步，通过字面义谜底（那还用盐），得到与之谐音的引申义谜底（那还用言），即“心里早已明白，用不着别人多说”。其中，第二步的推理用到了谐音，即例子中的“盐”和“言”。	Step 2: Through the literal answer, one derives the homophonous figurative answer (Needless to say). Step 2 utilizes the homophone, i.e., "Salt" and "Say".
任务：你的任务是根据“咸菜拌豆腐——那还用言”，写出一个与之相似的新造的“谐音歇后语”，并按照上述格式，写出“歇后语”、“谜面”、“字面义谜底”、“引申义谜底”、“解释”、“例句”。	**Task:** Your task is to write a *newly coined* "Homophonic Xiehouyu" similar to "Pickles mixed with tofu — Needless to say," based on that example. Follow the format above to write the "Xiehouyu," "Riddle," "Literal Answer," "Figurative Answer," "Explanation," and "Example Sentence."
新造示例（**1-shot**）：	**New Creation Example (1-shot):**
歇后语：咸菜汤里加酱油——察言观色	Xiehouyu: Adding soy sauce to pickle soup — Observing words and expressions
谜面：咸菜汤里加酱油	Riddle: Adding soy sauce to pickle soup
字面义谜底：察盐观色	Literal Answer: Checking salt and observing color (*Cha yan guan se*)
引申义谜底：察言观色	Figurative Answer: Observing words and expressions (*Cha yan guan se*)
谐音字：盐：言	Homophone: Salt (*Yan*) : Words (*Yan*)
解释：盐：与“言”同音相谐。指揣度对方的话语，观察对方的脸色。	Explanation: "Salt" is homophonous with "Words". Refers to speculating on someone’s words and observing their facial expressions.
例句：谈判时他不停试探对方底线，就像咸菜汤里加酱油——察言（盐）观色，生怕哪句话过了火候。	Example Sentence: During negotiation, he kept testing the other party’s bottom line, just like adding soy sauce to pickle soup — observing words (salt) and expressions, afraid that any sentence might cross the line.
注意：确保不能输出已经存在的歇后语。确保谐音准确，即声母韵母必须都相同。	**Note:** Ensure you do not output an existing Xiehouyu. Ensure the homophony is accurate (initials and finals must be identical).

Table 15: The prompt for the Homophonic *xiehouyu* generation task in Experiment 3. The model is instructed to generate a new *xiehouyu* by analogy, following the logic of the provided example.

D Example Errors from LLM output in Experiment 2

Model response (excerpt)	Claimed Homophones (Error)	Model Name
在汉语中，“胡”与“虚”在某些方言或语境下读音相近（尤其在口语或谐音游戏中），而“须”（胡子的“须”）与“假”组合成“虚假”。	胡 (hú) 虚 (xū)	DS-R1
在中文中，“不鸣”与“无名”发音相似（“不鸣”读作bù míng，“无名”读作wú míng），但通过歇后语的谐音技巧，“不鸣”被转化为“无名”，意思是“没有名声”或“不出名”。	不鸣 (bù míng) 无名 (wú míng)	DS-R1
在中文中，“听”（tīng）和“瘾”（yǐn）发音相近，尤其是在一些方言或口语中，容易产生谐音。	听 (tīng) 瘾 (yǐn)	DS-R1
而“书宅”与“输家”谐音，听起来像“输家”，	书宅 (shū zhái) 输家 (shū jiā)	Kimi-k2
“补”与“候”谐音（尤其在方言或口语中可能更接近）	补 (bǔ) 候 (hòu)	Kimi-k2
如某些方言中“弦”读若“xián”，与“功”的韵母“ong”形成谐音	弦 (xián) 功 (gōng)	Kimi-k2
“杏核儿”读作**xìng húr**... 但“杏核儿”在口语中快速连读时... 近似“人儿苦”（rénr kǔ）...	杏核儿 (xìng húr) 人儿苦 (rénr kǔ)	Qwen3-max

Table 16: Hallucinated Homophones in Model Explanations

Dictionary Explanation: 掸子：一种除尘工具，多用鸡毛绑在小竹棍上做成。掸：与“胆”同音相谐。常用于惊叹或指责人的胆子太大。 English translation for illustrative purposes: Duster (Dǎnzi): A type of tool used for dusting, typically made by tying chicken feathers to a small bamboo stick. The Pun: The character for "duster" (掸, dǎn) creates a homophonic pun with the character for "gallbladder/guts" (胆, dǎn). Meaning/Usage: This saying is commonly used to express astonishment at someone's audacity or to accuse someone of being overly bold or reckless (having "huge guts").
Model Explanation: “旗杆上插鸡毛——好大的胆子”是一句形象生动、富有幽默讽刺意味的歇后语。它的后半部分“好大的胆子”是对前半部分“旗杆上插鸡毛”这一荒诞行为的点评，其逻辑和寓意可以从以下几个方面理解：一、字面荒谬性制造反差。旗杆：通常是高大、庄严的象征，古代用于悬挂旗帜，代表国家、军队或权威，不可轻慢。鸡毛：轻飘无足轻重之物，常被用来象征琐碎、不值一提，甚至“鸡毛蒜皮”就是形容小事。把“鸡毛”插在“旗杆”上——本应挂旗帜的庄严位置，却插了个轻飘飘、毫无分量的东西，显得极其荒唐、不合规矩。这种行为本身就需要“胆大包天”才敢做，所以自然引申出“好大的胆子”。..... English translation for illustrative purposes: The xiehouyu “Sticking chicken feathers on a flagpole — what a huge guts!” is a vivid folk saying characterized by humor and satire. The second part, “what a huge guts,” comments on the absurd action described in the first part, “sticking chicken feathers on a flagpole.” Its logic and implied meaning can be understood from the following aspects. First, the literal absurdity creates a strong contrast. A flagpole is typically tall and solemn; in traditional contexts, it is used to hoist flags representing the state, the military, or authority, and therefore should not be treated lightly. A chicken feather, by contrast, is light and insignificant, often symbolizing something trivial or worthless. Placing a chicken feather on a flagpole—where a flag is supposed to be raised—appears highly inappropriate and ridiculous. Such an act would require extraordinary audacity, which naturally leads to the judgment “what enormous nerve.” .....

Table 17: Dictionary and model explanations of the *xiehouyu*: 旗杆上插鸡毛——好大的掸子(胆子) *Sticking chicken feathers on a flagpole — what a huge duster (guts)!*.

E Measuring novelty of created *xiehouyu*

Given a new *xiehouyu* (r, a), where r is the *riddle* and a is the *answer-intended*, we concatenate them into a single string s. For each existing *xiehouyu* (r_i, a_i) in the dictionary $\mathcal{D}_{\text{old}}$, we similarly form s_i. Let $\text{set}(\cdot)$ map a string to the set of its *distinct characters*. We compute overlap score between two *xiehouyu* using the Jaccard index:

$$\text{Overlap}(s, s_i) \;=\; \frac{|\text{set}(s) \cap \text{set}(s_i)|}{|\text{set}(s) \cup \text{set}(s_i)||}.$$

We compute this score for all i and take the maximum overlap

$$m \;=\; \max_i \text{Overlap}(s, s_i).$$

If $m < 0.5$ (with $\tau = 0.5$ in our implementation), we label the new *xiehouyu* as `original`, and return no candidates. Otherwise, we collect the set of *xiehouyu* achieving the same maximum overlap $\mathcal{T} = \left\{ i \ : \ \text{Overlap}(s, s_i) = m \right\}$ and label the new *xiehouyu* as `possible_dup`.