VIDAI: VIDukathAI Interpretation Through Analysis of In-context Reasoning in Tamil using LLMs

R S Mughil Srinivasan, Kesavan T, Abhijith Balan, Abhinav P M, Parameswari Krishnamurthy, Oswald C

Proceedings of the 39th Pacific Asia Conference on
Language, Information and Computation (PACLIC 39)

Emmanuele Chersoni, Jong-Bok Kim (eds.)

2025

VIDAI: VIDukathAI Interpretation Through Analysis of In-context Reasoning in Tamil using LLMs

R S Mughil Srinivasan[1]
106122104@nitt.edu

Kesavan T[1]
106122068@nitt.edu

Abhijith Balan[1]
406123001@nitt.edu

Abhinav P M[2]
abhinav.pm@research.iiit.ac.in

Parameswari Krishnamurthy[2]
param.krishna@iiit.ac.in

Oswald C[1]
oswald@nitt.edu

[1]National Institute of Technology, Tiruchirappalli
[2]International Institute of Information Technology, Hyderabad

Abstract

This work investigates VIDAI, a study on Tamil vidukathai, where *vidai* means 'answer' and *vidukathai* means 'riddle' in the Tamil language, focusing on the challenges Large Language Models (LLMs) face in solving them. Tamil, a morphologically rich and culturally embedded classical Dravidian language of India, with over 2000 years of history, is spoken officially in three countries and across global diaspora communities. Although models such as LLaMA, Phi, Gemma, and Qwen excel in general NLP tasks, they struggle with Tamil riddles due to their reliance on next-token prediction and limited reasoning ability. Tamil riddles frequently use metaphors, puns, cultural references, and abstract logic, posing difficulties for models trained primarily on generic corpora. We curated a dataset of 2,283 riddles[1] and evaluated the models under various prompting strategies. The highest performance achieved was a BERTScore of 0.846 with a random 1-shot no-CoT prompt. VIDAI's findings highlight riddles as a promising benchmark for testing reasoning in LLMs.

Keywords: CoT, In-Context Learning, LLM, Question Answering, Riddle, Tamil

1 Introduction

Large Language Models (LLMs) such as GPT, LLaMA, Phi, and Gemma have advanced natural language processing, excelling in summarization, translation, question answering, and text generation (Sanchez-Bayona and Agerri, 2025; Tong et al., 2024; Giadikiaroglou et al., 2024; Jiang et al., 2023). These strengths stem from large-scale pretraining and fine-tuning on diverse datasets. However, LLMs struggle with tasks that require symbolic abstraction, associative reasoning, and cultural grounding, such as solving a riddle (Liu et al., 2024).

[1]https://anonymous.4open.science/r/TamilRiddlesDataset-48C3

LLMs are trained on next token prediction objectives (Lin et al., 2025), favoring high-frequency, contextually probable outputs. Although effective for factual tasks, this biases models against riddles, which rely on ambiguity, metaphor, and wordplay. Solving riddles requires inference, analogy, and creative abilities that were not explicitly learned during training.

These challenges are amplified in morphologically rich and culturally grounded languages such as Tamil. With more than 2000 years of literary history, Tamil has agglutinative morphology, rich inflections, eight grammatical cases (Lehmann, 1993), and distinctive phonology documented in the *Tolkāppiyam* (Tolkāppiyar, Ancient). Its oral traditions feature riddles (*Vidukathai* in Tamil language) using metaphors, puns, idioms, and poetic conventions such as *venpa*, *ullurai*, and *iraicchi* (Schiffman, 1999; Iyyanarithanar, 9th Century CE). These cultural markers are rarely present in large corpora in English, causing models to default to generic or plausible answers.

VIDAI introduces the first dedicated Tamil riddles dataset, providing a benchmark to evaluate LLM performance by inferring and understanding Tamil riddles, and assesses open-source models using structured prompting, semantic example selection, and multi-metric evaluation. VIDAI's study highlights reasoning gaps and conditions that improve model performance, offering insight into handling metaphor-rich, culturally specific tasks.

The paper is organized as follows: §1 introduces the motivation, contributions, and problem statement. Related riddle-solving and multilingual LLM reasoning work are reviewed in §2. §3 explains the creation of the Tamil Dataset. The details of the design of VIDAI and the prompting strategies are explained in §4. §5 discusses the evaluation, metrics, and observations of the experiments. §6 presents the key takeaways. The conclusion and directions for future work are presented

in §7.

1.1 Contributions

- **Dataset Creation:** Compiled a cleaned dataset of 2,250+ Tamil riddles from books and public sources, categorized into Objects, Nature, Actions, People, and Abstract Concepts.
- **Few-Shot Prompting and CoT:** Designed random and semantically similar few-shot settings (1, 2, 3, 5 shots) and generated CoT explanations using ChatGPT for selected examples and manually curated to ensure alignment with human interpretations in solving a riddle.
- **Multi-Metric Evaluation:** Used Exact Match, Levenshtein Distance, and BERTScore to assess both literal and semantic correctness.
- **Local LLM Execution:** Ran experiments on Phi-4, Gemma2: 9b, Gemma3.1: 12b, LLaMA3: 8b and Qwen2.5: 7b via Ollama[2].
- **Riddle Reasoning in Tamil:** Focused on metaphor-rich riddles in a low-resource language to test cultural and linguistic reasoning.

1.2 Problem Description

LLMs excel at next-token prediction, favoring common, contextually likely continuations, which suits tasks like summarization or dialogue. However, the riddles are based on misdirection, wordplay, and layered meanings, demanding abstraction and creative reasoning. Tamil riddles add challenges with cultural references, phonetic puns, and idioms, rare in mainstream corpora, reducing model effectiveness, and highlighting a gap in multi-step, metaphorical reasoning.

The problem statement is defined as follows: Let $\mathcal{R} = \{(r_1, a_1), (r_2, a_2), \ldots, (r_n, a_n)\}$ be a collection of Tamil riddles, where each r_i represents a riddle, a_i - its corresponding answer, and $n = |\mathcal{R}|$, for $1 \leq i \leq n$. The task is to curate $\mathcal{R}$, then evaluate a series of Large Language Models to answer the riddles by predicting an answer b_i and computing the similarity of (b_i, a_i) with the help of various metrics.

[2] https://ollama.com/

2 Related Work

Recent studies have begun to treat riddle solving as a test of LLM reasoning. (Panagiotopoulos et al., 2025) proposed a context-reconstructed augmentation method that improves performance by providing structurally similar riddles as few-shot prompts. However, their work focuses on multiple-choice formats rather than VIDAI's method of QA style. (Giadikiaroglou et al., 2024) survey puzzles, classifying them into rule-based and rule-less forms, and note that the riddles remain challenging for LLMs. Although they call for better datasets and hybrid methods, their analysis is largely language-agnostic and not tailored to low-resource languages. (Lin et al., 2021) and (Jiang et al., 2023) address commonsense and linguistic creativity in English riddles and lateral puzzles. However, their methods are language-specific and do not address metaphor in non-English contexts. Similarly, (Tan et al., 2016) tackles the Chinese character riddles using language-specific fine-tuning, limiting cross-lingual applicability. Few-shot learning studies (Brown et al., 2020; Agarwal et al., 2025) informed VIDAI's shot size design, while (Wei et al., 2022b) inspired VIDAI's CoT approach. However, these works primarily test English tasks. (Zhang and Wan, 2022) and (Xu et al., 2023) confirm riddles as multilingual challenges but focus on multiple-choice formats, unlike VIDAI's QA format. The work of (Liu et al., 2022) validates the benefit of semantically similar examples, aligning with VIDAI's sampling strategy. Reasoning-oriented prompting research (Kojima et al., 2022; Fu et al., 2023) offers useful insights, but is based on mathematical and logic puzzles. Theoretical studies (Han et al., 2024; Wei et al., 2022a) explore general reasoning but neglect the resolution of culturally rich metaphors. In general, no major study goals are open in any Indian language, such as Tamil, which requires deep cultural and symbolic reasoning.

3 Dataset Creation

3.1 Dataset Sources

We collected the Tamil riddle dataset from various public sources, including books and online repositories. The sources are from classical Tamil riddle books such as (Muthaiah, 1987) and (Manivasan, 2018). We incorporate riddles from educational websites such as (Dheivegam, 2023),

(FreshTamil.com, 2020), (FreshTamil.com, 2024), and quiz portals such as (Vinaval, 2025) and (Vidukathaigal, 2025). The VIDAI's final dataset consisted of 2,283 unique Tamil riddles, each paired with a single ground-truth answer. This represents the largest available collection of Tamil riddles from which we could extract the best from online sources.

3.2 Riddle Structure

Riddles ranged from one to five poetic lines in colloquial Tamil, with answers as single words or short phrases. While some mapped to concrete concepts, others required abstract or symbolic interpretation.

3.3 Categorization of Riddles

We classified VIDAI's dataset based on the Tamil riddle answers into five broad semantic categories.

- **Natural Elements and Weather**: Answers related to natural phenomena and cycles, often expressed through environmental metaphors. Examples: சூரியன் - (Sūriyaṉ, Sun), காற்று - (Kāṯru, Wind).
- **Human Body and Senses**: Refers to body parts or sensory actions, typically using anatomical or functional metaphors. Examples: மூச்சு - (Mūccu, Breath), நாக்கு - (Nākku, Tongue).
- **Objects and Tools**: Man-made items are described through their form, function, or purpose. Examples: நாற்காலி - (Nāṟkāli, Chair), கடிகாரம் - (Kaṭikāram, Clock).
- **Food and Plants**: Edible items or plants, often described using taste, texture, or appearance. Examples: வெங்காயம் - (Veṅkāyam, Onion), கரும்பு - (Karumpu, Sugarcane).
- **Animals and Insects**: Creatures referenced through behavior, sounds, or cultural associations. Examples: சிலந்தி - (Silanti, Spider), நாய் - (Nāy, Dog).

4 Design

Figure 1 shows the architecture of VIDAI, using five open-source LLMs: LLaMA 3.1 (8B), Phi-4, Gemma 2 (9B), Gemma 3.1 (12B), and Qwen 2.5 (7B) with zero, one, and few-shot prompting. In few-shot settings, the models received 2, 3, and 5 riddle-answer examples.

We employ two sampling strategies: Random Sampling and Semantic Similarity Sampling. In Random Sampling, examples were selected arbitrarily from the training set, with each example likely originating from a different class. In contrast, Semantic Similarity Sampling involved first selecting one of the five predefined categories, established through manual classification of the dataset based on the answers, as described in Section *Categorization of Riddles*, and then randomly drawing all examples from that category. For instance, in a three-shot prompt, Random Sampling would yield three examples from potentially different categories, whereas Semantic Similarity Sampling would select a single category at random and then draw three examples exclusively from it.

4.1 Chain-of-Thought

Initially, the examples included only riddle-answer pairs. In a second phase, we extended the prompts using **Chain-of-Thought (CoT)** reasoning, adding explanatory reasoning to each example. CoT explanations were presented in a consistent deductive format in all examples. Each riddle is first restated and explained in English, followed by the revelation of the answer. The explanation then proceeds by mapping each segment of the riddle to its underlying meaning, interpreting phrases in relation to the proposed answer, and the justification explicitly shows how the answer satisfies each part of the riddle. This ensures that the meanings embedded in figurative, metaphorical, or cultural contexts are extracted and clarified. The length of the explanations naturally varies according to the complexity of the riddle.

4.2 Direct Prompting

In this setup, the model was given only the riddle, without any additional guidance, and tasked with producing an answer. This zero-shot approach relies entirely on the internal knowledge and reasoning of the model to interpret and solve the riddle.

An example prompt given without CoT explanation :

```
Provide only the final answer in
Tamil without any translations or
explanations in English for the given
Tamil riddle below.
Question : பறக்கும் ஆனால் பறந்து போகாது,
```

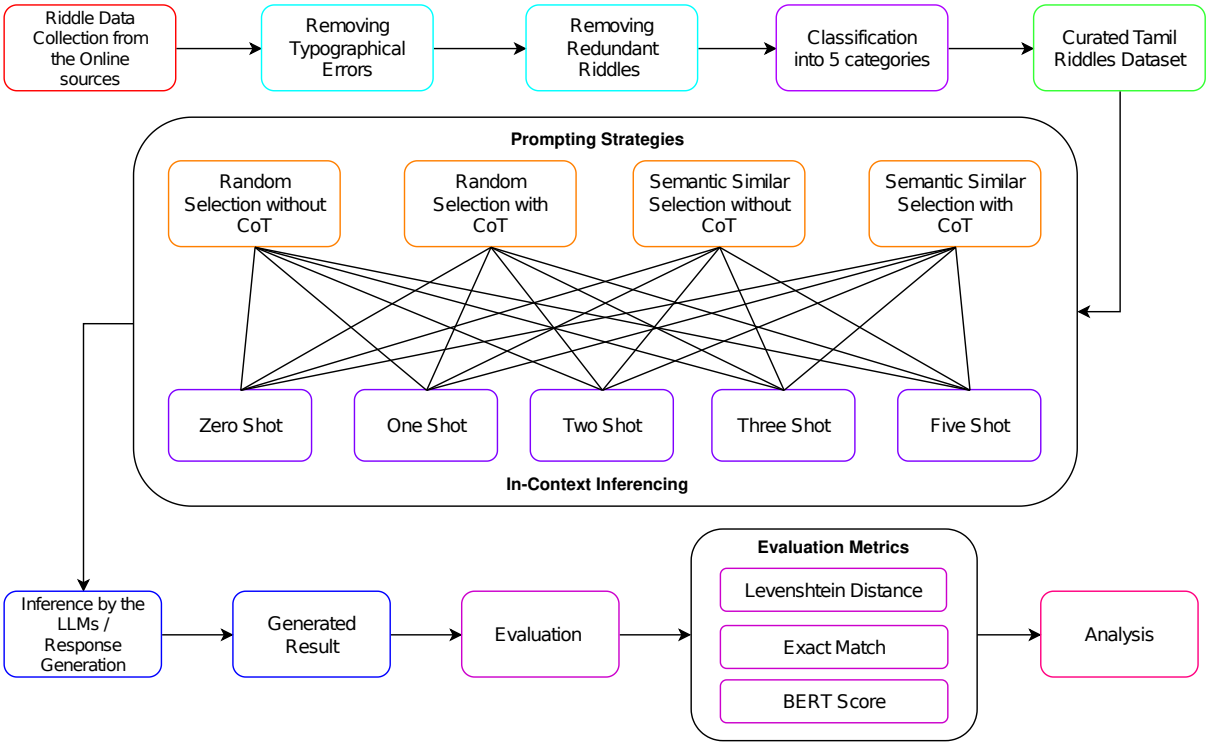


Figure 1: Architecture diagram of VIDAI: VIDukathAI Interpretation Through Analysis of In-context Reasoning in Tamil using LLMs

அது என்ன?

Answer:

4.3 Few-Shot Prompting

We adopt a few-shot strategy with Chain-of-Thought (CoT) reasoning, where each example includes the riddle, its answer, and step-by-step reasoning to interpret metaphors, recognize symbols, and eliminate unlikely options. These CoT explanations, generated using ChatGPT-4o with the riddle and its known answer, were then manually curated to ensure coherence, contextual relevance, logical foundation and alignment with human interpretations.

An example of a 1-shot prompt given with CoT explanations is:

Question: வீட்டுக்குள்ளே இருப்பாள், விந்தையாகப் பேசுவாள், வந்தவரை வா என்பாள், வாசல் தாண்டிப் போகமாட்டாள். அவள் யார்?

Answer: நாக்கு

Explanation: The riddle describes something that stays inside the house, speaks mysteriously, invites people, and never crosses the threshold. The answer is நாக்கு (tongue). வீட்டுக்குள்ளே இருப்பாள் refers to the tongue that remains inside the mouth (the house). விந்தையாகப் பேசுவாள் highlights its role in speaking, வந்தவரை வா என்பாள் shows how it welcomes speech, and வாசல் தாண்டிப் போகமாட்டாள் means it never leaves the mouth.

Provide only the final answer in Tamil without any translations or explanations in English for the given Tamil riddle below.

Question: பாதுகாப்பான பெட்டிக்குள்ளே பலரும் விரும்பும் கடிகாரம், அது என்ன?

Answer:

Enriched examples are especially useful for solving riddles, where metaphorical and cultural understanding is the key. Comparing these two prompting approaches helps evaluate how in-context learning and example-based reasoning affect LLM performance on complex, ambiguous queries. All experiments were run locally using models downloaded via Ollama, depending on available resources.

5 Evaluation, Results and Observations

To assess model performance in solving Tamil riddles, we employed four key evaluation metrics, each capturing different aspects of answer quality:

- **Exact Match:** Checks if the model answer exactly matches the ground truth. It is simple but rigid, often penalizing correct synonyms or alternate Tamil wordings. It is defined as:

  $$\text{Exact Match (EM)} = \frac{\sum_{i=1}^{N} 1[\hat{y}_i = y_i]}{N}$$

- **Levenshtein Distance:** Measures similarity via minimal character edits, normalized as [0, 1]. It detects typos and near matches, but misses semantically correct words with different spellings. It is defined as:

  $$\text{Levenshtein Similarity} = 1 - \frac{d_{\text{lev}}(\hat{y}, y)}{\max(|\hat{y}|, |y|)}$$

- **BERTScore:** Uses transformer embeddings to compare tokens, capturing semantic meaning and paraphrases, effective for Tamil riddles with subtle contextual nuances. It is defined as:

  $$\text{BERTScore}_{F1} = \frac{1}{N} \sum_{i=1}^{N} F1(\hat{y}_i, y_i)$$

Tamil riddles, rich in metaphor and cultural nuance, require evaluation beyond strict string matching. We use BERTScore with surface metrics to capture semantic equivalence across varying wordings, analyzing the effects of context, model size, and explanation style across multiple prompting setups.

5.1 Algorithm Trace Example

This section presents a sample execution of Algorithm 1 for one configuration. The model is Phi-4, using the *Semantic similarity* prompting strategy with 3 shots. The test riddle is மேலே மேலே போகும், கீழே கீழே போகாது. என்ன? (Mēlē mēlē pōkum, kīḻē kīḻē pōgātu. Enna?) – It keeps going upward, but never goes downward. The ground truth answer is வயது (Vayatu) – Age. The evaluation begins with `GetFewShotExamples(model, riddle, shot=3, strategy="semantic")`, which selects three semantically related examples, focusing on abstract concepts from *Human Body & Senses* category. Then, `LoadTestSet()` loads the riddle and its ground truth answer. Next, `ConstructPrompt(few_shot_examples, test_riddle)` builds the prompt by combining the examples with the target riddle in QA format. Passing this on to `GenerateAnswer(model, prompt)` yields புகை (Pogaei) – Smoke. Although incorrect, this highlights the model's reliance on surface-level cues rather than abstract temporal reasoning. Finally, `Evaluate()` computes metrics such as *BERTScore*, reflecting partial semantic proximity in the metaphorical interpretation between smoke and age.

Algorithm 1 Evaluate Models on Riddle QA with Few-Shot Prompts

Input: *TestSet*, *FewShotData*, *ShotCounts*, *Models*, *PromptModes*
for all Model ∈ Models **do**
  Load model via local or remote interface
  for all PromptMode ∈ PromptModes **do**
    for all ShotCount ∈ ShotCounts **do**
      if ShotCount > 0 **then**
        Examples ← `GetFewShotExamples(` FewShotData, PromptMode, ShotCount)
      else
        Examples ← ∅
      end if
      (Questions, Answers) ← `LoadTestSet` (TestSet)
      Predictions ← [ ]
      for all Question ∈ Questions **do**
        Prompt ← `ConstructPrompt(` Examples, Question, PromptMode)
        Response ← `GenerateAnswer(` Model, Prompt)
        Append Response to Predictions
      end for
      Metrics ← `Evaluate`(Predictions, Answers)
      Evaluation Metrics: *Exact Match*, *Cosine Similarity*, *BERTScore*, *Levenshtein Similarity*
    end for
  end for
end for

5.2 Results and Discussions

We ran the five open-source LLMs on 14th Gen Intel Core i7 CPU, NVIDIA T1000 8 GB GPU with 16 GB Main Memory and 512 GB Hard Disk on Linux platform using Python code. The models LLaMA 3.1 (8B), Phi-4, Gemma 2 (9B), Gemma 3.1 (12B), and Qwen 2.5 (7B) were evaluated across four setups: semantically similar or random riddles, with or without CoT explanations, under zero, one, and few-shot (2, 3, 5) conditions. Performance was measured using exact match, Levenshtein distance, and BERTScore. The results are tabulated in Table 1 and Table 2 and visualized in Figures 3, 4, 5 and 6.

The direct comparison of State-of-the-art was not relevant because previous work differs substantially; for example, RiSCORE (Panagiotopoulos et al., 2025) treats riddles as multiple choice with ranking metrics, while RiddleSense (Lin et al.,

2021) and BRAINTEASER (Jiang et al., 2023) address English riddles using knowledge-based solvers in different formats. VIDAI's focus was on raw inference in natural QA. In this setup, each prompt contained only a Tamil riddle, and the model was asked to generate the answer directly without any additional clues or multiple choice options.

5.3 LLaMA 3.1 (8B)

LLaMA 3.1 showed steady results: Best Exact Match 0.007 (5-shot CoT, semantic), Levenshtein 0.140 (3-shot CoT, semantic), BERTScore 0.775 (1-shot no-CoT). The results highlight its strong multilingual pretraining yet limited adaptability to the Tamil riddle's metaphorical style. CoT slightly improved generalization, but the modest gains suggest a token prediction bias rather than true abstract reasoning.

5.4 Phi-4

Phi-4 peaked with 5-shot CoT semantic runs: Exact Match 0.007, Levenshtein 0.141, BERTScore 0.767. Its compact, reasoning-oriented training made it suitable for structured prompts. CoT aligned well with the modeling intermediate steps. Although absolute scores remain modest, consistent improvements show adaptability. Still, the lack of deep contextual embeddings limits its figurative understanding of Tamil riddles.

5.5 Gemma 2 (9B)

Gemma 2 varied widely. Best Exact Match 0.022 (5-shot CoT random), BERTScore 0.846 (1-shot no-CoT random), Levenshtein 0.303 (5-shot no-CoT random). These spikes suggest reliance on lexical overlap rather than reasoning, aided by token similarity. CoT often reduced performance, implying a mismatch with training. The results show that size alone does not guarantee reasoning capacity.

5.6 Gemma 3.1 (12B)

Gemma 3.1 produced consistent results: BERTScore 0.785 (5-shot no-CoT semantic), Exact Match 0.018 (zero-shot CoT), Levenshtein 0.170 (3-shot CoT random). Its larger size likely supports better semantic generalization and metaphor detection. However, CoT offered little benefit, suggesting that internal reasoning suffices. In general, Gemma 3.1 balances surface accuracy and semantic similarity, adapting to prompting conditions.

5.7 Qwen 2.5 (7B)

Qwen 2.5 scored lowest overall: Exact Match 0.004, Levenshtein 0.149, BERTScore 0.768 (all no-CoT, both prompts). The results show a weakness in multi-step reasoning and poor CoT handling, reflecting multilingual pretraining not tuned for Tamil. Its training favors fluency and factual QA over metaphorical reasoning. Low scores highlight difficulty with analogy, figurative interpretation, and inference.

5.8 Cross-Model Observations

- **Few-shot prompting** (3–5 shots) consistently outperforms **zero/one-shot**, mainly in **exact match** and **Levenshtein**.
- **Semantically similar example selection** outperforms **random sampling** in smaller models by providing more relevant **contextual alignment**.
- **CoT explanations** improve mid-sized models but sometimes reduce performance in larger ones due to **prompt-structure mismatch**.
- **Embedding-based metrics** better capture **semantic understanding** than **exact match**, reflecting the high linguistic diversity of valid **Tamil riddle answers**.

In summary, **architecture**, **training data**, and **reasoning alignment** significantly influence performance in decoding **Tamil riddles**.

6 Key Takeaways

- **Riddle solving needs more than facts:** Tamil riddles rely on metaphor, symbolism, and cultural cues that LLMs often miss.
- **Cultural grounding matters:** Most models overlook idioms and poetic clues, reducing accuracy.
- **Bigger models aren't always better:** Large LLMs are stable, but smaller ones can sometimes outperform them.
- **Current metrics fall short:** The exact match is too strict, and even BERTScore cannot fully assess understanding.

Model - Shots	Semantically Similar Riddles			Randomly Selected Riddles		
	Exact Match	Leven-shtein	BERT Score	Exact Match	Leven-shtein	BERT Score
Llama3.1:8b - 0	0.31	13.73	77.12	0.13	13.53	76.94
Llama3.1:8b - 1	0.48	13.17	77.57	0.48	13.44	77.30
Llama3.1:8b - 2	0.35	12.54	77.16	0.44	13.69	77.40
Llama3.1:8b - 3	0.35	13.24	77.13	0.39	13.31	77.30
Llama3.1:8b - 5	0.66	13.44	77.24	0.57	13.22	77.15
Phi4 - 0	0.13	13.00	74.97	0.18	12.93	74.91
Phi4 - 1	0.26	11.61	72.62	0.44	13.47	75.36
Phi4 - 2	0.13	10.38	71.84	0.22	11.30	73.06
Phi4 - 3	0.44	11.91	73.04	0.44	11.51	73.90
Phi4 - 5	0.39	12.64	74.45	0.53	12.83	74.61
Gemma2:9b - 0	1.40	16.72	77.60	0.00	20.54	79.48
Gemma2:9b - 1	1.36	16.50	77.63	0.00	20.00	84.61
Gemma2:9b - 2	1.80	16.56	77.78	0.00	25.00	82.39
Gemma2:9b - 3	1.62	16.59	77.89	0.00	21.25	81.17
Gemma2:9b - 5	1.88	16.56	77.93	0.00	30.29	82.91
Gemma3.1:12b - 0	1.88	15.80	78.19	1.80	15.78	78.17
Gemma3.1:12b - 1	1.53	15.74	78.39	1.71	15.47	78.30
Gemma3.1:12b - 2	1.53	15.99	78.18	1.31	15.66	78.38
Gemma3.1:12b - 3	1.49	15.72	78.31	1.31	15.78	78.47
Gemma3.1:12b - 5	1.75	16.12	78.53	1.05	16.61	77.77
Qwen2.5:7b - 0	0.00	14.00	75.88	0.04	14.07	75.93
Qwen2.5:7b - 1	0.00	14.18	75.73	0.31	14.59	76.80
Qwen2.5:7b - 2	0.18	13.49	76.19	0.22	14.16	75.86
Qwen2.5:7b - 3	0.26	15.00	76.04	0.22	14.29	76.34
Qwen2.5:7b - 5	0.39	14.84	76.24	0.35	14.42	75.89

Table 1: Performance of various LLMs on Tamil riddles using Exact Match, Levenshtein Distance, and BERTScore without CoT explanations (values are scaled by a factor of 100)

Model - Shots	Semantically Similar Riddles			Randomly Selected Riddles		
	Exact Match	Leven-shtein	BERT Score	Exact Match	Leven-shtein	BERT Score
Llama3.1:8b - 0	0.18	13.10	77.04	0.31	13.17	76.92
Llama3.1:8b - 1	0.39	13.71	77.21	0.39	13.55	77.22
Llama3.1:8b - 2	0.35	12.85	77.00	0.48	13.65	77.31
Llama3.1:8b - 3	0.66	14.09	77.37	0.26	13.21	77.15
Llama3.1:8b - 5	0.79	13.61	77.11	0.48	13.51	77.33
Phi4 - 0	0.09	13.38	75.13	0.13	13.13	74.66
Phi4 - 1	0.39	13.37	75.44	0.44	12.86	75.19
Phi4 - 2	0.26	12.47	74.76	0.44	13.32	75.60
Phi4 - 3	0.48	13.96	76.31	0.53	13.14	75.56
Phi4 - 5	0.70	14.10	76.72	0.35	12.89	74.96
Gemma2:9b - 0	1.31	16.46	77.68	1.49	17.00	77.55
Gemma2:9b - 1	1.45	16.16	77.50	1.45	15.98	77.61
Gemma2:9b - 2	1.18	15.68	77.70	1.58	16.24	77.80
Gemma2:9b - 3	1.66	16.51	77.90	1.80	16.79	78.00
Gemma2:9b - 5	1.71	16.21	78.11	2.23	16.99	77.98
Gemma3.1:12b - 0	1.88	16.06	78.19	1.84	15.77	78.20
Gemma3.1:12b - 1	0.96	15.35	77.82	1.40	13.90	78.18
Gemma3.1:12b - 2	1.84	16.09	78.27	1.05	15.77	77.76
Gemma3.1:12b - 3	1.23	15.96	77.76	1.58	16.97	77.69
Gemma3.1:12b - 5	1.27	15.95	78.36	0.88	16.59	77.48
Qwen2.5:7b - 0	0.04	13.97	75.95	0.09	13.82	76.01
Qwen2.5:7b - 1	0.04	13.65	75.68	0.09	14.02	76.06
Qwen2.5:7b - 2	0.13	14.02	76.25	0.18	14.45	75.64
Qwen2.5:7b - 3	0.22	14.47	76.28	0.22	14.26	75.94
Qwen2.5:7b - 5	0.39	14.65	76.47	0.18	14.39	75.28

Table 2: Performance of various LLMs on Tamil riddles using Exact Match, Levenshtein Distance, and BERTScore with CoT explanations (values are scaled by a factor of 100)

- **Riddles are a strong benchmark:** They challenge creative and cultural reasoning beyond the generation of fluent text.

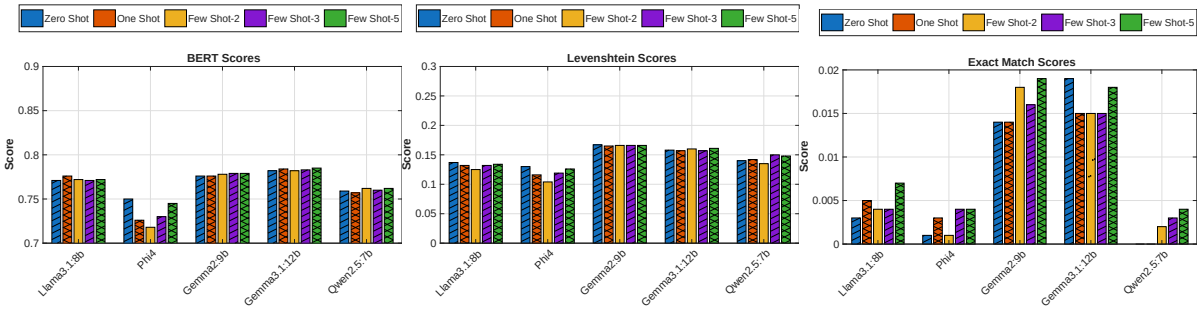
7 Conclusions and Future Work

This study evaluated modern open-source LLMs on Tamil riddles, highlighting their difficulty in handling metaphor, cultural nuance, and symbolic reasoning. Although structured prompts and CoT offered slight gains, the model relied primarily on surface patterns rather than deep understanding.

Future work shall focus on culturally rich datasets with annotated reasoning, explore hybrid neuro-symbolic methods, and incorporate cross-modal and retrieval-augmented approaches. Richer evaluation metrics and human-in-the-loop assessments are essential to push LLM toward genuine cognitive and cultural comprehension beyond fluent language generation.

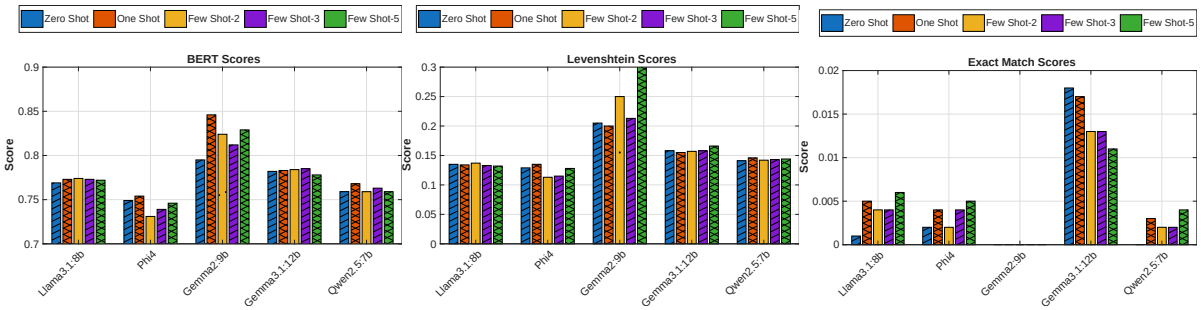
Acknowledgment

We gratefully acknowledge the authors of the online resources from which the riddles were scraped to create our dataset and the use of generative AI tools for help in paraphrasing and small edits.



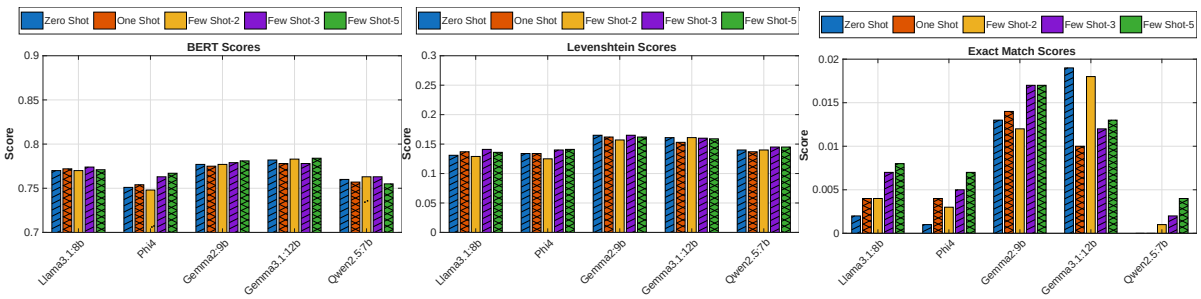
(a) BERT Score (b) Levenshtein Score (c) Exact Match Score

Figure 3: Graphical Representation of the Metric Comparisons for Semantic Similar Selection without CoT



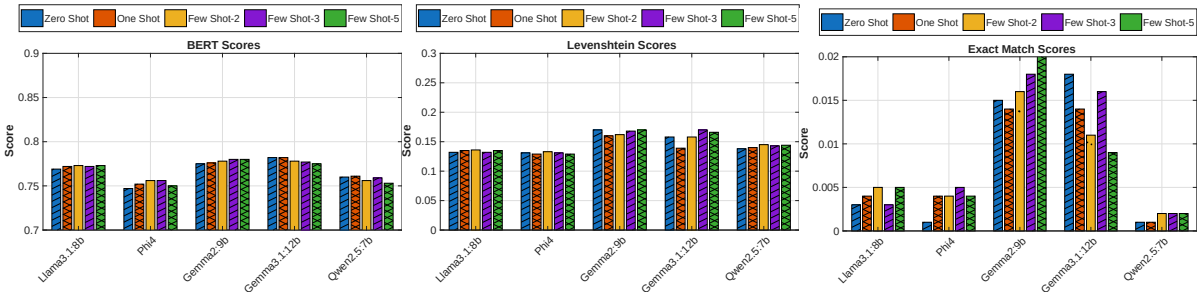
(a) BERT Score (b) Levenshtein Score (c) Exact Match Score

Figure 4: Graphical Representation of the Metric Comparisons for Random Selection without CoT



(a) BERT Score (b) Levenshtein Score (c) Exact Match Score

Figure 5: Graphical Representation of the Metric Comparisons for Semantic Similar Selection with CoT



(a) BERT Score (b) Levenshtein Score (c) Exact Match Score

Figure 6: Graphical Representation of the Metric Comparisons for Random Selection with CoT

References

Rishabh Agarwal, Avi Singh, Lei Zhang, Bernd Bohnet, Luis Rosias, Stephanie Chan, Biao Zhang, Ankesh Anand, Zaheer Abbas, Azade Nova, John D. Co-Reyes, Eric Chu, Feryal Behbahani, Aleksandra Faust, and Hugo Larochelle. 2025. Many-shot in-context learning. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24, Red Hook, NY, USA. Curran Associates Inc.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Dheivegam. 2023. Tamiḻ viṭukataikaḷ 300.

FreshTamil.com. 2020. Tamiḻ viṭukataikaḷ | 100+ vidukathaigal in tamil | tamil riddles.

FreshTamil.com. 2024. Tamiḻ viṭukataikaḷ viṭaiyuṭaṉ 2024 | tamil vidukathaigal.

Yao Fu, Hao Peng, Ashish Sabharwal, Peter Clark, and Tushar Khot. 2023. Complexity-based prompting for multi-step reasoning.

Panagiotis Giadikiaroglou, Maria Lymperaiou, Giorgos Filandrianos, and Giorgos Stamou. 2024. Puzzle solving using reasoning of large language models: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 11574–11591, Miami, Florida, USA. Association for Computational Linguistics.

Simon Jerome Han, Keith J. Ransom, Andrew Perfors, and Charles Kemp. 2024. Inductive reasoning in humans and large language models. volume 83, page 101155.

Iyyanarithanar. 9th Century CE. *Purapporul Venbamaalai*. Classical Tamil grammatical treatise on the *puram* genre.

Yifan Jiang, Filip Ilievski, Kaixin Ma, and Zhivar Sourati. 2023. BRAINTEASER: Lateral thinking puzzles for large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14317–14332, Singapore. Association for Computational Linguistics.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Thomas Lehmann. 1993. *A Grammar of Modern Tamil*. Pondicherry Institute of Linguistics and Culture, Pondicherry, India.

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

Pengxiao Lin, Zhongwang Zhang, and Zhi-Qin John Xu. 2025. Reasoning bias of next token prediction training.

Chen Cecilia Liu, Fajri Koto, Timothy Baldwin, and Iryna Gurevych. 2024. Are multilingual llms culturally-diverse reasoners? an investigation into multicultural proverbs and sayings.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2022. What makes good in-context examples for GPT-3? In *Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 100–114, Dublin, Ireland and Online. Association for Computational Linguistics.

N. Manivasan. 2018. *500 Viṭukataikaḷum ataṉ aṟputa patilkaḷum*. aedahamlibrary.

Mullai Muthaiah. 1987. *1000 Viṭukataikaḷ*, 2 edition. New Century Book House Private Limited, 41-B. CITCO Industrial Estate, Chennai - 600098.

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

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

Harold F. Schiffman. 1999. *A Reference Grammar of Spoken Tamil*. Cambridge University Press, Cambridge, UK.

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, Austin, Texas. Association for Computational Linguistics.

Tolkāppiyar. Ancient. *Tolkāppiyam*. Earliest extant Tamil grammatical text, dating between 500 BCE–500 CE depending on tradition.

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

Vidukathaigal. 2025. Tamil molikal: Vidukathaigal.

Vinaval. 2025. Vidukathai vina vidaighal | vinaval.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022a. Emergent abilities of large language models.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022b. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Fan Xu, Yunxiang Zhang, and Xiaojun Wan. 2023. Cc-riddle: A question answering dataset of chinese character riddles.

Yunxiang Zhang and Xiaojun Wan. 2022. Birdqa: A bilingual dataset for question answering on tricky riddles.