From Phonemes to Meaning: Evaluating Large Language Models on Tamil

Jeyarajalingam Varsha Menan Velayuthan Sumirtha Karunakaran
Rasan Nivethiga Kengatharaiyer Sarveswaran
Department of Computer Science, University of Jaffna, Sri Lanka.
{varshajeyaraj, vmenan95, sumirthakarunakaran96,rasanniksha}@gmail.com, sarves@univ.jfn.ac.lk

Abstract

Large Language Models (LLMs) have shown strong generalization across tasks in high-resource languages; however, their linguistic competence in low-resource and morphologically rich languages such as Tamil remains largely unexplored. Existing multilingual benchmarks often rely on translated English datasets, failing to capture the linguistic and cultural nuances of the target language. To address this gap, we introduce *ILAKKANAM*, the first Tamil-specific linguistic evaluation benchmark manually curated using 820 questions from Sri Lankan school-level Tamil subject examination papers. Each question is annotated by trained linguists under five linguistic categories and a factual knowledge category, spanning Grades 1–13 to ensure broad linguistic coverage. We evaluate both closed-source and open-source LLMs using a standardized evaluation framework. Our results show that Gemini 2.5 achieves the highest overall performance, while open-source models lag behind, highlighting the gap in linguistic grounding. Category- and grade-wise analyses reveal that all models perform well on lower-grade questions but show a clear decline as linguistic complexity increases. Further, no strong correlation is observed between a model's overall performance and its ability to identify linguistic categories, suggesting that performance may be driven by exposure rather than genuine understanding.

Keywords: Tamil, Linguistic Benchmark, Linguistic diagnostics, LLM

1. Introduction

Since the public release of ChatGPT in 2022 (OpenAI, 2022), Large Language Models (LLMs) have drawn significant public attention and rapidly integrated into everyday life. This growing interest has attracted substantial investment and funding toward companies developing these systems, resulting in a proliferation of models from different vendors. Closed-source models such as GPT-5 (OpenAI, 2025), Claude Sonnet 4.5 (Claude, 2025), and Gemini 2.5 (Comanici et al., 2025), as well as open-source counterparts like LLaMA 4 (Llama4Herd, 2025), DeepSeek-V3 (DeepSeek-AI et al., 2025), Qwen 2.5 (Qwen et al., 2025), and Grok 4 (Grok4, 2025), represent this expanding ecosystem. As LLMs become integrated into human workflows, including their use as evaluators for complex tasks (LLM-as-a-Judge) (Gu et al., 2025; Fu and Liu, 2025), the responsibility lies with the research community to evaluate these models, understand their capabilities, and identify their limitations (Chang et al., 2023).

The GLUE benchmark (Wang et al., 2018) and its extended version, SuperGLUE (Wang et al., 2019), established a standardized framework for evaluating language understanding across lexical semantics, logic, and grammar. BLiMP (Warstadt et al., 2020) extended this direction by introducing minimal pair evaluations for core syntactic phenomena such as subject–verb agreement and filler–gap dependencies. HELM (Liang et al., 2023) broadened the evaluation scope to incorporate ethical and demographic dimensions while emphasizing the limited multilingual and typological coverage of existing benchmarks. The MMLU dataset (Hendrycks et al., 2021) introduced a multitask evaluation across 57 academic and professional subject domains assessing both factual knowledge and reasoning ability. Despite measurable gains from larger models such as GPT-3, performance remains below expert level, indicating persistent limitations in knowledge depth and reliability.

While several efforts have attempted to create multilingual benchmarks, most are direct translations of their English counterparts (Singh et al., 2025; Bandarkar et al., 2024). Such approaches often fail to capture the cultural and linguistic nuances of the target languages (Ji et al., 2023). Following the design of MMLU, comparable benchmarks have been developed for other languages, including Sinhala (Pramodya et al., 2025), Arabic (Koto et al., 2024), Chinese (Li et al., 2024), Turkish (Yüksel et al., 2024), Indonesian (Koto et al., 2023), Korean (Son et al., 2025), and Persian (Ghahroodi et al., 2024). These efforts demonstrate the importance of language-specific benchmarks developed by native-speaking communities, ensuring that linguistic and cultural characteristics are represented, which are otherwise often absent in large-scale multilingual settings.

While SEA-HELM[1] (Susanto et al., 2025) eval-

[1] Previously known as BHASHA

uates the linguistic capabilities of LLMs through its LINDSEA suite for Tamil, its coverage remains limited. To address this gap, we introduce the *ILAKKANAM*, a manually curated dataset designed for Tamil linguistic assessment. Inspired by MMLU (Hendrycks et al., 2021), we compile questions from Sri Lankan school-level Tamil language examination papers.

ILAKKANAM comprises 820 questions spanning Grades 1–13, each annotated by trained linguists under five linguistic categories. This paper outlines the procedures used to collect, clean, annotate, and evaluate these questions against both closed- and open-source LLMs. To summarize, our work makes the following core contributions:

- We introduce *ILAKKANAM*, a manually curated Tamil linguistic benchmark consisting of 820 questions from Sri Lankan school-level Tamil language examination papers, annotated across five linguistic categories and a factual knowledge category.
- We design a structured evaluation pipeline to assess both closed-source and open-source LLMs, enabling fine-grained comparison across linguistic dimensions and grade-level complexity.
- Through comprehensive analysis, we reveal a consistent performance gap between closed- and open-source models, limited correlation between linguistic accuracy and category classification, and highlight the need for deeper linguistic grounding in Tamil language modeling.

2. Background

This section provides a brief introduction to the Tamil language and the Sri Lankan education system to contextualise and support the work presented in this paper.

2.1. The Tamil language

Tamil (tam) is a member of the South Dravidian branch of the Dravidian language family and is spoken by approximately 90 million people worldwide[2]. It is an agglutinative language with rich morphological and syntactic constructions. Tamil has a documented history of over two millennia, evolving through distinct historical stages. It holds official language status in Sri Lanka, Singapore, and the Indian state of Tamil Nadu, and recognised as a second language in many other countries.

From a computational perspective, Tamil is classified as a low-resource language (Abirami et al., 2024) according to Joshi's typology (Joshi et al., 2020). Although widely spoken, Tamil lacks comprehensive, high-quality, and error-free resources—particularly annotated datasets and standardized benchmarks—which limits the development and evaluation of robust NLP tools for the language.

2.2. The Sri Lankan education system

Sri Lanka provides 13 years of free general education across four cycles: Primary (Grades 1–5, ages 5–10), Junior Secondary (Grades 6–9, ages 11–14), Senior Secondary (Grades 10–11, ages 15–16), and Advanced Level (Grades 12–13, ages 17–18)(Liyanage, 2014). Schooling is compulsory from Grade 1 to 13, with three major national examinations: the Grade 5 Scholarship Exam (for merit-based school access and financial aid), the GCE O/L at Grade 11 (stream selection for A/L), and the GCE A/L at Grade 13 (university admission)(Liyanage, 2014).

Tamil is taught both as a subject and a medium of instruction for Tamil-speaking students from Grade 1 to 11, and as a specialization in Grades 12–13. This extended exposure, reinforced by national assessments, builds strong linguistic competence.

By Grade 5[3], students are expected to acquire foundational skills in listening, reading, speaking, and writing. By Grade 11[4], they are expected to master grammar and apply Tamil in academic and social contexts. Though not always explicitly labeled, the curriculum gradually introduces key linguistic domains: phonetics, phonology, morphology, syntax, and semantics. In Grades 12–13, students specializing in Tamil engage with advanced material, including literature, history, and poetry.

3. Related Work

3.1. Tamil in Multilingual Benchmarks

A few multilingual benchmark evaluation datasets include Tamil among other Indic and Southeast Asian languages.

SEA-HELM[5] (Susanto et al., 2025; Leong et al., 2023), through its LINDSEA suite for Tamil, assesses morphology, syntax, and semantics using minimal pair structures and handcrafted diagnostics. In their work, the authors show that both GPT-4

[2] https://www.worlddata.info/languages/tamil.php

[3] https://nie.lk/pdffiles/tg/tGR05TGTAMILLANGUAGE.pdf

[4] https://nie.lk/pdffiles/tg/t11tim159.pdf

[5] Previously known as BHASHA

and GPT-3.5-Turbo perform poorly on Tamil morphological analysis, with scores of 16.43% and 41.43%, respectively, even when prompted in English. Issues were particularly acute in processing gender, person, and tense agreement, as well as case marking in Question Answering tasks. This highlights that even prominent commercial models still have progress to make in achieving true multilinguality. SEA-HELM further revealed that Tamil's non-Latin script introduces complications in prompt design, such as the use of capitalization or punctuation conventions that do not align with Tamil's orthographic norms.

While PARIKSHA (Watts et al., 2024) scored Tamil-generated content highly in terms of fluency and grammaticality, it remains unclear whether this reflects deep language modeling or surface-level token fluency. Collectively, these studies provide strong evidence for the creation and evaluation of LLMs on Tamil in a more fine-grained manner.

3.2. Categorization of Linguistic Evaluation

Linguistic evaluation can span several axes. However, we use the five key axes: Phonetics, Phonology, Morphology, Syntax and Semantics for our evaluation[6].

3.2.1. Phonetics & Phonology

Studies outside Tamil, such as Begus et al. (2025), evaluated metalinguistic phonological abilities and showed that LLMs generalize over patterns like intervocalic gemination using synthetic words. PolyBench (Suvarna et al., 2024) further introduced tasks on grapheme-to-phoneme conversion and syllable counting, noting that phonological competence remains underexplored even for English. These approaches provide useful reference points for designing Tamil-specific phonological evaluations.

Model evaluations have yet to examine Tamil phonetics and phonology in detail, although some existing studies offer transferable insights. In SEA-HELM's LINDSEA suite, models such as GPT-4 struggled with verbal reduplication and other syllabic patterns unique to Tamil, producing incoherent translations in modally complex sentences. The same study reported that Tamil's script posed challenges for grapheme-level prompt formatting, adding further difficulty for phonology-related tasks.

3.2.2. Morphology

Tamil's morphological richness continues to pose difficulties for current LLMs. Results from the LINDSEA tests in SEA-HELM showed frequent errors in morphological agreement, including mismatches in case, gender, and number (Leong et al., 2023; Susanto et al., 2025). Such problems appeared consistently across Question Answering (QA) and sentence completion tasks, indicating that the models lack a stable inflection mechanism for agglutinative languages.

PARIKSHA (Watts et al., 2024) reported a perfect acceptability score for GPT-4-Turbo on Tamil texts, but this self-assessment measure does not necessarily capture linguistic depth. The gap between surface acceptability and internal representation highlights the need for systematic probing of morphological understanding in Tamil.

3.2.3. Syntax

Tamil syntax, with its free word order and nested predicate structures, presents additional challenges for current models. The syntactic diagnostics in LINDSEA evaluate phenomena such as argument structure and filler–gap dependencies, areas where LLMs often underperform (Leong et al., 2023). Following the approach of BLiMP (Warstadt et al., 2020), these tests use minimal pairs to probe specific syntactic contrasts and expose weaknesses in handling long-distance dependencies or embedded clause scrambling.

Benchmarks like PAWS (Zhang et al., 2019) also inform Tamil syntax evaluation by showing that models tend to rely on surface word overlap rather than syntactic structure—a concern particularly relevant for Tamil, where flexible word order is grammatical. Begus et al. (2025) reported similar limitations, noting that LLMs struggle with recursion and structural ambiguity, consistent with the syntactic difficulties observed for Tamil.

3.2.4. Semantics

Semantic understanding remains one of the most challenging aspects for Tamil, both in isolation and in inference-based settings. SEA-HELM evaluated Tamil through the IndicXNLI benchmark and reported that GPT-4o reached a score of 64.7, notably lower than its performance on Vietnamese and Indonesian (Susanto et al., 2025). The earlier BHASHA study observed even lower semantic comprehension (33.43%) for GPT-4 when English prompts were used (Leong et al., 2023). The model performed better, around 70%, on salient elements such as the emphatic particle *taan*, suggesting that high-salience, language-specific tokens are more reliably captured.

[6] https://linguistics.ucla.edu/undergraduate/what-is-linguistics/

4. Data Creation

ILAKKANAM comprises 820 questions spanning Grades 1–13, each annotated by two trained linguists under five linguistic categories (see Table 1). Questions targeting factual knowledge are labeled as Facts to evaluate Tamil world knowledge in LLMs. This resource is designed to systematically assess both linguistic competence and culturally grounded knowledge in Tamil.

To build this evaluation resource, we developed a structured pipeline to convert school-level examination materials into a machine-readable dataset. The workflow involves three key stages: (1) collecting question papers from open educational archives, (2) digitizing and cleaning scanned documents, and (3) structuring and filtering the finalized data for experimental use.

4.1. Data Collection

The dataset was built using school-level Tamil language examination questions from Grades 1–13 in Sri Lanka. The papers were sourced from the `Noolaham School`[7] section of `Noolaham.org`[8]. Two latest exam papers were selected per grade during the curation phase to ensure grade-wise coverage and topic diversity. In the digitization process, essay-type and structured questions were excluded, focusing instead on items that yield concrete, verifiable answers.

4.2. Digitization & Cleaning

The exam papers were available only as scanned PDFs because most were typed using non-Unicode Tamil fonts. To obtain machine-readable text, Optical Character Recognition (OCR) was applied using Google Docs[9], chosen for its accessibility and ease of use. Each file was opened directly in Google Docs, which extracted text through its built-in OCR system.

Automated conversion introduced character, spacing, and formatting errors, which were manually corrected. A simple web interface was used to input cleaned questions and ensure the formatting consistency.

4.3. Data Structure & Filtering

To ensure broader task diversity, the dataset includes nine question types, extending beyond traditional Multiple-Choice Questions (MCQs). After automated filtration and manual correction, a final inspection was performed to remove both exact and near-duplicate questions.

Each datapoint was described using six key fields to ensure comprehensive information capture. A detailed description of these fields and their subfields is provided in Table 1. Questions requiring lengthy written responses were adapted into multiple-choice format with non-obvious answer options to facilitate automated evaluation.

It should also be noted that questions related to *Pragmatics*, *Discourse*, *Stylistics* and other higher order were incorporated into the *Semantics* category (L5; see Table 1).

Additionally, the assigned marks for each question were incorporated, as higher marks indicate greater complexity. These weighted scores will be useful for model evaluation.

After the extraction and validation phases were completed, the finalized dataset was exported in JSON format for experimental use.

5. Large Language Model Evaluation

This section outlines the setup, configuration, and methodology used to evaluate multiple LLMs on the curated Tamil question–answer dataset.

5.1. Evaluation Setup and Configuration

We utilized `Abacus.AI`[10] as a unified interface to access both open-source and closed-source models. The list of evaluated models can be found in Table 2.

To guarantee consistency and factuality in responses, the generation temperature was fixed at 0. All other hyperparameters were left at their default values provided by the platform. To prevent information leakage, ground-truth answers were stored in a separate JSON file, locally, leaving only the questions accessible to the models. Each question file was passed systematically to every LLM under evaluation, and the generated responses were stored in separate files for later analysis. This setup allowed multiple models to be evaluated in parallel while preserving data integrity throughout the process.

5.2. Evaluation and Analysis

Model performance was evaluated by comparing each model's responses with the validated ground-truth answers, in the our local machines. Scores were measured at several levels of detail to capture both overall and category-specific performance. In addition, we conducted a classification task where models were prompted to assign each question to one of six predefined categories (L1–L5 and F) in Zero-short settings.

[7] https://noolaham.school/

[8] A digital archive of open Tamil educational resources.

[9] https://docs.google.com/document/

[10] https://chatllm.abacus.ai

Table 1: Design fields and their detailed structure.

Field	Subfield	Description
Paper ID	—	A composite identifier containing the school or institution name, grade level, and year of examination, enabling precise tracking of question origins.
Question Type ID	QT01	Fill in the blanks
	QT02	Provide answer based on the given set of letters/words
	QT03	Order the words/letters
	QT04	Question and Answer
	QT05	Sentence completion
	QT06	Rewrite with punctuation marks
	QT07	Multiple Choice Questions (MCQ)
	QT08	Question and Answer based on given paragraph
	QT09	True or False
Linguistic Category ID	L1	Phonetics
	L2	Phonology
	L3	Morphology
	L4	Syntax
	L5	Semantics
	F	Fact (questions testing knowledge of Tamil cultural, historical, or factual information)
Question Text	—	Question from the examination paper.
Answer	—	The ground truth or correct answer, validated by professional linguists.
Score	—	The point value assigned to each question in the original examination paper, preserved to maintain the weighted importance of different questions.

Table 2: Model catalog grouped by Access and Provider.

Access	Provider	Model
Closed-Source	OpenAI	GPT-5
	Anthropic	Claude Sonnet 4.5
	Google	Gemini 2.5
	xAI	Grok 4
Open-Source	Meta	Llama 4
	DeepSeek	DeepSeek-V3
	Alibaba	Qwen 2.5 72B

Table 3: We report the overall results of each model through Score Percentage (SP). The numbers in parentheses indicate the actual count of correct answers out of 820 questions. Results of the best performing model is made bold.

Model	SP (/820)
Claude Sonnet 4.5	71.09 (579)
DeepSeek-V3	58.04 (491)
Gemini 2.5	**79.55 (659)**
Llama 4	60.67 (501)
OpenAI GPT5	75.94 (633)
Qwen 2.5	37.93 (320)
xAI Grok 4	78.15 (638)

Since not all questions from the original examination papers were included, the number of items differed across grades. To allow fair comparison, the grade-level scores were normalized to a 100-point scale before analysis. We used the following equation to obtain the Score Percentage for each analysis.

$$SP = \frac{S_o}{S_t} \times 100 \tag{1}$$

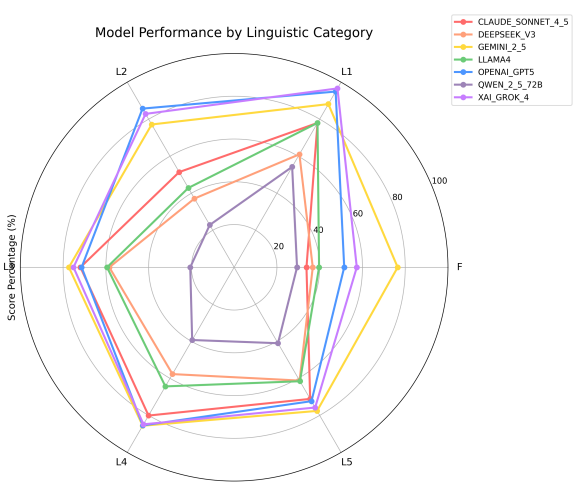
where S_o denotes the total score obtained by the model and S_t denotes the total attainable score.

5.3. Manual Evaluation and Validation

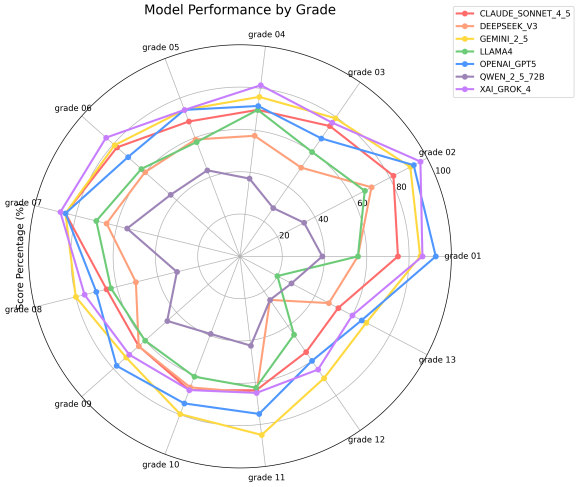
Responses marked as incorrect in the automated evaluation were separated and manually reviewed by trained linguists. The review focused on two main goals: identifying cases where model outputs differed from the reference but were still linguistically acceptable, and capturing valid alternative answers that were not present in the original ground truth. All such verified alternative responses were subsequently incorporated into the final evaluation metrics, ensuring that the reported accuracy more accurately reflected true model performance rather than strict lexical matching.

6. Results and Discussion

We evaluate model performance across four complementary dimensions to obtain a comprehensive understanding of their linguistic and task-level behavior:



(a) Score Percentage (SP) of each model across linguistic categories.



(b) Score Percentage (SP) of each model across grade levels.

Figure 1: Performance comparison of LLMs. (a) Category-wise results showing linguistic variation. (b) Grade-wise results showing variation across difficulty levels.

Table 4: Linguistic category-wise performance of each model through Score Percentage (SP). The numbers in parentheses indicate the actual count of correct answers. The best score in each linguistic category is made bold.

Model	**L1 (20)**	**L2 (32)**	**L3 (75)**	**L4 (169)**	**L5 (512)**	**F (12)**
Claude Sonnet 4.5	77.97 (15)	44.74 (15)	75.34 (53)	79.28 (129)	69.68 (362)	37.50 (5)
DeepSeek-V3	61.02 (13)	36.84 (13)	59.36 (42)	57.55 (99)	59.55 (319)	37.50 (5)
Gemini 2.5	88.14 (17)	71.05 (23)	**79.91 (57)**	**85.51 (143)**	**77.69 (410)**	**75.00 (9)**
Llama 4	77.97 (15)	28.95 (9)	62.10 (45)	63.78 (105)	61.12 (323)	32.50 (4)
OpenAI GPT5	94.92 (19)	78.95 (25)	75.80 (54)	84.51 (142)	72.62 (386)	57.50 (7)
Qwen 2.5 (72B)	54.24 (10)	31.58 (10)	21.92 (17)	40.44 (69)	39.22 (210)	35.00 (4)
xAI Grok 4	**96.61 (19)**	**84.21 (27)**	79.00 (56)	83.90 (136)	75.56 (393)	57.50 (7)
Average score	78.70	53.76	64.78	70.71	65.06	47.50

1. examine overall performance across the full dataset

2. analyze results by linguistic category (L1–L5 and F) to capture category-specific variations (refer Table 4 and Figure 1a)

3. report grade-wise performance to observe how models handle questions of varying complexity (refer Table 5 and Figure 1b)

4. present results from the linguistic category classification task, which assesses the models' ability to identify the underlying linguistic phenomenon in each question.

6.1. Overall Performance

As presented in Table 3, the overall evaluation reveals clear variation in performance across models, with Gemini 2.5 achieving the highest score percentage of approximately 80%, reflecting superior linguistic understanding and factual precision—likely supported by Google's extensive multilingual and high-quality training data. Among all models, the closed-source group consistently occupies the top three ranks, highlighting their advantage in optimization, alignment, and dataset diversity. Among the open-source models, LLaMA 4 performed comparatively well with a score percentage of 60.67%, demonstrating strong generalization ability despite limited access to proprietary data. In contrast, Qwen 2.5, despite being a large-scale 72B model, recorded the lowest score (39.02%), reinforcing that model size alone does not guarantee better performance without effective linguistic grounding and diverse, representative training corpora.

Table 5: We report the Score Percentage (SP) for each grade, with the number of correct answers shown in parentheses. The best result in each grade is highlighted in bold.

Grade	Claude Sonnet 4.5	DeepSeek-V3	Gemini 2.5	Llama 4	OpenAI GPT5	Qwen 2.5	xAI Grok 4
Gr1 (60)	74.74 (41)	55.79 (29)	85.26 (50)	55.79 (29)	**92.63 (55)**	38.95 (21)	86.32 (51)
Gr2 (58)	81.98 (46)	70.27 (41)	90.99 (53)	66.67 (39)	92.79 (54)	34.23 (20)	**96.40 (56)**
Gr3 (90)	74.84 (68)	50.97 (46)	**79.35 (70)**	60.00 (55)	67.74 (62)	27.74 (26)	76.77 (71)
Gr4 (120)	69.75 (91)	57.41 (76)	75.93 (98)	69.75 (89)	71.60 (94)	37.04 (48)	**81.48 (102)**
Gr5 (111)	68.18 (79)	59.09 (73)	**74.03 (89)**	57.79 (71)	**74.03 (88)**	43.51 (52)	**74.03 (88)**
Gr6 (37)	77.69 (29)	60.00 (22)	79.23 (29)	62.31 (23)	70.77 (26)	43.85 (16)	**84.62 (31)**
Gr7 (40)	85.00 (34)	65.00 (26)	85.00 (34)	70.00 (28)	85.00 (34)	55.00 (22)	**87.50 (35)**
Gr8 (40)	65.00 (26)	50.71 (21)	**80.00 (32)**	62.86 (25)	70.00 (28)	30.71 (13)	75.71 (30)
Gr9 (50)	64.00 (32)	64.00 (32)	72.00 (36)	60.00 (30)	**78.00 (39)**	46.00 (23)	70.00 (35)
Gr10 (74)	67.57 (50)	66.22 (49)	**79.73 (59)**	60.81 (45)	74.32 (55)	39.19 (29)	67.57 (50)
Gr11 (80)	63.75 (51)	65.00 (52)	**85.00 (68)**	62.50 (50)	75.00 (60)	42.50 (34)	65.00 (52)
Gr12 (20)	55.00 (11)	25.00 (5)	**70.00 (14)**	45.00 (9)	60.00 (12)	25.00 (5)	65.00 (13)
Gr13 (40)	52.50 (21)	47.50 (19)	**67.50 (27)**	20.00 (8)	65.00 (26)	27.50 (11)	60.00 (24)

6.2. Linguistic-wise Evaluation

Models were also evaluated across linguistic categories to assess how effectively they capture different aspects of linguistic understanding. The results are presented in Table 4 and Figure 1. The best-performing model overall, Gemini 2.5, demonstrated consistent performance across all categories, indicating a balanced grasp of Tamil language structure. In contrast, most other models showed weaker performance in the factual (F) category, which does not fall under linguistic analysis but assesses a model's Tamil world knowledge—including familiarity with poem authors, cultural references, literary works, and historical facts. When considering the linguistic dimensions alone, none of the models exceeded 80%, with Gemini 2.5 achieving the highest scores of 79.91% in L3 and 77.69% in L5, showing its relative strength in linguistic comprehension. The highest score within a linguistic category was observed in L1 (phonetics)(see Figure 1a), achieved by Grok 4 (96.61%), closely followed by GPT-5 (94.92%), reflecting their superior performance in this aspect.

In addition, as noted in previous studies, models continue to perform poorly on phonology and morphology tasks, likely due to the complex and rich morphological structure of Tamil. Although better performance is generally expected on semantic tasks—since models can leverage contextual information to infer meaning—the overall scores remain low. This may also be attributed to the fact that the semantic test set also contains pragmatic questions, which introduce additional challenges for Tamil.

6.3. Grade-wise Evaluation

Table 5 presents the grade-wise evaluation results of all LLMs, while Figure 1b illustrates the overall performance trend across grades. The grade-wise score percentage analysis shows that all models performed relatively well in Grades 1 and 2, which is expected since the questions at these levels are simpler and focus more on basic language skills that can be easily captured from training data. As the grade level increases, particularly from Grade 5 onwards, a noticeable decline in performance is observed across models. This corresponds to the increasing linguistic and conceptual complexity of questions that require a stronger command of Tamil grammar, vocabulary, and linguistic structure. The lowest scores appear around Grades 5 and 13, which align with national-level examinations in Sri Lanka, where the questions are more challenging and require precise linguistic understanding. Even though Grade 11 (G.C.E. O/L) is also a national exam, it is less competitive, which is reflected in slightly better scores. We see that xAI Grok 4 performing best for lower grades 1-7, while Gemini 2.5 outperforms all he models at higher grades (8-13). Among all models, Gemini 2.5 maintained consis-

tently high performance across grades, achieving above 67% even at the higher levels, reflecting its stronger adaptability to linguistic variation and question complexity compared to other models.

Table 6: Overall Accuracy Percentage (AP) for the linguistic classification task. Numbers in parentheses show correctly classified tags out of 820. Best-performing model is in bold.

Model	SP (/820)
Claude Sonnet 4.5	44.02 (361)
DeepSeek V3	27.93 (229)
Gemini 2.5	52.07 (427)
LLaMA 4	51.59 (423)
OpenAI GPT-5	**65.61 (538)**
Qwen 2.5 72B	52.07 (427)
xAI Grok 4	61.59 (505)

6.4. Linguistic Category Classification

In addition to the primary evaluation task, an additional experiment was conducted to better understand each model's ability to capture linguistic awareness (refer Table 6). In this setup, the models were instructed to assign an appropriate linguistic tag (from L1 to L5) to each question, with any item not fitting a linguistic category explicitly directed to be tagged as F. A notable observation emerged from this experiment: while Gemini 2.5 performed exceptionally well in answering questions during the main evaluation, it ranked second in this linguistic categorization task, falling behind GPT-5, with 111 fewer correctly tagged questions. This finding reinforces the argument that Gemini 2.5's strong performance may not stem from genuine linguistic understanding, but rather from its extensive training data coverage. Since the evaluation questions are based on Tamil school examination papers, they are likely to follow predictable patterns and exhibit limited novelty, making them easier for Gemini to match with seen data. In contrast, the linguistic tagging task requires deeper analytical ability and true understanding of linguistic structures—skills that cannot be derived from memorized or surface-level data. In this respect, GPT-5 demonstrated stronger linguistic awareness and interpretive precision.

7. Copyrights

8. Conclusion

We introduce *ILAKKANAM*, a Tamil linguistic benchmark dataset consisting of 820 manually curated questions from Sri Lankan school-level Tamil grade-wise examination papers. We perform an extensive evaluation on both closed-source and open-source LLMs and show that there is a clear gap between them in terms of Tamil linguistic performance. We analyze the overall results, discuss model performance across linguistic categories and grade levels, and present observations from the linguistic categorization task. Our studies show that Gemini 2.5 performs best on our benchmark dataset. We find no clear relationship between a model's performance on the linguistic tasks and it's ability to classify these questions into their respective linguistic categories, further suggesting that LLM performance may not stem from linguistic understanding but rather from broad exposure to training data. We hope that both the *ILAKKANAM* dataset and our analyses help researchers better understand the limitations of current models and encourage further efforts toward evaluating and benchmarking LLMs for the Tamil language. Dataset will be provided upon request to ensure no data leakage.

9. Limitation

First, only two recent papers were selected from each grade level, which may not fully capture the breadth and diversity of Tamil linguistic phenomena. As a result, certain aspects of language use and structure may not be adequately represented in the current dataset. We are actively expanding the question bank to improve linguistic coverage and representation.

Second, our analysis focused on five core linguistic dimensions: phonetics, phonology, morphology, syntax, and semantics. Extended linguistic areas, such as pragmatics and stylistics, were grouped under semantics because of their limited presence in the dataset. This grouping may reduce the granularity of analysis for these higher-level aspects.

10. Bibliographical References

A M Abirami, Wei Qi Leong, Hamsawardhini Rengarajan, D Anitha, R Suganya, Himanshu Singh, Kengatharaiyer Sarveswaran, William Chandra Tjhi, and Rajiv Ratn Shah. 2024. Aalamaram: A large-scale linguistically annotated treebank for the Tamil language. In *Proceedings of the 7th Workshop on Indian Language Data: Resources and Evaluation*, pages 73–83, Torino, Italia. ELRA and ICCL.

Lucas Bandarkar, Davis Liang, Benjamin Muller, Mikel Artetxe, Satya Narayan Shukla, Donald Husa, Naman Goyal, Abhinandan Krishnan, Luke Zettlemoyer, and Madian Khabsa. 2024. The belebele benchmark: a parallel reading comprehension dataset in 122 language variants. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 749–775, Bangkok, Thailand. Association for Computational Linguistics.

Gasper Begus, Maksymilian Dabkowski, and Ryan Rhodes. 2025. Large linguistic models: Investigating llms' metalinguistic abilities. *IEEE Transactions on Artificial Intelligence*, page 1–15.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Wei Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qiang Yang, and Xing Xie. 2023. A survey on evaluation of large language models.

Claude. 2025. Introducing Claude Sonnet 4.5 — anthropic.com. https://www.anthropic.com/news/claude-sonnet-4-5. Accessed 2025-10-03.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, Luke Marris, Sam Petulla, Colin Gaffney, ..., and Wesley Helmholz. 2025. Gemini 2.5: Pushing the Frontier with Advanced Reasoning, Multimodality, Long Context, and Next Generation Agentic Capabilities.

DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, ..., and Zizheng Pan. 2025. Deepseek-v3 technical report.

Xiyan Fu and Wei Liu. 2025. How Reliable is Multilingual LLM-as-a-Judge?

Omid Ghahroodi, Marzia Nouri, Mohammad V. Sanian, Alireza Sahebi, Doratossadat Dastgheib, Ehsaneddin Asgari, Mahdieh Soleymani Baghshah, and Mohammad Hossein Rohban. 2024. Khayyam Challenge (PersianMMLU): Is Your LLM Truly Wise to The Persian Language? *ArXiv*, abs/2404.06644.

Grok4. 2025. Grok 4. https://x.ai/news/grok-4. Accessed 2025-10-05.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Saizhuo Wang, Kun Zhang, Yuanzhuo Wang, Wen Gao, Lionel Ni, and Jian Guo. 2025. A Survey on LLM-as-a-Judge.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*.

Meng Ji, Meng Ji, Pierrette Bouillon, and Mark Seligman. 2023. *Cultural and Linguistic Bias of Neural Machine Translation Technology*, Studies in Natural Language Processing, page 100–128. Cambridge University Press.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Fajri Koto, Nurul Aisyah, Haonan Li, and Timothy Baldwin. 2023. Large language models only pass primary school exams in Indonesia: A comprehensive test on IndoMMLU. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12359–12374, Singapore. Association for Computational Linguistics.

Fajri Koto, Haonan Li, Sara Shatnawi, Jad Doughman, Abdelrahman Sadallah, Aisha Alraeesi, Khalid Almubarak, Zaid Alyafeai, Neha Sengupta, Shady Shehata, Nizar Habash, Preslav Nakov, and Timothy Baldwin. 2024. ArabicMMLU: Assessing massive multitask language understanding in Arabic. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5622–5640, Bangkok, Thailand. Association for Computational Linguistics.

Wei Qi Leong, Jian Gang Ngui, Yosephine Susanto, Hamsawardhini Rengarajan, Kengatharaiyer Sarveswaran, and William Chandra Tjhi. 2023. BHASA: A Holistic Southeast Asian Linguistic and Cultural Evaluation Suite for Large Language Models.

Haonan Li, Yixuan Zhang, Fajri Koto, Yifei Yang, Hai Zhao, Yeyun Gong, Nan Duan, and Timothy Baldwin. 2024. CMMLU: Measuring massive multitask language understanding in Chinese. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 11260–11285, Bangkok, Thailand. Association for Computational Linguistics.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang

Yuan, Bobby Yan, Ce Zhang, Christian Alexander Cosgrove, Christopher D Manning, Christopher Re, Diana Acosta-Navas, Drew Arad Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue WANG, Keshav Santhanam, Laurel Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri S. Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Andrew Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. 2023. Holistic evaluation of language models. *Transactions on Machine Learning Research*. Featured Certification, Expert Certification.

IM Kamala Liyanage. 2014. Education system of sri lanka: strengths and weaknesses. *Educ Syst Sri Lanka*, pages 116–40.

Llama4Herd. 2025. The llama 4 herd: The beginning of a new era of natively multimodal ai innovation. https://ai.meta.com/blog/llama-4-multimodal-intelligence/. Accessed 2025-10-01.

OpenAI. 2022. ChatGPT: Optimizing Language Models for Dialogue. Accessed: 2024-08-15.

OpenAI. 2025. OpenAI Introducing GPT5. https://openai.com/index/introducing-gpt-5/. Accessed 2025-10-22.

Ashmari Pramodya, Nirasha Nelki, Heshan Shalinda, Chamila Liyanage, Yusuke Sakai, Randil Pushpananda, Ruvan Weerasinghe, Hidetaka Kamigaito, and Taro Watanabe. 2025. SinhalaMMLU: A Comprehensive Benchmark for Evaluating Multitask Language Understanding in Sinhala.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2025. Qwen2.5 technical report.

Shivalika Singh, Angelika Romanou, Clémentine Fourrier, David Ifeoluwa Adelani, Jian Gang Ngui, Daniel Vila-Suero, Peerat Limkonchotiwat, Kelly Marchisio, Wei Qi Leong, Yosephine Susanto, Raymond Ng, Shayne Longpre, Sebastian Ruder, Wei-Yin Ko, Antoine Bosselut, Alice Oh, Andre Martins, Leshem Choshen, Daphne Ippolito, Enzo Ferrante, Marzieh Fadaee, Beyza Ermis, and Sara Hooker. 2025. Global MMLU: Understanding and addressing cultural and linguistic biases in multilingual evaluation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 18761–18799, Vienna, Austria. Association for Computational Linguistics.

Guijin Son, Hanwool Lee, Sungdong Kim, Seungone Kim, Niklas Muennighoff, Taekyoon Choi, Cheonbok Park, Kang Min Yoo, and Stella Biderman. 2025. KMMLU: Measuring massive multitask language understanding in Korean. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4076–4104, Albuquerque, New Mexico. Association for Computational Linguistics.

Yosephine Susanto, Adithya Venkatadri Hulagadri, Jann Railey Montalan, Jian Gang Ngui, Xianbin Yong, Wei Qi Leong, Hamsawardhini Rengarajan, Peerat Limkonchotiwat, Yifan Mai, and William Chandra Tjhi. 2025. SEA-HELM: Southeast Asian holistic evaluation of language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 12308–12336, Vienna, Austria. Association for Computational Linguistics.

Ashima Suvarna, Harshita Khandelwal, and Nanyun Peng. 2024. PhonologyBench: Evaluating phonological skills of large language models. In *Proceedings of the 1st Workshop on Towards Knowledgeable Language Models (KnowLLM 2024)*, pages 1–14, Bangkok, Thailand. Association for Computational Linguistics.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. *Advances in neural information processing systems*, 32.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. 2020. BLiMP: The benchmark of linguistic minimal pairs for English. *Transactions of the Association for Computational Linguistics*, 8:377–392.

Ishaan Watts, Varun Gumma, Aditya Yadavalli, Vivek Seshadri, Manohar Swaminathan, and Sunayana Sitaram. 2024. PARIKSHA: A large-scale investigation of human-LLM evaluator agreement on multilingual and multi-cultural data. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7900–7932, Miami, Florida, USA. Association for Computational Linguistics.

Arda Yüksel, Abdullatif Köksal, Lütfi Kerem Senel, Anna Korhonen, and Hinrich Schuetze. 2024. TurkishMMLU: Measuring massive multitask language understanding in Turkish. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7035–7055, Miami, Florida, USA. Association for Computational Linguistics.

Yuan Zhang, Jason Baldridge, and Luheng He. 2019. PAWS: Paraphrase adversaries from word scrambling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1298–1308, Minneapolis, Minnesota. Association for Computational Linguistics.