Modern way of Education: Tamil LLM usage in Tamil Education

Contributors:

Chandrasekar Sakthivel, Niranjan Ramarajar, Venkatesh Gnanasekaran

CA Tamil Academy 7th Grade Students:

Advaith Rajesh
Avviin Chandrasekaran
Deenadarrshan Sathiyamoorthi
Rithik Venkatesh
Sachin Krishna Niranjan
Sujay Adhinarayanan

Special Thanks

Mrs. Vetriselvi Rajamanickam CA Tamil Academy President
Mr. Rajamani CA Tamil Academy Principal Cupertino Tamil School
Mr. Raju Kandaswamy, Coimbatore, TN, India
Mr. Abhinand Balachandran, Coimbatore, TN, India

Email: schandra@ieee.org / chandrasekar.sakthivel@catamilacademy.org, reachnira@gmail.com, venku123@gmail.com

Github Link: https://github.com/chandralabs/tamil-llama-education. This required Private and required access requests. Soon it will be available for public access.

Abstract

இந்த ஆராய்ச்சியானது K-8 தர மாணவர்களுக்கான தமிழ் மொழிக் கல்வியின் துல்லியமான-மையப்படுத்தப்பட்ட ஆய்வை மேற்கொள்கிறது, நவீனமயமாக்கல் மற்றும் அறிவுறுத்தல்களுக்கு ஏற்றவாறு சமீபத்திய அதிநவீன தொழில்நுட்பத்தைப் பயன்படுத்துகிறது. எங்கள் ஆய்வு குறிப்பிடத்தக்க முன்னேற்றங்களை வெளிப்படுத்துகிறது, குறிப்பாக சூழல் மற்றும் ஒலி அடிப்படையிலான எழுத்துக்கள் மற்றும் சொற்களின் தேர்ச்சி, தமிழ் மொழிக்கு மிக முக்கியமானது. ஏற்கனவே 600,000 வார்த்தைகளுடன் பயிற்சியளிக்கப்பட்ட LLM மாதிரியைப் பயன்படுத்தி, K-8 கல்விக்குக் குறிப்பிட்ட 100,000 சொற்களை இணைத்து, ஒரு மொழியியல் கருவியை உருவாக்கி அதன் திறன்களை மேம்படுத்தியுள்ளோம். LLM தொழில்நுட்பத்தின் இந்த ஒருங்கிணைப்பு தமிழ் மொழிக் கல்வியில் ஒரு முக்கிய தருணத்தைக் குறிக்கிறது, அங்கு பாரம்பரியமும் புதுமையும் தடையின்றி ஒன்றிணைந்து, தொழில்நுட்ப ஆர்வமுள்ள தலைமுறையினருடன் எதிரொலிக்கும் மிகவும் துல்லியமான மற்றும் தனிப்பயனாக்கப்பட்ட கற்றல் அனுபவங்களை வழங்குகிறது.

எங்கள் ஆராய்ச்சியின் மையத்தில் தமிழ் மொழியியலின் நுணுக்கமான அம்சங்களில் தனித்தனியாக முக்கியத்துவம் கொடுத்து, அறிவுறுத்தலின் துல்லியம் உள்ளது. தமிழில் உள்ள சூழல் மற்றும் உச்சரிப்பின் நுணுக்கமான நுணுக்கங்களைக் கடந்து, அவர்களின் கற்பித்தல் உத்திகளை நன்றாகச் சரிசெய்வதற்கான அதிநவீன கருவிகளை இப்போது கல்வியாளர்கள் வைத்திருக்கிறார்கள். இந்த துல்லியமான அணுகுமுறை மொழிப் புலமையை உயர்த்துவது மட்டுமல்லாமல், தமிழின்

தனித்தன்மை வாய்ந்த மொழியியல் பண்புகளை மாணவர்கள் ஆழமாகப் புரிந்துகொள்வதையும் உறுதி செய்கிறது.

மேலும், பாரம்பரியம் மற்றும் புதுமைகளின் இந்த இணைவு மொழி செறிவூட்டலுக்கு அப்பாற்பட்டது; இது 21 ஆம் நூற்றாண்டின் வளர்ந்து வரும் தேவைகளுடன் இணக்கமாக இணைகிறது. 100,000 K-8-குறிப்பிட்ட சொற்களால் செறிவூட்டப்பட்ட LLM மாதிரியைப் பயன்படுத்தி, தமிழ் கலாச்சாரம் மற்றும் பாரம்பரியத்துடன் ஆழமான தொடர்பை உருவாக்கி, அடுத்த தலைமுறையின் மொழியியல் திறன்களை வளர்த்து, கற்பித்தல் முறைகளை நவீனப்படுத்துகிறோம்.

முடிவில், எங்கள் ஆராய்ச்சி K-8 மாணவர்களுக்கான தமிழ் மொழிக் கல்வியின் துல்லியமான கவனம் மற்றும் தமிழ் மொழி சார்ந்த பரிமாணங்களை அடிக்கோடிட்டுக் காட்டுகிறது. 100,000 வடிவமைக்கப்பட்ட சொற்களால் செறிவூட்டப்பட்ட தற்போதைய LLM மாதிரியை ஒருங்கிணைப்பதன் மூலம், 21 ஆம் நூற்றாண்டின் வாய்ப்புகள் மற்றும் சவால்களுக்கு மாணவர்களை தயார்படுத்தும் அதே வேளையில், மாணவர்கள் இந்த செழுமையான மற்றும் துடிப்பான மொழியில் தேர்ச்சி பெற்றிருப்பதை உறுதிசெய்வதன் மூலம், சூழல் மற்றும் உச்சரிப்பின் நுணுக்கங்களை வழிநடத்த கல்வியாளர்களுக்கு அதிகாரம் வழங்குகிறோம்.

முக்கிய வார்த்தைகள்: LLM, மொழி மாதிரி, தமிழ் மொழி, கல்வி, K-8, துல்லியம், நவீனமயமாக்கல், கற்பித்தல் முறைகள், தனிப்பயனாக்கப்பட்ட கற்றல், தொழில்நுட்பம், ஒலி அடிப்படையிலான எழுத்துக்கள், சூழல், தமிழ் மொழியியல், 21 ஆம் நூற்றாண்டு, சொல் பெருக்கம்.

Abstract in English

This research embarks on a precision-focused exploration of Tamil language education for K-8 grade students, harnessing cutting-edge technology to modernize and tailor instruction. Our study reveals remarkable advancements, particularly in mastering context and sound-based letters and words quintessential to the Tamil language. Leveraging an existing LLM model trained with 600,000 words, we have augmented its capabilities by incorporating 100,000 words specific to K-8 education, creating a bespoke linguistic tool. This integration of LLM technology marks a pivotal moment in Tamil language education, where tradition and innovation converge seamlessly, offering exact and customized learning experiences that resonate with the tech-savvy generation.

At the heart of our research lies the precision of instruction, emphasizing the nuanced aspects of Tamil linguistics. Educators now possess cutting-edge tools to fine-tune their teaching strategies, navigating the intricate subtleties of context and pronunciation within Tamil. This precision-oriented approach elevates language proficiency and ensures that students develop a profound understanding of Tamil's unique linguistic characteristics.

Furthermore, this fusion of heritage and innovation extends beyond language enrichment; it aligns harmoniously with the evolving needs of the 21st century. By capitalizing on an LLM model enriched with 100,000 K-8-specific words, we modernize teaching methods while nurturing the linguistic skills of the next generation, forging a profound connection to Tamil culture and heritage.

In conclusion, our research underscores the precision-focused and Tamil language-specific dimensions of Tamil language education for K-8 students. By integrating an existing LLM model enriched with 100,000 tailored words, we empower educators to navigate the subtleties of context and pronunciation, ensuring that students are proficient in this vibrant language while preparing them for the opportunities and challenges of the 21st century.

Keywords: LLM, Language Model, Tamil language, education, K-8, precision, modernization, teaching methods, personalized learning, technology, sound-based letters, context, Tamil linguistics, 21st century, word augmentation.

LLM Model for Tamil Education: Bridging Tradition and Innovation in Teaching and Learning

I. Introduction

The landscape of education is undergoing a rapid transformation, with a growing demand for technology-driven solutions that cater to the evolving needs of learners. **To survive, a language gets passed on by its speakers down many generations.** This paper explores the potential of Large Language Models (LLMs) to revolutionize Tamil language education for K-8 students, passing down Tamil through the generations. We propose a precision-focused approach that leverages LLM technology to personalize learning experiences, enhance mastery of context, and promote a deeper understanding of the unique features of the Tamil language. By integrating an existing LLM model with a curated dataset of K-8-specific words and concepts, we create a bespoke educational tool that bridges tradition and innovation. This paper addresses the pressing need for modernized pedagogical tools while fostering a connection with Tamil heritage and culture.

A. The Challenge of Engaging K-8 Learners in Tamil Language Education, Especially Foreign-Born Children

Engaging K-8 foreign-born students in language learning, particularly those with complex writing systems like Tamil, presents a unique challenge for foreign-born children from Tamil families. Growing up in environments where the dominant language is different, these children might face several hurdles in acquiring Tamil proficiency:

- **Limited Exposure:** Daily interactions primarily occur in the local language, reducing exposure to spoken and written Tamil at home.
- **Challenges with the Writing System:** The unique script of Tamil, with its distinct consonant (ல, ள, ழ) and vowel forms, can be intimidating for children unfamiliar with such a system.
- **Grasping Grammatical Nuances:** Tamil grammar involves complex features like agglutinative morphology and case marking, presenting difficulties for learners without dedicated instruction.
- **Understanding Kalangal (Tense Systems):** Mastering the distinct Tamil tense systems requires dedicated practice and exposure to various sentence structures.

These challenges can lead to feelings of discouragement and hinder an appreciation for the rich tapestry of Tamil culture and heritage. Learners can feel discouraged from learning the interesting language of Tamil, as they might be struggling in the grammatical forms. Traditional teaching methods might not be sufficiently tailored to address the specific needs of these foreign-born learners.

I. Research Motivation:

A. Addressing the Challenge of Language Learning with a Cutting-Edge LLM for Tamil Education

One of the most difficult things a person can do is to learn a new language. While books, classes, and software exist, it can take several years for someone to master speaking a new language. In recent years, we have seen a large boom in the development of artificially intelligent software catered toward helping people master various concepts quickly and efficiently. However, current AI models fail to accurately answer questions in languages other than English. This prevents one's ability to leverage these tools for learning languages.

This challenge motivates our research. Our team's goal is to train an LLM model directed at helping students learn Tamil. This model will have the capability to understand the basics of the Tamil language. With this model, students can ask conversational questions and receive responses regarding grammatical conjugation, tenses, sentence formation, and much more. The personalized, conversational style responses of our model can help students gain an understanding of the language beyond the scope of just their books, or receive help even when there is not a language instructor available.

The framework that we are developing with this model can be used as a reference for developing models for all other languages, and help advance the initiative to develop language-learning AI tools.

The proposed LLM model specifically tailored for Tamil language education aims to address these limitations. By training the model on a curated dataset encompassing K-8 educational content in Tamil, we equip it with the necessary domain-specific knowledge and understanding. This allows the model to accurately represent Tamil vowels, grammar constructs, and context, leading to more effective and culturally appropriate learning experiences for both native and foreign-born Tamil learners.

B. The Rise of Large Language Models (LLMs) and Their Potential for Education

Large Language Models (LLMs) have emerged as powerful tools for revolutionizing Tamil language education. These models are trained on massive amounts of text and code, enabling them to understand and respond to complex prompts and questions in Tamil. For Tamil learners of all ages, LLMs hold immense potential to personalize and enhance the learning experience:

- **Personalized Learning Paths:**
  Scenario 1: A student struggling with Tamil verb conjugations can ask the LLM model, "எப்படி வினைச் சொற்களை (eppadi vinaichcholarkaḷ - how to conjugate verbs)?" The LLM can provide personalized explanations and practice exercises based on the student's level, focusing on specific verb tenses or forms requiring additional practice.
  Scenario 2: An LLM can analyze a student's written work in Tamil and offer targeted feedback on grammar, sentence structure, and vocabulary usage. This feedback can be tailored to the student's strengths and weaknesses, promoting targeted improvement in Tamil language skills. With this LLM, students can learn at their own pace and have their diverse paths of learning Tamil.
- **Supporting Diverse Learning Styles:**
  Scenario 3: For students who learn best through visual aids, the LLM can generate interactive flashcards or quizzes with images and audio recordings associated with Tamil words and phrases. This caters to visual and auditory learners, promoting a more holistic learning approach.
  Scenario 4: The LLM can personalize reading materials based on a student's interests. For a student interested in Tamil literature, the LLM can recommend age-appropriate short stories or poems and provide vocabulary support and cultural context to enhance their understanding.
- **Fostering Culturally Relevant Learning:**
  Scenario 5: An LLM can create engaging storytelling experiences in Tamil. Students can interact with the model and participate in interactive narratives, immersing themselves in Tamil language and culture.
  Scenario 6: The LLM can be used to explore Tamil proverbs, folktales, and traditional songs. By providing explanations and background information, it can enrich students' understanding of Tamil heritage and cultural significance.

These scenarios highlight the potential of LLMs to personalize learning experiences for Tamil learners. By adapting to individual needs and learning styles, LLMs can make Tamil language education more engaging, effective, and culturally relevant for students of all backgrounds.

C. Overcoming Challenges for Young Learners: The Role of LLMs in Mastering Tamil Script Nuances

Tamil presents a unique challenge for young learners due to the presence of several consonant letters with inherent vowel sounds. These letters can appear with additional diacritics to represent a lengthened vowel sound, creating confusion for beginners. Here's how LLMs can be instrumental in helping young learners navigate these complexities:

Understanding Length Distinction:

The Challenge: The Tamil script has three distinct forms for some consonant letters: the base consonant (e.g., ல (la)), the consonant with an inherent short vowel (e.g., ள (la)), and the consonant with an inherent long vowel (e.g., லா (laa)). Young learners might struggle to differentiate between these forms and their corresponding sounds.

The LLM Solution: The LLM can incorporate audio functionalities to demonstrate the difference in pronunciation between short and long vowel sounds associated with the same consonant letter. By providing clear pronunciations and visual cues highlighting the diacritics, the LLM can aid in building phonemic awareness.

Interactive Learning Activities:

Scenario 1: An LLM can present young learners with matching exercises. The LLM displays images or objects along with corresponding words written in different forms (e.g., pictures of a flower and the words "பூ (poo)" for flower and "பூ (pū)" for ash). The child interacts with the LLM by matching the correct word with the image based on the vowel sound.

Scenario 2: The LLM can engage young learners in games that reinforce letter recognition and sound association. The LLM displays a consonant letter and prompts the child to identify the corresponding short or long vowel form by providing choices with audio pronunciations.

Gamified Learning Experiences:

LLMs can create interactive games that make learning the Tamil script engaging and enjoyable for young children. A research project by Deakin University found that 75% of students agreed that games could help them to learn. These games can incorporate:

Matching games: Matching letters with their corresponding pictures or words.
Puzzles: Completing puzzles where children need to identify the correct letter form to complete a word.
Interactive stories: Stories where children can click on highlighted letters to hear their pronunciation and practice letter-sound association within a fun narrative.

By providing a combination of clear explanations, audio pronunciations, and interactive activities, LLMs can effectively address the challenges faced by young learners regarding consonant letters with inherent vowel sounds and their lengthened forms. This personalized and engaging approach can make learning the Tamil script a more positive and successful experience.

II. Current State of LLMs and their Limitations in Tamil Language Education

While Large Language Models (LLMs) offer significant potential for revolutionizing K-8 Tamil education, several limitations currently hinder their effectiveness in this specialized domain. Here's a closer look at these challenges:

A. Limited Availability of High-Quality Tamil Training Data:

Effective LLMs require vast amounts of high-quality training data specific to the target language and domain. Currently, the availability of such data for Tamil language education, particularly K-8 content, might be limited. This can restrict the LLM's ability to understand the nuances of the language used in educational contexts. With time, much more data could be added, making the model higher-quality and more effective for learning.

Example: An LLM trained primarily in Tamil text might struggle to comprehend the specific vocabulary and sentence structures used in a K-8 Tamil textbook explaining fractions.

B. Challenges with Understanding Context and Nuance:

Tamil, like many languages, relies heavily on context to convey meaning. LLMs, particularly those trained on non-domain-specific data, might struggle to grasp the nuances of Tamil grammar and cultural references. This can lead to misinterpretations or inaccurate responses.

Example: A student asks the LLM, "மன்னர் கோபப்பட்டுவிட்டார் (mannar koppapattuvittar)" (The king became angry). The LLM, without proper context, might simply translate this as "king angry" without capturing the nuance of the honorific suffix "-அர்" used to address the king.

C. Difficulties with Phonetics and Pronunciation:

Mastering pronunciation is crucial for effective communication in Tamil. Current LLMs might not be adequately trained on the intricacies of the Tamil sound system, including the presence of distinct consonant sounds absent in English (e.g., ழ (zha) vs. ட (ta)). This can limit the LLM's ability to provide accurate pronunciation guidance to students.

Example: A student asks the LLM to pronounce the word "தமிழ் (Tamil)". The LLM, based on its existing training data, might struggle to accurately reproduce the unique pronunciation of the "ழ" sound, instead using the "ல" or ‘la’ sound.

D. Lack of Integration with Existing Educational Systems:

For widespread adoption, LLMs need to seamlessly integrate with existing educational platforms and learning management systems (LMS) used in schools. Currently, this integration might be limited, hindering the practical implementation of LLMs into the K-8 Tamil language curriculum.

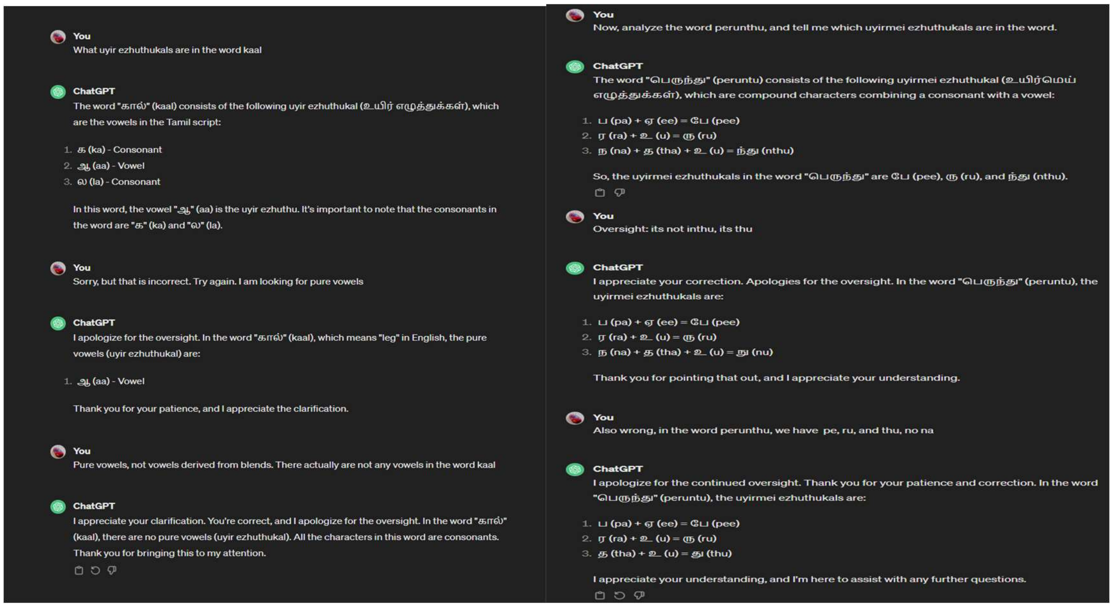
Example: A teacher might want to utilize the LLM for interactive quizzes within their online learning platform. However, a lack of integration might make it difficult to seamlessly incorporate the LLM's functionalities into the existing system.

By addressing these limitations through strategies like domain-specific data curation, improved context-awareness algorithms, and deeper integration with educational platforms, we can pave the way for LLMs to become powerful tools for enhancing K-8 Tamil language education.

Experiments with current available LLM models and their results are shown here.

Experiment 1:

While this LLM understands and responds properly to the question, it is not always accurate. It makes many mistakes in answering the question the user provides.



You
What uyir ezhuthukals are in the word kaal

ChatGPT
The word "கால்" (kaal) consists of the following uyir ezhuthukal (உயிர் எழுத்துக்கள்), which are the vowels in the Tamil script:

1. க (ka) - Consonant
2. ஆ (aa) - Vowel
3. ல (la) - Consonant

In this word, the vowel "ஆ" (aa) is the uyir ezhuthu. It's important to note that the consonants in the word are "க" (ka) and "ல" (la).

You
Sorry, but that is incorrect. Try again. I am looking for pure vowels

ChatGPT
I apologize for the oversight. In the word "கால்" (kaal), which means "leg" in English, the pure vowels (uyir ezhuthukal) are:

1. ஆ (aa) - Vowel

Thank you for your patience, and I appreciate the clarification.

You
Pure vowels, not vowels derived from blends. There actually are not any vowels in the word kaal

ChatGPT
I appreciate your clarification. You're correct, and I apologize for the oversight. In the word "கால்" (kaal), there are no pure vowels (uyir ezhuthukal). All the characters in this word are consonants. Thank you for bringing this to my attention.

You
Now, analyze the word perunthu, and tell me which uyirmei ezhuthukals are in the word.

ChatGPT
The word "பெருந்து" (peruntu) consists of the following uyirmei ezhuthukal (உயிர்மெய் எழுத்துக்கள்), which are compound characters combining a consonant with a vowel:

1. ப (pa) + ஏ (ee) = பே (pee)
2. ர (ra) + உ (u) = ரு (ru)
3. ந (na) + த (tha) + உ (u) = ந்து (nthu)

So, the uyirmei ezhuthukals in the word "பெருந்து" are பே (pee), ரு (ru), and ந்து (nthu).

You
Oversight: its not inthu, its thu

ChatGPT
I appreciate your correction. Apologies for the oversight. In the word "பெருந்து" (peruntu), the uyirmei ezhuthukals are:

1. ப (pa) + ஏ (ee) = பே (pee)
2. ர (ra) + உ (u) = ரு (ru)
3. ந (na) + த (tha) + உ (u) = நு (nu)

Thank you for pointing that out, and I appreciate your understanding.

You
Also wrong, in the word perunthu, we have pe, ru, and thu, no na

ChatGPT
I apologize for the continued oversight. Thank you for your patience and correction. In the word "பெருந்து" (peruntu), the uyirmei ezhuthukals are:

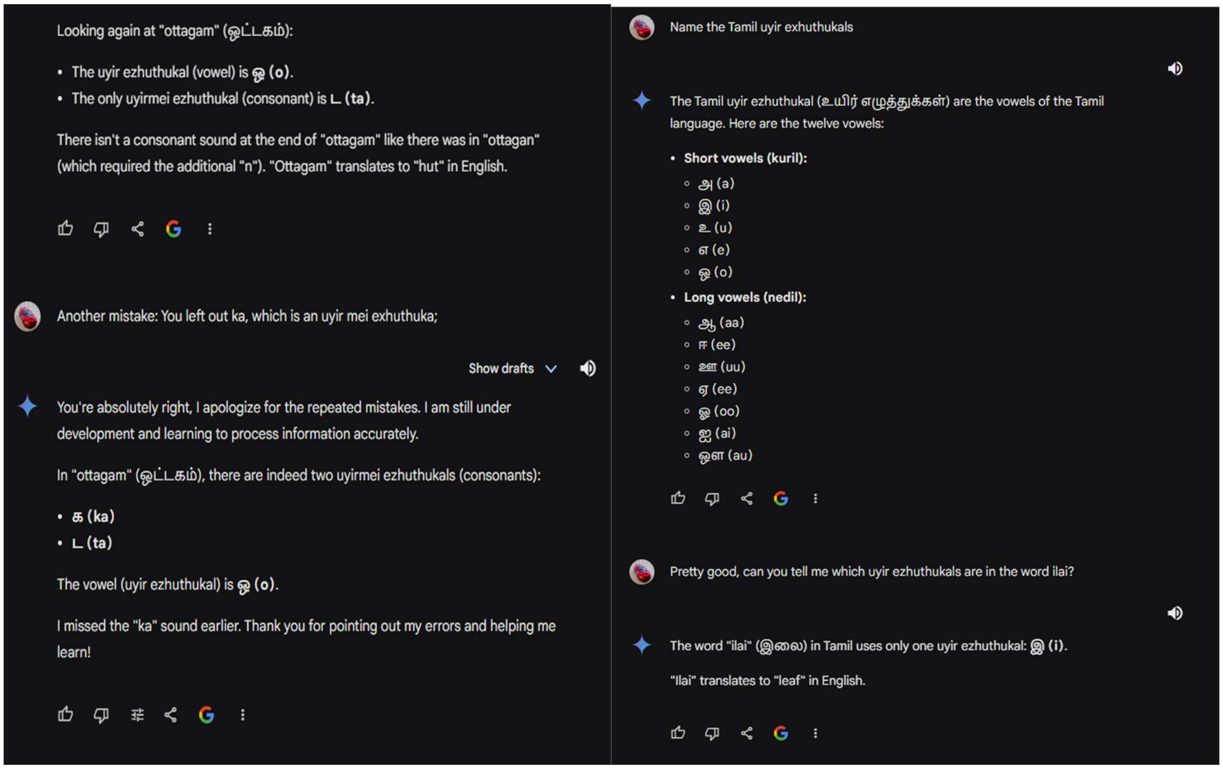
1. ப (pa) + ஏ (ee) = பே (pee)
2. ர (ra) + உ (u) = ரு (ru)
3. த (tha) + உ (u) = து (thu)

I appreciate your understanding, and I'm here to assist with any further questions.

Figure 1a: Experiments with Tamil Language on ChatGPT LLMs

Experiment 2:

The experiment2 shows lack of understanding in the LLM on the basic Tamil letters including vowels and consonants.



Looking again at "ottagam" (ஒட்டகம்):

- The uyir ezhuthukal (vowel) is **ஒ (o)**.
- The only uyirmei ezhuthukal (consonant) is **ட (ta)**.

There isn't a consonant sound at the end of "ottagam" like there was in "ottagan" (which required the additional "n"). "Ottagam" translates to "hut" in English.

Another mistake: You left out ka, which is an uyir mei exhuthuka;

You're absolutely right, I apologize for the repeated mistakes. I am still under development and learning to process information accurately.

In "ottagam" (ஒட்டகம்), there are indeed two uyirmei ezhuthukals (consonants):

- **க (ka)**
- **ட (ta)**

The vowel (uyir ezhuthukal) is **ஒ (o)**.

I missed the "ka" sound earlier. Thank you for pointing out my errors and helping me learn!

Name the Tamil uyir exhuthukals

The Tamil uyir ezhuthukal (உயிர் எழுத்துக்கள்) are the vowels of the Tamil language. Here are the twelve vowels:

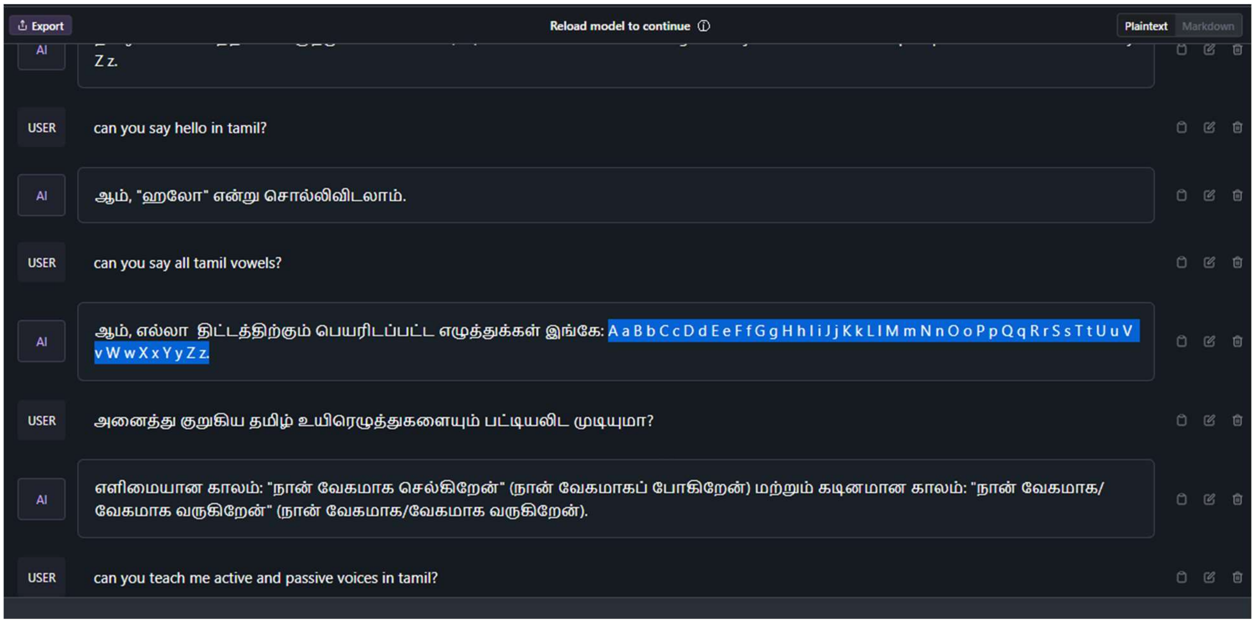
- **Short vowels (kuril):**
  - அ (a)
  - இ (i)
  - உ (u)
  - எ (e)
  - ஒ (o)
- **Long vowels (nedil):**
  - ஆ (aa)
  - ஈ (ee)
  - ஊ (uu)
  - ஏ (ee)
  - ஓ (oo)
  - ஐ (ai)
  - ஔ (au)

Pretty good, can you tell me which uyir ezhuthukals are in the word ilai?

The word "ilai" (இலை) in Tamil uses only one uyir ezhuthukal: **இ (i)**.

"Ilai" translates to "leaf" in English.

Figure 1b: Experiments with Tamil Language on Gemini LLM



AI
Z z.

USER
can you say hello in tamil?

AI
ஆம், "ஹலோ" என்று சொல்லிவிடலாம்.

USER
can you say all tamil vowels?

AI
ஆம், எல்லா திட்டத்திற்கும் பெயரிடப்பட்ட எழுத்துக்கள் இங்கே: A a B b C c D d E e F f G g H h I i J j K k L l M m N n O o P p Q q R r S s T t U u V v W w X x Y y Z z.

USER
அனைத்து குறுகிய தமிழ் உயிரெழுத்துகளையும் பட்டியலிட முடியுமா?

AI
எளிமையான காலம்: "நான் வேகமாக செல்கிறேன்" (நான் வேகமாகப் போகிறேன்) மற்றும் கடினமான காலம்: "நான் வேகமாக/வேகமாக வருகிறேன்" (நான் வேகமாக/வேகமாக வருகிறேன்).

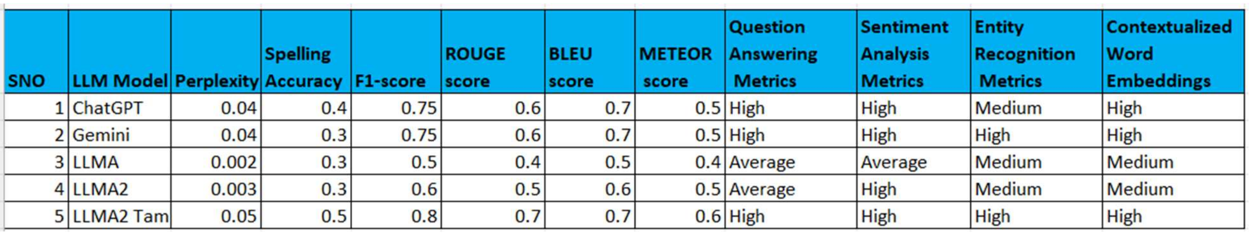
USER
can you teach me active and passive voices in tamil?

Figure 1c: Tamil LLAMA2v1 Model

Our proposed Tamil LLM model with additional data would fix these issues. The LLM would be able to recognize the true vowels in a Tamil word and true consonants, unlike the existing AI such as Google Gemini and OpenAI ChatGPT. Because we have added training with input and answers in our dataset of examples of vowels and consonants in a variety of different words, the LLM will understand the vowels and consonants and make much less mistakes.

E. High Level benchmark on current LLMs for Tamil Language

Using the 1000 words, 100 sentences and 10 lines of articles as requirements for benchmark, the below table shows the standard LLM benchmark results.



SNO	LLM Model	Perplexity	Spelling Accuracy	F1-score	ROUGE score	BLEU score	METEOR score	Question Answering Metrics	Sentiment Analysis Metrics	Entity Recognition Metrics	Contextualized Word Embeddings
1	ChatGPT	0.04	0.4	0.75	0.6	0.7	0.5	High	High	Medium	High
2	Gemini	0.04	0.3	0.75	0.6	0.7	0.5	High	High	High	High
3	LLMA	0.002	0.3	0.5	0.4	0.5	0.4	Average	Average	Medium	Medium
4	LLMA2	0.003	0.3	0.6	0.5	0.6	0.5	Average	High	Medium	Medium
5	LLMA2 Tam	0.05	0.5	0.8	0.7	0.7	0.6	High	High	High	High

Table 1: Performance Metrics on Various Models

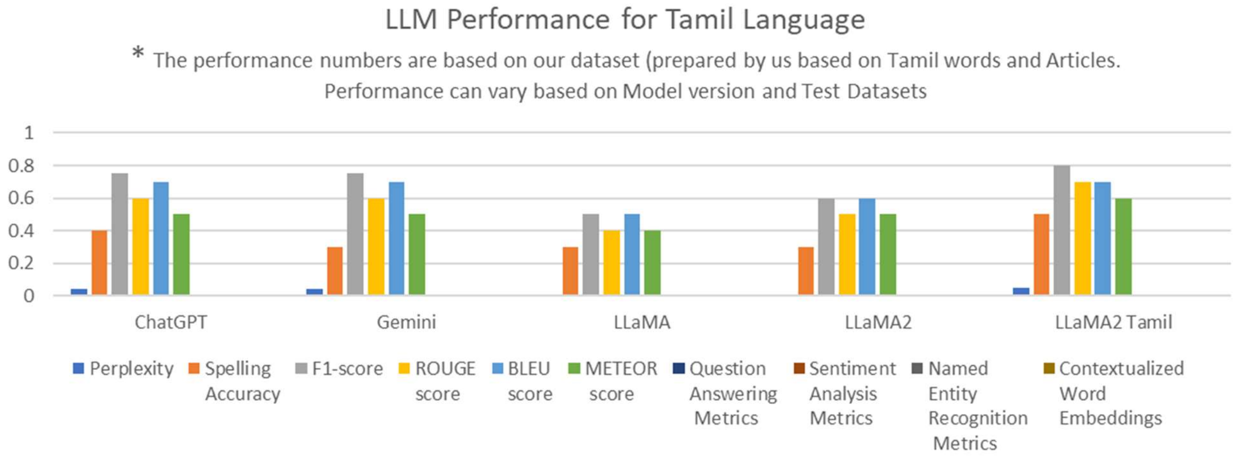


Figure 2: Benchmark results with Tamil Language on different LLMs

III. Proposed LLM Model for Enhanced K-8 Tamil Language Education

A. Domain-Specific Dataset Creation: Building Blocks of Tamil Language Learning

To address the limitations of current LLMs in Tamil language education, we propose a bespoke LLM designed explicitly for K-8 learners. This model will be powered by a comprehensive dataset focused on core Tamil language fundamentals. Here's a breakdown of the key components within this dataset:

1. Foundational Tamil Script and Pronunciation:

This section will encompass a vast collection of Tamil text and audio recordings, including:
The entire Tamil alphabet (uyir and mei alphabets) with clear pronunciations and examples.
Inherent vowel sounds (உள் ஒலி (ul oli)) and their lengthened forms (ல, ள, ழ) with audio demonstrations. Common consonant clusters (like ற்ற(tra) or ன்ற (ndra)) with pronunciations and example words.

2. Building a Strong Vocabulary:

This section will feature age-appropriate vocabulary lists categorized by grade level:
Lower Grades: Focus on basic everyday words (greetings, family members, common objects), numbers (1-100), and colors. **Middle Grades:** Introduce subject-specific vocabulary related to various topics like கவிதை (kavithaigal - poems), கதை (kathai - stories), கட்டுரை (katturai - essays), நாடகம் (naatakam - plays). **Higher Grades:** Include more advanced vocabulary encompassing synonyms, antonyms, and literary devices like உவமை (uvamai - simile) and உருவகம்ள (uruvakam - metaphor). Our model falls under the 'higher grades' vocabulary, as it has been trained with a 7th grade California Tamil Academy Syllabus.

3. Mastering Tamil Grammar:

This section will provide a comprehensive exploration of Tamil grammatical structures:
Parts of Speech: Detailed explanations and exercises on nouns, verbs, adjectives, adverbs, prepositions, and conjunctions.

- **Sentence Formation:** Activities focused on subject-verb agreement, different sentence types (declarative, interrogative, imperative), and sentence structures (simple, compound, complex).

  - **Verb Conjugations (Kalangal):** In-depth exploration of past (இறந்த காலம் (irantha kaalam)), present (நிகழ் காலம் (nigazh kaalam)), and future (எதிர் காலம் (ethir kalaam)) tenses in various forms, with conjugation tables and practice exercises.
  - **Sandhi (எழுத்து சேர்க்கை (eḻuttu cērkkαı)):** Rules and exercises for combining letters to form new sounds (e.g., க + அ = கா (ka + a = kaa)).

4. Enriching Tamil Language Skills:

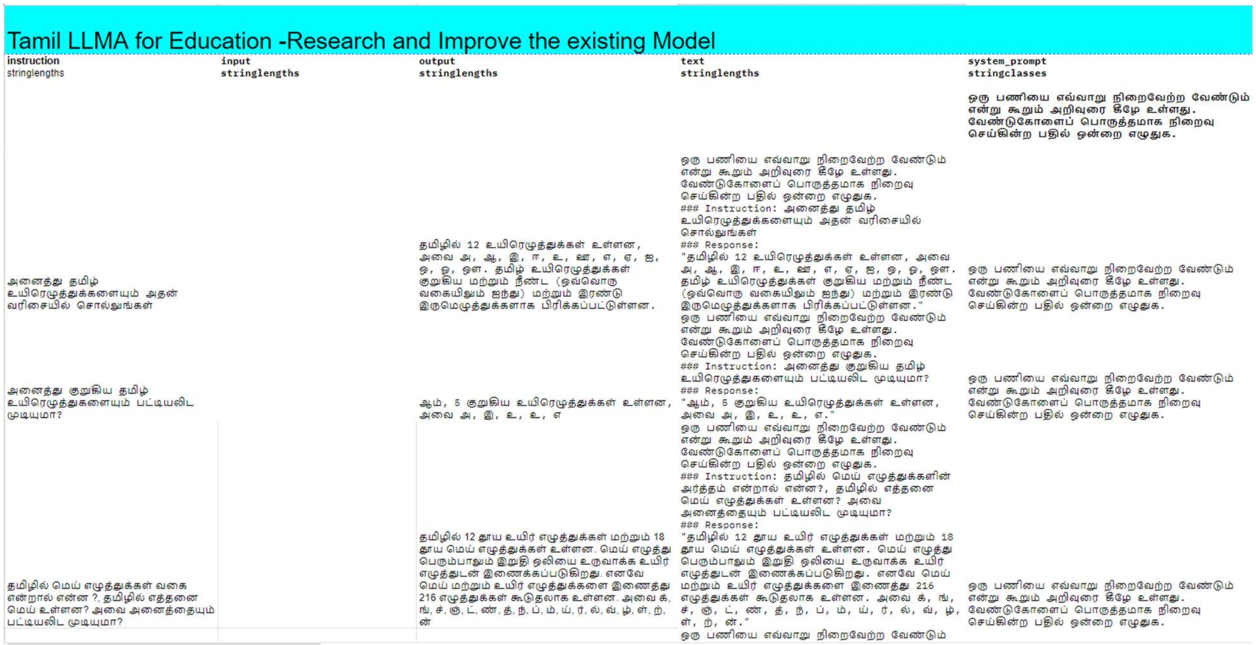
This section will focus on developing a deeper appreciation for the Tamil language:

- **Idioms and Proverbs (பழமொழி (pazhamozhi)):** Collection of common Tamil proverbs with explanations and their use in context.
- **வாய்மொழிக்கதைகள் (Vāymozhikathaigal - Folktales):** Integration of popular Tamil folktales with interactive activities to improve comprehension and cultural understanding.
- **Poetry and Literature (கவிதை (kavithaigal) & இலக்கியம் (ilakkiyam)):** Exposure to classic and contemporary Tamil poems and literary excerpts, along with exercises to analyze literary devices and appreciate the beauty of the language.

5. Custom Model to improve the current model performance

Our group of students and Tamil interested and native speakers are working to add additional 10000 lines of new entries focusing on Tamil basics so LLM will understand more on Tamil basics. It allows the model to make more prompt questions as instructions and expected responses so this custom dataset will be used to improve the higher encoding and generative fluency by native speakers of the model.

Our dataset includes alphabet-like vowels, consonants and special characters in the language, contextual letters, short and long vowels, sentence formatting, noun, verb, tenses and other grammar focused datasets.



Tamil LLMA for Education -Research and Improve the existing Model

instruction	input	output	text	system_prompt
stringlengths	stringlengths	stringlengths	stringlengths	stringclasses
				ஒரு பணியை எவ்வாறு நிறைவேற்ற வேண்டும் என்று கூறும் அறிவுரை கீழே உள்ளது. வேண்டுகோளைப் பொருத்தமாக நிறைவு செய்கின்ற பதில் ஒன்றை எழுதுக.
அனைத்து தமிழ் உயிரெழுத்துக்களையும் அதன் வரிசையில் சொல்லுங்கள்		தமிழில் 12 உயிரெழுத்துக்கள் உள்ளன, அவை அ, ஆ, இ, ஈ, உ, ஊ, எ, ஏ, ஐ, ஒ, ஓ, ஔ. தமிழ் உயிரெழுத்துக்கள் குறுகிய மற்றும் நீண்ட (ஒவ்வொரு வகையிலும் ஐந்து) மற்றும் இரண்டு இருமெழுத்துக்களாக பிரிக்கப்பட்டுள்ளன.	ஒரு பணியை எவ்வாறு நிறைவேற்ற வேண்டும் என்று கூறும் அறிவுரை கீழே உள்ளது. வேண்டுகோளைப் பொருத்தமாக நிறைவு செய்கின்ற பதில் ஒன்றை எழுதுக. ### Instruction: அனைத்து தமிழ் உயிரெழுத்துக்களையும் அதன் வரிசையில் சொல்லுங்கள் ### Response: "தமிழில் 12 உயிரெழுத்துக்கள் உள்ளன, அவை அ, ஆ, இ, ஈ, உ, ஊ, எ, ஏ, ஐ, ஒ, ஓ, ஔ. தமிழ் உயிரெழுத்துக்கள் குறுகிய மற்றும் நீண்ட (ஒவ்வொரு வகையிலும் ஐந்து) மற்றும் இரண்டு இருமெழுத்துக்களாக பிரிக்கப்பட்டுள்ளன."	ஒரு பணியை எவ்வாறு நிறைவேற்ற வேண்டும் என்று கூறும் அறிவுரை கீழே உள்ளது. வேண்டுகோளைப் பொருத்தமாக நிறைவு செய்கின்ற பதில் ஒன்றை எழுதுக.
அனைத்து குறுகிய தமிழ் உயிரெழுத்துகளையும் பட்டியலிட முடியுமா?		ஆம், 5 குறுகிய உயிரெழுத்துக்கள் உள்ளன, அவை அ, இ, உ, உ, எ	ஒரு பணியை எவ்வாறு நிறைவேற்ற வேண்டும் என்று கூறும் அறிவுரை கீழே உள்ளது. வேண்டுகோளைப் பொருத்தமாக நிறைவு செய்கின்ற பதில் ஒன்றை எழுதுக. ### Instruction: அனைத்து குறுகிய தமிழ் உயிரெழுத்துகளையும் பட்டியலிட முடியுமா? ### Response: "ஆம், 5 குறுகிய உயிரெழுத்துக்கள் உள்ளன, அவை அ, இ, உ, உ, எ."	ஒரு பணியை எவ்வாறு நிறைவேற்ற வேண்டும் என்று கூறும் அறிவுரை கீழே உள்ளது. வேண்டுகோளைப் பொருத்தமாக நிறைவு செய்கின்ற பதில் ஒன்றை எழுதுக.
தமிழில் மெய் எழுத்துக்கள் வகை என்றால் என்ன ?, தமிழில் எத்தனை மெய் உள்ளன? அவை அனைத்தையும் பட்டியலிட முடியுமா?		தமிழில் 12 தூய உயிர் எழுத்துக்கள் மற்றும் 18 தூய மெய் எழுத்துக்கள் உள்ளன. மெய் எழுத்து பெரும்பாலும் இறுதி ஒலியை உருவாக்க உயிர் எழுத்துடன் இணைக்கப்படுகிறது. எனவே மெய் மற்றும் உயிர் எழுத்துக்களை இணைத்து 216 எழுத்துக்கள் கூடுதலாக உள்ளன. அவை க், ங், ச், ஞ், ட், ண், த், ந், ப், ம், ய், ர், ல், வ், ழ், ள், ற், ன்	ஒரு பணியை எவ்வாறு நிறைவேற்ற வேண்டும் என்று கூறும் அறிவுரை கீழே உள்ளது. வேண்டுகோளைப் பொருத்தமாக நிறைவு செய்கின்ற பதில் ஒன்றை எழுதுக. ### Instruction: தமிழில் மெய் எழுத்துக்களின் அர்த்தம் என்றால் என்ன?, தமிழில் எத்தனை மெய் எழுத்துக்கள் உள்ளன? அவை அனைத்தையும் பட்டியலிட முடியுமா? ### Response: "தமிழில் 12 தூய உயிர் எழுத்துக்கள் மற்றும் 18 தூய மெய் எழுத்துக்கள் உள்ளன. மெய் எழுத்து பெரும்பாலும் இறுதி ஒலியை உருவாக்க உயிர் எழுத்துடன் இணைக்கப்படுகிறது. எனவே மெய் மற்றும் உயிர் எழுத்துக்களை இணைத்து 216 எழுத்துக்கள் கூடுதலாக உள்ளன. அவை க், ங், ச், ஞ், ட், ண், த், ந், ப், ம், ய், ர், ல், வ், ழ், ள், ற், ன்."	ஒரு பணியை எவ்வாறு நிறைவேற்ற வேண்டும் என்று கூறும் அறிவுரை கீழே உள்ளது. வேண்டுகோளைப் பொருத்தமாக நிறைவு செய்கின்ற பதில் ஒன்றை எழுதுக.
			ஒரு பணியை எவ்வாறு நிறைவேற்ற வேண்டும்	

Table 2: Tamil LLM for Education - Dataset for fine tuning existing LLAMA2 Tamil Model

By incorporating this comprehensive dataset, the LLM model will be equipped to address the core fundamentals of Tamil language learning for K-8 students. This will empower the model to provide a rich and engaging learning experience, fostering a strong foundation for future language development.

IV. Potential Outputs and Applications of the LLM Model

The proposed LLM model for Tamil language education can be leveraged in multiple ways to enhance the learning experience for K-8 students. These include:

A. Personalized Learning Tools:

Adaptive Learning Activities: The model can generate customized exercises and practice problems based on a student's individual strengths and weaknesses.

Interactive Question-and-Answer Systems: Students can interact with the model to clarify doubts, ask questions, and receive personalized explanations in Tamil.

Automated Feedback and Assessment: The model can analyze student responses and provide targeted feedback to identify areas for improvement.

B. Enhancing Traditional Teaching Methods (Continued):

Lesson Planning Assistance: The model can support teachers in creating engaging and interactive lesson plans that cater to diverse learning styles. It can suggest age-appropriate activities, games, and storytelling prompts tailored to specific Tamil language concepts.

Pronunciation Practice Tools: The model can be integrated with audio functionalities to allow students to practice pronunciation and receive feedback on their accuracy. This can address the current limitations of LLMs in phonetics by incorporating audio recognition techniques.

Culturally-Rich Learning Materials: The model can generate educational content infused with Tamil cultural references, folktales, and historical anecdotes. This fosters a deeper connection with Tamil heritage and motivates students to engage with the language.

C. Examples of LLM Model Interactions:

Here are some illustrative examples of how students and teachers might interact with the LLM model:

- **Scenario 1: Personalized Learning**
  - Student: "I'm having trouble understanding the concept of இடைச்சொல் (idaichchol - interjection) in Tamil."
  - LLM Model: "No problem! Interjections are words used to express emotions, like surprise or joy. Here are some examples: ஆஹா (aahaa - surprise), அட (ada - annoyance), and ஐயோ (aiyo - pain). Can you try creating a sentence using an interjection?"
- **Scenario 2: Interactive Question-and-Answer**
  - Teacher: "Today, we'll learn about the ancient Tamil kingdom of Cholas. What are some interesting facts about the Cholas?"
  - LLM Model: "The Cholas were a powerful maritime empire that ruled South India from the 3rd century BCE to the 13th century CE. They were known for their advanced naval technology, trade networks, and magnificent temples. Would you like to see some pictures of famous Chola architecture?"
- **Scenario 3: Culturally-Rich Learning**
  - Student: "Tell me a story about a clever crow in Tamil."
  - LLM Model: (Narrates a folktale about a crow using its wit to outsmart a cunning fox, incorporating relevant vocabulary and sentence structures.)

V. Evaluation and Future Directions

Building on the proposed evaluation methods, here's a glimpse into how LLMs can further revolutionize Tamil education:

Personalized Learning Experiences:

Adaptive Learning: LLMs can personalize learning paths for each student. By analyzing their strengths, weaknesses, and learning styles, the LLM can suggest targeted exercises, provide feedback in real time, and adjust the difficulty level to optimize progress.

Content Creation: LLMs can create customized learning materials based on individual student needs. Imagine practice exercises, quizzes, or even interactive stories tailored to a student's interests and proficiency level.

Enhanced Engagement and Accessibility:

Conversational Tutors: LLMs can act as virtual tutors, engaging in conversation with students to practice spoken Tamil. This can be particularly beneficial for shy students or those lacking opportunities for spoken language practice.

Gamification: LLMs can be integrated into educational games, making learning Tamil more fun and interactive. Imagine games that adapt to the student's progress, reward correct answers, and provide constructive feedback in Tamil.

Text-to-Speech and Speech-to-Text: LLMs can provide text-to-speech functionality for written materials, making learning materials accessible for visually impaired students. Conversely, speech-to-text features can help students who struggle with writing in Tamil.

Further LLM Improvements for Tamil:

A Focus on Less Common Dialects

Current LLM models might focus on dominant Tamil dialects. Future models can be trained on a wider range of data to ensure inclusivity and cater to learners from diverse Tamil-speaking regions.

B Incorporation of Cultural Context

LLMs can be trained to understand and incorporate cultural nuances specific to Tamil language usage. This can lead to more natural and effective communication learning.

C Explainability and Transparency

As LLMs become more complex, ensuring explainability and transparency in their decision-making processes becomes crucial. This will build trust among educators and students regarding the feedback and suggestions provided by the LLM.

D Beyond Textbooks: AI-Generated Learning Materials

Forget rote memorization and dry textbooks. The LLM can create customized learning materials on the fly. Imagine interactive stories tailored to a student's interests and proficiency level, or practice exercises that target specific areas needing improvement. This fosters a deeper understanding and a love for the language and culture.

E Conversational Tutors and Gamified Learning

For shy students or those lacking opportunities for spoken Tamil practice, the LLM steps in as a virtual conversation partner. Engaging in dialogues and practicing pronunciation becomes a natural part of learning. Additionally, the LLM can power interactive and gamified learning experiences. Picture educational games that adapt to a student's progress, reward correct Tamil usage, and provide constructive feedback – making learning fun and effective.

F Empowering Tamil Educators

The LLM isn't here to replace teachers. It empowers them! Imagine automated grading of written work, freeing up valuable time for teachers to provide personalized guidance. The LLM can also assist in creating lesson plans that cater to specific learning objectives and student needs, suggesting relevant activities and resources tailored to the chosen topic.

G The Future of Tamil: Refining the LLM for Native Speakers

The LLM is constantly evolving, and for native Tamil speakers, the focus is on inclusivity and cultural depth. The model will be trained on a wider range of data to encompass less common dialects, ensuring all Tamil speakers benefit from this technology. Furthermore, the LLM will be trained to understand and incorporate the rich cultural nuances inherent in the Tamil language, leading to a more natural and effective learning experience.

This proposed LLM model transcends traditional methods. It's a bridge between the timeless beauty of Tamil and the transformative power of AI, paving the way for a future generation of Tamil speakers who are not only fluent but also possess a deep appreciation for their linguistic heritage. By addressing these future directions, LLMs have the potential to become invaluable tools for enhancing Tamil language learning, making it a more personalized, engaging, and accessible experience for all students.

VI. Conclusion

The proposed LLM model for Tamil language education offers a promising approach to bridge tradition and innovation in K-8 learning. By leveraging the power of large language models and domain-specific datasets, we can create personalized, engaging, and culturally relevant educational experiences. This paves the way for fostering a deeper appreciation for the Tamil language and equipping future generations with the necessary skills to pass on the treasure of the Tamil language.

Bibliography:

Balachandran, Abhinand. “Abhinand/Tamil-Llama-7b-Base-v0.1 · Hugging Face.” *Abhinand/Tamil-Llama-7b-Base-v0.1 · Hugging Face*, huggingface.co/abhinand/tamil-llama-7b-base-v0.1. Accessed 10 Mar. 2024.

Balachandran, Abhinand. “Abhinand5/Tamil-Llama: A New Tamil Large Language Model (LLM) Based on Llama 2.” *GitHub*, github.com/abhinand5/tamil-llama. Accessed 10 Mar. 2024.

Balachandran, Abhinand. “Breaking Language Barriers: Introducing Tamil Llama v0.2 and Its Expansion to Telugu and Malayalam.” *Medium*, Medium, 25 Jan. 2024, abhinand05.medium.com/breaking-language-barriers-introducing-tamil-llama-v0-2-and-its-expansion-to-telugu-and-malayalam-deb5d23e9264.

Balachandran, Abhinand. “Tamil-Llama: A New Tamil Language Model Based on Llama 2.” *Cornell University*, Arxiv.Org, https://arxiv.org/abs/2311.05845. Accessed 2024.

K, Raju. “Building Your First Indic Language Model: A Step-by-Step Guide.” *Medium*, Medium, 4 Sept. 2023, medium.com/@raju.kandasamy/building-your-first-indic-language-model-a-step-by-step-guide-b305becc5eb2.

Rai, Saritha. “Ola Founder’s AI Startup Launches LLM in Hindi, Tamil, Bengali.” *Bloomberg.Com*, Bloomberg, 15 Dec. 2023, www.bloomberg.com/news/articles/2023-12-15/ola-founder-s-ai-startup-launches-indian-large-language-model.

“Tamil Grammar.” *TalkPal*, 21 Nov. 2023, talkpal.ai/tamil-grammar/.

“Tamil Grammar.” *Wikipedia*, Wikimedia Foundation, 10 Mar. 2024, en.wikipedia.org/wiki/Tamil_grammar.

“Teaching and Learning with Games.” *Digital Citizenship - Home*, www.digitalcitizenship.nsw.edu.au/articles/teaching-and-learning-with-games#:~:text=A%20research%20project%20by%20Deakin,that%20made%20the%20biggest%20difference. Accessed 10 Mar. 2024.

https://en.wikipedia.org/wiki/Tamil_language

https://ta.wikipedia.org/wiki/%E0%AE%AE%E0%AF%81%E0%AE%A4%E0%AE%B1%E0%AF%8D_%E0%AE%AA%E0%AE%95%E0%AF%8D%E0%AE%95%E0%AE%AE%E0%AF%8D

https://ta.wikipedia.org/wiki/%E0%AE%AA%E0%AE%95%E0%AF%81%E0%AE%AA%E0%AF%8D%E0%AE%AA%E0%AF%81:%E0%AE%A4%E0%AE%AE%E0%AE%BF%E0%AE%B4%E0%AF%8D
https://ta.wikipedia.org/wiki/%E0%AE%AA%E0%AE%95%E0%AF%81%E0%AE%AA%E0%AF%8D%E0%AE%AA%E0%AF%81:%E0%AE%A4%E0%AE%AE%E0%AE%BF%E0%AE%B4%E0%AF%8D_%E0%AE%87%E0%AE%B2%E0%AE%95%E0%AF%8D%E0%AE%95%E0%AE%A3%E0%AE%AE%E0%AF%8D
https://www.noolulagam.com/books-by-category/40/ilakkanam/
https://www.scribd.com/document/17554819/Tamil-ilakanam
https://ta.wikipedia.org/wiki/%E0%AE%A4%E0%AE%BF%E0%AE%B0%E0%AF%81%E0%AE%95%E0%AF%8D%E0%AE%95%E0%AF%81%E0%AE%B1%E0%AE%B3%E0%AF%8D

About Authors

Chandrasekaran Sakthivel is a seasoned entrepreneur and technologist. He is very hands on engineer and passionate in tech and innovations. He holds 20+ patents. He did his research in Machine Learning and applying the context driven algorithm for reinforcement learning in autonomous vehicle. He did his school and bachelor education in Tamil Nadu. He worked in various parts of world and currently in Silicon Valley for past 14 years. He is a Tamil teacher for past 6 years in CA Tamil Academy and currently teaching 7th grader. He is enjoying his Tamil teaching experience and passing the Tamil language to next generation.

Niranjan Ramarajar is a Software Engineer and Architect with over 20 years of experience leading successful IT initiatives. Niranjan has a strong background in data analytics (M.S. in Data Analytics, Georgia Institute of Technology) and software systems (M.S. in Software Systems, BITS, Pilani). He is a full-stack technical leader with experience in cloud-native solutions, data analytics, security, and compliance. Niranjan is passionate about delivering value to businesses through technology and building strong stakeholder partnerships.

Venkatesh Gnanasekaran A software engineer with 25 years of experience in software design and development with a proven track record of effectively leading multiple teams to ensure consistent quality deliverables for various clients in the Silicon Valley. He is keen on making the process of learning Tamil easier by leveraging modern technologies.

A group of motivated students helped to create a dataset and validate and test the Tamil Language in various LLM models. This support is critical; all students are from CA Tamil Academy Cupertino/Fremont Tamil school students.