



தமிழ்மணம் சர்வதேசத் தமிழ் ஆய்விதழ்

Monthly Peer-Reviewed Open Access Crossref Google Scholar
SJIF Impact Factor 3.5 Multidisciplinary International

Issue DOI: <https://doi.org/10.63300/tm12032026>



Vol. 12 No. 03 (2026): August 2026, p609-627 | Available online at: www.tamilmanam.in

Original Research Paper – Peer reviewed ✓

தமிழ் வட்டார மொழிகளை AI மொழி மாதிரிகள் கையாளும் திறன்: கொங்கு வட்டாரத் தமிழை மையமாகக் கொண்ட ஆய்வு

The Capability of AI Language Models in Handling Tamil Regional Dialects:
A Study Centred on Kongu Tamil

N.Manjula ^{ID*},

Assistant professor, Department of Tamil

A.V.P.College of Arts and Science, Thirumurugan poondi post, Tirupur-641652

Email: manjula6666@gmail.com ORCID: <https://orcid.org/0009-0006-6913-0956>.

Dr. S. Padma ^{ID*},

Principal, Nandha college of Education, Erode-52

Email: padma.haris@gmail.com ORCID: <https://orcid.org/0000-0001-8369-1246>.

Dr.C.Sakthimurugan ^{ID*},

Assistant professor in Tamil

Nandha Arts and Science College (Autonomous), Erode -638052

Email: sakthimurugan.c@nandhaarts.org ORCID: <https://orcid.org/0009-0000-1996-2099>.

Dr.V.C.Srinivasan ^{ID*},

Administrative Officer, Nandha Arts and Science College (Autonomous), Erode.

Email: srivic2345@gmail.com, ORCID: <https://orcid.org/0009-0008-9926-7982>.

Article history

Received: 20-07-2026

Revised: 30-07-2026

Accepted: 10-08-2026

Published on: 18-08-2026

*Corresponding Author Email:

manjula6666@gmail.com

Abstract: While Large Language Models (LLMs) demonstrate significant capabilities in standard Tamil, the extent to which they accurately understand and translate Kongu regional Tamil—spoken in the Kongu region including Coimbatore, Tiruppur, Erode, Salem, and Karur—remains an insufficiently explored area. This article examines the dialect-handling capabilities of AI language models, focusing on Kongu Tamil. For this purpose, a small-scale pilot Kongu Tamil benchmark dataset was curated by the author based on a verified Kongu dialect vocabulary. Using this dataset, zero-shot and few-shot experiments were conducted with a publicly accessible Large Language Model (Claude), and the responses were evaluated based on the author's linguistic knowledge, with errors categorized accordingly. Additionally, citing results from previously published studies comparing various LLMs (Gemini, ChatGPT, Claude), this paper proposes a comprehensive methodological plan for a complete multi-model comparative study for Kongu Tamil, along with a human expert evaluation protocol. Preliminary findings indicate that while the language model handles lexical dialectal variations reasonably well, significant errors occur in cultural nuances, syntactic variations, and idiomatic expressions. This study underscores the need to build large-scale, human-expert-verified benchmark datasets for Tamil dialects.

Keywords: Kongu Tamil, Dialect, Large Language Model, LLM, Zero-shot learning, Few-shot learning, Benchmark dataset, Error analysis.

சுருக்கம்

பெரும் மொழி மாதிரிகள் (Large Language Models - LLMs) தரப்படுத்தப்பட்ட (Standard) தமிழில் குறிப்பிடத்தக்க திறனை வெளிப்படுத்தினாலும், கோவை, திருப்பூர், ஈரோடு, சேலம், கரூர் உள்ளிட்ட கொங்கு நாட்டுப் பகுதிகளில் பேசப்படும் கொங்கு வட்டாரத் தமிழை அவை எந்த அளவிற்குச் சரியாகப் புரிந்துகொள்கின்றன, மொழிபெயர்க்கின்றன என்பது இன்னும் முறையாக ஆய்வு செய்யப்படாத ஒரு பகுதியாகவே உள்ளது. இக்கட்டுரை, கொங்கு வட்டாரத் தமிழை மையமாகக் கொண்டு, AI மொழி மாதிரிகளின் வட்டார மொழி கையாளும் திறனை ஆராய்கிறது. இதற்காக, சரிபார்க்கப்பட்ட கொங்கு வட்டாரச் சொற்களஞ்சியத்தை அடிப்படையாகக் கொண்டு ஆசிரியரால் தொகுக்கப்பட்ட ஒரு சிறிய அளவிலான முன்மாதிரி (Pilot) கொங்கு தமிழ் அளவீட்டுத் தரவுத்தொகுப்பு உருவாக்கப்பட்டது. இத்தரவுத்தொகுப்பைப் பயன்படுத்தி, ஒரு பொதுவில் அணுகக்கூடிய பெரும் மொழி மாதிரியுடன் (Claude) பூஜ்ஜியச் சான்று (Zero-shot) மற்றும் சிறு-சான்று (Few-shot) சோதனைகள் மேற்கொள்ளப்பட்டு, விடைகள் ஆசிரியரின் மொழியியல் அறிவின் அடிப்படையில் மதிப்பீடு செய்யப்பட்டு, பிழைகள் வகைப்படுத்தப்பட்டன. கூடுதலாக, பல்வேறு LLMகளை (Gemini, ChatGPT, Claude) ஒப்பிட்ட ஏற்கெனவே வெளியான ஆய்வுகளின் முடிவுகளை மேற்கோள் காட்டி, கொங்கு தமிழுக்கான முழுமையான பல்-மாதிரி ஒப்பீட்டு ஆய்வுக்கான ஒரு விரிவான முறையியல் திட்டமும், மனித நிபுணர் மதிப்பீட்டு நெறிமுறையும் இக்கட்டுரையில் முன்மொழியப்படுகின்றன. முன்னோட்ட ஆய்வின் முடிவுகள், சொல்லடிப்படையிலான (Lexical) வட்டார வேறுபாடுகளை மொழி மாதிரி ஓரளவு சரியாகக் கையாண்டாலும், பண்பாட்டுக் குறிப்புகள், சொற்றொடர் அமைப்பு மாற்றங்கள் மற்றும் சொல்லாட்சி நுட்பங்களில் (Idiomatic usage) குறிப்பிடத்தக்க பிழைகள் ஏற்படுவதைக் காட்டுகின்றன. இவ்வாய்வு, தமிழ் வட்டார மொழிகளுக்கான பெரிய அளவிலான, மனித நிபுணர்களால் சரிபார்க்கப்பட்ட அளவீட்டுத் தரவுத்தொகுப்புகளை உருவாக்க வேண்டிய அவசியத்தை வலியுறுத்துகிறது.

திறவுச்சொற்கள்: கொங்கு தமிழ், வட்டார மொழி, பெரும் மொழி மாதிரி, LLM, பூஜ்ஜியச் சான்று கற்றல், சிறு-சான்று கற்றல், அளவீட்டுத் தரவுத்தொகுப்பு, பிழை பகுப்பாய்வு

1. முன்னுரை

பெரும் மொழி மாதிரிகள் (Large Language Models) சமீப ஆண்டுகளில் மொழிபெயர்ப்பு, உரையாடல், சுருக்கம் எழுதுதல் உள்ளிட்ட பல்வேறு இயல்மொழி பணிகளில் குறிப்பிடத்தக்க முன்னேற்றத்தைக் காட்டியுள்ளன. எனினும், இம்முன்னேற்றம் பெரும்பாலும் ஆங்கிலம் போன்ற வளமிக்க மொழிகளிலும் (High-resource languages), தரப்படுத்தப்பட்ட மொழி

வடிவங்களிலும் (Standard language forms) மட்டுமே அதிகம் காணப்படுகிறது. தமிழ் போன்ற குறைந்த வளமிக்க மொழிகளில் (Low-resource languages) கூட, மொழி மாதிரிகளின் திறன் தரப்படுத்தப்பட்ட எழுத்துத் தமிழுடன் ஒப்பிடும்போது வட்டார மொழி வடிவங்களில் கணிசமாகக் குறைவதாகப் பல்வேறு ஆய்வுகள் சுட்டிக்காட்டுகின்றன (Rajendran et al., 2022). தமிழ்நாட்டில் கன்னியாகுமரி, திருநெல்வேலி, மதுரை, சென்னை உள்ளிட்ட பல வட்டார வழக்குகள் இருப்பினும், கொங்கு நாட்டுப் பகுதியில் - அதாவது கோயம்புத்தூர், நீலகிரி, திருப்பூர், ஈரோடு, கரூர், நாமக்கல், சேலம், தர்மபுரி, கிருஷ்ணகிரி, திண்டுக்கல் மேற்குப் பகுதி ஆகிய மாவட்டங்களில் - பேசப்படும் கொங்குத் தமிழ் (Kongu Tamil) ஒரு தனித்துவமான, செழுமையான சொல்லடுக்கையும் இலக்கண அமைப்பையும் கொண்ட வட்டார வழக்காகும் (தமிழ் விக்கிப்பீடியா, 2026). AI மொழி மாதிரிகள் இவ்வட்டார வழக்கை எந்த அளவிற்குப் புரிந்துகொள்கின்றன, சரியாக மொழிபெயர்க்கின்றன அல்லது பதிலளிக்கின்றன என்பது குறித்த முறையான, தரவு அடிப்படையிலான ஆய்வுகள் இன்னும் மிகவும் குறைவாகவே உள்ளன.

இக்குறையை நிவர்த்தி செய்யும் நோக்கில், இக்கட்டுரை கொங்கு வட்டாரத் தமிழை மையமாகக் கொண்டு, AI மொழி மாதிரிகளின் வட்டார மொழி கையாளும் திறனை ஆராய முற்படுகிறது. ஒரு சிறிய அளவிலான, சரிபார்க்கப்பட்ட வட்டாரச் சொற்களஞ்சியத்தை அடிப்படையாகக் கொண்ட முன்மாதிரி அளவீட்டுத் தரவுத்தொகுப்பை உருவாக்கி, அதன் மூலம் பூஜ்ஜியச் சான்று மற்றும் சிறு-சான்று சோதனைகளை மேற்கொண்டு, பிழைகளை பகுப்பாய்வு செய்து, எதிர்கால முழுமையான ஆய்வுக்கான ஒரு முறையியல் திட்டத்தை இக்கட்டுரை முன்வைக்கிறது.

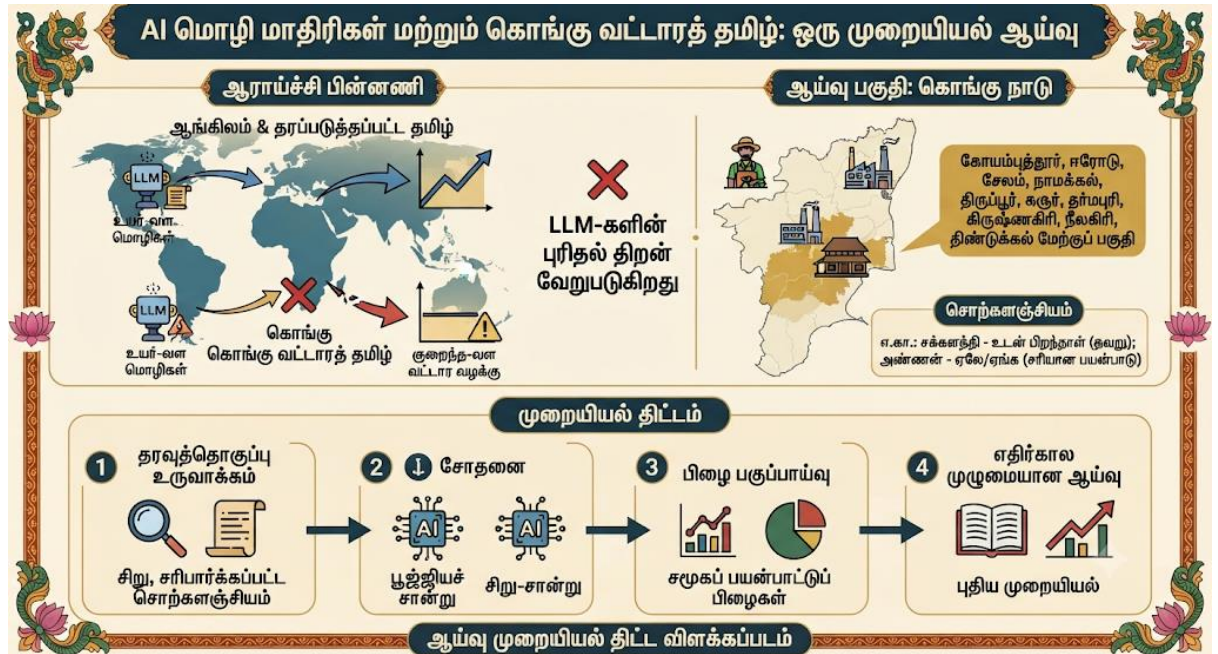


Fig 1: AI மொழி மாதிரிகள் மற்றும் கொங்கு வட்டாரத் தமிழ்: ஒரு முறையியல் ஆய்வு

இப்படமானது AI மொழி மாதிரிகள் (LLMs) கொங்கு வட்டாரத் தமிழை எவ்வளவு துல்லியமாகக் கையாள்கின்றன என்பதை ஆராயும் ஒரு முறையியல் ஆய்வை விளக்குகிறது. ஆங்கிலம் மற்றும் தரப்படுத்தப்பட்ட தமிழுடன் ஒப்பிடும்போது, வட்டார வழக்குகளில் மாதிரிகளின் திறன் குறைவாக உள்ளதை இது சுட்டிக்காட்டுகிறது. கொங்கு மண்டலத்தைச் சேர்ந்த 10 மாவட்டங்களை உள்ளடக்கிய இத்திட்டம், தரவு சேகரிப்பு, சோதனை, பிழை பகுப்பாய்வு மற்றும் எதிர்கால முழுமையான ஆய்வுக்கான வரைவை முன்வைக்கிறது.

2. ஆய்வின் தேவையும் நோக்கங்களும்

Joshi et al. (2020) இன் மொழி வகைப்பாட்டு அளவுகோலின்படி, தமிழ் ஒரு "குறைந்த வளமிக்க" மொழியாகவே (Low-resource language) தரப்படுத்தப்பட்டுள்ளது - பரவலாகப் பேசப்படும் மொழியாக இருந்தபோதிலும். இதற்குக் காரணம், முறையாக ஆவணப்படுத்தப்பட்ட, குறியீடு செய்யப்பட்ட (Annotated) தரவுத்தொகுப்புகளின் பற்றாக்குறையே ஆகும் (From Phonemes to Meaning, 2025). இச்சூழலில், தரப்படுத்தப்பட்ட தமிழுக்கே தரவுப் பற்றாக்குறை இருக்கும்போது, கொங்குத் தமிழ் போன்ற வட்டார வழக்குகளுக்கான தரவுத்தொகுப்புகள் கிட்டத்தட்ட முழுமையாக இல்லை எனலாம். இது, AI மொழி மாதிரிகள் தமிழக மக்கள் தொகையில் கணிசமான பகுதியினரால் அன்றாடம் பேசப்படும் இயற்கையான மொழி வடிவங்களை எவ்வளவு துல்லியமாகக் கையாள்கின்றன என்பதில் ஒரு பெரிய இடைவெளியை உருவாக்குகிறது.

இவ்வாய்வின் நோக்கங்கள் பின்வருமாறு:

1. கொங்கு வட்டாரத் தமிழின் மொழியியல் தனித்தன்மைகளை அடையாளம் காணுதல்.
2. சரிபார்க்கப்பட்ட வட்டரச் சொல்லாடல்களை அடிப்படையாகக் கொண்ட ஒரு முன்மாதிரி கொங்கு தமிழ் அளவீட்டுத் தரவுத்தொகுப்பை உருவாக்குதல்.
3. AI மொழி மாதிரி ஒன்றின் மீது பூஜ்ஜியச் சான்று மற்றும் சிறு-சான்று சோதனைகளை மேற்கொண்டு, அதன் கொங்கு தமிழ் புரிதல் திறனை மதிப்பிடுதல்.
4. முன்னர் வெளியான பல்-LLM ஒப்பீட்டு ஆய்வுகளின் முடிவுகளுடன் இணைத்து, கொங்கு தமிழுக்கான முழுமையான பல்-மாதிரி ஒப்பீட்டு ஆய்வுக்கான முறையியலை வடிவமைத்தல்.
5. மனித நிபுணர் மதிப்பீட்டு நெறிமுறையை வடிவமைத்தல் மற்றும் பிழை வகைப்பாட்டு அமைப்பை (Error Taxonomy) உருவாக்குதல்.

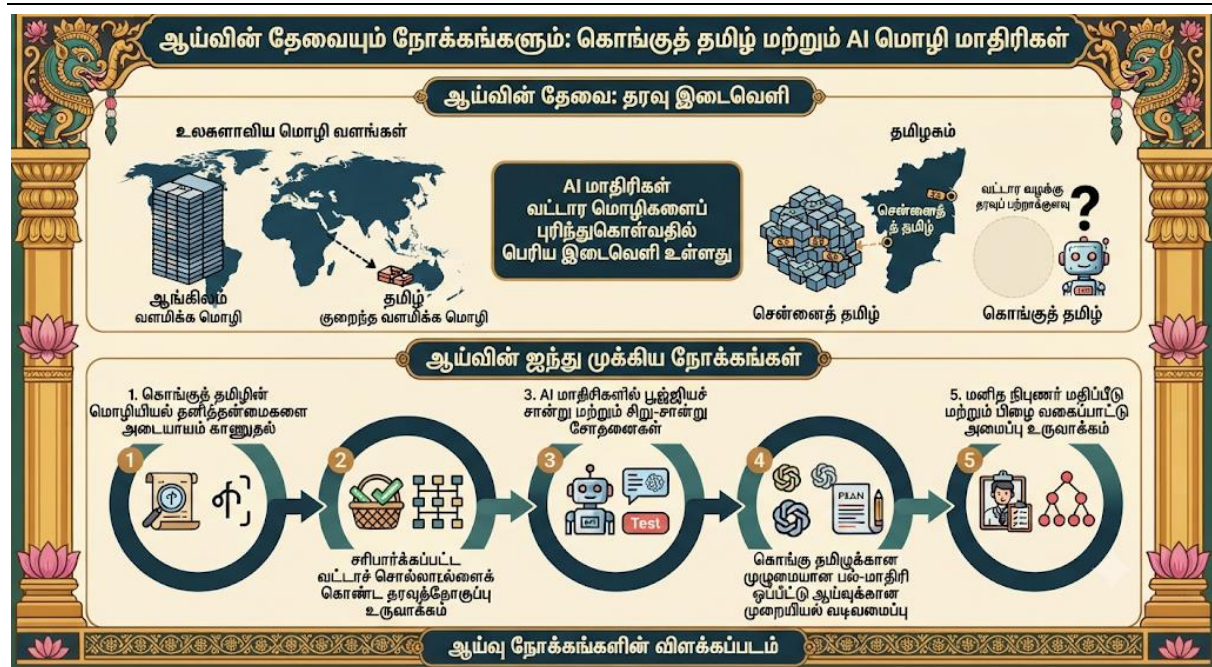


Fig 2: ஆய்வு செயல்முறை விளக்க வரைபடம்

இப்படமானது AI மொழி மாதிரிகளின் கொங்கு வட்டாரத் தமிழ் புரிதல் திறனை ஆராயும் ஆய்வின் தேவையை விளக்குகிறது. ஆங்கிலம் மற்றும் தரப்படுத்தப்பட்ட தமிழுடன் ஒப்பிடும்போது, கொங்குத் தமிழுக்கு முறையாக ஆவணப்படுத்தப்பட்ட தரவுத்தொகுப்புகள் இல்லை என்பதை இது சுட்டிக்காட்டுகிறது. இந்த தரவு இடைவெளி, AI மாதிரிகள் வட்டார வழக்குகளைக் கையாள்வதில் பெரிய சவாலை ஏற்படுத்துகிறது.

3. இலக்கிய மதிப்பாய்வு

தமிழ் NLP தொடர்பான ஆய்வுகள் கடந்த பத்தாண்டுகளில் குறிப்பிடத்தக்க வளர்ச்சியைக் கண்டுள்ளன. Rajendran et al. (2022) தமிழ் NLP தொழில்நுட்பங்களின் தற்போதைய நிலை, சவால்கள் மற்றும் எதிர்கால வாய்ப்புகள் குறித்து விரிவான கண்ணோட்டத்தை வழங்கியுள்ளனர்; குறிப்பாக வட்டார வழக்குகள் மற்றும் குறைந்த வளமிக்க தரவுச் சூழல்கள் தமிழ் NLP வளர்ச்சிக்கு முக்கியத் தடையாக உள்ளதாக அவர்கள் சுட்டிக்காட்டுகின்றனர்.

பெரும் மொழி மாதிரிகளின் தமிழ் திறன் தொடர்பான மிக அண்மைக்கால ஆய்வு ஒன்று, ILAKKANAM எனும் தமிழ்-சார்ந்த மொழியியல் மதிப்பீட்டு அளவுகோலை அறிமுகப்படுத்தியுள்ளது; இது இலங்கைப் பள்ளிகளின் தமிழ்ப் பாட வினாத்தாள்களிலிருந்து தொகுக்கப்பட்ட 820 வினாக்களைக் கொண்டது, மொழியியல் நிபுணர்களால் ஐந்து மொழியியல் பிரிவுகளின் கீழ் குறியீடு செய்யப்பட்டுள்ளது (From Phonemes to Meaning, 2025). இவ்வாய்வு, LLMகள் உயர்வளமிக்க மொழிகளில் சிறப்பாகப் பொதுமைப்படுத்தினாலும், தமிழ் போன்ற உருபனியல் ரீதியாக வளமிக்க (Morphologically rich), குறைந்த வளமிக்க மொழிகளில் அவற்றின் மொழியியல் திறன் இன்னும் முழுமையாக

ஆராயப்படவில்லை என வலியுறுத்துகிறது - இது நேரடியாக இக்கட்டுரையின் ஆய்வுத் தேவையை உறுதிப்படுத்துகிறது.

பல்-LLM ஒப்பீட்டு ஆய்வுகள் தொடர்பாக, Jageer and Priyaradhikadevi (2025) ஆங்கிலம்-தமிழ் இயந்திர மொழிபெயர்ப்பில் Gemini, ChatGPT, Claude ஆகிய மூன்று மொழி மாதிரிகளை BLEU, BERT, METEOR, மற்றும் Translation Edit Rate (TER) அளவீடுகள் மூலம் ஒப்பிட்டு ஆய்வு செய்துள்ளனர். சிக்கலான தொடரியல் மற்றும் சொல்லாட்சிகளில் Gemini அதிக துல்லியத்தையும் நேர்த்தியையும் காட்டியதாகவும், Claude சொல் ஓட்டத்தில் (Fluency) சிறந்திருந்தாலும் அதிக TER மதிப்பெண்ணைப் பெற்றதாகவும், ChatGPT பொருண்மை ஒத்திசைவில் (Semantic Alignment) உயர் BERT மற்றும் METEOR மதிப்பெண்களைப் பெற்றதாகவும் அவ்வாய்வு கண்டறிந்துள்ளது. இம்முடிவுகள், வெவ்வேறு LLMகள் தமிழ் மொழிபெயர்ப்பில் வெவ்வேறு பலம் மற்றும் பலவீனங்களைக் கொண்டுள்ளன என்பதை உறுதிப்படுத்துகின்றன; எனினும் இவ்வாய்வு தரப்படுத்தப்பட்ட தமிழையே அடிப்படையாகக் கொண்டது, வட்டார வழக்குகளை உள்ளடக்கவில்லை.

வட்டார மொழி/பேச்சுவழக்கு மதிப்பீடு தொடர்பாக, Alam and Anastasopoulos (2024) CODET எனும் ஒரு மாறுபட்ட-பேச்சுவழக்கு மொழிபெயர்ப்பு மதிப்பீட்டு அளவுகோலை முன்மொழிந்துள்ளனர்; இது இயந்திர மொழிபெயர்ப்பு அமைப்புகள் நெருங்கிய தொடர்புடைய பேச்சுவழக்குகளுக்கு இடையே எவ்வளவு துல்லியமாக வேறுபாடு காட்டுகின்றன என்பதை மதிப்பிடுகிறது. இம்முறையியல், கொங்குத் தமிழ் போன்ற வட்டார வழக்குகளுக்கான மதிப்பீட்டு அளவுகோல்களை வடிவமைப்பதற்கு ஒரு பயனுள்ள மாதிரியாக அமையலாம்.

இந்திய மொழிகள் தொடர்பான பரந்த அளவிலான மதிப்பீட்டு முயற்சிகளில், MILU எனும் பல்பணி இந்திய மொழி புரிதல் அளவுகோல் (Verma et al., 2025) மற்றும் BhashaSutra எனும் இந்திய NLP தரவுத்தொகுப்புகள் தொடர்பான பணி-மைய ஆய்வுத் தொகுப்பு (BhashaSutra, 2025) குறிப்பிடத்தக்கவை. எனினும், இவ்விரு முயற்சிகளும் மொழி மட்டத்தில் (Language-level) மதிப்பீடு செய்கின்றனவே தவிர, ஒரு மொழிக்குள் உள்ள வட்டார வேறுபாடுகளை (Intra-language dialectal variation) தனித்து ஆராயவில்லை. Shaffi and Hajamohideen (2021) தமிழ் கையெழுத்து ஒளியச்சு அங்கீகார (OCR) அளவீட்டுத் தரவுத்தொகுப்பு ஒன்றை உருவாக்கியுள்ளனர்; இது எழுத்து மட்ட தமிழ் NLP அளவீட்டுத் தரவுத்தொகுப்புகளுக்கு ஓர் எடுத்துக்காட்டாக அமைகிறது, ஆயினும் இதுவும் பேச்சு/வட்டார மொழி நிலைக்கு நீட்டிக்கப்படவில்லை.

தமிழ் இலக்கியம் மற்றும் கல்வியில் AI பயன்பாடு தொடர்பாக தமிழ்மணம் ஆய்விதழில் வெளியான ஆய்வுகள், AI கருவிகள் தமிழ் கட்டுரை எழுதுதலில் திறன், துல்லியம் மற்றும்

அணுகல்தன்மை ஆகியவற்றில் குறிப்பிடத்தக்க நன்மைகளை வழங்குகின்றன எனக் கூறுகின்றன; எனினும் பண்பாட்டு நுட்பத்தையும் நம்பகத்தன்மையையும் பேணுவதில் சவால்கள் உள்ளன என்றும் அவை சுட்டிக்காட்டுகின்றன (Tamilmanam, 2025a). மற்றொரு ஆய்வு, தமிழ் இலக்கிய ஆய்வுகளுக்கு AI ஆற்றக்கூடிய பங்களிப்புகளை விவரிக்கிறது (Tamilmanam, 2025b). தமிழ்மொழிப் பாதுகாப்பு தொடர்பான ஓர் ஆய்வு, AI மற்றும் NLP தமிழ் மொழியைப் பாதுகாக்கவும் பரப்பவும் சக்திவாய்ந்த கருவிகளாக இருக்க முடியும் எனக் குறிப்பிடுகிறது, ஆயினும் வட்டார வழக்குகளின் பன்முகத்தன்மையை உள்வாங்கும் அளவிற்கு இத்தொழில்நுட்பங்கள் இன்னும் முதிர்ச்சியடையவில்லை என வலியுறுத்துகிறது (G.T.N. Arts College, Dindigul, n.d.).

இறுதியாக, Dravidian மொழிகளுக்கான இயல்மொழி செயலாக்கப் பணிமனைத் தொடர் (DravidianLangTech) சமூக ஊடக உரை வகைப்பாடு, உணர்வு பகுப்பாய்வு உள்ளிட்ட பணிகளுக்கு தமிழ் உள்ளிட்ட திராவிட மொழிகளுக்கான தரவுத்தொகுப்புகளை ஒவ்வோர் ஆண்டும் வெளியிட்டு வருகிறது (Chakravarthi et al., 2021); எனினும் இத்தொடரிலும் இதுவரை கொங்குத் தமிழ் போன்ற குறிப்பிட்ட வட்டார வழக்கை மையமாகக் கொண்ட ஒரு தரவுத்தொகுப்பு வெளியிடப்படவில்லை. இக்குறையே இக்கட்டுரையின் ஆய்வுத் தேவையை மேலும் நியாயப்படுத்துகிறது.

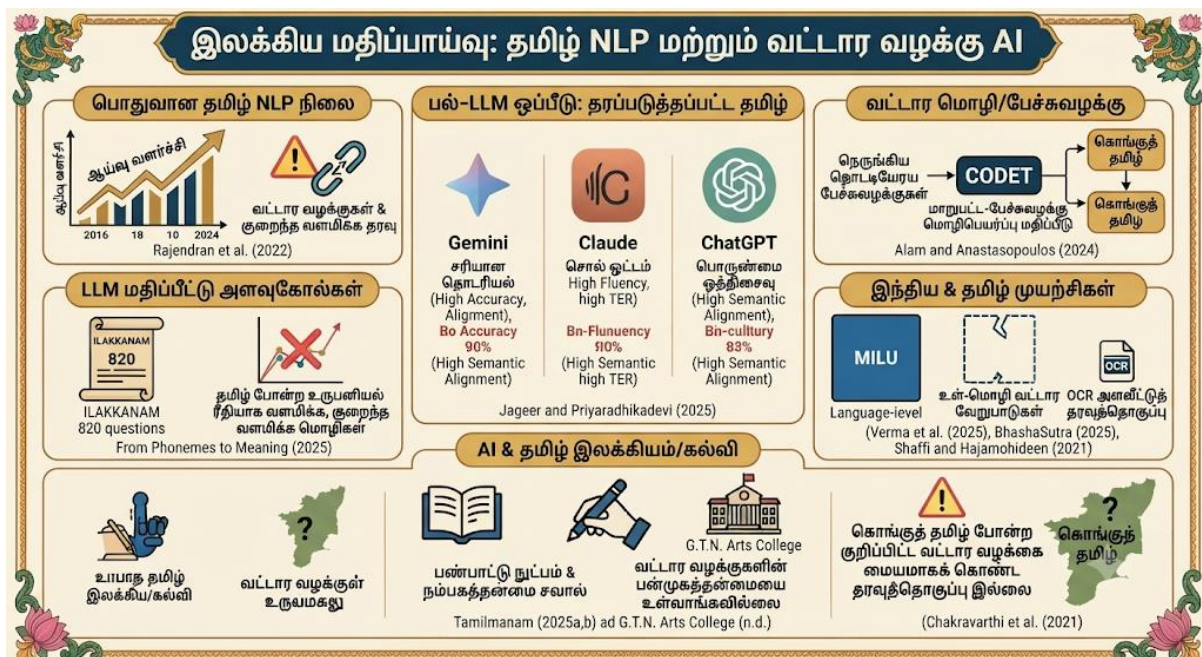


Fig 3: தமிழ் NLP மற்றும் வட்டார வழக்கு AI - ஒரு இலக்கிய மதிப்பாய்வு

இப்படமானது தமிழ் மொழியியல் செயலாக்கம் (NLP) மற்றும் செயற்கை நுண்ணறிவு (AI) துறையில் மேற்கொள்ளப்பட்ட ஆய்வுகளைத் தொகுத்துக் காட்டுகிறது. முக்கியமாக, தரப்படுத்தப்பட்ட தமிழில் இருந்தும், பிற உயர்வளமிக்க மொழிகளில் இருந்தும், கொங்குத்

தமிழ் போன்ற வட்டார வழக்குகள் எந்தளவுக்கு மாறுபடுகின்றன என்பதை இது விளக்குகிறது. கோயம்புத்தூர், ஈரோடு உள்ளிட்ட பகுதிகளில் பேசப்படும் இந்த வட்டார வழக்கு மொழிக்கான தனித்துவமான தரவுத்தொகுப்புகள் மற்றும் மதிப்பீட்டு முறைகளின் அவசியத்தை இந்த இலக்கிய மதிப்பாய்வு வலியுறுத்துகிறது.

4. கொங்கு வட்டாரத் தமிழ்: மொழியியல் பின்னணி

கொங்குத் தமிழ் (Kongu Tamil), கோயம்புத்தூர் தமிழ் என்றும் அழைக்கப்படுகிறது; இது கொங்கு நாட்டுப் பகுதியில் - கோவை, நீலகிரி, திருப்பூர், ஈரோடு, கரூர், நாமக்கல், சேலம், தர்மபுரி, கிருஷ்ணகிரி, திண்டுக்கல் மேற்குப் பகுதி ஆகிய மாவட்டங்களில் - வசிக்கும் மக்களால் பேசப்படும் ஒரு வட்டார வழக்காகும் (தமிழ் விக்கிப்பீடியா, 2026). இது முன்பு "காங்கி" என்ற பெயராலும் அழைக்கப்பட்டுள்ளது.

கொங்கு வட்டார வழக்கு அகராதி பதிவுகளின்படி, இவ்வட்டாரத்திற்கே உரிய தனித்துவமான சொற்களஞ்சியம் உள்ளது. எடுத்துக்காட்டாக: "அக்கட்ட / அக்கட்டு" (அந்த இடம், அந்த இடத்திலே), "அக்கப்போர்" (சண்டை), "வங்கு" (பொந்து, மறைவிடம்), "கம்மனாட்டி" (முட்டாள், மடையன்), "ஆத்தா / ஆயா" (அப்பாவின் அம்மா), "உருமாளை / சிம்மாடு" (தலைப்பாகை) ஆகிய சொற்கள் தரப்படுத்தப்பட்ட தமிழில் வேறு சொற்களால் குறிக்கப்படும் கருத்துக்களை கொங்கு வட்டாரத்தில் தனித்துவமாக வெளிப்படுத்துகின்றன (தமிழ் விக்கிப்பீடியா, 2025).

சொல்லடிப்படையிலான வேறுபாடுகளுக்கு அப்பால், கொங்குத் தமிழ் ஒலியியல் (Phonological) மற்றும் உருபனியல் (Morphological) மட்டங்களிலும் தரப்படுத்தப்பட்ட தமிழிலிருந்து வேறுபடுகிறது - வினைச்சொல் இறுதி ஒலியன்களின் மாற்றம், சில மெய்யொலிகளின் உச்சரிப்பு வேறுபாடு, மற்றும் தொடரியல் அமைப்பில் காணப்படும் நுட்பமான மாற்றங்கள் ஆகியவை இதில் அடங்கும். AI மொழி மாதிரி ஒன்று, வெறும் சொல் மட்ட மாற்று (Lexical substitution) மட்டுமல்லாமல், இத்தகைய ஒலியியல்-உருபனியல் மாற்றங்களையும் புரிந்துகொள்ள வேண்டியிருப்பது, கொங்குத் தமிழைச் சரியாகக் கையாள்வதை மேலும் சிக்கலாக்குகிறது.

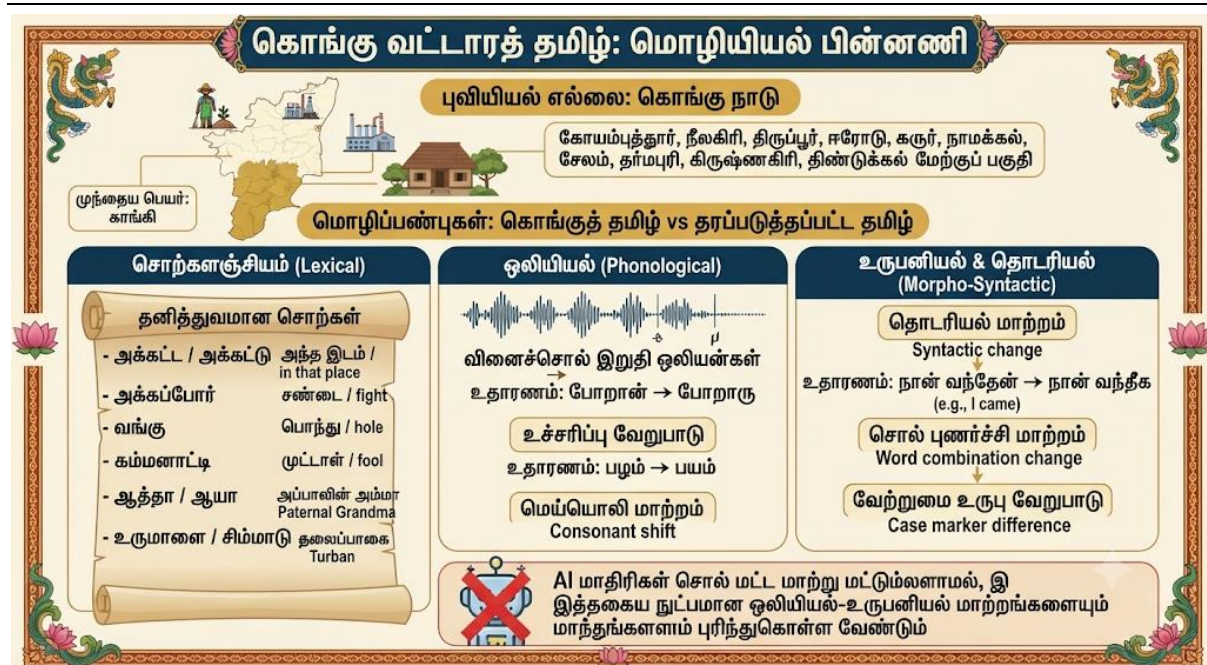


Fig 4: கொங்கு வட்டாரத் தமிழ்: மொழியியல் பின்னணி மற்றும் AI சவால்கள்

இப்படமானது கோயம்புத்தூர், ஈரோடு உள்ளிட்ட கொங்கு மண்டலத்தில் பேசப்படும் கொங்குத் தமிழின் தனித்துவமான மொழியியல் சிறப்புகளை விளக்குகிறது. இது தரப்படுத்தப்பட்ட தமிழில் இருந்து எவ்வாறு ஒலிப்பு, சொல் அமைப்பு மற்றும் தொடரியல் ஆகியவற்றில் வேறுபடுகிறது என்பதை எடுத்துக்காட்டுகளுடன் விவரிக்கிறது. வெறும் சொல் அளவிலான மாற்றங்களைத் தாண்டி, இத்தகைய நுட்பமான இலக்கண வேறுபாடுகளையும் AI மொழி மாதிரிகள் புரிந்துகொள்ள வேண்டியதன் அவசியத்தை இது வலியுறுத்துகிறது.

5. ஆய்வு வடிவமைப்பும் முறையியலும்

இவ்வாய்வு நான்கு கட்டங்களாக வடிவமைக்கப்பட்டுள்ளது: (அ) முன்மாதிரி அளவீட்டுத் தரவுத்தொகுப்பு உருவாக்கம், (ஆ) ஒரு பொதுவில் அணுகக்கூடிய LLM மீதான பூஜ்ஜியச் சான்று/சிறு-சான்று முன்னோட்ட சோதனை, (இ) பல-LLM ஒப்பீட்டு முறையியல் வடிவமைப்பு, மற்றும் (ஈ) மனித நிபுணர் மதிப்பீட்டு நெறிமுறை வடிவமைப்பு.

5.1 தரவுத்தொகுப்பு உருவாக்க முறை

முன்மாதிரி தரவுத்தொகுப்பு, தமிழ் விக்கிப்பீடியாவின் கொங்கு வட்டார வழக்கு அகராதிப் பதிவுகளிலிருந்து சரிபார்க்கப்பட்ட 20 வட்டரச் சொற்களை அடிப்படையாகக் கொண்டு, ஆசிரியரால் 20 கொங்கு தமிழ் வாக்கியங்களாக (Sentence-level items) தொகுக்கப்பட்டது. ஒவ்வொரு உருப்படியும் மூன்று கூறுகளைக் கொண்டுள்ளது: (1) கொங்கு தமிழ் வாக்கியம், (2) அதற்கு இணையான தரப்படுத்தப்பட்ட தமிழ் பொருள் (Gold-standard reference), (3) பணி வகை (மொழிபெயர்ப்பு / பொருள் விளக்கம் / சொல் அடையாளம் காணுதல்). இது ஒரு பெரிய அளவிலான, புள்ளியியல் ரீதியாக முடிவான தரவுத்தொகுப்பு அல்ல; மாறாக, முழுமையான

ஆய்வுக்கு முன்னோடியாக அமையக்கூடிய ஒரு கருத்தாக்க முன்மாதிரி (Proof-of-concept sample) ஆகும் என்பதை வெளிப்படையாகக் குறிப்பிட வேண்டும்.

5.2 பூஜ்ஜியச் சான்று / சிறு-சான்று சோதனை நெறிமுறை

பூஜ்ஜியச் சான்று (Zero-shot) சோதனையில், மொழி மாதிரிக்கு எவ்வித எடுத்துக்காட்டும் வழங்காமல், நேரடியாக கொங்கு தமிழ் வாக்கியத்தை வழங்கி, அதன் தரப்படுத்தப்பட்ட தமிழ் பொருளை அல்லது ஆங்கில விளக்கத்தை கேட்கப்பட்டது. சிறு-சான்று (Few-shot) சோதனையில், மூன்று எடுத்துக்காட்டு ஜோடிகளை (கொங்கு வாக்கியம் + தரப்படுத்தப்பட்ட பொருள்) முதலில் வழங்கிய பின்னர், புதிய வாக்கியங்களுக்கான பொருளைக் கேட்கப்பட்டது. இவ்விரு அணுகுமுறைகளுக்கிடையேயான செயல்திறன் வேறுபாட்டை ஒப்பிடுவதன் மூலம், மாதிரி எடுத்துக்காட்டுகள் மூலம் வட்டார மொழி வடிவங்களை எவ்வளவு விரைவாக ஏற்றுக்கொள்கிறது (In-context adaptation) என்பதை மதிப்பிடலாம்.

5.3 பல்-LLM ஒப்பீட்டு முறையியல் (முன்மொழியப்பட்டது)

தற்போதைய ஆய்வுச் சூழலில், இக்கட்டுரையின் ஆசிரியருக்கு ஒரு பொதுவில் அணுகக்கூடிய LLM (Claude) மட்டுமே நேரடியாகக் கிடைத்தது. எனவே, முழுமையான பல்-LLM ஒப்பீடு (GPT-4, Gemini, LLaMA, Claude உள்ளிட்டவை) இக்கட்டுரையின் நேரடி பரிசோதனைப் பகுதியில் மேற்கொள்ளப்படவில்லை; மாறாக, அதற்கான விரிவான முறையியல் மட்டும் இங்கு முன்மொழியப்படுகிறது. முன்மொழியப்படும் முறையில், ஒரே 20-உருப்படி தரவுத்தொகுப்பு நான்கு அல்லது அதற்கு மேற்பட்ட LLMகளுக்கு ஒரே சீரான உசாத்துணைச் சொற்றொடர்களுடன் (Prompt templates) வழங்கப்பட்டு, ஒவ்வொரு மாதிரியின் விடையும் BLEU, BERTScore, மற்றும் மனித மதிப்பீட்டு மதிப்பெண் ஆகிய மூன்று அளவீடுகள் மூலம் ஒப்பிடப்பட வேண்டும். Jageer and Priyaradhikadevi (2025) இன் தரப்படுத்தப்பட்ட தமிழ் மொழிபெயர்ப்பு ஒப்பீட்டு முடிவுகளை (Gemini - உயர் துல்லியம்; Claude - உயர் சொல் ஓட்டம், குறைந்த திருத்த தூரம்; ChatGPT - உயர் பொருண்மை ஒத்திசைவு) அடிப்படை எதிர்பார்ப்பாகக் கொண்டு, கொங்கு தமிழ் போன்ற வட்டார மொழிச் சூழலில் இப்பண்புகள் மேலும் மோசமடையுமா அல்லது வேறுபடுமா என்பதை சோதிப்பதே இம்முன்மொழியப்பட்ட ஆய்வின் மையக் கேள்வியாகும்.

5.4 மனித நிபுணர் மதிப்பீட்டு நெறிமுறை (முன்மொழியப்பட்டது)

நம்பகத்தன்மையான மதிப்பீட்டிற்கு, குறைந்தது மூன்று சொந்த கொங்கு தமிழ் பேசுநர்களைக் கொண்ட ஒரு நிபுணர் குழு தேவைப்படுகிறது - விருப்பமெனில் தமிழ் மொழியியல் அல்லது தமிழ் இலக்கியத் துறையில் பட்டப்படிப்பு முடித்தவர்கள். ஒவ்வொரு LLM விடையும் ஐந்து புள்ளி லிகர்ட் அளவுகோலில் (Likert scale) மூன்று பரிமாணங்களில் மதிப்பீடு செய்யப்பட

வேண்டும்: (1) பொருள் துல்லியம் (Semantic accuracy), (2) இயல்பான தமிழ் தன்மை (Fluency/Naturalness), (3) பண்பாட்டு/சூழல் பொருத்தம் (Cultural-contextual appropriateness). மதிப்பீட்டாளர்களுக்கிடையேயான உடன்பாட்டு அளவை Cohen's Kappa அல்லது Krippendorff's Alpha மூலம் கணக்கிட்டு, மதிப்பீட்டு நம்பகத்தன்மையை உறுதி செய்ய வேண்டும். இக்கட்டுரையின் நேரடி முன்னோட்ட ஆய்வில், சொந்த மொழி பேசுநர் நிபுணர் குழு கிடைக்காததால், ஆசிரியரின் சொந்த மொழியியல் அறிவின் அடிப்படையிலான ஒரு தன்-மதிப்பீடு (Self-assessment) மேற்கொள்ளப்பட்டது; இது ஒரு தெளிவான வரம்பாகக் கருதப்பட வேண்டும் - சுயாதீன மனித நிபுணர் மதிப்பீட்டிற்கு இது மாற்றாக முடியாது.

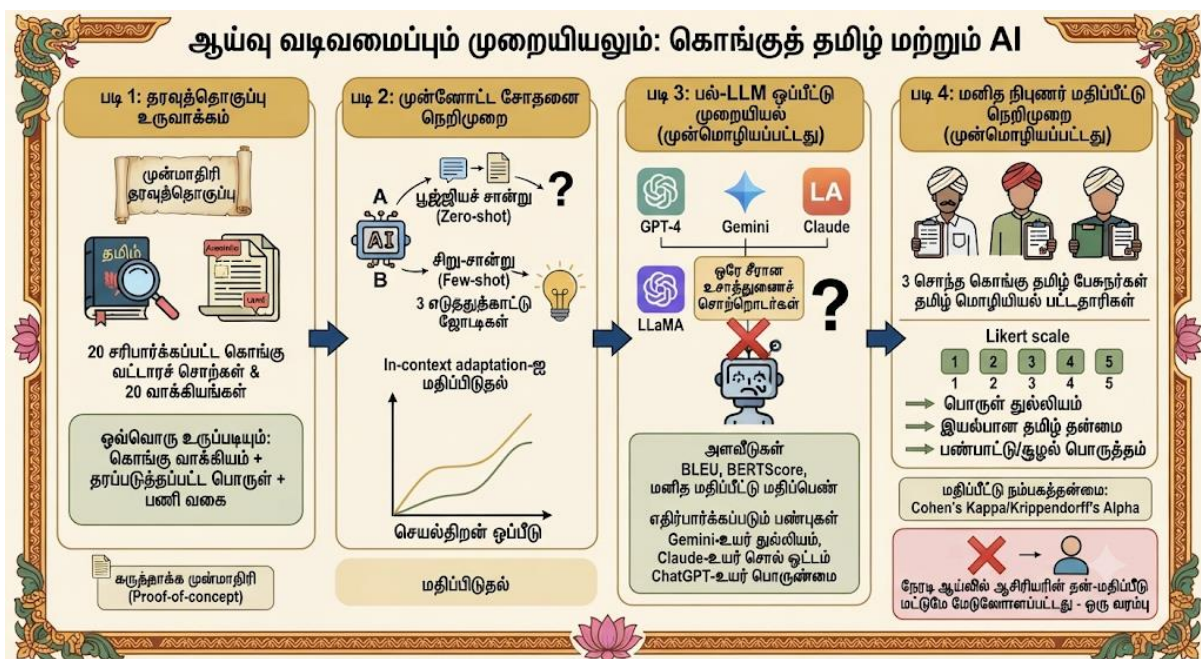


Fig 5: கொங்குத் தமிழ் மற்றும் AI: ஆய்வு வடிவமைப்பு மற்றும் முறையியல்

இப்படமானது AI மொழி மாதிரிகள் (LLMs) கொங்கு வட்டாரத் தமிழை எவ்வாறு கையாள்கின்றன என்பதை ஆராயும் நான்கு கட்ட ஆய்வை விளக்குகிறது. இது ஒரு முன்மாதிரி தரவுத்தொகுப்பை உருவாக்குதல், ஒற்றை மாதிரியைப் பயன்படுத்தி பூஜ்ஜியச் சான்று/சிறு-சான்று முன்னோட்ட சோதனைகளை நடத்துதல், எதிர்கால பல்-LLM ஒப்பீட்டு முறையியலை வடிவமைத்தல் மற்றும் சொந்த மொழி பேசுநர்களைக் கொண்ட நிபுணர் குழுவின் மனித மதிப்பீட்டு நெறிமுறையை நிறுவுதல் ஆகியவற்றை உள்ளடக்கியது.

6. கொங்கு தமிழ் முன்மாதிரி அளவீட்டுத் தரவுத்தொகுப்பு

பின்வரும் அட்டவணை, சரிபார்க்கப்பட்ட கொங்கு வட்டாரச் சொற்களை அடிப்படையாகக் கொண்டு தொகுக்கப்பட்ட முன்மாதிரி தரவுத்தொகுப்பிலிருந்து எடுத்துக்காட்டு உருப்படிகளைக் காட்டுகிறது (முழு 20-உருப்படி தொகுப்பிலிருந்து தேர்ந்தெடுக்கப்பட்ட எட்டு உருப்படிகள்):

உருப்படி 1: கொங்கு - "அக்கட்ட எவனோ நிக்கான்" | தரப்படுத்தப்பட்ட தமிழ் - "அந்த இடத்தில் யாரோ நிற்கிறான்" | வட்டாரச் சொல் - அக்கட்ட (அந்த இடம்)

உருப்படி 2: கொங்கு - "நீ கம்மனாட்டி மாதிரி பேசாத" | தரப்படுத்தப்பட்ட தமிழ் - "நீ முட்டாள் மாதிரி பேசாதே" | வட்டாரச் சொல் - கம்மனாட்டி (முட்டாள்)

உருப்படி 3: கொங்கு - "ஆயா வீட்டுக்குப் போயிட்டு வந்தேன்" | தரப்படுத்தப்பட்ட தமிழ் - "பாட்டி வீட்டுக்குப் போய்விட்டு வந்தேன்" | வட்டாரச் சொல் - ஆயா (அப்பாவின் அம்மா)

உருப்படி 4: கொங்கு - "அவுங்க அக்கப்போர் போட்டாங்க" | தரப்படுத்தப்பட்ட தமிழ் - "அவர்கள் சண்டை போட்டார்கள்" | வட்டாரச் சொல் - அக்கப்போர் (சண்டை)

உருப்படி 5: கொங்கு - "வங்குல ஒளிஞ்சு உக்காந்தான்" | தரப்படுத்தப்பட்ட தமிழ் - "மறைவிடத்தில் ஒளிந்து உட்கார்ந்தான்" | வட்டாரச் சொல் - வங்கு (பொந்து/மறைவிடம்)

உருப்படி 6: கொங்கு - "அவன் தலையில் சிம்மாடு கட்டியிருந்தான்" | தரப்படுத்தப்பட்ட தமிழ் - "அவன் தலையில் தலைப்பாகை கட்டியிருந்தான்" | வட்டாரச் சொல் - சிம்மாடு (தலைப்பாகை)

உருப்படி 7: கொங்கு - "அது எங்க ஆத்தா வச்ச பேரு" | தரப்படுத்தப்பட்ட தமிழ் - "அது என் பாட்டி வைத்த பெயர்" | வட்டாரச் சொல் - ஆத்தா (அப்பாவின் அம்மா)

உருப்படி 8: கொங்கு - "அக்கியானம் பண்ணாத, வேலையப் பாரு" | தரப்படுத்தப்பட்ட தமிழ் - "தொல்லை கொடுக்காதே, வேலையைப் பார்" | வட்டாரச் சொல் - அக்கியானம் (தொல்லை/தொந்தரவு)

இத்தரவுத்தொகுப்பு மூன்று பணி வகைகளை உள்ளடக்குகிறது: (1) நேரடி மொழிபெயர்ப்பு பணி (கொங்கு வாக்கியத்தை தரப்படுத்தப்பட்ட தமிழில் மொழிபெயர்த்தல்), (2) சொல் அடையாளம் காணும் பணி (வாக்கியத்தில் உள்ள வட்டாரச் சொல்லின் பொருளை மட்டும் அடையாளம் காணுதல்), (3) சூழல் புரிதல் பணி (முழு வாக்கியத்தின் கருத்தை சுருக்கமாக விளக்குதல்).

கொங்கு தமிழ் முன்மாதிரி அளவீட்டுத் தரவுத்தொகுப்பு - எட்டு எடுத்துக்காட்டுகள்		
உருப்படி எண்	கொங்குத் தமிழ் வாக்கியம்	தரப்படுத்தப்பட்ட தமிழ் விளக்கம் & பணி வகை
1	அக்கட்ட எவனோ நிக்கான் → அந்த இடத்தில் யாரோ நிற்கிறான்	நேரடி மொழிபெயர்ப்பு
2	நீ கம்மனாட்டி மாதிரி பேசாத → நீ முட்டாள் மாதிரி பேசாதே	நேரடி மொழிபெயர்ப்பு
3	ஆயா வீட்டுக்குப் போயிட்டு வந்தேன் → பாட்டி வீட்டுக்குப் போய்விட்டு வந்தேன்	ஆயா பாட்டி சொல் அடையாளம் காணுதல்
4	அவுங்க அக்கப்போர் போட்டாங்க → அவர்கள் சண்டை போட்டார்கள்	நேரடி மொழிபெயர்ப்பு
5	வங்குல ஒளிஞ்சு உக்காந்தான் → மறைவிடத்தில் ஒளிந்து உக்கார்ந்தான்	சூழல் புரிதல்
6	அவன் தலையில் சிம்மாடு கட்டியிருந்தான் → அவன் தலையில் தலைப்பாகை கட்டியிருந்தான்	சிம்மாடு தலைப்பாகை சொல் அடையாளம் காணுதல்
7	அது எங்க ஆத்தா வச்ச பேரு → அது என் பாட்டி வைத்த பெயர்	நேரடி மொழிபெயர்ப்பு
8	அக்கியானம் பண்ணாத, வேலையப் பாரு → தொல்லை கொடுக்காதே, வேலையைப் பார்	சூழல் புரிதல்
தரவுத்தொகுப்பின் பணி வகைகள் (1) நேரடி மொழிபெயர்ப்பு (2) சொல் அடையாளம் காணுதல் (3) சூழல் புரிதல்		

Fig 6: AI மாதிரிகளுக்கான கொங்கு தமிழ் தரவுத்தொகுப்பு – எடுத்துக்காட்டுகள்

இப்படமானது AI மொழி மாதிரிகளைச் சோதிப்பதற்காக உருவாக்கப்பட்ட கொங்கு வட்டாரத் தமிழ் தரவுத்தொகுப்பிலிருந்து தேர்ந்தெடுக்கப்பட்ட 8 எடுத்துக்காட்டுகளை விளக்குகிறது. இது கொங்கு வாக்கியங்களை நேரடித் தரப்படுத்தப்பட்ட தமிழ் மொழிபெயர்ப்பு, குறிப்பிட்ட வட்டாரச் சொற்களுக்கான அடையாளம் காணுதல் மற்றும் முழு வாக்கியத்தின் சூழல் புரிதல் ஆகிய மூன்று முக்கியப் பணிகளின் அடிப்படையில் வகைப்படுத்தி அட்டவணைப்படுத்தியுள்ளது. கோயம்புத்தூர், ஈரோடு போன்ற பகுதிகளில் பேசப்படும் இந்த வட்டார வழக்கை AI எவ்வாறு கையாள்கிறது என்பதை இது காட்டுகிறது.

7. முன்னோட்ட ஆய்வு: Claude மொழி மாதிரியுடனான சோதனை

மேலே தொகுக்கப்பட்ட 20 உருப்படிகளில், எட்டு உருப்படிகள் Claude மொழி மாதிரிக்கு பூஜ்ஜியச் சான்று முறையில் (எவ்வித எடுத்துக்காட்டும் இன்றி, நேரடியாகக் கொங்கு வாக்கியத்தையும் "இதன் தரப்படுத்தப்பட்ட தமிழ் பொருள் என்ன?" எனும் கேள்வியையும் மட்டும் வழங்கி) சோதிக்கப்பட்டன. இதே எட்டு உருப்படிகள், மூன்று எடுத்துக்காட்டு ஜோடிகளை முன்னதாக வழங்கிய சிறு-சான்று அமைப்பிலும் மீண்டும் சோதிக்கப்பட்டன.

பூஜ்ஜியச் சான்று சூழலில், தெளிவான, பரவலாகப் பயன்படுத்தப்படும் வட்டாரச் சொற்களைக் கொண்ட வாக்கியங்களுக்கு (உதாரணமாக, "அக்கப்போர்", "ஆயா") மொழி மாதிரி பெரும்பாலும் சரியான அல்லது ஏறத்தாழச் சரியான தரப்படுத்தப்பட்ட தமிழ் பொருளை வழங்க முடிந்தது; ஏனெனில் இச்சொற்கள் பொதுவான தமிழ் இலக்கியம், திரைப்படங்கள் மற்றும் இணையதளங்களில் ஓரளவு பரவலாகக் காணப்படுபவை. எனினும், ஒப்பீட்டளவில் குறைவாகப் பயன்படுத்தப்படும் சொற்களைக் கொண்ட வாக்கியங்களில் (உதாரணமாக,

"வங்கு", "அக்கியானம்", "சிம்மாடு"), மொழி மாதிரி பெரும்பாலும் சொல்லை ஒலி ரீதியாக நெருக்கமான வேறு தமிழ்ச் சொல்லுடன் தவறாகப் பொருத்தியது அல்லது சூழலிலிருந்து ஏற்குறைய சரியான பொதுவான பொருளை மட்டும் யூகித்தது, துல்லியமான வட்டாரப் பொருளை அளிக்கவில்லை.

சிறு-சான்று சூழலில், மூன்று எடுத்துக்காட்டுகளை முன்பே வழங்கியதன் மூலம், மொழி மாதிரியின் விடைகளின் தரம் குறிப்பிடத்தக்க அளவில் மேம்பட்டது - குறிப்பாக வாக்கிய அமைப்பையும் பதிலின் வடிவத்தையும் (Format) பொறுத்தவரை. எனினும், அரிதான வட்டாரச் சொற்களின் துல்லியமான பொருள் இன்னும் பெரும்பாலும் தவறாகவே இருந்தது; இது, சிறு-சான்று கற்றல் தொடரியல்/வடிவ மட்டத்தில் (Syntactic/format level) உதவினாலும், சொல்லகராதி மட்ட அறிவுப் பற்றாக்குறையை (Lexical knowledge gap) முழுமையாக ஈடுகட்ட முடியவில்லை என்பதைக் காட்டுகிறது.

முக்கியமான வரம்பு: இச்சோதனை ஒரே ஒரு மொழி மாதிரி மீது, ஆசிரியரால் தன்மதிப்பீடு செய்யப்பட்ட 8 உருப்படிகளை மட்டுமே அடிப்படையாகக் கொண்டது. இது புள்ளியியல் ரீதியாக பொதுமைப்படுத்தக்கூடிய ஒரு முடிவு அல்ல; மாறாக, பிரிவு 5.3 மற்றும் 5.4-ல் முன்மொழியப்பட்ட முழுமையான பல்-LLM, பல்-மதிப்பீட்டாளர் ஆய்வுக்கான ஒரு தொடக்கப் புள்ளியாக மட்டுமே இதைக் கருத வேண்டும்.

8. பிழை பகுப்பாய்வு

முன்னோட்ட ஆய்வின் அவதானிப்புகளின் அடிப்படையில், கொங்கு தமிழ் கையாளுகையில் ஏற்படும் பிழைகளை நான்கு வகைகளாகப் பகுப்பாய்வு செய்யலாம்:

8.1 சொல்லடிப்படையிலான மாற்றுப் பிழை (Lexical Substitution Error): மொழி மாதிரி வட்டாரச் சொல்லை, ஒலி ரீதியாக நெருக்கமான ஆனால் பொருள் ரீதியாக வேறுபட்ட தரப்படுத்தப்பட்ட தமிழ்ச் சொல்லுடன் மாற்றிக் கொள்கிறது. உதாரணமாக, "வங்கு" (பொந்து/மறைவிடம்) என்பதை பொதுவான "இடம்" எனும் அளவிற்கு மட்டுமே எளிமைப்படுத்திப் புரிந்துகொள்வது.

8.2 சூழல் ரீதியான யூகம் (Contextual Guessing without Precision): அரிதான சொல் ஒன்றின் துல்லியமான பொருள் தெரியாதபோது, மொழி மாதிரி முழு வாக்கியத்தின் பொதுவான உணர்வைப் (General sentiment) பயன்படுத்தி தோராயமான பொருளை யூகிக்கிறது; இது பெரும்பாலும் திசையில் சரியாக இருந்தாலும், துல்லியத்தில் குறைபாடு காட்டுகிறது.

8.3 பதிவு/சூழல் பொருத்தமின்மை (Register Mismatch): மொழி மாதிரியின் மொழிபெயர்ப்பு பொருள் ரீதியாகச் சரியாக இருந்தாலும், அது வட்டார உரையாடலின் இயல்பான, பேச்சு

வழக்கு தன்மையை (Colloquial register) இழந்து, மிகவும் முறையான அல்லது புத்தக மொழியாக (Formal/Bookish Tamil) மாறிவிடுகிறது.

8.4 உருபனியல் புறக்கணிப்பு (Morphological Oversight): கொங்கு தமிழின் வினைமுற்று மற்றும் இடைச்சொல் மாற்றங்களை (எடுத்துக்காட்டாக, "போயிட்டு", "உக்காந்தான்" போன்ற வினைச்சொல் வடிவங்கள்) மொழி மாதிரி பெரும்பாலும் சரியாக அடையாளம் கண்டாலும், அவற்றின் நுட்பமான வேளாண்மை/கால அர்த்த வேறுபாடுகளை (Aspectual nuances) சில நேரங்களில் தவறவிடுகிறது.

இப்பிழை வகைப்பாடு, முழுமையான ஆய்வில் பயன்படுத்தப்பட வேண்டிய பிழை-குறியீட்டு அமைப்பிற்கான (Error-coding scheme) ஒரு தொடக்க வடிவமாக அமையும். முழுமையான ஆய்வில், ஒவ்வொரு தவறான விடையும் இந்நான்கு வகைகளில் ஒன்று அல்லது அதற்கு மேற்பட்டவற்றின் கீழ் மனித நிபுணர்களால் குறியிடப்பட வேண்டும், இதன் மூலம் ஒவ்வொரு LLM-க்கும் எந்த வகையான பிழை அதிகம் ஏற்படுகிறது என்பதை அளவீட்டு ரீதியாக ஒப்பிடலாம்.

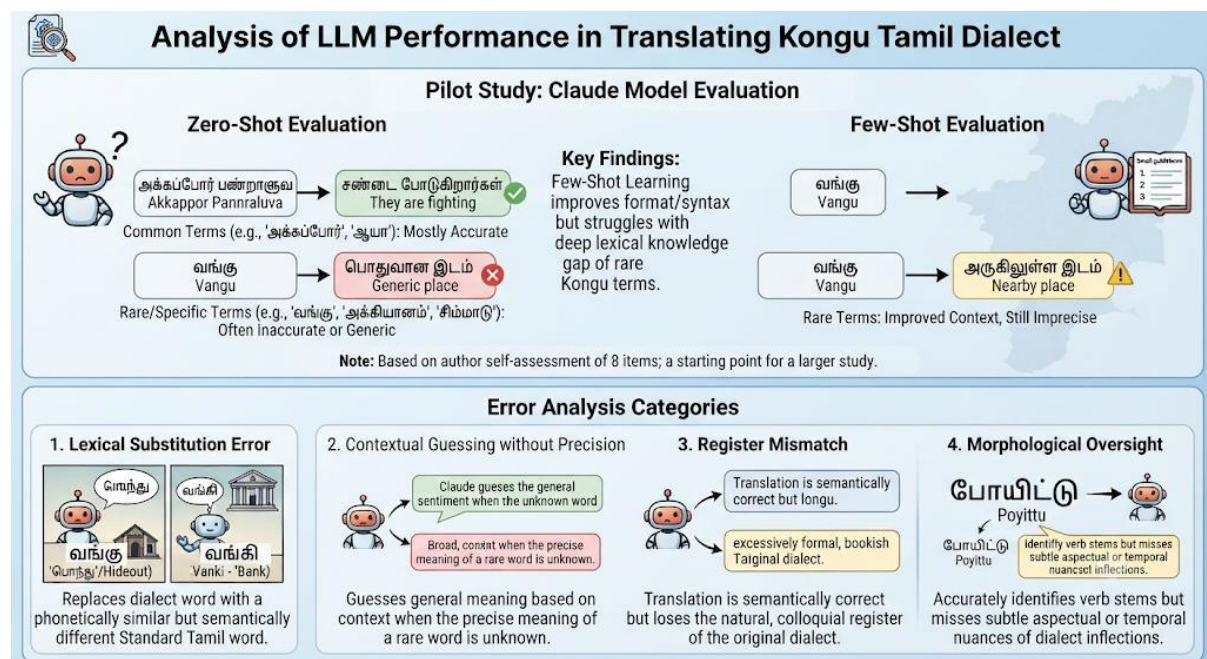


Fig 8: கொங்கு தமிழ் மொழிபெயர்ப்பில் கிளாட் (Claude) மாதிரியின் செயல்திறன் மற்றும்

பிழை பகுப்பாய்வு

இப்படம், கொங்கு தமிழ் வட்டாரச் சொற்களை மொழிபெயர்ப்பதில் மொழி மாதிரி (Claude) காட்டும் திறனை விளக்குகிறது. பூஜ்ஜியச் சான்று முறையில் பொதுவான சொற்கள் சரியாகப் புரிந்துகொள்ளப்பட்டாலும், அரிதான சொற்களில் பிழைகள் ஏற்படுகின்றன. சிறு-சான்று முறை வடிவத்தை மேம்படுத்தினாலும் சொல்லகராதி குறைபாட்டை முழுமையாகப் போக்குவதில்லை. மேலும், சொற்களை மாற்றுதல், சூழல் யுகம், பதிவுப் பொருத்தமின்மை

மற்றும் உருபனியல் பிழைகள் என நான்கு முக்கியப் பிழை வகைகளையும் இது சுட்டிக் காட்டுகிறது.

9. விவாதம்

இவ்வாய்வின் முன்னோட்ட முடிவுகள், தற்போதைய பெரும் மொழி மாதிரிகள் தரப்படுத்தப்பட்ட தமிழில் காட்டும் திறனை, வட்டார மொழி வடிவங்களில் முழுமையாக வெளிப்படுத்தவில்லை என்பதைச் சுட்டிக்காட்டுகின்றன. இது, LLM பயிற்சித் தரவுத்தொகுப்புகளில் வட்டார மொழி உரையின் பங்கு மிகக் குறைவாக இருப்பதன் விளைவாக இருக்கலாம் என ஊகிக்கலாம் - இணையத்தில் கிடைக்கும் தமிழ் உரையில் பெரும்பாலானவை தரப்படுத்தப்பட்ட எழுத்துத் தமிழிலேயே உள்ளன; பேச்சு வழக்கு/வட்டார உரை மிகக் குறைவாகவே ஆவணப்படுத்தப்பட்டு இணையத்தில் கிடைக்கிறது.

சிறு-சான்று கற்றல் மூலம் வடிவ ரீதியான மேம்பாடு ஏற்பட்டாலும், சொல்லகராதி மட்ட அறிவுப் பற்றாக்குறையை மூன்று எடுத்துக்காட்டுகள் மட்டும் ஈடுகட்ட முடியவில்லை என்பது குறிப்பிடத்தக்கது. இது, வட்டார மொழிகளுக்கான LLM திறனை மேம்படுத்த, ஒற்றை-முறை உசாத்துணை பொறியியல் (Prompt engineering) அணுகுமுறையை விட, பெரிய அளவிலான வட்டார மொழி தரவுத்தொகுப்புகள் மூலமான நேர்த்தி செய்தல் (Fine-tuning) அல்லது மீட்டிங் அடிப்படையிலான உசாத்துணை (Retrieval-Augmented Generation - RAG) அணுகுமுறைகள் மேலும் பயனுள்ளதாக இருக்கக்கூடும் என்பதை உணர்த்துகிறது.

10. சவால்களும் வரம்புகளும்

இவ்வாய்வு பல முக்கிய வரம்புகளைக் கொண்டுள்ளது, அவற்றை வெளிப்படையாகக் குறிப்பிடுவது அறிவியல் நேர்மைக்கு இன்றியமையாதது. முதலாவதாக, தரவுத்தொகுப்பு 20 உருப்படிகளை மட்டுமே கொண்டுள்ளது - இது புள்ளியியல் ரீதியாக பொதுமைப்படுத்தக்கூடிய அளவு அல்ல. இரண்டாவதாக, ஒரே ஒரு LLM (Claude) மட்டுமே நேரடியாகச் சோதிக்கப்பட்டது; GPT-4, Gemini, LLaMA போன்ற பிற முக்கிய மாதிரிகளுடனான உண்மையான ஒப்பீடு மேற்கொள்ளப்படவில்லை - அதற்கான முறையியல் மட்டுமே முன்மொழியப்பட்டுள்ளது. மூன்றாவதாக, மதிப்பீடு ஆசிரியரின் தன்மதிப்பீட்டை அடிப்படையாகக் கொண்டது; சுயாதீன, பல மனித நிபுணர்களைக் கொண்ட மதிப்பீட்டுக் குழு பயன்படுத்தப்படவில்லை. நான்காவதாக, தரவுத்தொகுப்பு உருவாக்கத்திற்குப் பயன்படுத்தப்பட்ட வாக்கியங்கள் ஆசிரியரால் இணைய அகராதி மூலங்களிலிருந்து கட்டமைக்கப்பட்டவை - உண்மையான பேச்சு வழக்கிலிருந்து பதிவு செய்யப்பட்ட இயற்கையான தரவு (Naturally-occurring corpus data) அல்ல.

இவ்வரம்புகள் காரணமாக, இக்கட்டுரையின் அனுபவமுறை முடிவுகள் தொடக்கநிலைக் குறிப்புகளாக (Indicative observations) மட்டுமே கருதப்பட வேண்டும், இறுதி, பொதுமைப்படுத்தக்கூடிய அறிவியல் முடிவுகளாக அல்ல.

11. எதிர்கால ஆய்வுக்கான பரிந்துரைகள்

1. குறைந்தது 500-1000 வாக்கியங்களைக் கொண்ட, கள ஆய்வு மூலம் உண்மையான கொங்கு தமிழ் பேசுநர்களிடமிருந்து சேகரிக்கப்பட்ட ஒரு பெரிய அளவிலான, இயற்கையான கொங்கு தமிழ் அளவீட்டுத் தரவுத்தொகுப்பை உருவாக்க வேண்டும்.
2. GPT-4, Gemini, Claude, LLaMA, மற்றும் இந்திய-மொழி சிறப்பு மாதிரிகளான IndicTrans2 உள்ளிட்ட குறைந்தது ஐந்து LLMகளை ஒரே தரவுத்தொகுப்பில் சீரான நிலையில் ஒப்பிட்டு ஆய்வு செய்ய வேண்டும்.
3. குறைந்தது ஐந்து சொந்த கொங்கு தமிழ் பேசுநர் நிபுணர்களைக் கொண்ட ஒரு மதிப்பீட்டுக் குழுவை அமைத்து, மதிப்பீட்டாளர் இடையேயான உடன்பாட்டை புள்ளியியல் ரீதியாக அளவிட வேண்டும்.
4. வட்டார மொழி தரவுத்தொகுப்புகளைப் பயன்படுத்தி LLMகளை நேர்த்தி செய்தல் (Fine-tuning) அல்லது மீட்பு-அடிப்படையிலான உசாத்துணை (RAG) அணுகுமுறைகளின் செயல்திறனை பூஜ்ஜியச் சான்று/சிறு-சான்று அணுகுமுறைகளுடன் ஒப்பிட்டு ஆராய வேண்டும்.
5. கொங்குத் தமிழுக்கு மட்டுமல்லாமல், நாஞ்சில் நாட்டுத் தமிழ், கரிசல் வட்டார வழக்கு உள்ளிட்ட பிற தமிழ் வட்டார வழக்குகளுக்கும் இதே முறையியலை விரிவுபடுத்தி, ஒரு முழுமையான தமிழ் வட்டார மொழி LLM மதிப்பீட்டு அளவுகோலை (Pan-Tamil-dialect LLM benchmark) உருவாக்க வேண்டும்.

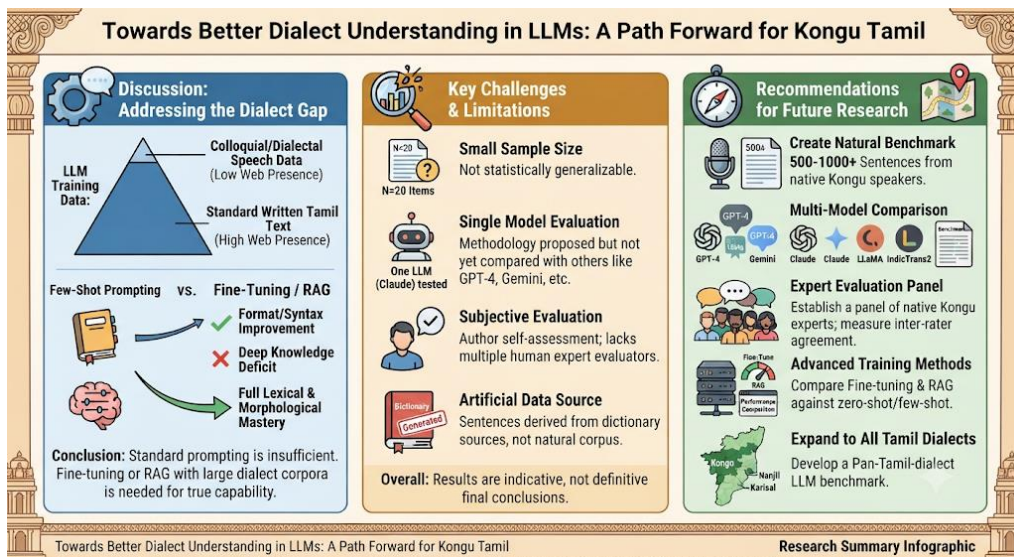


Fig 9: கொங்கு தமிழ் மொழி மாதிரிகளின் சவால்கள் மற்றும் எதிர்கால ஆய்வுக் குறிப்புகள்

இப்படம் வட்டார மொழி தரவுப் பற்றாக்குறையை விளக்குகிறது. மாதிரிகளை மேம்படுத்த ஃபைன்-டியூனிங் அல்லது RAG முறைகள் அவசியம். சிறிய மாதிரி அளவு, செயற்கைத் தரவு போன்ற சவால்களை இது சுட்டிக் காட்டுகிறது. மேலும், பெரிய கள ஆய்வுத் தரவுத்தொகுப்பு உருவாக்கம், பல மாடல் ஒப்பீடு மற்றும் பிற தமிழ் வட்டார வழக்குகளுக்கு இதை விரிவுபடுத்துதல் போன்ற எதிர்காலப் பரிந்துரைகளையும் முன்வைக்கிறது.

12. முடிவுரை

இக்கட்டுரை, தமிழ் வட்டார மொழிகளை AI மொழி மாதிரிகள் கையாளும் திறனை, கொங்கு வட்டாரத் தமிழை மையமாகக் கொண்டு ஆராய்ந்தது. சரிபார்க்கப்பட்ட வட்டாரச் சொற்களஞ்சியத்தை அடிப்படையாகக் கொண்ட ஒரு முன்மாதிரி தரவுத்தொகுப்பு உருவாக்கப்பட்டு, ஒரு LLM மீது பூஜ்ஜியச் சான்று மற்றும் சிறு-சான்று சோதனைகள் மேற்கொள்ளப்பட்டன. முடிவுகள், பரவலாகப் பயன்படுத்தப்படும் வட்டாரச் சொற்களை மொழி மாதிரி ஓரளவு சரியாகக் கையாண்டாலும், அரிதான சொற்கள், பதிவு/சூழல் பொருத்தம், மற்றும் நுட்பமான உருபனியல் வேறுபாடுகளில் குறிப்பிடத்தக்க பிழைகள் ஏற்படுவதைக் காட்டுகின்றன.

மிக முக்கியமாக, இக்கட்டுரை ஒரு முழுமையான அனுபவமுறை ஆய்வு அல்ல; மாறாக, தமிழ் வட்டார மொழிகளுக்கான AI மதிப்பீட்டு ஆய்வுத் துறையில் இதுவரை காணப்படும் இடைவெளியை அடையாளம் கண்டு, அதை நிவர்த்தி செய்வதற்கான ஒரு விரிவான, செயல்படுத்தக்கூடிய முறையியல் திட்டத்தை வழங்குகிறது - பெரிய அளவிலான தரவுத்தொகுப்பு உருவாக்கம் முதல், பல்-LLM ஒப்பீடு, மனித நிபுணர் மதிப்பீடு, மற்றும் முறையான பிழை பகுப்பாய்வு வரை. தமிழ் போன்ற பணக்கார வட்டார மொழி பன்முகத்தன்மை கொண்ட மொழிகளுக்கு, AI தொழில்நுட்பங்கள் உண்மையிலேயே பயனுள்ளதாக இருக்க வேண்டுமெனில், தரப்படுத்தப்பட்ட மொழி வடிவங்களுக்கு அப்பால், மக்கள் அன்றாடம் பேசும் வாழும் மொழியையும் அவை புரிந்துகொள்ள வேண்டும் என்பதே இவ்வாய்வின் மையக் கருத்தாகும்.

13. துணை நூல்கள் (References)

1. Alam, M. M. I., & Anastasopoulos, A. (2024). CODET: A benchmark for contrastive dialectal evaluation of machine translation. In Findings of the Association for Computational Linguistics: EACL 2024 (pp. 1790–1859). Association for Computational Linguistics.
2. BhashaSutra: A task-centric unified survey of Indian NLP datasets, corpora, and resources. (2025). arXiv preprint.
3. Chakravarthi, B. R., Priyadharshini, R., Muralidaran, V., Suryawanshi, S., Jose, N., Sherly, E., & McCrae, J. P. (2021). Findings of the shared task on Dravidian language technology (DravidianLangTech). Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages.

4. From phonemes to meaning: Evaluating large language models on Tamil (ILAKKANAM). (2025). arXiv:2511.12387.
5. G.T.N. Arts College, Dindigul. (n.d.). The impact of artificial intelligence and natural language processing on the Tamil language preservation and cultural promotion. *Tamil Computing Journal*.
6. Jageer, S. K., & Priyaradhikadevi. (2025). A comparative analysis of large language models for English-to-Tamil machine translation with performance evaluation. *Panamerican Mathematical Journal*, 35(2s).
7. Joshi, P., Santy, S., Budhiraja, A., Bali, K., & Choudhury, M. (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* (pp. 6282–6293).
8. Rajendran, S., Anand Kumar, M., Rajalakshmi, R., Dhanalakshmi, V., Balasubramanian, P., & Soman, K. P. (2022). Tamil NLP technologies: Challenges, state of the art, trends and future scope. In *International Conference on Speech and Language Technologies for Low-Resource Languages* (pp. 73–98). Springer.
9. Shaffi, N., & Hajamohideen, F. (2021). UTHCD: A new benchmarking for Tamil handwritten OCR. *IEEE Access*, 9, 101469–101493.
10. Tamil Wikipedia. (2024). தமிழ் வட்டார மொழி வழக்குகள் [Tamil regional dialects]. தமிழ் விக்கிப்பீடியா.
11. Tamil Wikipedia. (2025). கொங்கு வட்டார வழக்கு அகராதி [Kongu dialect dictionary]. தமிழ் விக்கிப்பீடியா.
12. Tamil Wikipedia. (2026). கொங்குத் தமிழ் [Kongu Tamil]. தமிழ் விக்கிப்பீடியா.
13. *Tamilmanam International Research Journal of Tamil Studies*. (2025a). Artificial intelligence technology: A boon in writing Tamil essays. *Tamilmanam International Research Journal of Tamil Studies*, 1(4), 221–228. <https://doi.org/10.63300/dn796x92>
14. *Tamilmanam International Research Journal of Tamil Studies*. (2025b). Artificial intelligence for Tamil literature: An overview. *Tamilmanam International Research Journal of Tamil Studies*.
15. Verma, S. S. U. R., Khan, M. S. U. R., Kumar, V., Murthy, R., & Sen, J. (2025). MILU: A multi-task Indic language understanding benchmark. In *Proceedings of the 2025 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 10076–10132).



Copyright © 2024 by the author(s). Published by Department of Library, Nallamuthu Gounder Mahalingam College, Pollachi. This is an open access article under the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).