VakyArth: Evaluating Pragmatic Competence in LLMs across Indic Languages

Usneek Singh[1*] **Poorvaja Veera Balaji Kumar**[1] **Parth Nanda**[1]
Anand Madhusoodanan[1] **Geyang Guo**[1] **Wei Xu**[1] **Junyi Jessy Li**[2]

[1]Georgia Institute of Technology, [2]University of Texas at Austin
* usingh68@gatech.edu

Abstract

Real-world communication often requires pragmatic reasoning: interpreting meanings implied through context and cultural convention rather than stated literally. Existing pragmatic evaluation remains largely limited to English and high-resource languages, leaving Indic languages unexplored despite their linguistic and cultural diversity. We introduce **VakyArth**, the first pragmatic benchmark for Indic languages, designed as a diagnostic evaluation covering Hindi, Punjabi, Tamil, and Malayalam. VakyArth evaluates models across five phenomena: deixis, speech acts, implicature, social pragmatics, and coherence; through multiple-choice questions, natural language inference, and translation, with all items authored by native speakers. Across multilingual large language models (LLMs) of varying families and sizes, we find consistent failures on pragmatic meanings rooted in Indic linguistic and cultural conventions. Our analysis shows systematic differences across languages and tasks: MCQ accuracy exceeds NLI accuracy in all model-language combinations, translation performance does not reliably track pragmatic understanding, and Indo-Aryan languages show a translation advantage over Dravidian languages. We further show that automatic translation metrics can miss fluent but pragmatically unfaithful outputs, especially for implicature and deixis.

1 Introduction

As large language models (LLMs) become widely used, their impact depends not only on supporting professional tasks but also on helping users navigate everyday communication across diverse linguistic and cultural contexts. India offers such a setting: it has the second-largest AI user base globally, yet usage remains concentrated in professional contexts such as software development, with much less adoption for everyday and personal tasks (Appel, 2026). Broadening AI's impact in India, therefore, requires moving beyond work-related applications toward systems that can support daily usage.

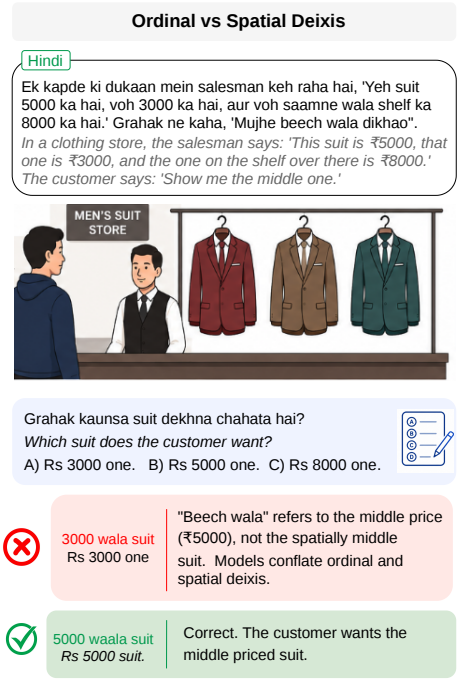

Figure 1: A Hindi deixis item from VAKYARTH. The customer asks for *beech wala* (the middle one), referring to the middle-priced suit (Rs 5000). Any Hindi speaker resolves this by price ranking, not by the order of mention. Evaluated models pick Rs 3000 missing that *beech wala* here anchors to price, not utterance order.

Everyday communication in Indic contexts often depends on culturally grounded meanings that go beyond literal form. For example, as shown in Figure 1, when a Hindi-speaking customer asks for "*beech wala*" in a clothing-store dialogue, the phrase literally means "*the middle one*". In context, however, the phrase refers to the middle-priced suit rather than the spatially middle item, a distinction missed by multiple existing LLMs (e.g., Gemma-4-31B). Pragmatic reasoning is needed here: recovering meaning from context, cultural convention, and speaker intent (Ma et al., 2025).

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 43027–43041
October 24-29, 2026 ©2026 Association for Computational Linguistics

Existing pragmatic evaluation mainly focuses on English or high-resource languages (Ma et al., 2025; Park et al., 2024b). To the best of our knowledge, no pragmatic evaluation exists for Indic languages. Furthermore, Indic pragmatics is highly diverse: languages across India span multiple families, scripts, and sociolinguistic communities, with pragmatic cues encoded differently.

We introduce **VakyArth**[1], the first pragmatic benchmark for Indic languages. VakyArth is a diagnostic benchmark that covers Hindi, Punjabi, Tamil, and Malayalam, which represent diverse linguistic and regional contexts across India. It evaluates multilingual LLMs along five pragmatic phenomena: deixis, speech acts, implicature, social pragmatics, and coherence (see examples in Figure 2). Each phenomenon is constructed in the corresponding language and evaluated under three task formats: multiple-choice questions, natural language inference, and translation, allowing us to assess pragmatic understanding at different levels of explicitness. All items are authored by native speakers and include naturally code-mixed instances to reflect real usage patterns.

We evaluate five multilingual LLMs across model families and sizes (ranging from 8B to 111B), and find consistent failure patterns rooted in Indic linguistic and cultural conventions. Models struggle with culturally grounded sarcasm and hyperbole, indirect refusals and requests, shifting deictic anchors such as "kal" and "parso", and social cues involving kinship address and gender norms. Our quantitative analysis further reveals systematic differences across tasks, phenomena, and languages. Indo-Aryan languages show a translation advantage over Dravidian languages, likely reflecting differences in pretraining coverage (Naous and Xu, 2025). MCQ accuracy exceeds NLI accuracy in all model-language combinations, suggesting pragmatic interpretation benefits from answer-choice scaffolding. Translation performance does not track MCQ or NLI scores, indicating that recognizing pragmatic meaning and expressing it faithfully across languages are distinct capabilities. Interestingly, Gemma-4-31B achieves the highest performance despite being smaller and lacking Indic specialization. Automatic metrics also fail to penalize fluent but pragmatically unfaithful outputs for implicature and deixis.

[1] *VakyArth* combines *Vakya* (utterance) and *Arth* (meaning) from Sanskrit, reflecting the benchmark's focus on the meaning beneath the utterance.

We release our dataset and code at `https://github.com/Usneek1/VakyArth`.

2 Related Work

Pragmatic evaluation of LLMs. The evaluation of pragmatic competence in LLMs has attracted growing attention. Ma et al. (2025) provide a comprehensive survey, noting that most resources cover a narrow set of phenomena and are heavily skewed toward high-resource languages. Early diagnostic work includes IMPPRES (Jeretic et al., 2020), a targeted NLI dataset for implicature and presupposition, and Ruis et al. (2023), who showed instruction-tuned models substantially outperform base models on implicature resolution. More comprehensive benchmarks have since emerged: PUB (Sravanthi et al., 2024) covers implicature, presupposition, reference, and deixis across fourteen multiple-choice tasks, and Yu et al. (2026) focus on alternative-based pragmatic reasoning. All of these benchmarks are in English.

Multilingual and cross-cultural pragmatic benchmarks. Extending pragmatic evaluation beyond English has been limited. MultiPragEval (Park et al., 2024b) evaluates models across English, German, Korean, and Chinese using Gricean maxims, but covers only European and East Asian languages. Park et al. (2024a) and Koo et al. (2025) introduce pragmatic benchmarks for Korean, with latter reporting substantial gaps between human and model performance on indirect speech acts. A related line of work evaluates cultural knowledge rather than pragmatic competence (Chiu et al., 2024; Shaikh et al., 2023). Unlike these studies, which test whether models know cultural facts, our work evaluates whether models can reason about implied meaning in context. No prior work evaluates pragmatic competence for any Indic language.

Evaluation of LLMs on Indic languages. Indic NLP evaluation has progressed from foundational benchmarks to increasingly culturally grounded ones, yet pragmatic competence remains unaddressed. Early work such as IndicNLPSuite (Kakwani et al., 2020) and IndicXTREME (Doddapaneni et al., 2023) established baselines for classification and named entity recognition. More recent work covers generation: IndicGenBench (Singh et al., 2024) evaluates translation and summarization across 29 languages, and MILU (Verma et al., 2025) introduces culturally specific knowl-

edge tasks. The most culturally grounded effort is PARIKSHA (Watts et al., 2024), which evaluates 30 models across 10 Indic languages and finds that LLM evaluators align poorly with humans on culturally nuanced responses a finding that resonates with our own human-COMET divergence on pragmatic translation. Despite this progression, none of these benchmarks require pragmatic inference: understanding implied meaning, resolving deictic reference, or interpreting indirect speech acts. VakyArth is the first to address this gap.

3 Evaluation Design

We design our evaluation around three guiding principles: **phenomena coverage** (the testing should span the core dimensions of pragmatic competence, not a single category), **task diversity** (evaluation should probe understanding at multiple levels of explicitness), and **cultural authenticity** (items should reflect how Indic languages are actually used, including code-mixing and culturally specific pragmatic conventions).

3.1 Pragmatic Phenomena

We evaluate five phenomena that collectively capture the range of non-literal, context-dependent meaning central to everyday communication (see Figure 2). These categories follow the taxonomy of Ma et al. (2025), but are instantiated through distinctively Indic examples drawn from cultural knowledge of native speakers.

(1) Deixis refers to words whose reference depends entirely on the discourse context, who is speaking, where, and when (Webber, 1988). Indic languages present particular deictic challenges: temporal words like *kal* (yesterday/tomorrow) and *parso* (day before yesterday/day after tomorrow) are bidirectional, resolving to past or future only through context; proximal-distal contrasts like *intha-antha* carry meaning that translation flattens; and cultural calendar references. *For example*,

> A Tamil item presents a speaker comparing two pumpkins: “intha poosani paaka pinjaa irukku, antha poosani nallaa irukku” (“this pumpkin looks less ripe, that one looks better”) illustrates how the proximaldistal contrast (*intha* vs. *antha*) carries the entire communicative point. A translation that flattens both to “the pumpkin” is grammatically fine but loses the comparison entirely.

We also add multi-party dialogues for more complexity. For example, in a Punjabi station exchange, speaker B says “tussi edher aa jao” (“you come over here”) while speaker A is already walking toward the wrong gate. Correctly resolving “edher” (“here”) requires tracking which gate each speaker is at - a reference chain that models often collapse.

(2) Speech Acts are utterances that perform actions such as requests, offers, refusals, and warnings, whose illocutionary force often diverges substantially from their surface form (Austin, 1975). South Asian communicative norms make particularly heavy use of indirect speech acts as politeness strategies. *For example,*

> A mother who enters her son’s room at 11 AM and asks (in Hindi) *"Beta, kya aaj chutti hai kya?"* ("Son, is it a holiday today?") is not seeking calendar information. She is delivering an indirect command to wake up and get to work.

Likewise, a Punjabi host who says “*kuch khaas nahi baneya*” (“nothing special was made”) while placing food in front of guests is performing a ritual offer, not a sincere self-criticism.

(3) Implicature refers to meaning conveyed but not explicitly stated, requiring the listener to reason beyond the literal content (Bach, 2006). In Indic languages, implicature is frequently expressed through culturally embedded expressions that are opaque without cultural knowledge. *For example,*

> A teacher who tells Rahul "Tumne toh iss pareeksha mein pass hokar kila jeet liya" ("You have conquered a fort by passing this exam") is using a common Hindi hyperbole to express praise. A model that processes the surface meaning predicts that Rahul won a real fort; the correct inference is that the teacher is simply congratulating him on passing.

Similarly, in Tamil, a mother telling departing guests *“kai nenachu po”* (“wet your hands and go”) is not issuing a hygiene instruction but extending a warm, face-saving invitation to stay and eat. The ritual indirectness is the social point of the expression.

(4) Social Pragmatics captures how social factors such as kinship, gender norms, power dynamics, and cultural conventions around face shape language use (Ma et al., 2025). These items require models to bring cultural world knowledge to bear

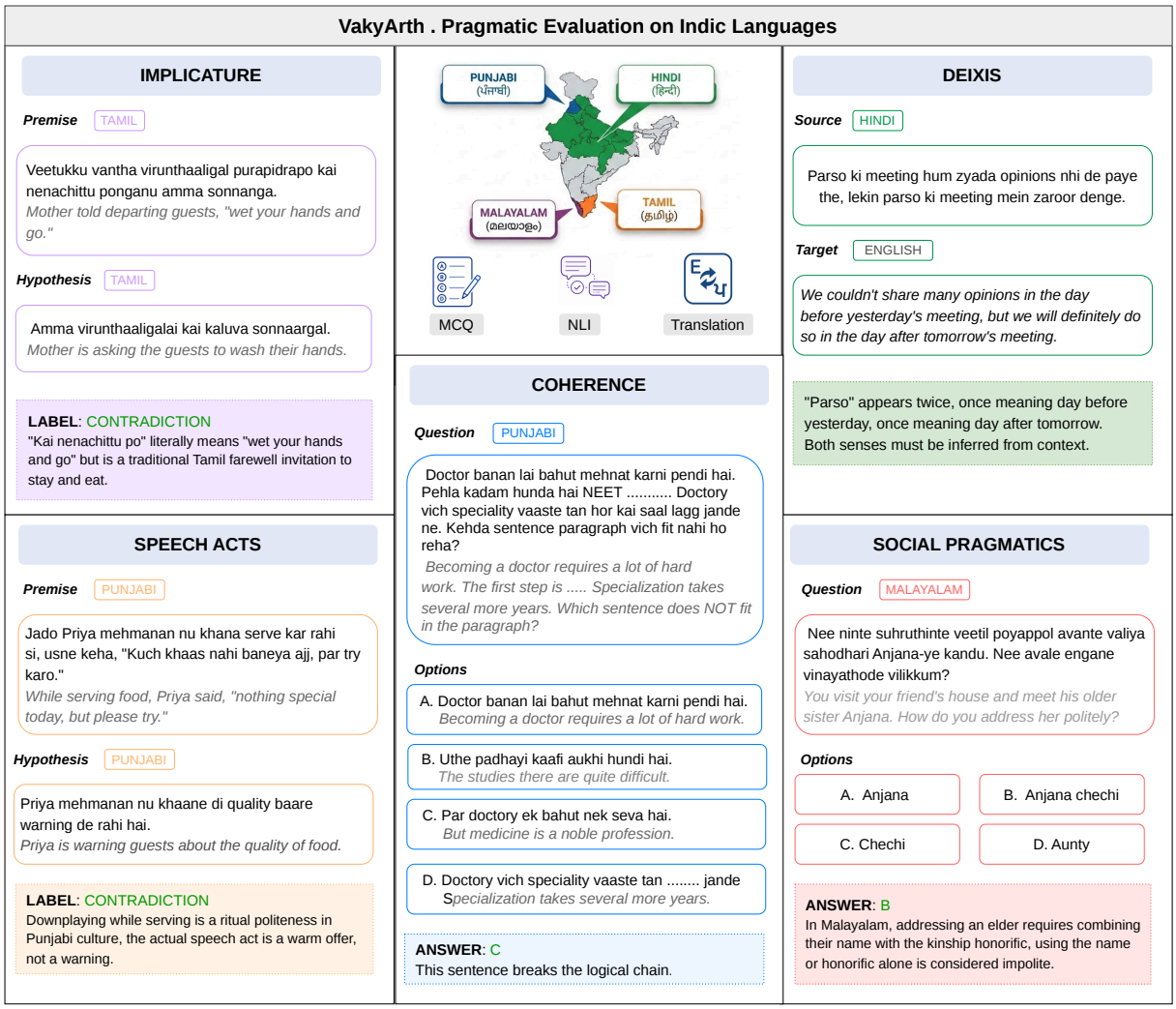

Figure 2: VakyArth benchmark overview. Items span five pragmatic phenomena : implicature, deixis, speech acts, social pragmatics, and coherence across Hindi, Punjabi, Tamil, and Malayalam, evaluated under NLI, translation, and MCQ formats. In each case, the correct label or translation cannot be recovered from surface form alone.

in ways that go beyond linguistic competence. *For example,*

> In Punjabi, greeting a male friend with a hug while greeting his wife with folded hands (*hath jodna*) and *Sat Sri Akal* is the culturally appropriate norm for mixed-gender visits. A model without cultural grounding might incorrectly read this a sign of differential regard toward the wife.

In Malayalam, addressing a friend's older sister requires combining her name with the kinship honorific *chechi*: "Anjana chechi" is the polite form, while "Anjana" alone signals disrespect and "Chechi" alone is too impersonal.

(5) Coherence requires reasoning about how utterances relate within a larger discourse unit, beyond what any single sentence conveys (Van Dijk et al., 1997). Existing work draws extensively on sentence-ordering or narrative cloze tasks (Barzilay and Lapata, 2008; Mostafazadeh et al., 2016; Beyer et al., 2021). In contrast, our items test discourse-level reasoning within Indic language paragraphs, where topically plausible but discourse-incoherent distractors require models to track argumentative structure rather rely on keyword overlap. *For example,*

> A Tamil paragraph describes a traditional banana leaf meal in sequence.. hot rice, lentils, ghee, kootu, poriyal ending with payasam. The most appropriate response the accompaniments like vadai and appalam that complete the feast. Options like "banana leaf meals are no longer common" or "it is difficult to serve on a leaf" are topically related to banana leaf meals but break the sequence the paragraph has built up.

3.2 Task Types

Each phenomenon is evaluated across three task formats, chosen to probe pragmatic understanding at different levels of explicitness:

Multiple-Choice Question (MCQ). The model picks the most appropriate interpretation from four options. We deliberately construct distractors that reflect the literal meaning of the utterance, so models that do not reason pragmatically will be misled.

Natural Language Inference (NLI). The model labels the relationship between a premise and a hypothesis as entailment, contradiction, or neutral. This is intended to be harder than MCQ since no options to guide the model. We include more contradiction cases, especially for implicature, since models tend to treat figurative premises as entailing hypotheses that they actually contradict.

Translation. The model translates an Indic-language utterance into English. Translating pragmatic content is challenging because implied meanings, cultural references, and speaker intent do not always have a direct equivalent in English. A model that does not understand the pragmatic meaning will produce a translation that is fluent but wrong.

3.3 Dataset Construction

All items were written from scratch by a team of four native speakers, one per language, each with graduate-level NLP training. Annotators drew on personal linguistic knowledge, cultural familiarity, and grammar references to write items. Each item was reviewed by the full team to check for ambiguity, label correctness, and translation accuracy. Items where the correct answer was recoverable from surface cues alone were removed. Most items are code-mixed with English. This reflects how people actually use these languages in daily life, especially in Hindi and Punjabi. Tamil and Malayalam items have less code-mixing, consistent with real usage patterns in those communities.

Table 1 summarizes the language statistics. VakyArth contains **578** items across four languages: Hindi (185), Punjabi (125), Tamil (134), Malayalam (134), spanning a combined speaker population exceeding 800 million. The corpus totals 14,921 words of source text and 424 human-authored English reference sentences in the translation subcorpus. The token/word figures reflect a fundamental typological difference between the two language families: Tamil and Malayalam are agglutinative, requiring significantly more tokens per word than the morphologically simpler Hindi & Punjabi, consistently across all three tokenizers.

	Hindi	**Punjabi**	**Tamil**	**Malay.**
Family	Indo-Aryan	Indo-Aryan	Dravidian	Dravidian
Script	Devanagari	Gurmukhi	Tamil	Malay.
Items	185	125	134	134
TR sen.	147	48	86	143
T/W (Llama)	1.95	1.90	3.15	4.44
T/W (Gemma)	1.76	1.73	2.84	3.82
T/W (Qwen)	1.96	1.87	3.13	4.32

Table 1: Language statistics for VakyArth. T/W (Tokens/Words) is the average tokens per word under each model's tokenizer for Latin-transliterated source text. Dravidian languages consistently require 1.6–2.5× more tokens per word than Indo-Aryan languages across all three tokenizers. TR sentences = total sentences in the translation subcorpus.

4 Evaluation Setup

Models. We evaluate five instruction-tuned models ranging from 8B to 111B parameters. Three are open-weight models run locally on an NVIDIA H100 GPU via the HuggingFace Transformers library (Wolf et al., 2020): `Llama-3-8B-Instruct`, `Gemma-4-31B`, and `Qwen3-32B`. The remaining two are open-weight multilingual models accessed via their respective APIs rather than run locally: `Sarvam-105B`, which has dedicated Indic language pretraining, and `Command-A-03-2025` (111B parameters), Cohere's flagship multilingual model. All models are run with a temperature of 0.2 and a maximum of 128 new tokens.

Prompting. We use a three-shot prompting setup for all models and tasks. Each prompt contains three fixed examples written in the same language and script as the test item. Examples are shared across phenomena within a language but are task-specific. We use few-shot prompting rather than zero-shot because pilot runs showed that models frequently failed to follow the required output format under zero-shot conditions, making automatic parsing unreliable. More details on prompts can be found in Appendix A.

Input format. All items are presented in Latin transliteration. This choice reflects everyday usage: Roark et al. (2020) documents informal romanization as the standard way South Asian languages are typed in digital communication, which is the setting our benchmark targets. Methodologically,

Jaavid et al. (2024) show that romanized Indic text reduces token fertility by 2–4× and matches or outperforms native-script representation across NLU, NLG, and MT tasks, supporting our choice to minimize tokenization artifacts. We additionally show the effect of using native script vs. Latin script on pragmatic understanding (see §5.5).

Evaluation metrics. For MCQ and NLI, we report accuracy, computed by exact match between the model's predicted label and the gold answer. For translation, we report two scores: (1) automatic evaluation using COMET (Rei et al., 2020) with the `wmt22-comet-da` checkpoint, and (2) human evaluation scores on a 1–5 scale covering adequacy, fluency, and pragmatic adequacy, collected from four native speaker raters, one per language. We chose `wmt22-comet-da` over newer checkpoints after empirically validating that it achieves the highest correlation with human judgements in our setting compared to XCOMET (Guerreiro et al., 2024) and MetricX (Juraska et al., 2023); we present a detailed comparison of all three metrics in §5.4.

5 Results and Analysis

Table 2 presents results across all four models, four languages, and five phenomena for MCQ accuracy, NLI accuracy, and COMET score. We organize our analysis around the pragmatic failure patterns that VakyArth uncovers. We discuss what is specifically challenging about Indic pragmatics for current models. We also present some quantitative patterns and insights in §5.2.

5.1 Pragmatic Failure Patterns

Deictic Reference Failure. Models frequently struggle to resolve deictic expressions that shift fluidly across discourse contexts. In particular, models make mistakes on bidirectional temporal expressions in Hindi and Punjabi. For instance, if a speaker says on Monday, *"kal mera interview hai"* ("My interview is tomorrow"), and is asked on Tuesday how it went, Llama, Command-A and Gemma fail to correctly anchor the event timeline. Beyond absolute shifts, models also fail on relative calendar references conventionalized among native speakers: when presented with a Friday utterance like *"agle hafte Somvar ko exam hai"* ("The exam is next week on Monday"), models compute the target date as seven days from the utterance rather than three.

Culturally Embedded Expressions. Models make mistakes in understanding idiomatic expressions rooted in regional socio-religious traditions. For instance, when a mother in Tamil tells departing guests *"kai nenachittu ponganu"* ("wet your hands and go"), Llama, Gemma, and Qwen predict an instruction to return home and eat, while Sarvam and Command-A classify it as a hygiene command. No model recovers the correct pragmatic meaning which is a conventionalized invitation to stay for a meal. Similarly, when someone in Punjabi states *"bas meri kudi da viyah karke mai ganga nha lavan"* ("once my daughter is married, I will bathe in the Ganga"), all models predict a literal travel intention, missing the culturally embedded vow signifying relief at fulfilling a major parental duty.

Indirect Speech Acts and Soft Refusals. South Asian communicative norms strongly disprefer direct refusals and favor indirect speech acts that models misinterpret at face value. For instance, when a speaker declines an invitation with the standard Hindi hedge *"dekhte hain, try karungi"* ("Let's see, I'll try"), models predict she intends to attend. Native speakers recognize this as a polite soft declination. When a Tamil host asks a departing guest *"enna avasaram"* ("What is the rush?"), models interpret it as an inquiry into an actual emergency rather than a standard hospitality request, urging the guest to stay longer.

Social and Kinship Norms. Models fail to interpret utterances whose meaning depends on regional social conventions, or ritual politeness scripts. For instance, in Punjabi culture, an elder telling a younger relative *"tu taan bada kamzor lag reha hai"* (You look quite thin) is a conventionalized expression of care; all five models misread it as mockery. Models also fail to parse inverted politeness scripts: when a young relative accepts cash at a Malayalam wedding saying *"ithonnum vendaayirunnu, chetta/chechi"* (This wasn't necessary, brother/sister), models predict a genuine rejection, failing to recognize the socially obligatory script of verbal refusal during physical acceptance. The ritual farewell *"poyittu varam"* (I will go and come back) is read as a literal promise of immediate return by the evaluated models.

Sarcasm, Hyperbole, and Implied Criticism. Models misinterpret non-literal assertions, such as detecting hyperbole or decoding sarcasm used

Tasks: MCQ (Accuracy) NLI (Accuracy) Translation (COMET)

Lang.	Model	Deixis			Speech Acts			Implicature			Social Prag.			Coherence		
		MCQ	NLI	Trans.	MCQ	NLI	Trans.	MCQ	NLI	Trans.	MCQ	NLI	Trans.	MCQ	NLI	Trans.
Hindi	Llama-3-8B	0.64	0.38	0.80	**1.00**	0.45	0.78	0.73	0.20	0.76	**0.90**	0.22	0.71	0.79	0.50	0.80
	Gemma-4-31B	0.79	0.46	**0.85**	**1.00**	0.45	**0.84**	**0.91**	0.30	**0.86**	**0.90**	0.22	0.82	0.93	0.50	0.83
	Qwen3-32B	0.71	0.85	0.73	0.90	**0.73**	0.54	0.73	**0.80**	0.58	**0.90**	**0.44**	0.77	0.93	**0.92**	0.76
	Sarvam-105B	**0.93**	0.69	0.81	**1.00**	0.55	0.67	**0.91**	0.30	0.71	**0.90**	0.33	0.75	**1.00**	0.75	0.77
	Command-A-111B	0.86	**0.92**	0.84	**1.00**	**0.73**	0.84	**0.91**	**0.80**	0.83	**0.90**	**0.44**	**0.83**	0.93	**0.92**	**0.83**
Punjabi	Llama-3-8B	0.22	0.22	0.78	0.33	0.44	0.75	0.10	0.33	0.67	0.67	0.11	0.61	0.40	**0.50**	0.86
	Gemma-4-31B	**0.78**	0.33	**0.85**	**1.00**	**0.56**	**0.82**	**0.60**	0.33	**0.78**	**1.00**	**0.22**	**0.79**	0.60	0.33	0.89
	Qwen3-32B	0.56	0.11	0.58	0.78	0.11	0.49	0.50	0.33	0.35	0.83	**0.22**	0.43	0.40	0.33	0.52
	Sarvam-105B	0.67	**0.44**	0.84	0.67	0.33	0.80	0.50	0.33	0.55	0.83	**0.22**	0.71	**0.80**	**0.50**	0.89
	Command-A-111B	0.67	**0.44**	0.85	0.78	0.22	0.80	0.40	**0.44**	0.73	0.83	**0.22**	0.68	**0.80**	**0.50**	**0.89**
Tamil	Llama-3-8B	0.33	0.38	0.58	0.44	0.33	0.54	0.67	0.33	0.53	0.67	0.33	0.50	0.11	0.56	0.53
	Gemma-4-31B	**0.89**	**0.50**	**0.79**	**0.89**	**0.67**	0.82	**0.78**	0.44	0.71	**0.78**	**0.56**	**0.77**	0.56	**0.78**	0.76
	Qwen3-32B	0.67	0.25	0.59	0.33	0.44	0.57	0.67	**0.56**	0.47	0.56	0.22	0.52	**0.67**	**0.78**	0.51
	Sarvam-105B	0.67	0.25	0.79	0.44	0.44	0.75	**0.78**	0.00	**0.82**	**0.78**	0.33	0.75	0.56	0.56	0.66
	Command-A-111B	0.33	0.12	0.78	0.44	0.44	**0.82**	0.67	0.22	0.69	0.44	0.22	0.67	0.56	**0.78**	**0.76**
Malayalam	Llama-3-8B	0.22	0.22	0.58	0.56	**0.56**	0.61	0.50	**0.22**	0.55	0.44	0.22	0.58	0.56	0.44	0.59
	Gemma-4-31B	**0.89**	0.44	**0.83**	**1.00**	**0.56**	**0.80**	**1.00**	**0.22**	**0.78**	**1.00**	0.22	**0.78**	**1.00**	0.56	**0.75**
	Qwen3-32B	0.33	0.44	0.55	0.78	**0.56**	0.61	0.62	**0.22**	0.56	0.78	0.33	0.61	0.78	0.67	0.65
	Sarvam-105B	0.67	**0.67**	0.37	0.78	**0.56**	0.38	0.88	**0.22**	0.35	0.78	**0.44**	0.36	0.89	**0.78**	0.36
	Command-A-111B	0.56	0.44	0.77	0.89	**0.56**	0.77	0.88	0.11	0.75	**1.00**	0.22	0.74	0.89	0.44	0.74

Best score per column in **bold**; second best is underlined.

Table 2: MCQ accuracy, NLI accuracy, and translation COMET scores for all five models across four languages and five pragmatic phenomena. MCQ accuracy exceeds NLI accuracy in every model-language combination. Best score per column in bold; second-best underlined.

for implied criticism. For example, when a character arrives late and their aunt remarks in Punjabi, *"VIP aa gaye, hun program shuru kar dinde aa"* ("The VIP has arrived, now we can start the program"), all five models interpret the utterance as a genuine warm welcome. A native speaker instantly recognizes the pointed sarcasm about the lateness. Similarly, when a father in Malayalam exclaims *"ee veettil enikku shwaasam polum kittunnilla"* ("I cannot even breathe in this house"), all models predict literal respiratory distress, missing the culturally transparent hyperbolic expression of frustration.

5.2 Model and Task comparisons

Which model performs best? Gemma-4-31B is the strongest model overall in the benchmark (MCQ 0.86 NLI 0.43 COMET 0.81). Command-A-111B is the next strongest model, close behind Gemma on MCQ (0.74) and COMET (0.78), and edging ahead of it on NLI (0.46 vs. 0.43). Sarvam-105B is comparable to Command-A on MCQ (0.77) but noticeably weaker on translation (COMET 0.66). Despite being smaller than Sarvam-105B and Command-A-111B and lacking dedicated Indic pretraining, Gemma-4-31B is the most reliable model in our evaluation.

Do all tasks have equal performance? MCQ scores are higher than NLI scores in all 20 of 20 model-language aggregates. This mirrors the distinction between receptive and productive linguistic competence in psycholinguistics: recognizing a correct interpretation among options is cognitively easier than generating or inferring it without scaffolding (Ruis et al., 2023).

Do all languages have the same performance? The Dravidian languages, Tamil and Malayalam, lag behind on translation (though this gap is not consistent across MCQ and NLI) than the Indo-Aryan languages, Hindi and Punjabi. This is likely because of lower Dravidian language representation in model pretraining data rather than any structural difference in pragmatic difficulty. The translation gap is most pronounced for Coherence and Deixis, and smallest for Implicature.

5.3 Error patterns

In this section we look inside the errors, examining whether models fail in predictable directions, whether failures are shared across model families, and whether NLI errors reveal systematic biases.

Do models default to literal meaning? A design choice in VakyArth was to construct MCQ distractors that reflect the literal, surface reading of each utterance. If models are failing because they cannot reason pragmatically, they should fall into these traps. We test this directly on the MCQ subset where a literal-meaning distractor is unambiguously identifiable, and look at what models pick when they get the item wrong. Across all wrong-answer events, models chose the literal distractor **59.1%** of the time, nearly twice the 33.3% rate expected by chance ($p < 0.001$). This shows models are failing in a specific and predictable direction. For downstream applications, this means pragmatic failures are systematic and predictable rather than noisy, and can potentially be mitigated through targeted prompting or training.

Are failures model-specific or shared across models? For each item, we counted how many of the four models failed. If failures were driven by model-specific weaknesses, we would expect them scattered across different items. Instead, models fail on the same items. On NLI, **24.7%** of items are failed by all four models, rising to 41.7% for Social Pragmatics. This suggests that the items where all models fail together isolate cases where pragmatic reasoning breaks down regardless of the model's overall capability.

Do models show systematic label bias on NLI? We construct our NLI dataset more as contradiction labels (57.5%) because pragmatic reasoning often involves recognizing that a figurative or indirect utterance contradicts its literal reading. If models default to the literal reading, they should systematically mislabel these items as Entailment. Figure 3 shows this is exactly what happens. When the gold label is Entailment, models get it right 78.8% of the time. But when the gold label is Contradiction, models predict Entailment 53.8% of the time more often than they predict Contradiction (23.1%). The bias also compounds our earlier finding on literal distractors: when scaffolding is removed, models do not merely guess randomly among labels they systematically default to the reading where the utterance means what it literally says.

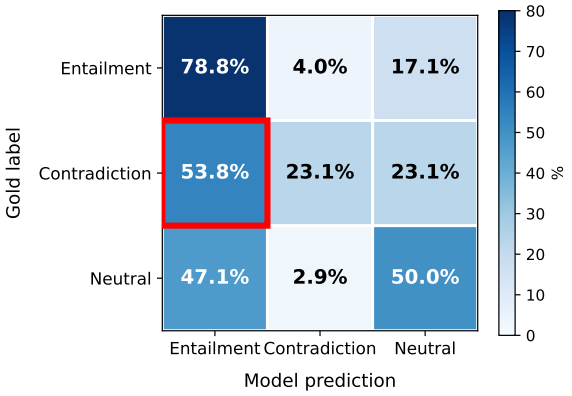

Figure 3: NLI confusion matrix pooled across all four models. Rows are gold labels, columns are model predictions. The highlighted cell marks the systematic bias toward Entailment (literal meaning in our case) when correct label is Contradiction.

5.4 Human Evaluation and COMET Correlation

We collect human translation scores on a 1–5 scale covering adequacy, fluency, and pragmatic adequacy from four native raters, one per language. Given the cost of native-speaker annotation, we sample a subset of 104 items (Hindi 31, Punjabi 26, Tamil 23, Malayalam 24) out of 204 total items, translated by four of the five evaluated models (Llama-3-8B, Gemma-4-31B, Qwen3-32B, Sarvam-105B), producing $4 \times 104 = 416$ scored outputs across all four languages and five phenomena. We compare them against three automatic metrics: COMET (Rei et al., 2020) with the `wmt22-comet-da` checkpoint, XCOMET (Guerreiro et al., 2024), and MetricX (Juraska et al., 2023).

As shown in Table 3, COMET achieves the highest overall correlation with human judgements (Pearson $r = 0.813$, Spearman $\rho = 0.860$), outperforming both XCOMET ($\rho = 0.792$) and MetricX ($\rho = 0.788$) despite being an older checkpoint. We hypothesize that XCOMET and MetricX, optimized on formal translation benchmarks, generalize less well to the colloquial, code-mixed, and culturally embedded text in our benchmark. Correlation varies across languages and phenomena. All three metrics agree most with human scores on Coherence and Social Pragmatics, where translation errors tend to be large and obvious. Agreement is weakest on Implicature and Deixis, where models produce fluent translations that sound correct but miss the pragmatic meaning, the kind of subtle error that automatic metrics struggle to detect. This

finding has a practical implication: automatic evaluation alone is insufficient for benchmarks where pragmatic faithfulness and surface fluency can diverge, and human judgment remains necessary for phenomena like Implicature and Deixis. We show a few such examples in §A.3.

	COMET		XCOMET		MetricX	
	r	ρ	r	ρ	r	ρ
All	**0.81**	**0.86**	0.80	0.79	0.73	0.79
By language						
Hindi	**0.85**	**0.77**	0.65	0.54	0.60	0.60
Punjabi	0.77	**0.80**	**0.78**	0.77	0.74	0.76
Tamil	**0.82**	**0.80**	0.72	0.61	0.62	0.62
Malayalam	0.68	**0.81**	**0.85**	0.76	0.71	0.73
By phenomenon						
Implicature	0.81	0.83	**0.80**	**0.82**	0.69	0.75
Deixis	0.80	0.81	**0.85**	**0.81**	0.72	**0.82**
Speech Acts	0.76	**0.87**	0.77	0.79	**0.74**	0.80
Social Prag.	**0.86**	**0.87**	0.85	**0.86**	0.77	0.83
Coherence	**0.87**	**0.88**	0.85	0.82	0.78	0.82

Table 3: Pearson (r) and Spearman (ρ) correlation of three automatic metrics with human judgements across all four languages and five phenomena ($n = 416$). COMET achieves the highest overall correlation ($\rho = 0.86$) despite being an older checkpoint compared to MetricX. XCOMET is weaker on Hindi and Tamil.

5.5 Script Experiment

We evaluate all MCQ items across all five phenomena, in both native script (Devanagari, Gurmukhi, Tamil, and Malayalam) and Latin transliteration, to study whether script choice affects pragmatic understanding. Pooled across all languages, phenomena, and models, native script yields a modest overall advantage (76.5% vs. 70.6% accuracy). However, the effect is neither universal nor uniform in size: Punjabi (Qwen3-32B) and Malayalam (Command-A) both perform slightly worse in native script, and the size of the advantage where it does appear varies widely, from negligible (Hindi, Command-A: no difference) to substantial (Malayalam, Qwen3-32B: +23.0 points; Tamil, Command-A: +17.8 points). No single model is uniformly script-invariant or uniformly script-sensitive across languages, and no consistent language-level pattern emerges either. We leave a deeper investigation of this effect to future work. Since VakyArth targets everyday communication, we adopt Latin transliteration as the standard script in our evaluation.

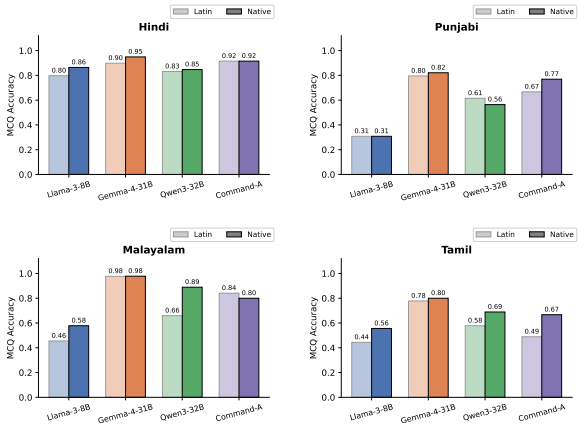

Figure 4: Script experiment: native script vs. Latin transliteration MCQ accuracy for all four languages, pooled across all five phenomena. No single model is uniformly script-invariant or uniformly script-sensitive across languages

6 Conclusion

We present **VakyArth**, the first diagnostic benchmark for pragmatic competence in Indic languages, covering Hindi, Punjabi, Tamil, and Malayalam across five phenomena and three task types. Evaluating five models ranging from 8B to 111B parameters, we find consistent, phenomenon-specific failure patterns that existing benchmarks cannot detect. Gemma-4-31B is the strongest model overall despite lacking Indic specialization, and a persistent MCQ-NLI gap suggests that recognizing and reasoning about pragmatic meaning are distinct competencies. When models fail, they do so in a systematic and predictable direction defaulting to literal readings rather than guessing randomly among labels. Standard automatic metrics, while broadly reliable, fail to penalize fluent but pragmatically unfaithful outputs for implicature and deixis. We hope VakyArth serves as a foundation for building more pragmatically capable and culturally aware language technologies for Indic language users.

7 Limitations

VakyArth has three main limitations. First, the script experiment covers only MCQ items; a fuller ablation and detailed investigation across all models, tasks, and languages remains future work. Second, all items are single-turn utterances or short passages and do not cover multi-turn pragmatics, which is an important dimension of real-world language use. Third, while our four languages repre-

sent a diverse sample of Indic language families, hundreds of other Indic languages remain uncovered.

Acknowledgments

This work was partially funded by NSF Awards 2107524 and 2145479, as well as Good Systems,[2] a UT Austin Grand Challenge to develop responsible AI technologies. We thank Kelly Marchisio for their valuable feedback. We also thank Cohere for providing API credits to support this work.

References

Ruth Appel. 2026. India country brief: The anthropic economic index.

John Langshaw Austin. 1975. *How to do things with words*. Harvard university press.

Kent Bach. 2006. Speech acts and pragmatics. *The Blackwell guide to the philosophy of language*, pages 147–167.

Regina Barzilay and Mirella Lapata. 2008. Modeling local coherence: An entity-based approach. *Computational Linguistics*, 34(1):1–34.

Anne Beyer, Sharid Loáiciga, and David Schlangen. 2021. Is incoherence surprising? targeted evaluation of coherence prediction from language models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4164–4173.

Yu Ying Chiu, Liwei Jiang, Bill Yuchen Lin, Chan Young Park, Shuyue Stella Li, Sahithya Ravi, Mehar Bhatia, Maria Antoniak, Yulia Tsvetkov, Vered Shwartz, and 1 others. 2024. Culturalbench: a robust, diverse and challenging benchmark on measuring (the lack of) cultural knowledge of llms.

Sumanth Doddapaneni, Rahul Aralikatte, Gowtham Ramesh, Shreya Goyal, Mitesh M Khapra, Anoop Kunchukuttan, and Pratyush Kumar. 2023. Towards leaving no indic language behind: Building monolingual corpora, benchmark and models for indic languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12402–12426.

Nuno M Guerreiro, Ricardo Rei, Daan van Stigt, Luisa Coheur, Pierre Colombo, and André FT Martins. 2024. xcomet: Transparent machine translation evaluation through fine-grained error detection. *Transactions of the Association for Computational Linguistics*, 12:979–995.

[2] https://goodsystems.utexas.edu/

J Jaavid, Raj Dabre, M Aswanth, Jay Gala, Thanmay Jayakumar, Ratish Puduppully, and Anoop Kunchukuttan. 2024. Romansetu: Efficiently unlocking multilingual capabilities of large language models via romanization. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15593–15615.

Paloma Jeretic, Alex Warstadt, Suvrat Bhooshan, and Adina Williams. 2020. Are natural language inference models imppressive? learning implicature and presupposition. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8690–8705.

Juraj Juraska, Mara Finkelstein, Daniel Deutsch, Aditya Siddhant, Mehdi Mirzazadeh, and Markus Freitag. 2023. Metricx-23: The google submission to the wmt 2023 metrics shared task. In *Proceedings of the Eighth Conference on Machine Translation*, pages 756–767.

Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, Gokul NC, Avik Bhattacharyya, Mitesh M Khapra, and Pratyush Kumar. 2020. Indicnlpsuite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for indian languages. In *Findings of the association for computational linguistics: EMNLP 2020*, pages 4948–4961.

Youngeun Koo, Jiwoo Lee, Dojun Park, Seohyun Park, and Sungeun Lee. 2025. Evaluating large language models on understanding korean indirect speech acts. *arXiv preprint arXiv:2502.10995*.

Bolei Ma, Yuting Li, Wei Zhou, Ziwei Gong, Yang Janet Liu, Katja Jasinskaja, Annemarie Friedrich, Julia Hirschberg, Frauke Kreuter, and Barbara Plank. 2025. Pragmatics in the era of large language models: A survey on datasets, evaluation, opportunities and challenges. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8679–8696.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. 2016. A corpus and cloze evaluation for deeper understanding of commonsense stories. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 839–849, San Diego, California. Association for Computational Linguistics.

Tarek Naous and Wei Xu. 2025. On the origin of cultural biases in language models: From pre-training data to linguistic phenomena. 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 6423–6443.

Dojun Park, Jiwoo Lee, Hyeyun Jeong, Seohyun Park, and Sungeun Lee. 2024a. Pragmatic competence

evaluation of large language models for the korean language. In *Proceedings of the 38th Pacific Asia Conference on Language, Information and Computation*, pages 256–266.

Dojun Park, Jiwoo Lee, Seohyun Park, Hyeyun Jeong, Youngeun Koo, Soonha Hwang, Seonwoo Park, and Sungeun Lee. 2024b. Multiprageval: Multilingual pragmatic evaluation of large language models. In *Proceedings of the 2nd GenBench Workshop on Generalisation (Benchmarking) in NLP*, pages 96–119.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. Comet: A neural framework for mt evaluation. In *Proceedings of the 2020 conference on empirical methods in natural language processing (emnlp)*, pages 2685–2702.

Brian Roark, Lawrence Wolf-Sonkin, Christo Kirov, Sabrina J Mielke, Cibu Johny, Isin Demirsahin, and Keith Hall. 2020. Processing south asian languages written in the latin script: the dakshina dataset. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2413–2423.

Laura Ruis, Akbir Khan, Stella Biderman, Sara Hooker, Tim Rocktäschel, and Edward Grefenstette. 2023. The goldilocks of pragmatic understanding: Fine-tuning strategy matters for implicature resolution by llms. *Advances in Neural Information Processing Systems*, 36:20827–20905.

Omar Shaikh, Caleb Ziems, William Held, Aryan Pariani, Fred Morstatter, and Diyi Yang. 2023. Modeling cross-cultural pragmatic inference with codenames duet. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6550–6569.

Harman Singh, Nitish Gupta, Shikhar Bharadwaj, Dinesh Tewari, and Partha Talukdar. 2024. Indicgenbench: A multilingual benchmark to evaluate generation capabilities of llms on indic languages. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11047–11073.

Settaluri Sravanthi, Meet Doshi, Pavan Tankala, Rudra Murthy, Raj Dabre, and Pushpak Bhattacharyya. 2024. Pub: A pragmatics understanding benchmark for assessing llms pragmatics capabilities. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12075–12097.

Teun A Van Dijk and 1 others. 1997. The study of discourse. *Discourse as structure and process*, 1(34):703–52.

Sshubam Verma, Mohammed Safi Ur Rahman Khan, Vishwajeet Kumar, Rudra Murthy, and Jaydeep Sen. 2025. Milu: A multi-task indic language understanding benchmark. 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 10076–10132.

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.

Bonnie Webber. 1988. Discourse deixis: Reference to discourse segments. In *26th Annual Meeting of the Association for Computational Linguistics*, pages 113–122.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and 1 others. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pages 38–45.

Kefan Yu, Qingcheng Zeng, Weihao Xuan, Wanxin Li, Jingyi Wu, and Rob Voigt. 2026. The pragmatic mind of machines: Tracing the emergence of pragmatic competence in large language models. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 192–213.

A Appendix

A.1 Prompt Templates

We use 3-shot prompts for all models and tasks. Instructions and few-shot examples are written in the source language to avoid introducing English-language bias into the evaluation. The prompt structure is identical across all four languages, only the instruction text and examples are adapted to each language by a native speaker. Below we show the templates for each task type using Hindi as a representative example.

Translation Prompt

```
Neeche diye gaye Hindi text ko English mein anuvad karo.
Sirf translated text hi output mein dikhayein.


  Source:
Unhone milkar kaam kiya.
Translation:
They worked together.


  Source:
Bazaar mein bahut saamaan milta hai.
Translation:
A wide variety of goods is available in the market.


  Source:
Usne apna vaada nibhaya aur samay par pahunch gaya.
Translation:
He kept his promise and arrived on time.


  Source:
[source utterance]
Translation:
```

MCQ Prompt

```
Is prashan ka sahi uttar kya hai?
Jabab mein sirf option ka akshar (A/B/C...) aur us option ka text likho.


  Context:
Priya ne kaha, “Aaj khana bahut achha bana hai.”
Options:
A. Priya ko khana pasand aaya
B. Priya ne khana nahi khaya
C. Priya ne khana banaya
D. Priya ko bhook nahi thi
Answer:
A. Priya ko khana pasand aaya


  Context:
Ramesh ne kaha, “Bahar bahut thand hai.” Lekin use bahar jaana tha.
Options:
A. Ramesh bahar nahi jayega
B. Ramesh bahar jayega
C. Ramesh ko thand nahi lagti
D. Ramesh ghar mein rehna chahta hai
Answer:
B. Ramesh bahar jayega


  Context:
Maa ne kaha, “Koi baat nahi, agli baar dhyan rakhna.”
Options:
A. Maa gusse mein hai
B. Maa ne maafi maangi
C. Maa ne tasalli di
D. Maa ne saza di
Answer:
C. Maa ne tasalli di


  Context:
[source utterance]
Question:
[question]
Options:
[A. option 1]
[B. option 2]
[C. option 3]
[D. option 4]
Answer:
```

NLI Prompt

```
Neeche diye gaye do vaakyon ke beech sambandh batao.
Jabab mein sirf 'Entailment', 'Contradiction', ya 'Neutral' likho.


  Premise:
A: Khana taiyaar hai?
B: Main abhi bazaar gaya tha.

Hypothesis:
Khana taiyaar hai.

Label:
Contradiction


  Premise:
A: Kya tujhe pata hai kal chutti hai?
B: Haan, main subah se plan kar raha hoon.

Hypothesis:
B kal kuch karna chahta hai.

Label:
Entailment


  Premise:
A: Bahar baarish ho rahi hai.
B: Theek hai.

Hypothesis:
B bahar jayega.

Label:
Neutral


  Premise:
[source premise]

Hypothesis:
[hypothesis]

Label:
```

A.2 Annotation Guidelines for Translation Rating

Human evaluation scores for translation outputs were collected from four native speaker annotators, one per language, each with graduate-level NLP training. Annotators rated each translation on a 1–5 scale assessing how well the model preserved the *pragmatic intent* of the source utterance including implied meaning, tone, cultural nuance, and illocutionary force rather than surface-level adequacy or fluency alone. Annotators were instructed to read the source utterance, its English gloss, and the model output, and assign a score based on the following rubric.

Annotators were explicitly instructed not to majorly penalise grammatical imperfections if the pragmatic meaning was preserved, and not to reward fluent outputs that missed the implied meaning. Each translation rated independently without access to scores from other models or annotators.

Score	Criterion
5	Pragmatic intent fully preserved, including tone, implicature, and cultural nuance.
4	Pragmatic intent mostly preserved; minor loss of nuance that does not affect overall meaning.
3	Core pragmatic intent present but cultural or tonal nuance is lost.
2	Pragmatic intent partially lost; translation feels overly literal.
1	Pragmatic intent completely lost; translation is misleading or pragmatically vacuous.

Table 4: Rubric used for human evaluation of translation outputs. Annotators assessed pragmatic fidelity rather than surface fluency.

A.3 COMET–Human Score Divergence Examples

We present four translation outputs where COMET assigns a high score but human raters assign 1 or 2 out of 5. These cases illustrate why automatic metrics are sometimes insufficient for evaluating pragmatic translation quality. COMET rewards lexical and structural similarity to the reference but cannot detect pragmatically inverted or temporally misanchored outputs.

Example 1: Deictic inversion (Malayalam, Deixis)

Source: *"Njan ippozhum stationinte pinnile gate-il aanu.* ***Ivide varenda****; nee avide thanne nilkku, njan athu vannu cherkam."*

Gloss: "I am still at the back gate of the station. **Don't come here**; stay where you are, I'll come there."

Model output (Qwen3-32B): "I am still at the station. **Come here**; you stand there, I will come there."

COMET: 0.78 **Human:** 1/5

The model inverts ***ivide varenda*** *(don't come here) to "Come here", producing a navigation instruction that is the exact opposite of the source. COMET scores the output highly because most content words are present; a human immediately recognises the instruction is backwards.*

Example 2: Subject-object inversion in implicature (Punjabi, Implicature)

Source: *"Surinder di salah leh ke na baithna, uh tah* ***kath da ullu*** *hai."* **Gloss:** "Don't take Surinder's advice; **he is a complete fool**."

Model output (Sarvam-105B): "**He made a fool of himself** by not listening to Surinder's advice."

COMET: 0.76 **Human:** 1/5

Kath da ullu *is a common Punjabi meaning "a complete fool." Sarvam mistranslates the subject of the insult — the output implies the speaker is foolish for ignoring Surinder, whereas the source calls Surinder the fool. The pragmatic meaning is exactly inverted.*

Example 3: Temporal deixis chain failure (Punjabi, Deixis)

Source: *"Seema ne Mangalvaar nu mainu dassya ki Priya ne pichhle Somvaar nu unhu keha si, 'pehla mera interview **parso** si...' "*

Gloss: "On Tuesday, Seema told me that Priya had told her last Monday, 'My interview was originally on **Wednesday** [the day after tomorrow relative to that Monday]...' "

Model output (Sarvam-105B): "My interview was the **day after tomorrow**..."

COMET: 0.86 **Human:** 2/5

*The word **parso** means the day two steps from today—past or future depending on context. The model fails to anchor it to the correct speech time (last Monday), producing a future reference where a past reference is required. COMET scores the output 0.86 because the words are nearly identical to the reference.*

Example 4: Completed vs ongoing action deixis (Tamil, Deixis)

Source: *"Shanthi, unnoda friends ellaarum **vanthutanga**, veliya saaptutu irukkanga, neeyum poi senthum saappudu."*

Gloss: "Shanthi, all your friends **have come** [already], they are eating outside, you also go and join them."

Model output (Qwen3-32B): "Shanthi, all your friends **are coming**, you are also going to join them."

COMET: 0.65 **Human:** 1/5

***Vanthutanga** is a perfective form indicating completed action — the friends have already arrived. The model translates it as an ongoing future event, losing the urgency that the friends are already there waiting. The deictic anchor shifts from present to future.*

A.4 Error Analysis

We present five representative failure cases, one per phenomenon. Each case illustrates a distinct failure mode that recurs across the benchmark.

Example 1: Face-saving invitation read literally (Tamil, Implicature, MCQ)

Context: *"Veetukku vantha virunthaaligal purapidrapo **kai nenachittu ponganu** amma sonnanga."*
Gloss: "As the guests who had come home were leaving, mother told them: **wet your hands and go**."

Question: What did mother tell the guests?

A. Kai kaluva sonnanga (Told them to wash their hands)
B. Inga saapida sonnanga (Invited them to eat here)
C. Veetukku poi saapida sonnanga (Told them to go home and eat)
D. Kulichutu poga sonnanga (Told them to bathe before leaving)

Model	Prediction	Gold
Llama-3.1-8B	C ✗	**B**
Gemma-4-31B	C ✗	**B**
Qwen3-32B	C ✗	**B**
Sarvam-105B	A ✗	**B**
Command-A	A ✗	**B**

***Kai nenachittu po** ("wet your hands and go") is a traditional Tamil farewell invitation to stay and eat — a face-saving formula that allows guests to decline gracefully. Llama, Gemma, and Qwen interpret it as telling guests to go home and eat (C); Sarvam reads it literally as a hygiene instruction (A). No model recovers the implied invitation to stay (B).*

Example 2: Referent confusion in family dialogue (Tamil, Deixis, MCQ)

Context: *"Unakku thevayaana thuni mani ellam eduthu vechachanu amma **unkitta** kekka sonnanga. Ne eduthuvechuttiyanu en **thangachi kitta** na ketten."*
Gloss: "Amma asked me to check with you [unkitta] whether you had taken all the cloth pieces you needed. I asked my younger sister [thangachi] whether she had taken them."

Question: Who does the speaker refer to with '**unakku**'?

A. Appa B. Amma
C. Thangachi D. Akka

Model	Prediction	Gold
Llama-3.1-8B	B (Amma) ✗	**C (Thangachi)**
Gemma-4-31B	C ✔	**C (Thangachi)**
Qwen3-32B	B (Amma) ✗	**C (Thangachi)**
Sarvam-105B	C ✔	**C (Thangachi)**
Command-A	B (Amma) ✗	**C (Thangachi)**

The utterance involves three participants (amma, speaker, thangachi) across two turns. Llama and Qwen conflate amma and thangachi, failing to track the referential chain across the two embedded clauses.

Example 3: Hyperbolic repetition read literally (Punjabi, Speech Acts, NLI)

Premise: *"Papa ne naraaz hoke Raj nu keha, 'Main tenu **100 vari** keha, ghar vich ball na kheliya kar.' "*
Gloss: "Papa angrily told Raj, 'I have told you **100 times**, don't play ball inside the house.' "

Hypothesis: Papa has already told Raj this same thing 99 times before.

Model	Prediction	Gold
Llama-3.1-8B	Entailment ✗	**Contradiction**
Gemma-4-31B	Contradiction ✔	**Contradiction**
Qwen3-32B	Entailment ✗	**Contradiction**
Sarvam-105B	Neutral ✗	**Contradiction**
Command-A	Entailment ✗	**Contradiction**

"100 vari" is a hyperbolic speech act expressing frustration, not a literal count. The hypothesis that papa said it exactly 99 times before contradicts the pragmatic meaning. Llama and Qwen treat it as a factual claim and predict Entailment.

Example 4: Veiled criticism read as neutral observation (Punjabi, Social Pragmatics, NLI)

Premise: *"Kal jado mainu meri maasi mili, unha ne keha, 'Tu taan bada **kamzor lag reha hai.**' "*
Gloss: "Yesterday when I met my aunt, she said, 'You look quite thin/weak.' "

Hypothesis: The aunt was mocking the speaker's health.

Model	Prediction	Gold
Llama-3.1-8B	Entailment ✗	**Contradiction**
Gemma-4-31B	Neutral ✗	**Contradiction**
Qwen3-32B	Neutral ✗	**Contradiction**
Sarvam-105B	Neutral ✗	**Contradiction**
Command-A	Neutral ✗	**Contradiction**

In Punjabi culture, an elder commenting on a younger person's appearance ("you look thin") is a conventional expression of concern and care, not mockery. The correct label is Contradiction — the aunt is being caring, not sarcastic. All four models miss this cultural norm.

Example 5: Topically related but discourse-incoherent (Tamil, Coherence, MCQ)

Context: *"Tamilnadu, India la irukurathulaaye south la irukkura state. Idhu areavula 10-avathu periya state, janathila 6-avathu periyadhu. Tamilnadula la iruka makkal tamil pesuvaanga..."*
Gloss: "Tamil Nadu is a southern state of India. It is the 10th largest by area and 6th most populous. People in Tamil Nadu speak Tamil..."

Question: What is the most appropriate next sentence?

A. Tamilnadukku naraya sutrula payanigal varuvaanga (Many tourists visit)
B. Chennai dhan thalainagaram (Chennai is the capital)
C. Tamil classica solappadum Thirukkural... (literary reference)
D. Tamil cinema industryah kollywoodnu solluvaanga (Kollywood)

Model	Prediction	Gold
Llama-3.1-8B	A ✗	**B**
Gemma-4-31B	B ✔	**B**
Qwen3-32B	C ✗	**B**
Sarvam-105B	D ✗	**B**
Command-A	B ✔	**B**

The paragraph introduces factual information about Tamil Nadu systematically (area, population, language). The next logical fact is the capital city. Options A, C, D are all topically related to Tamil Nadu but break the structured informational progression. Three models pick topically plausible but discourse-incoherent continuations.