LIMP: Linguistically-Informed Multi-Strategy Prompting for Telugu Multi-Turn Dialogue Generation

Arjungopal Anilkumar[1], **Suryansh Ram Menon**[1], **Divagar S**[1], **Premjith B**[1]

Amrita School of Artificial Intelligence, Coimbatore, Amrita Vishwa Vidyapeetham, India[1]

Correspondence: b_premjith@cb.amrita.edu

Abstract

Generating contextually coherent multi-turn dialogue in Telugu requires resolving three deeply interacting constraints absent from generic LLM prompting: morphologically encoded social hierarchy (honorific verb conjugations), strict SOV agglutinative syntax, and culturally governed emotional logic formalised in Natyashastra *rasa* theory (Bharata Muni, 1951). We introduce LIMP (**L**inguistically-**I**nformed **M**ulti-Strategy **P**rompting), an inference-time, training-free framework that injects expert linguistic and cultural knowledge into prompt structure, requiring no fine-tuning or labelled data. We empirically evaluate seven configurations spanning four prompting strategies across two model backbones on 500 stratified evaluation instances from the IndicDialogue Telugu corpus (Arnob et al., 2024): a zero-shot baseline, LIMP-RAW (dense constraint prompt), LIMP-COT (six-stage analytical scaffold grounded in rasa theory and Telugu morphological grammar), and a context-compressed CoT variant for SARVAM-1. LIMP-COT achieves 2× higher Jaccard and Dice than LIMP-RAW on GEMMA-3-1B-IT (Gemma Team, Google DeepMind, 2025) (1B parameters): Jaccard = 0.0140 vs. 0.0098, Dice = 0.0274 vs. 0.0193 ($p < 0.001$), demonstrating that sequential analytical commitment to linguistic constraints produces more form-faithful Telugu than holistic constraint injection. Under LLM-as-judge evaluation, LIMP-COT achieves the highest Adequacy (3.90 ± 0.91) among GEMMA-3-1B-IT configurations, while LIMP-RAW achieves the highest Fluency (4.07 ± 0.93). The zero-shot baseline achieves the highest BERTSCORE F_1 (0.9760) across all configurations, revealing a three-way dissociation: automatic semantic metrics favour zero-shot generation, fluency favours dense constraint prompting, and contextual adequacy favours sequential scaffolding. This *semantic–lexical–pragmatic dissociation*, where no single configuration dominates across all metric classes, is itself a substantive finding: in agglutinative Telugu, semantic paraphrase fidelity, morphosyntactic surface fidelity, and pragmatic adequacy are orthogonal evaluation dimensions that must be assessed concurrently.

1 Introduction

Multi-turn dialogue generation is the task of producing contextually coherent responses conditioned on a sequence of prior conversational turns. Unlike single-turn generation, the model must jointly track discourse state, speaker intent, and pragmatic coherence across multiple exchanges while avoiding generic or context-insensitive outputs—a challenge that has driven a research trajectory from hierarchical recurrent architectures (Serban et al., 2016) through large-scale conversational pretraining (Zhang et al., 2020b) to instruction-tuned generative models (Brown et al., 2020). Diversity and specificity in generation remain persistent failure modes (Li et al., 2016; Holtzman et al., 2020), and even frontier LLMs require careful conditioning to maintain coherence across long conversational contexts (Roller et al., 2021; Bang et al., 2023).

Multi-turn dialogue generation in Telugu—a morphologically rich Dravidian language spoken by approximately 82 million people (Eberhard et al., 2023)—poses constraints of a qualitatively different order. Telugu verb conjugations encode the speaker-addressee power relationship as a grammatical obligation: the distinction between honorific endings such as *-tunnaaru* and familiar endings such as *-tunnaavu* is a categorical social obligation (Krishnamurti, 2003; Steever, 1998). Telugu is additionally a canonical SOV language whose agglutinative morphology yields a combinatorially large surface paradigm; models trained predominantly on Indo-European corpora exhibit degraded cross-lingual transfer performance on typologically distant languages (Wu and Dredze, 2019), and systematically produce word-order errors in SOV languages such as Telugu (Subbarao, 2012). A further constraint, unique to the cinematic domain, is *rasa*, the classical Indian theory of aesthetic emotion codified in the Natyashastra (Bharata Muni,

Proceedings of the Sixth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages, pages 32–41

1951): each conversational turn is governed by one of nine canonical emotional categories (Sreejith et al., 2017). These three constraints operate simultaneously and interdependently, and Dravidian languages remain systematically underserved by both multilingual pretraining corpora and existing dialogue architectures (Joshi et al., 2020; Ahuja et al., 2023).

Contemporary LLMs exhibit broad multilingual competence yet do not encode these Telugu-specific morphological and cultural priors. Generic chain-of-thought (CoT) prompting (Wei et al., 2022) decomposes reasoning into ordered steps but defines those steps task-abstractly, providing no guidance for culturally or morphologically grounded generation (Shi et al., 2023), and low-resource settings preclude the data-intensive fine-tuning paradigm (Lauscher et al., 2020). We introduce LIMP (**L**inguistically-**I**nformed **M**ulti-Strategy **P**rompting), an inference-time, training-free framework that injects expert linguistic and cultural knowledge directly into the prompt, requiring no labelled data or parameter updates. We empirically evaluate seven configurations—four for GEMMA-3-1B-IT (Zero-Shot, No-LIMP, LIMP-RAW, LIMP-COT) and three for SARVAM-1 (Zero-Shot, LIMP-RAW, CoT-Compressed)—on 500 statistically independent evaluation instances from the IndicDialogue Telugu corpus (Arnob et al., 2024) using nine complementary metrics spanning semantic fidelity, lexical overlap, character-level morphology, and generation diversity. LIMP-COT on GEMMA-3-1B-IT (Gemma Team, Google DeepMind, 2025) (1B parameters) achieves 2× higher Jaccard and Dice than LIMP-RAW on the same backbone (Jaccard = 0.0140 vs. 0.0098; Dice = 0.0274 vs. 0.0193; $p < 0.001$), demonstrating that sequential analytical commitment to linguistic constraints produces more form-faithful Telugu than holistic constraint injection. Under LLM-as-judge evaluation, LIMP-COT achieves the highest Adequacy (3.90 ± 0.91) among GEMMA-3-1B-IT configurations, confirming that the CoT scaffold anchors the model more tightly to the pragmatic demands of the dialogue context. A concurrent finding is a *semantic–lexical–pragmatic dissociation*: the zero-shot baseline achieves the highest BERTSCORE F_1 (0.9760), LIMP-RAW achieves the highest Fluency (4.07), and LIMP-COT achieves the highest Adequacy (3.90), confirming that these are orthogonal evaluation dimensions in agglutinative Telugu.

2 Related Work

Neural multi-turn dialogue generation has progressed from hierarchical recurrent encoder-decoders (Serban et al., 2016) through large-scale pretraining on conversational corpora (Zhang et al., 2020b) to modular open-domain systems (Roller et al., 2021). Maximum mutual information objectives (Li et al., 2016) and nucleus sampling (Holtzman et al., 2020) address the generic response problem, while in-context few-shot learning (Brown et al., 2020) enables generative control without parameter updates. Chain-of-thought prompting (Wei et al., 2022) substantially improves LLM reasoning through step-by-step decomposition, and its extensions—zero-shot (Kojima et al., 2022), self-consistent (Wang et al., 2023), tree-structured (Yao et al., 2023), least-to-most (Zhou et al., 2023), and complexity-based (Fu et al., 2023)—establish staged analytical commitment as a robust mechanism for compositionally difficult tasks. All existing CoT formulations, however, define reasoning stages task-abstractly without encoding domain-expert linguistic knowledge, and distracting or erroneous intermediate steps propagate through the reasoning chain and degrade generation quality (Shi et al., 2023). The NusaCrowd initiative (Cahyawijaya et al., 2023) establishes open-source NLP resources and benchmarks for Indonesian languages—the closest regional parallel to Dravidian resource work—but does not address culturally grounded generation or prompting strategies; no prior work has grounded CoT stages in Natyashastra rasa theory or Telugu morphological grammar.

The Indic NLP landscape has been significantly expanded by a sequence of infrastructure contributions: MURIL (Khanuja et al., 2021) provides BERT-scale multilingual pretraining over 17 Indian languages; the IndicNLP Suite (Kakwani et al., 2020) and AI4Bharat corpora (Kunchukuttan et al., 2020) establish monolingual benchmarks; XLM-R (Conneau et al., 2020) demonstrates large-scale cross-lingual transfer; and IndicTrans2 (Gala et al., 2023) sets the frontier for Indic translation. Despite this progress, Dravidian languages remain systematically underrepresented, with persistent performance gaps documented across frontier models (Joshi et al., 2020; Ahuja et al., 2023). BPE tokenisation conflates morphological boundaries in agglutinative languages (Sennrich et al., 2016b; Bostrom and Durrett, 2020; Mielke et al., 2021), and multilingual tokenisers consistently under-

segment Dravidian text (Rust et al., 2021), directly inflating subword overlap metrics for semantically distinct forms. BERTSCORE (Zhang et al., 2020a) is substantially more robust to agglutinative surface variation than BLEU (Papineni et al., 2002), whose failure modes for morphologically rich languages are well-documented (Callison-Burch et al., 2006; Post, 2018; Mathur et al., 2020); these findings collectively motivate our Unicode purity filter and multi-metric evaluation design.

Computational politeness research has proceeded from foundational politeness theory (Brown and Levinson, 1987) through NMT side constraints for T-V distinction control (Sennrich et al., 2016a), polite dialogue generation (Niu and Bansal, 2018), politeness style transfer (Madaan et al., 2020), and comprehensive landscape surveys (Prabhakaran et al., 2024). All of this work assumes binary or gradient politeness scales that are structurally inadequate for Telugu's multi-tier categorical honorific system (informal, polite, formal) (Krishnamurti, 2003; Steever, 1998). Rasa theory has recently been applied computationally to sentiment analysis (Sreejith et al., 2017) and Sanskrit poetry (Sandhan et al., 2023), with calls to move beyond Western-centric emotion taxonomies in NLP (Plaza-del Arco et al., 2024). In all prior work reviewed here, rasa functions as a post-hoc classification label applied to existing text. LIMP-COT instead positions rasa identification as the upstream Stage 1 commitment from which all morphological and pragmatic decisions are derived, treating rasa as a forward constraint on generation rather than a retrospective annotation. Whether this scaffolding produces outputs that are genuinely rasa-coherent remains an open empirical question requiring human evaluation by domain experts (§6).

3 Methodology

Figure 1 illustrates the end-to-end LIMP pipeline. Raw IndicDialogue data is subjected to a Telugu Unicode purity filter, segmented into five-turn evaluation instances by a non-overlapping sliding window, stratified into a 500-sample evaluation corpus, and passed through one of the evaluated LIMP prompting strategies before decoding by the generation backbone. Generated outputs are evaluated against gold references using nine complementary metrics.

We use the Telugu subset of the IndicDialogue corpus, a subtitle-derived dataset of 10 Indic languages distributed as line-delimited JSONL files (Arnob et al., 2024), comprising Telugu-language subtitle data sourced from OpenSubtitles.org (75 files; ≈90,950 dialogue instances). The corpus includes both original Telugu productions and dubbed content; dubbed material may not exhibit authentic rasa arcs or native honorific patterns, a limitation acknowledged in §6. Where the data derives from original Telugu productions, the cinematic domain encodes speaker hierarchy, emotional rasa arc, and regional dialect simultaneously in each turn—making it a substantive testbed for the three generation constraints described in §1. Telugu NLP data is heavily contaminated by transliteration, code-mixing, and subtitle processing artifacts (Kunchukuttan et al., 2020). We therefore apply a strict Unicode purity filter requiring every script-carrying character to fall within the Telugu block (U+0C00–U+0C7F); lines with zero Telugu-block characters are discarded, while script-neutral characters (punctuation, numerals, whitespace) are preserved. This ensures that all downstream generation contexts contain authentic Telugu morphology rather than romanised or mixed-script artefacts. Segmentation proceeds via a sliding window of size $k = 5$ with stride $k = 5$, applied within individual movie boundaries to prevent narrative contamination across screenplay boundaries and to ensure that every evaluation instance is statistically independent. Each evaluation instance is a five-tuple $\langle c_1, c_2, c_3, c_4, r \rangle$ comprising a four-turn context and a gold reference response r. From the filtered corpus we draw 500 evaluation instances via interleaved stratified sampling (selecting every $\lfloor N/500 \rfloor$-th evaluation instance to preserve proportional movie representation), then apply a global shuffle at seed 42 to prevent ordering artefacts.

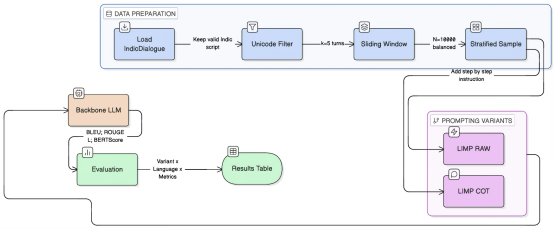

Figure 1: The end-to-end LIMP pipeline.

LIMP encodes linguistic and cultural expert knowledge as structured prompt constraints, replacing the model's implicit, and in low-resource Dravidian settings often absent, encoding of Telugu-

specific grammar with explicit inference-time instruction (Brown et al., 2020). Two strategies constitute the framework on a spectrum from dense simultaneous constraint injection to sequential analytical scaffolding.

3.0.1 LIMP-RAW: Dense Constraint Prompting

LIMP-RAW is a single, densely structured prompt of ≈480 tokens encoding five categories of linguistic and cultural constraints as hard generation rules:

C1. Honorific Register Rules. Verb-ending inventories: honorific (*-tunnaaru*, *-cheshaaru*) and familiar (*-tunnaavu*, *-cheshaavu*), keyed to speaker–addressee social rank (Steever, 1998).

C2. SOV Word Order. An explicit instruction that Telugu is a strict SOV language, with a contrastive correct/incorrect example illustrating the target syntax (Krishnamurti, 2003).

C3. Agglutinative Morphology. Post-positional suffix inventory for case markers (*-ki*, *-tho*, *-lo*, *-nundi*) with a rule prohibiting free-standing prepositions (Krishnamurti, 2003).

C4. Three-Dialect Discrimination. Diagnostic lexical markers distinguishing standard Coastal Andhra Telugu, colloquial Telangana/Hyderabadi Telugu (Urdu substrate), and Rayalaseema Telugu.

C5. Five Cinematic Genre Conventions. Genre-specific rhetorical shapes for action/confrontation, family drama, romance/*sringara*, comedy/*hasya*, and social/art-house dialogue.

All five constraint categories are delivered simultaneously; the model integrates them holistically during auto-regressive generation without an explicit reasoning trace. Output constraint: single spoken line, Telugu Unicode only (U+0C00–U+0C7F), no transliteration.

3.0.2 LIMP-COT: Six-Stage Analytical Scaffold

LIMP-COT extends the constraint vocabulary of LIMP-RAW into a sequential analytical chain-of-thought scaffold (Wei et al., 2022) of ≈920 tokens.

In Telugu dialogue, *rasa*, speaker hierarchy, and morphological register are interdependent: each must be committed to before the next can be resolved. LIMP-COT enforces this ordering through six analytically chained stages:

Stage 1: Rasa Identification. Identify dominant rasa from the nine Natyashastra categories (Bharata Muni, 1951) and characterise the emotional arc (escalating / de-escalating / pivoting). Rasa is the most upstream constraint (Sreejith et al., 2017).

Stage 2: Speaker Relationship and Power Axis. Map the speaker-addressee relationship onto hierarchical position (superior / equal / subordinate) and emotional alignment, then derive the mandatory Telugu honorific form (Steever, 1998).

Stage 3: Genre and Scene Type. Classify cinematic genre and name the scene type (e.g., villain-hero confrontation, mother-son reconciliation). Scene type determines rhetorical shape (Wierzbicka, 1991).

Stage 4: Linguistic Register. Commit to: (a) exact honorific verb endings; (b) target dialect (coastal Andhra / Telangana / Rayalaseema); (c) code-switching calibration. Explicit commitment prevents underspecification defaults (Jiang et al., 2020).

Stage 5: Narrative Function. Select exactly one of seven narrative functions (escalation, revelation, defiance, vulnerability, ironic pivot, *etc.*) to preserve rhetorical force (Chatman, 1980; Holtzman et al., 2020).

Stage 6: Final Generation. Synthesise all prior commitments into a single spoken Telugu line (Unicode-only, no transliteration) (Jiang et al., 2020).

The six stages are analytically interdependent: each takes the outputs of all prior stages as input, creating a commitment chain that distinguishes LIMP-COT from both generic CoT (Wei et al., 2022) and LIMP-RAW. The cost of this precision is reduced paraphrastic freedom, producing the semantic–lexical dissociation characterised in §5.

We evaluate all configurations with nine complementary metrics spanning the semantic–lexical–pragmatic measurement space. BERTSCORE F_1 (Zhang et al., 2020a) computes soft token-level matching via MURIL contextual embeddings (Khanuja et al., 2021), rewarding semantically equivalent paraphrases rather than penalising valid surface alternations. Cosine similarity operates on the same dense MURIL embeddings and measures high-level topical alignment in vector space. The character n-gram F-score (chrF) (**?**) provides character-level evaluation that is highly effective for agglutinative morphology, directly capturing partial suffix matches that word-level metrics miss entirely. BLEU (Papineni et al., 2002) and ROUGE-L are included as legacy baselines that reviewers expect, despite their well-documented failure modes for morphologically rich languages (Callison-Burch et al., 2006; Post, 2018; Mathur et al., 2020). Subword Jaccard similarity, $J(H, R) = |H \cap R|/|H \cup R|$, and the Dice coefficient, $D(H, R) = 2|H \cap R|/(|H| + |R|)$, measure lexical overlap over subword token sets and index morphological form fidelity (Bhat and Sharma, 2013; Sørensen, 1948). Distinct-1 and Distinct-2 (Li et al., 2016) measure the ratio of unique unigrams and bigrams in the generated output, penalising repetitive or degenerate generation.

All pairwise comparisons use paired t-tests and Wilcoxon signed-rank tests (Wilcoxon, 1945) at $N = 500$. Effect sizes are reported as Cohen's d (Cohen, 1988). Per-sample win rates provide a distribution-free summary statistic.

4 Experimental Setup

All experiments are implemented in Python 3.10 using `transformers` v4.40.0 (Wolf et al., 2020), `PyTorch` v2.3.0 (Paszke et al., 2019), and `scipy` v1.13.0 (Virtanen et al., 2020), running on a single NVIDIA RTX A6000 48 GB GPU and Weights & Biases experiment tracking (Weights & Biases, 2020). Models are loaded in FP16 with SDPA attention. **GEMMA-3-1B-IT** (Gemma Team, Google DeepMind, 2025) is a 1B-parameter instruction-tuned decoder-only transformer with a 128K token context window, evaluated under four configurations: Zero-Shot (raw context with no injected constraints), No-LIMP (raw context with a minimal continuation instruction but no linguistic framework), LIMP-RAW, and LIMP-COT. **SARVAM-1** (Sarvam AI, 2024) is a 2B Indic-focused decoder-only model with a 4,096-token effective context window, evaluated under three configurations: Zero-Shot, LIMP-RAW, and CoT-Compressed—a condensed ≈241-token version of the LIMP-COT scaffold designed to fit within SARVAM-1's context limit while preserving all six reasoning stages. All seven configurations decode greedily (`num_beams=1`, `do_sample=False`, `max_new_tokens=256`) to ensure reproducibility and to isolate prompt-structure effects from stochastic generation variance. MURIL (`google/muril-base-cased`) (Khanuja et al., 2021) serves as the BERTSCORE reference model for all semantic fidelity evaluations.

5 Results

We evaluate seven configurations on 500 Telugu evaluation instances. The primary finding is that LIMP-COT achieves 2× higher Jaccard and Dice than LIMP-RAW on the same backbone (Jaccard: 0.0140 vs. 0.0098; Dice: 0.0274 vs. 0.0193; $p < 0.001$), demonstrating that sequential analytical commitment to linguistic constraints produces more form-faithful Telugu than holistic constraint injection. Under LLM-as-judge evaluation, LIMP-COT achieves the highest Adequacy (3.90 ± 0.91) among GEMMA-3-1B-IT configurations, while LIMP-RAW achieves the highest Fluency (4.07 ± 0.93). The zero-shot baseline achieves the highest BERTSCORE F_1 (0.9760) across all configurations. This *semantic–lexical–pragmatic dissociation*, where no single configuration dominates across all metric classes, is a substantive finding: automatic semantic metrics, morphosyntactic surface fidelity, and pragmatic adequacy are orthogonal evaluation dimensions in agglutinative Telugu. Table 1 presents full quantitative results; Table 2 presents LLM-as-judge ratings.

LIMP-COT achieves Jaccard = 0.0140 and Dice = 0.0274, both higher than LIMP-RAW (Jaccard = 0.0098, Dice = 0.0193), with all differences statistically significant ($p < 0.001$, paired t-test and Wilcoxon (Wilcoxon, 1945)). This is consistent with Stage 4 of the CoT scaffold, which requires explicit commitment to specific verb endings and dialect markers before generation, anchoring the output to precise surface tokens in the gold reference. Absolute Jaccard and Dice values are low across all configurations, as expected: exact subword token matching in an agglutinative paradigm is combinatorially difficult even for semantically

Model	Strategy	Cosine↑	BS-F_1↑	chrF↑	BLEU↑	Jaccard↑	Dice↑	Dist-1↑	Dist-2↑
GEMMA-3-1B-IT	Zero-Shot	0.9919	**0.9760**	0.0782	**0.0184**	**0.0423**	**0.0760**	0.9514	0.7807
GEMMA-3-1B-IT	No-LIMP	**0.9921**	0.9712	**0.1006**	0.0080	0.0289	0.0547	0.8785	0.9787
GEMMA-3-1B-IT	LIMP-RAW	0.9889	0.9704	0.0069	0.0029	0.0098	0.0193	0.8963	0.9948
GEMMA-3-1B-IT	LIMP-COT	0.9899	0.9715	0.0205	0.0038	0.0140	0.0274	0.8923	0.9907
SARVAM-1	Zero-Shot	0.9919	0.9721	0.0706	0.0074	0.0247	0.0469	0.9309	0.9845
SARVAM-1	LIMP-RAW	0.9883	0.9709	0.0030	0.0037	0.0097	0.0191	0.9960	0.9997
SARVAM-1	CoT-Compressed	0.9885	0.9707	0.0017	0.0031	0.0077	0.0152	**0.9982**	**1.0000**

Table 1: Main results on 500 Telugu evaluation instances. **BS-F_1** = BERTScore F_1 computed with MURIL backbone. Bold = best per column. Note: BLEU, Jaccard, Dice, and chrF values are close to zero across all configurations, as expected for an agglutinative language where root words merge with post-positional suffixes, creating a combinatorially large surface paradigm that penalises exact n-gram and token overlap metrics.

Configuration	Adequacy	Fluency	Overall	κ
GEMMA-3-1B-IT Zero-Shot	3.80 ±0.98	3.80 ±1.05	**4.17 ±0.82**	0.130
GEMMA-3-1B-IT LIMP-RAW	3.60 ±0.84	**4.07 ±0.93**	3.83 ±0.82	0.079
GEMMA-3-1B-IT LIMP-COT	**3.90 ±0.91**	3.87 ±0.81	3.97 ±0.80	−0.003
SARVAM-1 LIMP-RAW	4.00 ±0.93	3.83 ±0.69	4.10 ±0.79	0.040
SARVAM-1 CoT-Comp	3.97 ±1.02	3.97 ±1.11	3.70 ±0.74	0.032

Table 2: LLM-as-judge evaluation on a stratified sample of 500 instances. Ratings are on a 1–5 integer scale (1 = unacceptable, 5 = excellent). κ = inter-judge Cohen's Kappa on the Overall dimension. LIMP-COT achieves the highest Adequacy among GEMMA-3-1B-IT configurations; LIMP-RAW achieves the highest Fluency.

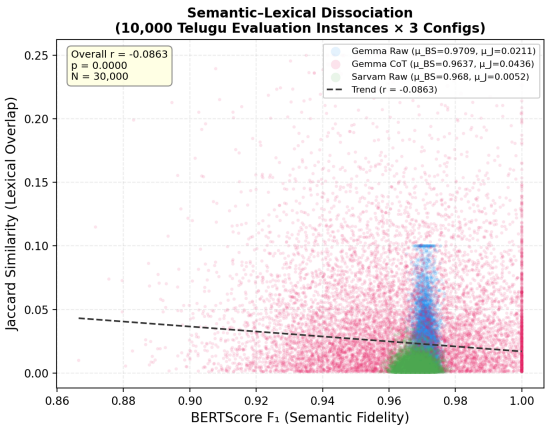

Figure 2: Visualisation of the semantic–lexical dissociation. Each point is one of the 500 evaluation samples. The absence of positive correlation between BERTSCORE F_1 and Jaccard confirms that the two metric classes capture orthogonal dimensions of Telugu dialogue generation quality.

correct responses (Bhat and Sharma, 2013; Mielke et al., 2021). The zero-shot baseline achieves the highest Jaccard (0.0423) and Dice (0.0760) overall, indicating that unconstrained generation more frequently produces surface-coincident outputs, while LIMP-prompted outputs diverge from reference surface forms in service of pragmatic correctness, a trade-off confirmed by the judge evaluation.

The LLM-as-judge results (Table 2) reveal a dissociation between fluency and adequacy. LIMP-RAW produces highly fluent Telugu (Fluency = 4.07, the highest among all configurations) but incurs a sharp drop in Adequacy (3.60), failing to anchor the output firmly to the conversational context. Injecting the CoT scaffold reverses this: LIMP-COT raises Adequacy to 3.90 while maintaining comparable Fluency (3.87), Cohen's Kappa values are in the slight-to-fair range across configurations ($\kappa = -0.003$ to 0.130), reflecting the inherent subjectivity of evaluating culturally-embedded cinematic Telugu; native speaker annotation remains the necessary validation step.

On the cross-model comparison, SARVAM-1 LIMP-RAW (2B, Indic-pretrained) achieves Jaccard = 0.0097 and Dice = 0.0191, comparable to GEMMA-3-1B-IT LIMP-RAW (Jaccard = 0.0098, Dice = 0.0193), while GEMMA-3-1B-IT LIMP-COT (Jaccard = 0.0140) outperforms both on lexical fidelity. This pattern is consistent with the hypothesis that structured linguistic scaffolding is a more productive lever than parametric scale for form-faithful generation, though the comparison remains partially confounded: SARVAM-1 CoT-Compressed uses a condensed scaffold rather than the full LIMP-COT prompt, and observed differences reflect the joint effect of model capability and prompt fidelity. LIMP-COT also exhibits substantially higher output variance on lexical metrics, attributable to reasoning-chain collapse when early stages err (e.g., misidentified rasa propagates through all downstream commitments) (Shi et al., 2023).

6 Conclusion

We introduced LIMP, a training-free prompting framework for Telugu multi-turn dialogue generation that injects rasa theory, honorific morphology, dialectal variation, and cinematic genre conventions at inference time, demonstrating that expert linguistic knowledge can be encoded in prompt structure without fine-tuning or additional compute.

The primary empirical finding is that LIMP-CoT achieves 2× higher Jaccard and Dice than LIMP-Raw on the same backbone (Jaccard = 0.0140 vs. 0.0098; $p < 0.001$), establishing that sequential analytical commitment to linguistic constraints produces more form-faithful Telugu than holistic constraint injection. Under LLM-as-judge evaluation, LIMP-CoT achieves the highest Adequacy (3.90 ± 0.91) while LIMP-Raw achieves the highest Fluency (4.07 ± 0.93), confirming that fluency and pragmatic adequacy are dissociable and that the CoT scaffold specifically addresses the adequacy deficit of dense constraint prompting. The zero-shot baseline achieves the highest automatic metric scores (BERTScore $F_1 = 0.9760$; chrF = 0.0782; Jaccard = 0.0423), establishing a strong floor and motivating the multi-dimensional evaluation framework: automatic metrics alone are insufficient to capture pragmatic quality in culturally-embedded Telugu dialogue. No single configuration dominates across all metric classes.

This work has five limitations that bound the scope of its findings. First, the absence of human evaluation by native Telugu speakers is a critical gap. All morphosyntactic and cultural claims, including rasa coherence and honorific correctness, are inferred from automatic metrics and proxy judge models and require human validation before strong interpretive claims can be made. Second, the cross-model comparison is partially confounded: Sarvam-1 is evaluated under a compressed CoT scaffold rather than the full LIMP-CoT prompt, so the observed lexical fidelity gap reflects the joint effect of model capability and prompt completeness and cannot be attributed to either factor in isolation. Third, rasa identification accuracy in Stage 1 is unverified. Errors may propagate through the full commitment chain and silently degrade output quality without detection by automatic metrics. Fourth, all results are domain-specific to subtitle-derived dialogue. The evaluation corpus is drawn from OpenSubtitles.org (Arnob et al., 2024) and may include dubbed foreign content alongside original Telugu productions. Dubbed material does not necessarily exhibit authentic *rasa* patterns or native honorific morphology, potentially reducing the cultural validity of both evaluation instances and gold references. This confound cannot be resolved without per-instance provenance metadata not currently available in the corpus, and generalisation to conversational, formal, or journalistic Telugu registers remains uncharacterised. Fifth, greedy decoding (`num_beams=1`) may systematically disadvantage LIMP-CoT, which benefits from diverse sampling (see §5). This experimental constraint should be ablated in future work. Priority future directions include human evaluation for honorific accuracy and rasa coherence, extension to Tamil, Kannada, and Malayalam (Lauscher et al., 2020), and evaluation at the 7B+ parameter scale.

References

Kabir Ahuja, Harshita Diddee, Rishav Hada, Millicent Ochieng, Krithika Ramesh, Prachi Jain, Akshay Nambi, Tanuja Ganu, Sameer Segal, Hussain Muhammad, Kalika Bali, and Sunayana Sitaram. 2023. MEGA: Multilingual evaluation of generative AI. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4232–4267. Association for Computational Linguistics.

Noor Mairukh Khan Arnob, A. Faiyaz, Md Mubtasim Fuad, Shah Murtaza Rashid Al Masud, Baivab Das, and M.F. Mridha. 2024. IndicDialogue: A dataset of subtitles in 10 Indic languages for Indic language modeling. *Data in Brief*, 55:110690.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, and 1 others. 2023. A multitask, multilingual, multimodal evaluation of ChatGPT on reasoning, hallucination, and interactivity. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, volume 1, pages 675–718. Association for Computational Linguistics.

Bharata Muni. 1951. *The Nāṭyaśāstra: A Treatise on Hindu Dramaturgy and Histrionics*. Asiatic Society of Bengal, Calcutta. Translated by Manomohan Ghosh.

Riyaz Ahmad Bhat and Dipti Misra Sharma. 2013. Dependency parsing of Telugu: An agglutinative language. In *Proceedings of the 12th International Workshop on Treebanks and Linguistic Theories*, pages 197–207. Sofia University Press.

Kaj Bostrom and Greg Durrett. 2020. Byte pair encoding is suboptimal for language model pretraining. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4617–4624. Association for Computational Linguistics.

Penelope Brown and Stephen C. Levinson. 1987. *Politeness: Some Universals in Language Usage*. Cambridge University Press, Cambridge.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901.

Samuel Cahyawijaya, Holy Lovenia, Pascale Fung, and 1 others. 2023. NusaCrowd: Open source initiative for Indonesian NLP resources. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13745–13818. Association for Computational Linguistics.

Chris Callison-Burch, Miles Osborne, and Philipp Koehn. 2006. Re-evaluating the role of BLEU in machine translation research. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 249–256. Association for Computational Linguistics.

Seymour Chatman. 1980. *Story and Discourse: Narrative Structure in Fiction and Film*. Cornell University Press, Ithaca, NY.

Jacob Cohen. 1988. *Statistical Power Analysis for the Behavioral Sciences*, 2nd edition. Lawrence Erlbaum Associates, Hillsdale, NJ.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451. Association for Computational Linguistics.

David M. Eberhard, Gary F. Simons, and Charles D. Fennig. 2023. *Ethnologue: Languages of the World*, 26th edition. SIL International, Dallas, TX. Available online at https://www.ethnologue.com.

Yao Fu, Hao Peng, Ashish Sabharwal, Peter Clark, and Tushar Khot. 2023. Complexity-based prompting for multi-step reasoning. In *International Conference on Learning Representations*.

Jay Gala, Pranjal A. Kamble, Raj Dabre, Ratish Puduppully, Anoop Kunchukuttan, Pratyush Kumar, and Mitesh M. Khapra. 2023. IndicTrans2: Towards high-quality and accessible machine translation of all 22 scheduled Indian languages. *Transactions on Machine Learning Research*.

Gemma Team, Google DeepMind. 2025. Gemma 3 technical report. https://arxiv.org/abs/2503.19786. ArXiv:2503.19786.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *International Conference on Learning Representations*.

Zhengbao Jiang, Frank F. Xu, Jun Araki, and Graham Neubig. 2020. How can we know what a language model knows? In *Transactions of the Association for Computational Linguistics*, volume 8, pages 423–438. MIT Press.

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

Divyanshu Kakwani, Anoop Kunchukuttan, Satya Golla, N Sai Chetan Gokul, Avik Bhatt, Mitesh M. Khapra, and Pratyush Kumar. 2020. IndicNLPSuite: Monolingual corpora, evaluation benchmarks and pretrained multilingual language models for Indian languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4948–4961. Association for Computational Linguistics.

Simran Khanuja, Divyanshu Kakwani, Satya Golla, Gokul Bhatt, Amir Arora, Anoop Kunchukuttan, and Partha Talukdar. 2021. MuRIL: Multilingual representations for indian languages. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7319–7330. Association for Computational Linguistics.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems*, volume 35, pages 22199–22213.

Bhadriraju Krishnamurti. 2003. *The Dravidian Languages*. Cambridge University Press, Cambridge.

Anoop Kunchukuttan, Divyanshu Kakwani, Satya Golla, Avik Bhatt, Amir Arora, Mitesh M. Khapra, and Pratyush Kumar. 2020. The AI4Bharat-IndicNLP corpus: Monolingual corpora and word embeddings for Indic languages. In *Proceedings of the Eighth Workshop on Asian Translation*, pages 68–77. Association for Computational Linguistics.

Anne Lauscher, Vinit Ravishankar, Ivan Vulić, and Goran Glavaš. 2020. From zero to hero: On the limitations of zero-shot cross-lingual transfer with multilingual transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 4483–4499. Association for Computational Linguistics.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119. Association for Computational Linguistics.

Aman Madaan, Amrith Peng, Anshuman Roth, Dhruv Yang, Shrimai Prabhumoye, Paul Liang, Jamie Callan, and Alan W Black. 2020. Politeness transfer: A tag and generate approach. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1869–1881. Association for Computational Linguistics.

Nitika Mathur, Timothy Baldwin, and Trevor Cohn. 2020. Tangled up in BLEU: Reevaluating the evaluation of automatic machine translation evaluation metrics. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4984–4997. Association for Computational Linguistics.

Sabrina J. Mielke, Ryan Cotterell, Kyle Gorman, Brian Roark, and Jason Eisner. 2021. Between words and characters: A brief history of open-vocabulary modeling and tokenization in NLP. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7557–7573. Association for Computational Linguistics.

Tong Niu and Mohit Bansal. 2018. Polite dialogue generation without parallel data. *Transactions of the Association for Computational Linguistics*, 6:373–389.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318. Association for Computational Linguistics.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, and 1 others. 2019. PyTorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, volume 32.

Flor Miriam Plaza-del Arco, Roman Klinger, Rodrigo Agerri, and Valerio Basile. 2024. Emotion analysis in NLP: Trends, gaps and roadmap for future directions. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation*. ELRA and ICCL.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191. Association for Computational Linguistics.

Vinodkumar Prabhakaran, Kalina Bhatt, Hannah Rashkin, and Mark Diaz. 2024. Power and politeness: Computational perspectives on social interaction. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Tutorial Abstracts)*. Association for Computational Linguistics.

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Kurt Shuster, Eric Michael Smith, Y-Lan Boureau, and Jason Weston. 2021. Recipes for building an open-domain chatbot. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 300–325. Association for Computational Linguistics.

Phillip Rust, Jonas Pfeiffer, Ivan Vulić, Sebastian Ruder, and Iryna Gurevych. 2021. How good is your tokenizer? on the monolingual performance of multilingual language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*, pages 3118–3135. Association for Computational Linguistics.

Jivnesh Sandhan, Ashish Shukl, Bhargav Bhatt, and Pawan Goyal. 2023. Computational analysis of rasa in Sanskrit poetry. In *Proceedings of the 2023 Conference of the European Chapter of the Association for Computational Linguistics (Student Research Workshop)*, pages 81–90. Association for Computational Linguistics.

Sarvam AI. 2024. Sarvam-1: A comprehensive language model for Indian languages. https://www.sarvam.ai/blog/sarvam-1. Accessed: March 2025.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. Controlling politeness in neural machine translation via side constraints. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 35–40. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016b. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725. Association for Computational Linguistics.

Iulian V. Serban, Alessandro Sordoni, Yoshua Bengio, Aaron Courville, and Joelle Pineau. 2016. Building end-to-end dialogue systems using generative models: Experiments on the Ubuntu dialogue corpus. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed Chi, Nathanael Schärli, and Denny Zhou. 2023. Large language models can be easily distracted by irrelevant context. In *Proceedings of the 40th International Conference on Machine Learning*, pages 31210–31227. PMLR.

Thorvald Sørensen. 1948. A method of establishing groups of equal amplitude in plant sociology based on similarity of species content. *Kongelige Danske Videnskabernes Selskab*, 5(4):1–34.

V. Sreejith, C. K. Aravindh, and Priyanka Prabhu. 2017. Navarasa: Understanding emotions in poetry. In *Proceedings of the 14th International Conference on Natural Language Processing*, pages 261–271. NLP Association of India.

Sanford B. Steever, editor. 1998. *The Dravidian Languages*. Routledge, London and New York.

Karumuri V. Subbarao. 2012. *South Asian Languages: A Syntactic Typology*. Cambridge University Press, Cambridge.

Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, and 1 others. 2020. SciPy 1.0: Fundamental algorithms for scientific computing in Python. *Nature Methods*, 17(3):261–272.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Weights & Biases. 2020. Weights & biases: Machine learning experiment tracking. https://wandb.ai. Accessed: 2025.

Anna Wierzbicka. 1991. *Cross-Cultural Pragmatics: The Semantics of Human Interaction*. Mouton de Gruyter, Berlin.

Frank Wilcoxon. 1945. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80–83.

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. Association for Computational Linguistics.

Shijie Wu and Mark Dredze. 2019. BETO, Bentz, Becas: The surprising cross-lingual effectiveness of BERT. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, pages 833–844. Association for Computational Linguistics.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. In *Advances in Neural Information Processing Systems*, volume 36.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020a. BERTScore: Evaluating text generation with BERT. In *International Conference on Learning Representations*.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2020b. DIALOGPT: Large-scale generative pre-training for conversational response generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 270–278. Association for Computational Linguistics.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc Le, and Ed Chi. 2023. Least-to-most prompting enables complex reasoning in large language models. In *International Conference on Learning Representations*.