# Shared Task on Prompt Style Recovery for Large Language Models in Telugu

**Premjith B[1], Jyothish Lal G[1], Bharathi Raja Chakravarthi[2], Saranya Rajiakodi[3], Durairaj Thenmozhi,[4], Ratnavel Rajalakshmi,[5], Rahul Ponnusamy[2], Chinthala Bhuvanesh[1],**

[1]Amrita School of Artificial Intelligence, Coimbatore,
Amrita Vishwa Vidyapeetham, India,
[2]Unit for Inclusive AI, Data Science Institute, University of Galway, Ireland,
[3]Central University of Tamil Nadu, Thiruvarur, Tamil Nadu, India,
[4]Department of Computer Science and Engineering,
Sri Sivasubramaniya Nadar College of Engineering,
Kalavakkam, Tamil Nadu, 603110, India,
[5]School of Computer Science and Engineering,
Vellore Institute of Technology, Chennai, India

**Correspondence:** b_premjith@cb.amrita.edu

## Abstract

This paper presents an overview of the Shared Task on Prompt Recovery for Large Language Models (LLMs) in Telugu, organized as part of DravidianLangTech @ ACL 2026. The task focuses on identifying the underlying communicative style of Telugu text excerpts, framed as a nine-class single-label classification problem covering Formal, Informal, Optimistic, Pessimistic, Humorous, Serious, Inspiring, Authoritative, and Persuasive tones. The dataset was constructed by collecting Telugu YouTube comments and generating style-modified variants using an LLM, resulting in 3,000 training instances, 300 validation samples, and 301 test samples. A total of 52 teams registered for the shared task, with 13 teams submitting valid system predictions. Systems explored diverse approaches, including transformer-based fine-tuning (IndicBERT, MuRIL, XLM-R), ensemble and stacking methods, pairwise modeling strategies, curriculum learning, and few-shot large language model prompting. Evaluation was conducted using Macro F1-score as the primary metric. The top-performing system achieved a Macro F1-score of 0.2987. Overall results indicate that Telugu prompt-style recovery remains a challenging problem, particularly due to stylistic overlap and high lexical similarity across classes.

## 1 Introduction

Large Language Models (LLMs) are capable of producing fluent, coherent, and contextually appropriate multilingual text across a wide range of domains. Typically, LLMs are general purpose models, which are pretrained on large-scale diverse corpora. This helps LLMs to perform tasks such as summarization, translation, question answering, dialogue generation and content creation with minimal task-specific supervision.

A key factor behind the effective use of LLMs is prompt engineering, which is an approach to design and structure the input prompts to guide the LLM to obtain desired results. unlike fine-tuning, Prompt engineering (Boonstra, 2025), (Marvin et al., 2024) doesn't update the parameters of the model, but leverages carefully crafted instructions, contextual cues, examples (few-shot prompting) and formatting strategies to control the model behavior. Therefore, it is possible to influence the style and intent of generated responses through variations in wording, constraints, or demonstrations (Liu et al., 2024a), (Zarra and Chiheb, 2025).

Prompt engineering for style transfer focuses specifically on controlling the communicative tone, register, or rhetorical framing of generated text (Luo et al., 2023), (Kong et al., 2025), (Lai et al., 2024). By explicitly specifying stylistic attributes—such as formal vs. informal tone, optimistic vs. pessimistic outlook, humorous vs. serious expression, or persuasive vs. authoritative voice—prompts can direct LLMs to rewrite or generate content in a target style while preserving the underlying semantic content (Liu et al., 2024b). Effective style-oriented prompting requires sensitivity to linguistic markers such as lexical choice, sentence structure, discourse patterns, and pragmatic cues. Consequently, understanding and recovering stylistic intent becomes essential not only for generating style-consistent outputs but also for



*Proceedings of the Sixth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 124–133

building systems capable of recognizing and modeling nuanced communicative variation across languages.

Building upon the growing research for style-aware prompting and generation, we introduce the Shared Task on Prompt Recovery for LLM in Telugu, organized as part of DravidianLangTech @ ACL 2026. The shared task focuses on Telugu Prompt-Style Recovery, where the objective is to automatically identify the communicative style underlying a Telugu transcript excerpt. Given an input sentence or short passage, participating systems must classify the text into one of nine stylistic categories: Formal, Informal, Optimistic, Pessimistic, Humorous, Serious, Inspiring, Authoritative, or Persuasive. Each category represents a distinct tonal, pragmatic, or rhetorical intention expressed by the speaker or writer (Gao et al., 2024), (Chen et al., 2024) .

The shared task attracted 52 registrations, of which 13 teams successfully submitted system predictions. We used macro F1 score for evaluating the predictions and the top performing team achieved an F1 score of 0.2987.

## 2 Task Description

This shared task focuses on identifying the communicative style embedded in a Telugu text excerpt. The primary objective of the task is multi-class style classification, where systems are required to predict the underlying prompt style used to generate or rewrite a given Telugu sentence or short paragraph.

Given an input text, participating systems must classify it into one of the following nine stylistic categories:

- **Formal**: polite, structured language; professional register; full sentences; minimal slang.
- **Informal**: conversational tone; colloquial expressions; slang, contractions, emojis; second-person address.
- **Optimistic**: positive outlook; encouragement; future-oriented success framing.
- **Pessimistic**: negative or doubtful tone; cautionary language; bleak outlook.
- **Humorous**: playful exaggeration; irony; jokes; light-hearted metaphors.
- **Serious**: sober and factual tone; grave subject matter; absence of humor.
- **Inspiring**: motivational language; calls to action; uplifting framing ("you can do it").
- **Authoritative**: directive or commanding voice; expert-like certainty; imperative guidance.
- **Persuasive**: convincing or benefit-oriented appeals; lobbying or sales-like language.

| Dataset | No. of Instances |
|---|---|
| Train | 3,000 |
| validation | 300 |
| Test | 301 |

Table 1: The distribution of the dataset across train, validation and test datasets

The problem is formulated as a single-label multi-class classification task, where each instance belongs to exactly one style category. Systems are evaluated using standard classification metrics to ensure fair comparison and reproducibility.

## 3 Dataset Description

The dataset was prepared by collecting comments in Telugu from YouTube. The comments were passed to ChatGPT[1] to generate style-modified versions. Each instance of the dataset consists of an original Telugu text paired with a style-modified version.

The dataset was split into three - train, test and validation. The split is given in Table 1. Each dataset contains four columns - ID (A unique identifier for each instance), original transcripts (Contains the original, unmodified Telugu text), Change style (Contains the modified version of the original transcript), and STYLE (Specifies the target stylistic category).

The *ORIGINAL TRANSCRIPTS* column contains the base Telugu sentences, while the *CHANGE STYLE* column provides modified versions of those sentences in a different stylistic tone. The *STYLE* column represents the target label, consisting of 9 distinct style categories. All nine styles are present in the training, validation, and test splits. In the training set, the most frequent style is Pessimistic with 347 samples; in the validation set,

[1] chatgpt.com

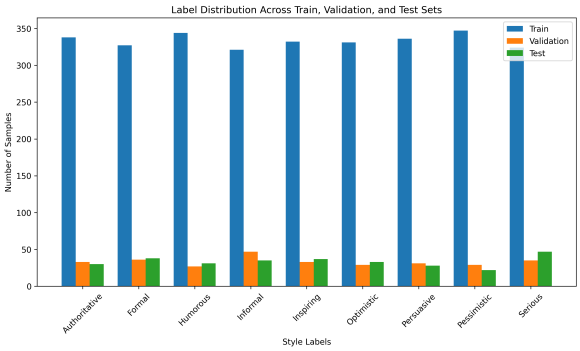


Figure 1: The distribution of data across nine classes in train, validation and test datasets

Informal appears most frequently with 47 samples; and in the test set, Serious is the most frequent with 47 samples. This suggests a reasonably distributed multi-class setting without severe imbalance.

The training set contains 395 unique original transcripts and 2545 unique style-modified sentences, indicating substantial paraphrasing variation in the CHANGE STYLE column. The validation and test sets also show strong diversity, with over 200 unique original transcripts in each split and nearly all style-modified sentences being unique. This demonstrates that the dataset supports robust modeling for paraphrasing, stylistic transformation, and embedding-based learning. The data are evenly distributed across nine classes in train, validation and test datasets. The label distribution is illustrated in Figure 1. Among all the classes, Pessimistic (347 samples), Informal (47 samples) and Serious (47 samples) are the most frequent classes in training, validation and test datasets.

Below is an example instance used for this task.

**Original text** — వంశీ అనేది ఒక విద్యార్థి అతనికి డిగ్రీ ఉంది ఉద్యోగాలు స్కిల్ స...ాలు కావాలంటే తెలపండి లేదా సారాంశం మాత్రమే అవసరమైతే అనుమతించండి.

**Changed style** — ఈ రోజుల్లో విద్యార్థులు విద్యతో పాటు వివిధ నైపుణ్యాలను అభ్యసి...టికల్ అనుభవం  ఇవే మీ ఉద్యోగ అవకాశాలను గణనీయంగా పెంచుతాయి.

**Prompt style** - Formal

## 4 System Description

This section discusses the descriptions of the systems submitted by the participating teams. There are 52 registrations for the shared task. However, 11 teams submitted their predictions through the provided Google form.

### 4.1 Still Loading

Team Stll Loading's submissions demonstrate a hybrid late-fusion ensemble that combines semantic grounding specific to Indic language with general multilingual capabilities. The system successfully captures both global semantic patterns and the unique morphosyntactic features of the Telugu script by fusing models like XLM-RoBERTa Large (Conneau et al., 2020) and mBERT (Devlin et al., 2019) with specialized frameworks like ai4bharat/IndicBERTv2-MLM-only (Doddapaneni et al., 2023) and MuRIL (Khanuja et al., 2021). For cross-segment style grounding, the method involves concatenating the styled target text with the original context. An optimization pipeline that combines label smoothing, cosine learning rate scheduling, and gradient accumulation achieves robust performance across a variety of Telugu "Rasas" while keeping an effective batch size of 32 constant.

### 4.2 Mano_sub

The MuRIL model was trained using a "frozen-backbone" technique, in which all transformer layers were frozen and only the classification head was updated. The AdamW optimizer (Loshchilov and Hutter, 2017) was used for 10 training epochs with a batch size of 16 and a learning rate of $1 \times 10^{-4}$. They used the Metal Performance Shaders backend for GPU acceleration on macOS. A sliding window with a 256-token sequence length was employed for inference. To ensure that the classifier produced results that were generally applicable, final predictions for the unlabeled test set were based on the epoch that obtained the highest Macro F1-score.

### 4.3 Medhastra

Team Medhastra uses XLM-RoBERTa-base to reformulate the Telugu prompt style classification task into a pairwise binary classification problem. By concatenating the input text with style labels, they produce nine candidate pairs for every training instance, increasing the training set from 3,000 to 27,000 instances in spite of the scarcity of labeled data. With one positive and eight negative examples per instance, this method produces a balanced binary task. The binary classifier evaluates

each style candidate during inference and chooses the one with the highest confidence score. After four epochs of fine-tuning, the model achieved a macro F1 score of 0.4915, with training and validation loss decreasing over time.

### 4.4 DLRG

The proposed system combines three different models - XLM-R, mT5 (Xue et al., 2021) and BiLSTM instead of relying on just one. By merging their prediction scores through a simple stacking method, the system uses each model's strengths, reduces errors, and achieves better and more stable accuracy. Stacked interaction and entropy features train a calibrated Logistic Regression meta-classifier, combining model strengths to improve robustness and macro-F1 performance.

### 4.5 JerinWarriors

They used a model that compares the original text and the rewritten text together. Both texts are processed at the same time so the model can learn how the style changes. The model focuses on the rewritten text and predicts the final style label based on the differences it learns. The model was trained using labeled examples of original and rewritten texts. Texts were limited to the first part to save memory and improve learning. During testing, when the original text was not available, the rewritten text was used in both inputs so the model could still make predictions.

### 4.6 TechNova

First, the training and validation datasets were combined, missing data was eliminated, and label formats were standardized across nine stylistic categories as part of a systematic preprocessing step for the proposed system. To guarantee proportional representation in both splits, the cleaned corpus was stratified. To find lexical patterns and stylistic cues in Telugu, feature extraction employed TF-IDF vectorizations based on word n-grams (one to four tokens). To deal with class imbalances and noisy annotations, an ensemble classifier that combined a Linear Support Vector machine, a Logistic regression model, and a Multinomial Naive Bayes model worked within a hard voting framework.

### 4.7 DeepScope

Using pretrained Indic language models, they tackled the Telugu Prompt-Style Recovery task by presenting it as a nine-class style classification problem. Through a custom PyTorch training loop, IndicBERTv2 was fine-tuned on a curated dataset. The model input consisted of task instructions and rewritten Telugu text that did not change or reveal target labels. Validation, the removal of samples with missing or invalid text, and intra-class consistency filtering to remove noise all improved data integrity. A curriculum-based approach was used for training, starting with high-confidence samples and progressively adding more ambiguous data.

### 4.8 PromptRecovery_Alchemists (PR_Alchemists)

The system uses GPT-5.2 (OpenAI, 2026) as its core model and employs a large language model-based approach through the DSPy framework. It builds 27 demonstrations from the training set using original Telugu texts, their style-transformed versions, and style labels, combining few-shot in-context learning and Chain-of-Thought reasoning. Untruncated text pairs and a DSPy signature (Khattab et al., 2024), (Khattab et al., 2022) directs the comparison of tone, punctuation, sentence structure, word choice, and emotional markers are given to the model during inference. Across nine categories—Authoritative, Formal, Humorous, Informal, Inspiring, Optimistic, Persuasive, Pessimistic, and Serious—the Chain-of-Thought module enables intermediate reasoning for more understandable style analysis.

### 4.9 Error_500

This submission explores multiple transformer-based approaches for Telugu style classification. Specifically, it utilizes IndicBERT v2 and XLM-RoBERTa base as foundational pretrained language models. The approaches incorporate varied training strategies, preprocessing techniques, and imbalance-handling mechanisms to enhance robustness and generalization. Each model is trained and validated independently, and predictions are generated from their respective best-performing checkpoints for comparative evaluation using macro-level metrics.

### 4.10 codecrackers

In this submission, they utilized the mT5-small transformer model from Hugging Face to predict one of nine predefined style labels. The methodology involved framing a prompt-based classification, where each text sample was fed to the model

with a structured instruction asking it to output the corresponding numeric style label. During inference, the model generated a number per input text, which was then validated and saved as the final label.

### 4.11 Semantica

They define the prompt-recovery task as style identification from pairs of texts: an Original Transcript and a Changed-Style Transcript. The training set is cleaned to enhance label consistency by removing duplicates, conflicting pairs with multiple style labels, and overlaps between training and validation datasets. A sliding-window strategy is employed for fine-tuning a Transformer classifier due to the encoder's context limit, where a fixed input template concatenates both texts, maximizing length constraints while generating overlapping windows. Training utilizes windowed examples inheriting their parent label and optimizing cross-entropy loss, with a weighted sampler for class balancing.

### 4.12 Cuet Yet Another Baseline (CYAB)

The submission investigates Telugu stylistic tone classification using transformer-based supervised learning. It fine-tunes a pretrained language model (l3cube-pune/telugu-bert) (Joshi, 2022) on a dataset with original and style-modified transcripts to explicitly capture stylistic changes. Enhanced encoder representations combine the [CLS] token with mean pooled token embeddings, processed through a multi-layer feed-forward classification head. The training applies weighted cross-entropy loss, cosine learning rate scheduling with warmup, and early stopping based on validation macro F1 score. In another run, MuRIL-base is fine-tuned on style-transferred Telugu text, concatenating [CLS] with surface-level linguistic features and classifying through a fusion layer using class-weighted cross-entropy loss. Final predictions are derived from soft voting across three seeds.

### 4.13 Axiom

The system utilizes a stacking ensemble of classical machine learning models, trained on TF-IDF features from the "Change Style" column, which encodes stylistic cues. It integrates character-level n-grams (2–6) and word-level n-grams (1–3) through a TF-IDF FeatureUnion with sublinear term frequency scaling, enhancing its sensitivity to Telugu morphological patterns and phrase-level style signals. A Logistic Regression meta-learner is employed on the probability outputs of these models, trained using 5-fold cross-validation to create meta-features. Evaluation involved a stratified 80/20 training data split for validation accuracy, followed by retraining on the full dataset for predictions on the unlabeled test set.

Table 2 summarizes the fusion approaches, embedding approaches and various notable features followed by the participating teams.

## 5 Analysis of the Submission

In this task, systems were evaluated on a held-out test set, with submissions assessed across four metrics: Macro F1-Score (primary ranking criterion), Accuracy, Precision, and Recall. Teams were permitted to submit three runs; the best-performing run per team was used for the official ranking.

Table 3 presents the official results for all 13 participating teams, sorted by Macro F1-Score (descending).

Figure 4 illustrates the Macro F1-scores across all 13 teams. The team Error_500 leads the ranklist with an F1 score of 0.2987, but the margin over 2nd place (JerinWarriors, 0.2588) is relatively small ($\approx$0.04). From the figure it is evident that a clear performance cliff exists between Rank 5 (0.2285) and Rank 6 (0.1745). The performance of bottom 3 teams (Axiom, Mano_Sub, DeepScope) score are well below 0.10, which may be due to the weaker approaches they designed for this task. Another interesting inference is most teams score below the average. It is evident from the results that Error_500 leads with an F1 of 0.2987, nearly $10\times$ that of last-place DeepScope (0.0289). Another inference from the Table 3 is that PromptRecovery_Alchemists (Rank 4) achieves the highest precision of any team (0.3346), surpassing even the 1st-place team, but this comes at the cost of lower recall (0.2472), which results in a lower overall F1.

Figure 2 maps each team in precision-recall space, colored by F1-score. Teams near or above the diagonal (P=R line) tend to have balanced metrics; systems far from the diagonal exhibit metric imbalance which may indicate different modeling assumptions or post-processing strategies. In the figure, the X-axis and Y-axis represent recall and precision, respectively. It is clear from the figure that PromptRecovery_Alchemists (PR_Alchemists) is the biggest outlier with a high-

| Team | Fusion / Approach | MuRIL | IndicBERT | XLM-R | mT5 | GPT | TF-IDF | Stacking | Sliding | Pairwise | Weighted | Curriculum | Notable Feature |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Still Loading | Late fusion (avg logits) | ✓ | ✓ | ✓ | × | × | × | × | × | × | ✓ | × | Label smoothing, cosine LR |
| Mano_sub | Frozen backbone + classifier | ✓ | × | × | × | × | × | × | × | × | × | × | Sliding window (256), MPS |
| Medhastra | Pairwise binary classification | × | × | ✓ | × | × | × | × | × | ✓ | × | × | Expands 3k→27k |
| DLRG | Stacking + LR meta | × | × | ✓ | ✓ | × | × | ✓ | × | × | × | × | Entropy features |
| JerinWarriors | Joint pair processing | × | × | × | × | × | × | × | × | × | × | × | Dual-input contrast |
| TechNova | Hard voting ensemble | × | × | × | × | × | ✓ | ✓ | × | × | × | × | Word n-grams (1–4) |
| DeepScope | Fine-tuned single model | × | ✓ | × | × | × | × | × | × | × | × | ✓ | Curriculum learning |
| PR_Alchemists | Few-shot CoT (DSPy) | × | × | × | × | ✓ | × | × | × | × | × | × | 27-shot demos |
| Error_500 | Multi-strategy ensemble | × | ✓ | ✓ | × | × | × | × | × | × | ✓ | × | Imbalance handling |
| codecrackers | Zero-shot prompting | × | × | × | ✓ | × | × | × | × | × | × | × | Numeric label gen. |
| Semantica | Sliding window + logit agg. | × | × | ✓ | × | × | × | × | ✓ | × | × | × | Weighted sampler |
| CYAB | Soft voting (3 seeds) | ✓ | × | × | × | × | × | × | × | ✓ | × | × | CLS + mean pool |
| Axiom | Stacking + LR meta | × | × | × | × | × | ✓ | ✓ | × | × | × | × | Char (2–6), word (1–3) |

Table 2: Comparison of methodologies and architectural choices across participating teams.

| Team Name | Run | F1 | Accuracy | Precision | Recall | Rank |
|---|---|---|---|---|---|---|
| Error_500 | Run 3 | 0.2987 | 0.2990 | 0.2979 | 0.3144 | 1 |
| JerinWarriors | Run 1 | 0.2588 | 0.2691 | 0.2926 | 0.2812 | 2 |
| DLRG | Run 1 | 0.2451 | 0.2425 | 0.2716 | 0.2496 | 3 |
| PromptRecovery_Alchemists | Run 1 | 0.2406 | 0.2525 | 0.3346 | 0.2472 | 4 |
| Cuet Yet Another Baseline | Run 2 | 0.2285 | 0.2193 | 0.2406 | 0.2356 | 5 |
| Semantica | Run 1 | 0.1745 | 0.1960 | 0.1619 | 0.2194 | 6 |
| Still Loading | Run 1 | 0.1703 | 0.1761 | 0.1679 | 0.1865 | 7 |
| Codecrackers | Run 1 | 0.1516 | 0.1761 | 0.1716 | 0.1735 | 8 |
| Medhastra | Run 1 | 0.1205 | 0.1196 | 0.1205 | 0.1231 | 9 |
| TechNova | Run 1 | 0.1106 | 0.1462 | 0.0974 | 0.1365 | 10 |
| Axiom | Run 1 | 0.0797 | 0.0797 | 0.0797 | 0.0805 | 11 |
| Mano_Sub | Run 1 | 0.0491 | 0.1462 | 0.0326 | 0.1161 | 12 |
| DeepScope | Run 1 | 0.0289 | 0.0498 | 0.0236 | 0.0674 | 13 |

Table 3: Official rank list for the Prompt Recovery for LLM in Telugu shared task.

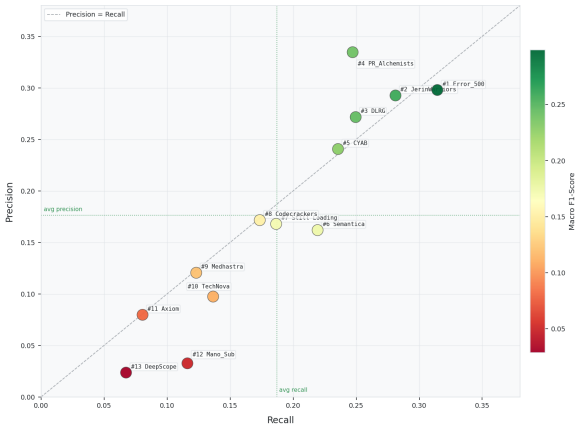


Figure 2: Precision vs Recall Scatter Plot

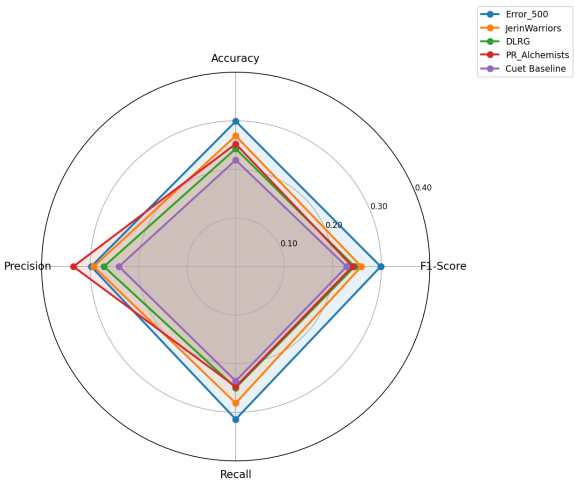


Figure 3: Radar chart for top 5 teams. A larger, more uniform polygon indicates better overall performance.

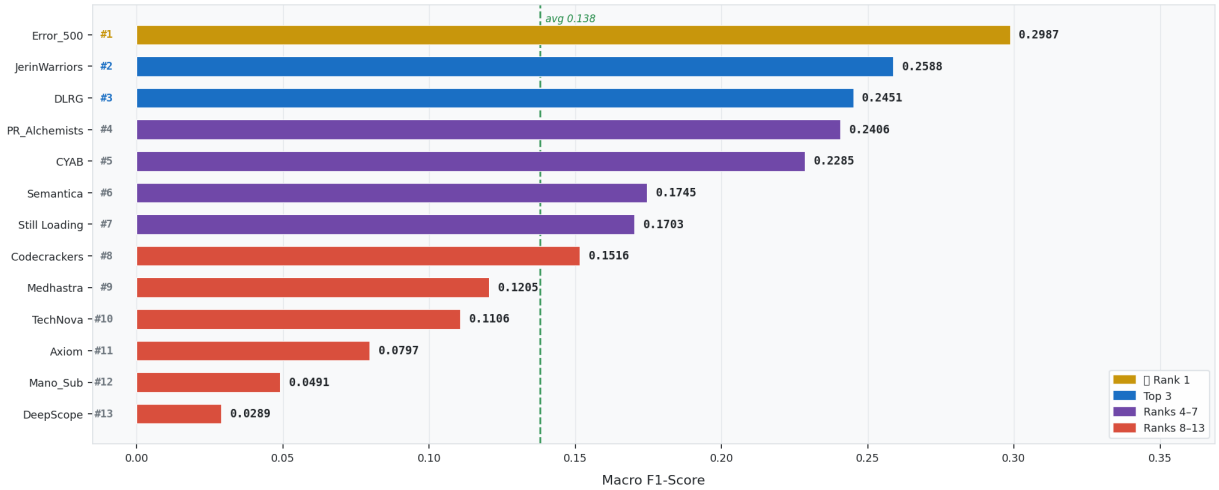


Figure 4: Rankings of the teams participated in the shared task

est precision (0.3346) but relatively low recall (0.2472). Semantica and TechNova are recall-heavy teams, who sit below the diagonal. Error_500, JerinWarriors, and DLRG cluster near the diagonal in the top-right, which indicates a balanced and well-tuned approach for the task. Mano_Sub and DeepScope appear in the bottom-left, which shows low scores on both axes, and it indicates weak overall performance from both teams.

Figure 3 presents a radar chart comparing the top 5 teams across all four metrics simultaneously, enabling a holistic comparison of their strengths and weaknesses. Error_500 has the most uniform and expansive polygon across all metrics. PromptRecovery_Alchemists has a distinctly elongated shape toward Precision, confirming its precision-biased strategy at the expense of recall.

Table 4 provides descriptive statistics for all four evaluation metrics across the 13 systems. Precision exhibits the largest range (0.3110), indicating the greatest variability in how teams trade off false positives. Recall is generally higher than precision across most teams, suggesting a mild recall bias in system design — possibly because systems over-generate candidate prompt segments to maximize recall.

All the top-performing teams scored low F1-score, which is due to the presence of same words and sentences in difference in the dataset. There are 2545 unique changed style texts in the training dataset, and 455 are repeated, which accounts to the 84.84% of the unique changed style text. The presence of the repeated contents generated by the LLMs for different style can be the reason for obtaining low scores for teams. Two examples of style modified text with similarities among multiple style tones are given in the appendix.

| Statistic | F1 | A | P | R |
|---|---|---|---|---|
| Mean | 0.1379 | 0.1656 | 0.1440 | 0.1688 |
| Median | 0.1205 | 0.1761 | 0.1205 | 0.1735 |
| Std Dev | 0.0820 | 0.0722 | 0.0844 | 0.0726 |
| Maximum | 0.2987 | 0.2990 | 0.3346 | 0.3144 |
| Minimum | 0.0289 | 0.0498 | 0.0236 | 0.0674 |
| Range | 0.2698 | 0.2492 | 0.3110 | 0.2470 |

Table 4: Descriptive statistics of evaluation metrics. (A) Accuracy, (P) Precision, (R) Recall

Most teams show near-identical F1 and Accuracy values, suggesting a reasonably balanced label distribution in the test set. Notable exceptions (TechNova, Mano_Sub) may reflect unusual or degenerate prediction patterns.

## 6 Conclusion

The Prompt Style Recovery Shared Task for Telugu provides one of the first systematic evaluations of style identification in LLM-generated Telugu text. Despite diverse modeling strategies—including transformer fine-tuning, ensemble stacking, pairwise contrastive formulations, curriculum learning, and LLM-based few-shot reasoning—the overall performance across systems remained modest, with the best Macro F1-score reaching 0.2987 . The relatively low scores highlight the intrinsic difficulty of distinguishing nuanced communicative styles, especially when lexical overlap exists across categories and when style-modified texts

share substantial semantic and structural similarity. Analysis of system behaviors revealed different optimization trade-offs, with some models favoring precision while others leaned toward recall. Transformer-based and ensemble approaches generally outperformed classical baselines, suggesting that contextual semantic modeling is crucial for capturing subtle stylistic cues in Telugu. However, repeated or near-identical LLM-generated style variations within the dataset likely limited discriminative learning.

## 7 Limitations

This section discusses the limitations of the shared task. The first challenge is the size of the dataset, which is relatively small, consisting of only 3,000 training samples with 300 validation and 301 test instances. Such a limited dataset may restrict the ability to generalize effectively and learn robust stylistic representations. Another limitation arises from the inherent overlap between some of the stylistic categories. Classes such as optimistic, inspiring, and persuasive, or serious and authoritative, may share similar lexical and semantic characteristics. This overlap introduces ambiguity in the classification process and increases the difficulty of clearly separating stylistic boundaries. Future work could address these limitations by expanding the dataset size, incorporating more naturally occurring style variations, and exploring richer contextual modeling approaches for multilingual prompt-style recovery.

## 8 Ethical Considerations

The dataset used in this shared task was constructed from publicly available Telugu YouTube comments and style-modified using a large language model. Care was taken to ensure that the dataset does not contain personally identifiable information. The task focuses solely on stylistic classification and does not attempt to infer sensitive personal attributes of individuals. Nevertheless, since the dataset contains user-generated content, there may be instances of informal or subjective language. Researchers using this dataset should ensure responsible use of the data and avoid deploying models in ways that could misrepresent or misuse stylistic interpretations.

## Acknowledgments

This work was conducted with the financial support from Science Foundation Ireland (SFI) under Grant Number SFI/12/RC/2289_P2(Insight_2), supported in part of Science Foundation Ireland Centre for Research Training in Artificial Intelligence under Grant No. 18/CRT/6223.

## References

Lee Boonstra. 2025. Prompt engineering. *Google, https://www. kaggle. com/whitepaper-prompt-engineering*.

Jianlong Chen, Wei Xu, Zhicheng Ding, Jinxin Xu, Hao Yan, and Xinyu Zhang. 2024. Advancing prompt recovery in nlp: A deep dive into the integration of gemma-2b-it and phi2 models.

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, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

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, Toronto, Canada. Association for Computational Linguistics.

Lirong Gao, Ru Peng, Yiming Zhang, and Junbo Zhao. 2024. DORY: Deliberative prompt recovery for LLM. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 10614–10632, Bangkok, Thailand. Association for Computational Linguistics.

Raviraj Joshi. 2022. L3cube-hindbert and devbert: Pre-trained bert transformer models for devanagari based hindi and marathi languages. *arXiv preprint arXiv:2211.11418*.

Simran Khanuja, Diksha Bansal, Sarvesh Mehtani, Savya Khosla, Atreyee Dey, Balaji Gopalan, Dilip Kumar Margam, Pooja Aggarwal, Rajiv Teja Nagipogu, Shachi Dave, Shruti Gupta, Subhash Chandra Bose Gali, Vish Subramanian, and Partha Talukdar. 2021. Muril: Multilingual representations for indian languages. *Preprint*, arXiv:2103.10730.

Omar Khattab, Keshav Santhanam, Xiang Lisa Li, David Hall, Percy Liang, Christopher Potts, and Matei Zaharia. 2022. Demonstrate-search-predict: Composing retrieval and language models for knowledge-intensive NLP. *arXiv preprint arXiv:2212.14024*.

Omar Khattab, Arnav Singhvi, Paridhi Maheshwari, Zhiyuan Zhang, Keshav Santhanam, Sri Vardhamanan, Saiful Haq, Ashutosh Sharma, Thomas T. Joshi, Hanna Moazam, Heather Miller, Matei Zaharia, and Christopher Potts. 2024. Dspy: Compiling declarative language model calls into self-improving pipelines.

Chaona Kong, Jianyi Liu, Yifan Tang, and Ru Zhang. 2025. Neuron activation modulation for text style transfer: Guiding large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 7735–7747.

Wen Lai, Viktor Hangya, and Alexander Fraser. 2024. Style-specific neurons for steering LLMs in text style transfer. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 13427–13443, Miami, Florida, USA. Association for Computational Linguistics.

Qingyi Liu, Jinghui Qin, Wenxuan Ye, Hao Mou, Yuxuan He, and Keze Wang. 2024a. Adaptive prompt routing for arbitrary text style transfer with pre-trained language models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(17):18689–18697.

Shenyang Liu, Yang Gao, Shaoyan Zhai, and Liqiang Wang. 2024b. Stylerec: A benchmark dataset for prompt recovery in writing style transformation. In *2024 IEEE International Conference on Big Data (BigData)*, pages 1678–1685.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Guoqing Luo, Yu Han, Lili Mou, and Mauajama Firdaus. 2023. Prompt-based editing for text style transfer. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5740–5750.

Ggaliwango Marvin, Nakayiza Hellen, Daudi Jjingo, and Joyce Nakatumba-Nabende. 2024. Prompt engineering in large language models. In *Data Intelligence and Cognitive Informatics*, pages 387–402, Singapore. Springer Nature Singapore.

OpenAI. 2026. Introducing gpt-5.2.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Taoufiq Zarra and Raddouane Chiheb. 2025. The influence of prompt politeness on response quality in large language models. In *2025 International Conference on Circuit, Systems and Communication (ICCSC)*, pages 1–7. IEEE.

## A Examples of Style Modified Text with Similarities among Multiple Style Tones

**Original text** — అందరికీ నమస్కారం అండి. అందరూ క్షేమంగా ఉన్నారని తలుస్తున్నాను. మనం ఏ తరగతి నూతన పాఠ్యపుస్తకంలో భాగంగా, ఐదవ పాఠ్యాంశం జలియన్ వాలాబాగ్ గురించి రెండు భాగాల్లో గా చెప్పుకున్నాం. ఇప్పుడు మూడో, నాలుగో భాగాలను కూడా పరిగణనలోకి తీసుకుని, మరికొన్ని ప్రతిపదార్థ పద్యాలను విశ్లేషించుకుందాం. ఈ పద్యాల నుండి మీరు గొప్ప జ్ఞానాన్ని పొందవచ్చు.జలియన్ వాలాబాగ్ అమృత్సర్ పట్టణంలోని ఒక ఉద్యానవనం. ఇక్కడ ఏప్రిల్ న జరిగిన దారుణమైన సంఘటన కారణంగా బ్రిటిష్ సైనికులు అమాయకులపై కాల్పులు జరిపి అనేక మంది ప్రాణాలు కోల్పోయారు. ఈ కాల్పుల్లో మంది ఆ తీరాం మరణించారు. చాలా మంది గాయపడ్డారు. ఉమర్ ఆలీషా గారు ఈ ఘటనను తాము రాసిన ఖండకావ్యాలలో సులభమైన పద్యాలతో వివరిస్తూ, భారతీయుల సరితక ధైర్యాన్ని, నిర్భయత్వాన్ని ఆస్వాదిస్తూ, ప్రభుత్వల పీడనలపై వ్యథ వ్యక్తం చేయడమే లక్ష్యం. ఈ చరిత్రను మీరు తప్పక తెలుసుకోవాలి.పద్య భాగం: ...

**Prompt styles** - Serious (Train ID: PR_TE_TR_0378) and Pessimistic (Train ID: PR_TE_TR_0379)

**Percentage of match** - 96.42

**Translation** - Hello everyone. I hope everyone is well. As part of the new textbook for class 6, we have discussed the fifth chapter of Jallianwala Bagh in two parts. Now, let us consider the third and fourth parts and analyze some more anti-material poems. You can gain great knowledge from these poems. Jallianwala Bagh is a park in the city of Amritsar. Due to the brutal incident that took place here in April, British soldiers opened fire on innocent people and many people lost their lives. Many people died in this firing. Many were injured. Umar Alisha has described this incident in simple poems in his Khandakavyas, enjoying the courage and fearlessness of Indians and expressing grief over the oppression of the government. You must know this history. Verse: ...

**Original text** — చరిత్రకి వినీ ఎరుగని విపత్తు ఇదే. కరీబ్ గంటలే తమిళనాడు మే ఆపత్ కి బారిష్ హో రహి హై. ఐసీ బారిష్ కి సాళ్లకా రికార్డ్ టూట్ గయా హై. చెంబరంబాకం ఏరియాలో అరసం మున్నోటియే తిరకాదదే. చెన్నై మరియు పూరనగర్‌కు లేటెస్ట్ న్యూస్ కమింగ్ ఇన్ ఫ్రమ్ చెన్నై, అక్కడ హెవీ రెయిన్స్ వల్ల బ్రీఫ్ రెస్పైట్ ఉంది. బంగాళాఖాతంలో ఏర్పడిన అల్పపీడనం కారణంగా తమిళనాడు రాష్ట్రవ్యాప్తంగా విస్తారంగా వర్షాలు కురుస్తున్నాయి. తెలుగు ఫిల్మ్ ఇండస్ట్రీకి చెందిన యాక్టర్లు అందరం ఒక గ్రూప్‌గా ఫామ్ అయ్యి మద్రాస్ కోసం ఒక క్యాంపెయిన్ స్టార్ట్ చేశాం. నవంబర్‌లో చెన్నై సిటీ జనాలందరూ గుంపులు గుంపులుగా ఫామ్ అయ్యి ఇళ్లలో ఉన్న మిగతావాళ్లిని కాపాడటానికి ట్రై చేస్తున్నారు. నీళ్లు నడుము వరకు వచ్చేస్తున్నాయి. కొన్ని చోట్ల మనిషి నిలుచుంటే తల మాత్రమే కనిపిస్తుంది. కొద్దిసేపటికే వేల మంది ఇల్లు లేని వాళ్ళు అయిపోయారు. వాతావరణం కనికరిస్తే బ్రతకాల్సిన పరిస్థితికి జనం వచ్చేసారు. హారిఫిక్ సిచ్యుయేషన్ వి హవ్ బీన్ ఫాల్ ఓవర్ తమిళనాడు అండ్ పుదుచ్చేరి. తమిళనాడులో వర్షాలు, వరదల బీభత్సం కొనసాగుతోంది.....

**Prompt styles** - Humorous (Train ID: PR_TE_TR_0338) and Pessimistic (Train ID: PR_TE_TR_0339)

**Percentage of match** - 100
**Translation**- This is a disaster that has never been heard of in history. It has been almost an hour since the Tamil Nadu floods. The annual record of IC rainfall has been broken. In the Chembarambakkam area, there is a brief respite from heavy rains. The latest news for Chennai and Puranagar is coming in from Chennai, where there is a brief respite due to heavy rains. Due to the low pressure formed in the Bay of Bengal, heavy rains are falling across the state of Tamil Nadu. All the actors from the Telugu film industry have formed a group and started a campaign for Madras. In November, all the people of Chennai city have formed groups and are trying to save the rest of the people who are in their houses. The water is reaching the waist. In some places, only the head is visible if a person is standing. In a short time, thousands of people have become homeless. People have come to a situation where they have to survive if the weather allows. Horrific situation has fallen over Tamil Nadu and Puducherry. Rains and floods continue to wreak havoc in Tamil Nadu. Northeast cyclones are lashing Chennai.....