Proceedings of CONF-SEML 2026 Symposium: Computational Analysis and Modeling in Complex Intelligent Systems
DOI: 10.54254/2755-2721/2026.GU32927

Template-Guided Prompting for Long-Tail Emotion Recognition

Yuqin Long

College of International Education, Hunan Institute of Engineering, Xiangtan, China
lyq_1930@163.com

Abstract. Emotion recognition (ER) poses a complex multi-class classification challenge, further complicated by significant class imbalances. In natural dialogue corpora, dominant emotions like neutral are prevalent, while minority emotions such as disgust and fear are notably scarce. This imbalance results in models consistently underperforming on less frequent categories. This paper investigates template-guided prompting as a method to improve long-tail emotion recognition using large language models (LLMs). We employ a unified evaluation framework on the MELD dataset to compare various methods: supervised baselines (TextCNN, BiLSTM), a fine-tuned pre-trained model (BERT-base), and training-free LLM inference (DeepSeek) using three structured prompt templates in both zero-shot and few-shot scenarios (K=1, 3, 5, 10). Our findings demonstrate that template-guided LLM prompting achieves the highest overall performance (Acc=0.6573, Macro-F1=0.5268) and significantly enhances minority-class F1 scores compared to all supervised baselines, without requiring parameter updates. A detailed analysis of hard-sample errors shows that 16.9% of test instances are misclassified by all five models, with minority emotions having hard-sample rates up to 48%. This bias remains even with balanced downsampling (Pearson r=0.986) and is linked to a systematic prediction bias toward the neutral class. These results imply that the difficulties in long-tail ER arise from intrinsic semantic ambiguity rather than just data imbalance, and that structured prompting offers a practical and effective solution for achieving more balanced emotion recognition.

Keywords: Emotion recognition, deep learning, large language models, long tail

1. Introduction

Emotion recognition (ER), which involves automatically identifying the emotional state conveyed in spoken or written communication, has become a key research challenge at the crossroads of natural language processing (NLP), affective computing, and human-computer interaction (HCI). As conversational AI systems rapidly expand, the capability to detect and respond to human emotions is becoming crucial in various applications, such as intelligent dialogue agents, mental health monitoring, educational tutoring systems, and social media analysis [1].

Emotion recognition (ER) is essentially a multi-class classification challenge. Unlike binary sentiment analysis, which classifies sentiments as either positive or negative, ER requires models to distinguish among a wide and qualitatively diverse range of emotional states, such as neutral, joy,

DOI: 10.54254/2755-2721/2026.GU32927

surprise, anger, sadness, disgust, and fear. Each state conveys distinct psychological and communicative signals [2]. This nuanced differentiation is further complicated by the context-dependent nature of emotional expression in real-world conversations. The same expression can convey entirely different emotional valence depending on factors like speaker intent, dialogue history, and pragmatic implications.

A persistent challenge in emotion recognition (ER) is the significant class imbalance. In natural human conversations, emotionally neutral utterances are overwhelmingly prevalent, while expressive emotions such as disgust, fear, and sadness occur much less frequently [3]. This long-tail distribution causes data-driven models to focus excessively on majority-class patterns, thereby diminishing performance on minority emotion categories. Previous research has addressed this issue using data augmentation, class-weighted loss functions, and re-sampling strategies [4]. However, these methods typically require access to training data and model parameters, limiting their flexibility in low-resource or deployment-constrained environments.

The emergence of large language models (LLMs) such as GPT-4, LLaMA, and DeepSeek has transformed NLP tasks by enabling training-free inference through natural language prompts [5]. These LLMs, pre-trained on vast datasets, develop a comprehensive understanding of world knowledge and human emotions, allowing them to classify emotions without task-specific fine-tuning. However, merely prompting the model to classify an utterance does not guarantee equitable treatment across all emotion categories. Minority classes may remain underrepresented due to the model's inherent biases, and without structured guidance, the model might default to predicting more common labels [6].

This gap brings us to our central research question: Can template-guided prompting systematically improve LLM performance on long-tail emotion categories? We hypothesize that carefully designed prompt templates—those that explicitly define label definitions, provide class-balanced in-context demonstrations, and constrain output format—can reduce majority-class bias and encourage more equitable classification behavior in LLMs.

In our study, we conducted a detailed comparative analysis using the Multimodal EmotionLines Dataset (MLED) benchmark, which is a naturalistic multi-speaker conversational corpus sourced from the TV series Friends [7]. The dataset comprises 13,708 utterances with 7-class emotion annotations, displaying a significant long-tail distribution (neutral: 47.0%; fear: 2.6%). We assessed five models within a unified framework: two architecture-based supervised baselines (TextCNN, BiLSTM), a fine-tuned pre-trained model (BERT-base), and DeepSeek LLM utilizing three structured prompt templates in both zero-shot and few-shot (K=1, 3, 5, 10) configurations [8-11]. This setup allows us to separate the effects of model capacity, task-specific training, and prompt structure on long-tail emotion recognition.

Our findings demonstrate that template-guided LLM prompting achieves superior overall performance (Acc=0.6573, Macro-F1=0.5268) and significantly higher minority-class F1 scores compared to all supervised baselines, without necessitating parameter updates. Under stable output conditions, zero-shot prompting either surpasses or matches most few-shot configurations, with prompt template design being crucial for maintaining performance consistency. Beyond benchmarking, we conduct a systematic hard-sample error analysis, defining hard samples as instances misclassified by all five models simultaneously, identifying 442 such cases (16.9% of the test set). The analysis reveals a strong negative correlation between class frequency and hard-sample rate (Pearson $r=-0.777$), a bias that persists even with balanced downsampling ($r=0.986$ between full and balanced profiles), and a systematic neutral-class prediction bias in confusion patterns (sadness→neutral: 235 cases; anger→neutral: 230 cases). These findings collectively illustrate that

DOI: 10.54254/2755-2721/2026.GU32927

the difficulty of long-tail ER is fundamentally rooted in intrinsic semantic and pragmatic ambiguity, rather than merely in training data imbalance, and that structured template prompting offers a practical and scalable strategy for addressing this challenge.

2. Methods

We evaluate four model families on the MELD 7-class ER within a unified framework, testing each on the same test split of 2,610 utterances. Due to the significant class imbalance in MELD, where neutral accounts for 47.2% and both disgust and fear represent only 2.7%, we prioritize Macro-F1 as our primary evaluation metric. Our analysis encompasses shallow supervised models like TextCNN and BiLSTM, a fine-tuned pre-trained model (BERT-base), and a training-free large language model (LLM) inference approach (DeepSeek), enabling a cross-paradigm comparison.

TextCNN. TextCNN effectively captures local n-gram patterns using parallel convolutional filters with varying kernel sizes. This capability makes it particularly well-suited for short, emotionally expressive utterances, where key sentiment cues are often concentrated in a few contiguous words. Its lightweight architecture and rapid training process make it a transparent and reproducible lower-bound baseline for this task.

BiLSTM. A bidirectional LSTM processes the token sequence in both forward and backward directions, effectively capturing long-range contextual dependencies that local convolutional models cannot handle. This bidirectional approach is particularly suitable for emotion recognition, as the sentiment of a word is frequently influenced by its surrounding context.

BERT. BERT-base serves as a robust supervised baseline due to its pre-training on extensive corpora, which equips it with comprehensive lexical and syntactic knowledge. This capability enables the model to detect subtle emotional nuances that surface-level features might overlook. Fine-tuning on task-specific labeled data further tailors these representations to the emotion recognition objective, exemplifying the current standard approach for text classification benchmarking.

LLM. DeepSeek-chat is assessed without updating any parameters to determine if a general-purpose large language model can equal or exceed the performance of supervised systems simply by following instructions. Two prompting strategies are evaluated: zero-shot, which depends entirely on the model's existing language comprehension, and few-shot, which provides a limited number of per-class demonstration examples to counteract the model's natural frequency bias toward the majority class.

3. Main results

3.1. Dataset description

This study utilizes the MELD dataset, focusing solely on the text modality, to perform a 7-class emotion classification task. The dataset is divided into three parts: the training set, used for model fitting; the validation set, employed for hyperparameter selection; and the test set, reserved for standardized comparative evaluation.

Table 1 reveals that the dataset comprises 13,708 utterances, divided into 9,989 samples for training, 1,109 for validation, and 2,610 for testing. To ensure comparability, all models were assessed using the identical test set of 2,610 examples.

DOI: 10.54254/2755-2721/2026.GU32927

Table 1. Statistics on the size of the MELD dataset

Category	Number of utterances
training set(train)	9,989
validation set(dev)	1,109
test set(test)	2,610
Total	13,708

This experiment categorizes emotions into seven groups: neutral, joy, surprise, anger, sadness, disgust, and fear. Table 2 illustrates a significant imbalance in class distribution: neutral comprises the largest portion with 4,710 samples (47.2%), whereas disgust (271 samples) and fear (268 samples) are the least represented. This long-tail distribution presents a fundamental challenge for model generalization, especially concerning minority emotion categories.

Table 2. Training set category distribution (emotion)

Category	Number of utterances
neutral	4,710
joy	1,743
surprise	1,205
anger	1,109
sadness	683
disgust	271
fear	268
Total	9,989

We evaluated multi-class performance using Accuracy, Precision, Recall, and F1-score. Accuracy indicates the proportion of correctly predicted samples. Precision is the ratio of true positives to predicted positives for each class. Recall measures the model's ability to identify actual positive instances. The F1-score, as the harmonic mean of Precision and Recall, offers a more reliable assessment in cases of class imbalance. In addition to overall Accuracy, we also report macro-averaged and micro-averaged metrics. Macro-averaging treats all classes equally, emphasizing performance on minority classes. In contrast, micro-averaging aggregates true positives, false positives, and false negatives across all samples to reflect overall classification performance.

3.2. Implementation details

TextCNN used max_len=50, embedding_dim=128, num_filters=128, kernel sizes (2, 3, 4, 5), dropout=0.5, batch_size=64, epochs=15, lr=1e-3. BiLSTM used max_len=50, embedding_dim=128, hidden_dim=256, num_layers=2, dropout=0.5, batch_size=64, epochs=15, lr=1e-3. BERT-base was fine-tuned using the bert-base-uncased model, with max_len=128, dropout=0.3, batch_size=16, epochs=5, lr=2e-5, and warmup_ratio=0.1. DeepSeek experiments used a classification-constrained prompt template; for zero-shot, a fixed-label-space prompt was used with temperature=0; For few-shot tasks, K examples were sampled from each emotion category (K=1,3,5,10), also using deterministic decoding (temperature=0), and the output was restricted to a single-label word.

DOI: 10.54254/2755-2721/2026.GU32927

3.3. Analysis

As demonstrated in Table 3, DeepSeek-ZeroShot attains the highest overall Accuracy (0.6573) and Macro-F1 (0.5268), outperforming both traditional supervised models and few-shot variants. Within the supervised baselines, BERT-base surpasses BiLSTM and TextCNN, highlighting the superior semantic modeling capabilities of pre-trained models.

In few-shot settings, DeepSeek consistently performs well across different shot numbers. Notably, ZeroShot outperforms few-shot models, achieving an almost 100% valid output rate by relying exclusively on pre-trained knowledge and prompt constraints. In contrast, few-shot examples may distract the model or conflict with its existing knowledge. The 5-shot scenario records the highest few-shot Macro-F1 score (0.5108) but has a low valid rate (0.4831), indicating inconsistent output stability. Further analysis of prompts is available in Section 4.

Table 3. Summary table of overall metrics for each model (including macro and micro)

Model	Acc	Macro-P	Macro-R	Macro-F1	Micro-P	Micro-R	Micro-F1	Weighted-F1
DeepSeek-ZeroShot	0.6573	0.5318	0.5334	0.5268	0.6573	0.6573	0.6573	0.6543
DeepSeek-FewShot-5	0.6257	0.5169	0.5277	0.5108	0.6257	0.6257	0.6257	0.6307
DeepSeek-FewShot-10	0.6209	0.4964	0.5292	0.5034	0.6209	0.6209	0.6209	0.6287
DeepSeek-FewShot-3	0.6103	0.4813	0.5310	0.4953	0.6103	0.6103	0.6103	0.6194
DeepSeek-FewShot-1	0.6052	0.4875	0.5273	0.4950	0.6052	0.6052	0.6052	0.6093
BERT-base	0.6077	0.4421	0.4709	0.4511	0.6077	0.6077	0.6077	0.6139
BiLSTM	0.4839	0.2939	0.2839	0.2840	0.4839	0.4839	0.4839	0.4680
TextCNN	0.4552	0.2949	0.2842	0.2743	0.4552	0.4552	0.4552	0.4467

Figure 1 demonstrates that DeepSeek and BERT-base perform exceptionally well in high-frequency categories such as neutral and joy. In more challenging categories—anger, sadness, disgust, and fear—the DeepSeek series generally outperforms traditional CNN and LSTM models. Notably, DeepSeek-FewShot-5 achieves particularly high scores in sadness (0.5385) and fear (0.3750).

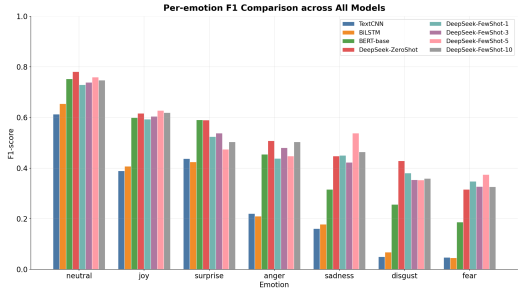

Figure 1. Bar chart showing the scores of each model across different emotion categories

DOI: 10.54254/2755-2721/2026.GU32927

4. Analysis of prompts

This section examines the impact of prompts on emotion recognition performance in large language models, divided into two parts: zero-shot and few-shot. In the zero-shot analysis, we compare three types of prompts: direct, definition, and contrast. In the few-shot analysis, we assess three prompt types: direct, definition, and JSON-instruction, while evaluating performance variations across four example sizes: K=1, 3, 5, and 10.

4.1. Zero-shot

• zs_v1_direct: You are an emotion classification assistant. You will be given one sentence from a TV show dialogue. Classify it into exactly one label: neutral, surprise, fear, sadness, joy, disgust, anger. Return ONLY one label word in lowercase.

• zs_v2_definition: Task: Emotion classification for dialogue text. Labels and rough meanings: neutral means no clear positive/negative emotion; surprise means unexpected reaction; fear means anxiety, worry, panic, or threat response; sadness means disappointment, grief, or low mood; joy means happiness, excitement, or relief; disgust means dislike, aversion, or repulsion; anger means irritation, blame, or rage. Pick the single best label for the sentence. Output exactly one word from the label set.

• zs_v3_contrast: You are a strict evaluator. Given one utterance, compare all seven emotion labels and choose the dominant one. Do not explain your reasoning. Allowed outputs only: neutral, surprise, fear, sadness, joy, disgust, anger. Output format: just one word.

Table 4. Performance comparison of three prompting methods in zero-shot learning

Prompt Style	Acc	Macro-P	Macro-R	Macro-F1	Weighted-F1
zs_v1_direct	0.6267	0.5067	0.5198	0.5006	0.6293
zs_v2_definition	0.6247	0.5029	0.4871	0.4861	0.6187
zs_v3_contrast	0.6221	0.4953	0.5071	0.4935	0.6214

The key information obtained from Table 4 is summarized as follows:

• The optimal zero-shot configuration is zs_v1_direct, yielding Acc=0.6267, Macro-F1=0.5006, and Weighted-F1=0.6293.

• Performance variations among the three zero-shot approaches are small, with an Accuracy variance of 3.63×10^{-6} and a Macro-F1 variance of 3.50×10^{-5}

• The zero-shot framework exhibits strong stability: prompt design affects prediction results but does not lead to severe performance fluctuations.

4.2. Few-shot

• fs_v1_direct: You are an emotion classification assistant. Classify into exactly one label: neutral, surprise, fear, sadness, joy, disgust, anger. Output ONLY one emotion word. Examples: Sentence: "{utt}" -> {label}.

• fs_v2_definition: Task: emotion classification. Label definitions: neutral = no obvious emotion; surprise = unexpected reaction; fear = anxiety/threat; sadness = loss/disappointment; joy = happiness/relief; disgust = repulsion; anger = irritation/rage. Choose exactly one label from: neutral, surprise, fear, sadness, joy, disgust, anger. Labeled examples: Input: "{utt}" | Label: {label}. Return only one label word.

DOI: 10.54254/2755-2721/2026.GU32927

• fs_v3_json_instruction: You are a strict classifier. First, infer the dominant emotion category, then output only the final label word. Allowed labels: neutral, surprise, fear, sadness, joy, disgust, anger. No explanation. Few-shot demonstrations: {"sentence": "{utt}", "emotion": "{label}"}. Now classify the next sentence and return only the label word.

Table 5. Results and stability of four few-shot prompting methods (K-factor analysis)

K	Prompt Style	Acc	Macro-P	Macro-R	Macro-F1
1	fs_v1_direct	0.6052	0.4874	0.5273	0.4949
	fs_v2_definition	0.6088	0.4726	0.4707	0.4681
	fs_v3_json_instruction	0.6149	0.4838	0.5068	0.4879
3	fs_v1_direct	0.6121	0.5026	0.5127	0.4957
	fs_v2_definition	0.6157	0.4954	0.4869	0.4872
	fs_v3_json_instruction	0.6042	0.4889	0.5018	0.4826
5	fs_v1_direct	0.6256	0.5169	0.5277	0.5108
	fs_v2_definition	0.6168	0.5019	0.4961	0.4928
	fs_v3_json_instruction	0.6045	0.4858	0.5258	0.4910
10	fs_v1_direct	0.6209	0.4964	0.5292	0.5034
	fs_v2_definition	0.6145	0.4820	0.4899	0.4819
	fs_v3_json_instruction	0.6141	0.4889	0.4925	0.4781

As can be seen from Table 5, hierarchical experiments conducted with K=1, 3, 5, and 10 demonstrate that few-shot performance exhibits a non-monotonic relationship with the number of examples, indicating that performance does not consistently improve as K increases. Stability varies significantly across different K values. Notably, K=5 achieves a higher performance ceiling but encounters considerable fluctuations in availability. Conversely, K=3 delivers the most stable Macro-F1 performance, whereas K=10 offers balanced performance when availability is high.

Few-shot approaches can surpass zero-shot methods in performance under specific conditions, such as when K=5. However, few-shot methods often exhibit considerable variability in output rate and stability. Conversely, zero-shot methods provide more consistent availability and a stronger baseline performance in this experiment.

5. Difficult samples analysis

5.1. Strict hard-sample definition and global statistics

A hard sample is defined as an utterance misclassified by all models at once. This criterion, spanning different architectures and paradigms, identifies instances that defy classification irrespective of model type, thus serving as a reliable indicator of intrinsic task difficulty.

Figure 2 depicts a clear gradient of difficulty within the complete test set of 2,610 samples. Among these, 258 samples (9.9%) are correctly classified by all five models, while 466 samples (17.9%) are misclassified by all five models.

DOI: 10.54254/2755-2721/2026.GU32927

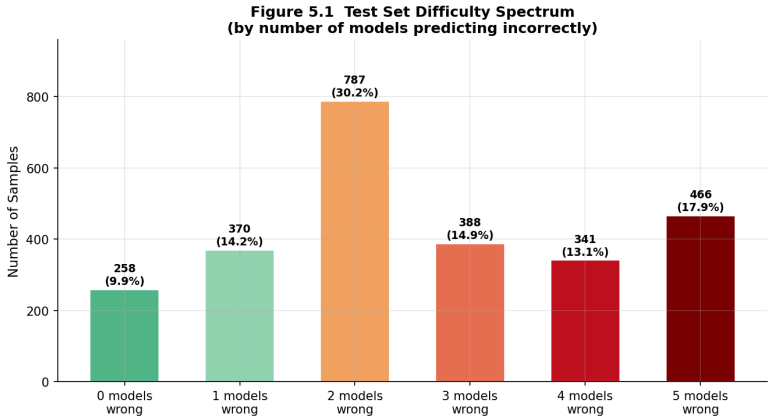

Figure 2. Difficulty spectral distribution of the test set (2,610 items)

As shown in Table 6, hard-sample rates vary significantly across emotion categories. In the original imbalanced training setup, the neutral category has the lowest hard rate at 5.2%. In contrast, minority emotions are disproportionately affected, with fear at 48.0%, sadness at 47.6%, and disgust at 41.2%, all exceeding 40%. The Pearson correlation between class frequency and hard rate is $r = -0.777$ ($p = 0.040$), indicating a strong negative relationship: the rarer the emotion, the more likely it is to become a hard sample. This long-tail hardness pattern aligns with overall performance results, as all models show significantly lower F1 scores for minority classes. Even the top-performing DeepSeek system struggles most with sadness, fear, and disgust.

Table 6. Emotion-wise hard sample rate (original imbalanced training)

Emotion	Hard / Total	Hard Rate
neutral	65 / 1256	5.2%
joy	59 / 402	14.7%
surprise	46 / 281	16.4%
anger	121 / 345	35.1%
sadness	99 / 208	47.6%
disgust	28 / 68	41.2%
fear	24 / 50	48.0%

5.2. Balanced training: decoupling class imbalance from intrinsic difficulty

The observed hardness pattern combines two distinct sources of difficulty: class imbalance, where models encounter too few minority examples during training, and intrinsic semantic complexity, where the emotion itself is inherently ambiguous. To disentangle these factors, we reconstructed the training set by downsampling each emotion class to 268 samples, aligning with the size of the smallest class (fear) in the original set. This adjustment results in a fully balanced training corpus of $7 \times 268 = 1{,}876$ samples. We then retrained TextCNN and BiLSTM on this balanced set and evaluated them on the unchanged original full test set of 2,610 samples. BERT-base and DeepSeek models remained unmodified; BERT-base because retraining would necessitate re-downloading pretrained weights not available offline, and DeepSeek because it is training-free and unaffected by training set

DOI: 10.54254/2755-2721/2026.GU32927

distribution. As shown in Table 7 and Figure 3, the two retrained models experienced significant accuracy drops (TextCNN: −0.180; BiLSTM: −0.241).

Table 7. Model accuracy: original vs balanced training

Model	Original Acc	Balanced-Train Acc	Drop	Note
TextCNN	0.4552	0.2755	−0.180	retrained
BiLSTM	0.4839	0.2433	−0.241	retrained
BERT-base	0.6077	0.6077	0.000	unchanged
DeepSeek-Zero	0.6571	0.6571	0.000	training-free
DeepSeek-Few	0.6103	0.6103	0.000	training-free

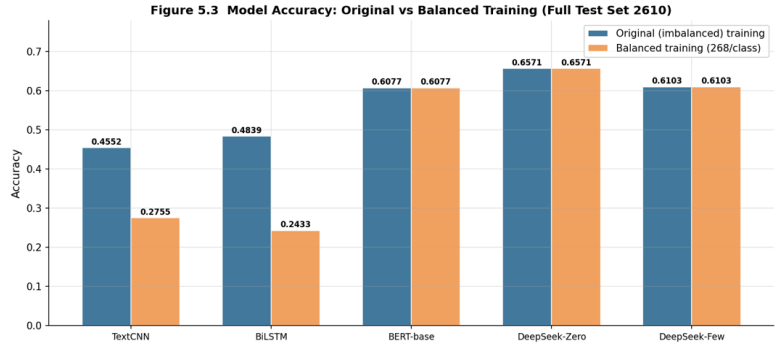

Figure 3. Comparison of the overall accuracy rates of each model on the complete test set under the conditions of original unbalanced training and balanced training

As shown in Table 8 and Figure 4, the distribution of hard-sample rates by emotion remains highly correlated across both conditions. The Pearson correlation coefficient between the original and balanced-training hard-rate profiles is r = 0.917 (p = 0.004). Minority emotions like sadness (38.9%), fear (36.0%), and anger (29.0%) continue to pose significant challenges, even after fully equalizing the training data. Conversely, the neutral category, now significantly under-resourced in training, sees its hard rate rise from 5.2% to 10.9%. This shift clearly illustrates the low-resource effect in isolation: reducing training data for the neutral category makes it more difficult for CNN/BiLSTM models, confirming that the quantity of training data affects performance. However, this does not address the existing bias against minority emotions.

Table 8. Emotion-wise hard rate: original vs balanced training

Emotion	Orig Hard / Total	Orig Rate	Bal Hard / Total	Bal Rate
neutral	65/1256	5.2%	137/1256	10.9%
joy	59/402	14.7%	72/402	17.9%
surprise	46/281	16.4%	43/281	15.3%
anger	121/345	35.1%	100/345	29.0%
sadness	99/208	47.6%	81/208	38.9%
disgust	28/68	41.2%	15/68	22.1%
fear	24/50	48.0%	18/50	36.0%

DOI: 10.54254/2755-2721/2026.GU32927

These results distinctly differentiate between two issues: (1) the low-resource effect, where overall accuracy declines as training data decreases; and (2) intrinsic emotion complexity, where minority emotions consistently pose challenges regardless of training balance.

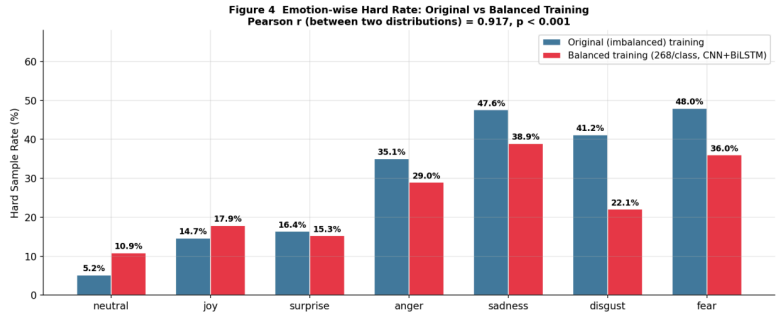

Figure 4. Comparison of the rate of difficult samples for each emotion category under the original unbalanced training and balanced training conditions

5.3. Linguistic signals in hard samples

Despite efforts to balance the training set, a hardness bias persists, leading us to explore whether surface-level linguistic features can distinguish hard samples from non-hard ones. As shown in Table 9, hard samples are, on average, slightly longer (9.28 words compared to 8.51). However, punctuation-based indicators do not exhibit a consistent pattern: hard samples contain fewer question marks (24.1% vs. 30.6%), ellipses (2.8% vs. 5.4%), and stutter expressions (1.9% vs. 2.9%) than non-hard samples. This absence of a reliable surface-level signal indicates that simple lexical features cannot effectively identify hard samples. Instead, the challenge lies in deeper semantic and pragmatic factors, such as implicit emotional cues, sarcasm, irony, and cross-utterance contextual dependencies.

Table 9. Linguistic features: hard vs non-hard samples

Feature	Hard (n=442)	Non-hard (n=2168)
Avg word count	9.28	8.51
Avg char count	47.94	43.05
Has ! (%)	38.7%	29.1%
Has ? (%)	26.7%	26.2%
Has ... (%)	2.5%	2.9%
Has stutter (%)	5.4%	4.2%

6. Conclusion

This study investigates the MELD (Multimodal Emotion Lines Dataset) emotion recognition task, focusing on verifying and enhancing the effectiveness of prompt-based large language models (LLMs) in handling long-tailed emotion categories. The experimental results show that the DeepSeek series outperforms traditional models like CNN and BiLSTM in overall performance, demonstrating superior semantic generalization capabilities. A detailed analysis of challenging samples reveals that long-tailed categories, such as fear, sadness, and anger, are primary sources of model failure, with a notable negative correlation between category frequency and difficulty rate. In

DOI: 10.54254/2755-2721/2026.GU32927

the undersampling balance experiment, the distribution structure of difficult samples remains highly consistent with the original data, indicating that the challenges of long-tailed emotions stem from both data imbalance and inherent semantic complexities.

References

[1] Voultsiou, E., & Moussiades, L. (2026). A Systematic Review of Large Language Models in Mental Health: Opportunities, Challenges, and Future Directions. Electronics, 15(3), 524.
[2] Poria, S., Majumder, N., at al. (2019). Emotion recognition in conversation: Research challenges, datasets, and recent advances. IEEE access, 7, 100943-100953.
[3] Hu, D., Wei, L., & Huai, X. (2021, August). Dialoguecrn: Contextual reasoning networks for emotion recognition in conversations. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers) (pp. 7042-7052).
[4] Shi, J., Wei, T., & Li, Y. (2024). Residual diverse ensemble for long-tailed multi-label text classification. Science China Information Sciences, 67(11), 212102.
[5] Bo-Hao, S., Upadhyay, S. G., & Chi-Chun, L. (2025, April). Toward zero-shot speech emotion recognition using llms in the absence of target data. In ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 1-5). IEEE.
[6] Zhang, Y., Wang, M., et al. (2025). Dialoguellm: Context and emotion knowledge-tuned large language models for emotion recognition in conversations. Neural Networks, 107901.
[7] Poria, S., Hazarika, D., et al. (2019, July). Meld: A multimodal multi-party dataset for emotion recognition in conversations. In Proceedings of the 57th annual meeting of the association for computational linguistics (pp. 527-536).
[8] Gong, L., & Ji, R. (2018). What does a TextCNN learn?. arXiv preprint arXiv: 1801.06287.
[9] Siami-Namini, S., Tavakoli, N., & Namin, A. S. (2019, December). The performance of LSTM and BiLSTM in forecasting time series. In 2019 IEEE International conference on big data (Big Data) (pp. 3285-3292). IEEE.
[10] Khadhraoui, M., Bellaaj, H., et al. (2022). Survey of BERT-base models for scientific text classification: COVID-19 case study. Applied Sciences, 12(6), 2891.
[11] Liu, A., Feng, B., et al. (2024). Deepseek-v3 technical report. arXiv preprint arXiv: 2412.19437.