United We Fine-Tune: Structurally Complementary Datasets for Hope Speech Detection

Priyadharshini Krishnaraj[1], Tulio Ferreira Leite da Silva[1,2], Gonzalo Freijedo Aduna[3], Samuel Chen[2], Farah Benamara[4,5], Alda Mari[3]
[1] CNRS@CREATE LTD, Singapore
[2] University of Sao Paulo, Brazil
[3] Institut Jean Nicod CNRS/ENS/EHESS/PSL University
[4] IRIT, Université de Toulouse, CNRS, Toulouse INP, Toulouse, France
[5] IPAL, CNRS-NUS-A*STAR, Singapore

Abstract

We propose a fine-tuning strategy for English Multiclass Hope Speech Detection using Mistral, leveraging two complementary datasets: PolyHope and CDB, a new unified framework for hope speech detection. While the former provides nuanced hope-related categories such as GENERALIZED, REALISTIC, and UNREALISTIC HOPE, the later introduces linguistically grounded dimensions including COUNTERFACTUAL, DESIRE, and BELIEF. By fine-tuning Mistral on both datasets, we enable the model to capture deeper semantic representations of hope. In addition to fine-tuning, we developed advanced prompting strategies which provide interpretable, zero-shot alternatives and further inform annotation and classification designs. Our approach achieved third place in the multiclass (Macro F1=71.77) and sixth in the binary (Macro F1=85.35) settings.

1 Introduction

Hope speech detection has recently evolved into a specialized area of classification within NLP, aimed at distinguishing constructive and future-oriented statements from neutral or negative content. While several datasets have been proposed to support this task (Goldberg et al., 2009; Palakodety et al., 2020; Chakravarthi, 2020; Balouchzahi et al., 2023a,b), their annotation schemas vary widely—ranging from affective taxonomies to structurally grounded categories—making generalization across label sets a persistent challenge.

This paper investigates whether structurally divergent but semantically related taxonomies can be combined to improve model performance on multiclass hope speech detection. We focus on two English-language datasets: *PolyHope* (Balouchzahi et al., 2023b), which classifies hope expressions into affective categories (GENERALIZED, REALISTIC, and UNREALISTIC HOPE), and *CDB* (Ferreira Leite da Silva et al., 2025), which bases its classification on the semantic notion of *modality* (Kratzer, 1991; Portner, 2009) in the broad sense, as encompassing propositional attitudes and speech acts (Giannakidou and Mari, 2021, 2026), and thus encoding the propositional structure of hope-related speech (COUNTERFACTUAL, DESIRE, BELIEF). Despite having disjoint label sets, both datasets target overlapping semantic phenomena. We treat them as complementary sources of supervision and fine-tune a Mistral-7B model on the merged corpus using a parameter-efficient strategy.

Our methodology is informed by recent findings in multi-task and cross-taxonomy learning. Prior work shows that combining tasks with high structural complementarity can produce synergistic gains in generalization, a phenomenon referred to as the "cocktail effect" (Brief et al., 2024). For example, Lai et al. (2024) proposed Multi-Task Implicit Sentiment Analysis (MT-ISA) which leverages auxiliary sentiment tasks to enhance main-task performance, while Ivison et al. (2023) Data-Efficient Fine-Tuning (DEFT) which shows that structural similarity between tasks is often a more reliable indicator of transfer effectiveness than surface-level alignment or data volume. Building on these insights, we treat PolyHope and CDB not as competing annotation schemes, but as complementary lenses on the semantic domain of hope. Instead of aligning or mapping between taxonomies, we fine-tune a generative model on both datasets simultaneously, alternating prompt formats within a single training pipeline. This setup enables the model to internalize both affective and propositional representations of hope.

In addition to supervised fine-tuning, and drawing on recent surveys of prompting strategies (Schulhoff et al., 2025; Fagbohun et al., 2023; White et al., 2023), we propose three zero-shot prompting methods tailored to the PolyHope tax-

Proceedings of the First Workshop on Comparative Performance Evaluation: From Rules to Language Models associated with RANLP 2025, pages 48–58, Varna, Bulgaria, Sep 11, 2025.
https://doi.org/10.26615/978-954-452-102-8-006

onomy for hope speech detection: *Confidence-Structured Output Prompting*, *Multiple Reasoning Path Prompting*, and *Decision Tree Prompting*. These strategies—developed specifically for this study—were designed to enhance model interpretability and decision consistency, particularly in multiclass scenarios where category boundaries are conceptually nuanced. By integrating structural supervision with reasoning-aware prompting, we evaluate both supervised and prompt-based approaches within a unified framework.

Our submission to the RANLP-2025 shared task ranked **3rd** in the English multiclass classification track and **6th** in the binary. Results suggest that cross-taxonomy fine-tuning without explicit task weighting, can yield competitive generalization performance. The codebase and prompt templates will be released publicly to support further research in structure-aware hope speech classification[1]. Our main contributions are:

1. A cross-taxonomy LLMs fine-tuning strategy leveraging structurally complementary datasets.

2. Proposal and evaluation of reasoning-aware zero-shot prompting strategies tailored to hope speech multiclass classification.

3. A qualitative error analysis of the best model, highlighting systematic confusion patterns in hope classification.

We begin in Section 2 with a review of previous work in the field. Section 3 details the datasets employed in the shared task. Our methodology is described in Section 4, and the corresponding results are reported in Section 5.

2 Related Work

2.1 Hope Speech Datasets

Prior research on hope speech has explored a range of perspectives, from peace-oriented discourse (Palakodety et al., 2020) to multilingual detection for promoting inclusion (Chakravarthi, 2020). Other works have examined expressions of regret and past-oriented hope (Balouchzahi et al., 2023a), and the expression of wish in products reviews and political discussions (Goldberg et al., 2009).

[1]https://github.com/Priyaaa-hub/Shared-task-prompts.git

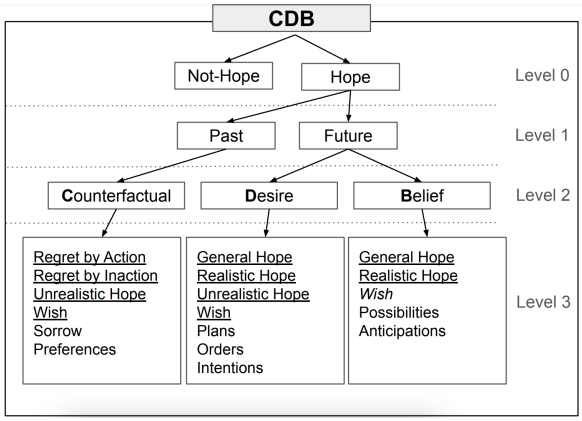

Figure 1: The COUNTERFACTUAL-DESIRE-BELIEF (CDB) model (Ferreira Leite da Silva et al., 2025).

The *PolyHope* dataset (Balouchzahi et al., 2023b) used in the shared task is annotated with four categories of future-oriented hope-related expressions: (a) NOT-HOPE, indicating the absence of hope; (b) GENERALIZED HOPE, referring to vague or non-specific statements of hope; (c) UNREALISTIC HOPE, denoting overly optimistic or implausible expectations; and (d) REALISTIC HOPE, capturing grounded and plausible expressions of hope. In contrast, the ReDDit dataset (Balouchzahi et al., 2023a) focuses exclusively on *past-oriented* hope, specifically targeting expressions of retrospective longing or regret.

Building on these prior works, the *CDB model* (Ferreira Leite da Silva et al., 2025) introduces a more fine-grained and linguistically grounded classification system. Unlike PolyHope and ReDDit, which each target a single temporal dimension of hope, the CDB model incorporates both: one class for past-oriented hope, two distinct classes for future-oriented hope, and one for the not-hope instances. This classification is grounded in the degree of speaker commitment implied by each expression, allowing for a more nuanced framework for annotation and classification. The model defines four core classes that subsume previous classification schemes, as illustrated in Figure 1: (a) NOT-HOPE: indicating the absence of any hope-related expression. (b) COUNTERFACTUAL: which captures expressions of regret and represents past-oriented hope. (c) DESIRE: encompassing future-oriented expressions of mere desire or wishful thinking that lack strong speaker commitment. (d) BELIEF: which also encodes future-oriented hope, but in this case grounded in epistemic or deontic considerations.

This taxonomy enables a more linguistically informed and temporally aware analysis of hope speech across discourse contexts. Two annotators achieved a Cohen's kappa of 74.88 for the binary classification task, and 70.46 for the multiclass classification, indicating substantial agreement given the subjectivity of the task.

2.2 Hope Speech Automatic Detection

Several shared tasks have also advanced the study of hope speech in multilingual and multicultural settings. At LT-EDI 2022 (Chakravarthi et al., 2022), LT-EDI 2023 (Kumaresan et al., 2023) and IberLEF 2023 (Jiménez-Zafra et al., 2023), the task was framed as binary classification in a variety of languages. These tasks laid the foundation for more detailed distinctions explored in subsequent years.

The IberLEF 2024 shared task (García-Baena et al., 2024) introduced a two subtasks reflecting distinct dimensions of hope: (1) *Hope for Equality, Diversity, and Inclusion* to detect supportive speech toward vulnerable groups, (2) *Hope as Expectations* that requires multi-class classification of generalized, realistic, and unrealistic expressions of future-oriented hope. Approaches based on transformer models, as well as those leveraging prompting with large language models (LLMs), have both demonstrated competitive performance. For instance, the top-ranked system in Subtask (1) (Thuy and Thin, 2024) employed a zero-shot prompting strategy using ChatGPT-3.5, incorporating class definitions into the prompt in both English and Spanish. Their solution explored multiple prompting techniques—zero-shot, one-shot, three-shot, and chain-of-thought (CoT)—combined with six different information strategies, including role-defining, class explanations, and task-specific concepts. The best performance was achieved using a one-shot prompt and an information-rich strategy that defined both class meanings and model roles, yielding a Macro F1-score of 0.7161 on the out-of-domain Spanish test set.

In Subtask (2), the winning team (Bui Hong et al., 2024) adopted a supervised approach, combining multilingual transformer models with rigorous data pre-processing and augmentation. Their method leveraged data combination across English and Spanish corpora and generated synthetic samples for minority classes using Gemini LLMs. Classification was performed via fine-tuned models such as XLM-R, mDeBERTa, and RoBERTuito, and predictions were aggregated using a max voting ensemble. This robust pipeline achieved the highest scores in the multiclass subtasks for both English (Macro F1 = 72.00) and Spanish (Macro F1 = 66.68), highlighting the effectiveness of multilingual augmentation and ensemble-based inference.

As we can see, the progression of hope speech detection methods—from traditional machine learning models to transformer architectures and, more recently, to prompt-based large language models—reflects a broader shift in NLP toward more flexible and powerful approaches, particularly for multilingual and cross-domain applications.

3 Datasets

We rely on two datasets, PolyHope and CDB. We first present them, then explain their complementarity.

3.1 PolyHope Dataset

The dataset consists of **8,256 tweets** collected in 2022, covering topics such as abortion rights, racial justice, religion, and politics. As illustrated in Table 2, the dataset exhibits moderate imbalance, with category NOT-HOPE comprising nearly half of the instances, while the remaining categories are notably less represented (GENERALIZED HOPE being more than twice as frequent as REALISTIC HOPE and nearly three times as frequent as UNREALISTIC HOPE). As the test set was not provided, we only report the statistics of the train and dev sets in the tables.

3.2 The CDB Dataset

The CDB dataset comprises **4,370 texts** in total, of which 3,092 were randomly selected and re-annotated from existing corpora (WISH (Goldberg et al., 2009), PolyHope (Balouchzahi et al., 2023b), and HopeEDI (Chakravarthi, 2020)), and 1,278 were newly collected from X (formerly Twitter) and Reddit (HopeDrone). As shown in Table 3, the dataset exhibits a slight class imbalance in the binary setting, with the HOPE category accounting for 57.28% of the total instances. In the multiclass setting, however, the distribution is more skewed: while NOT-HOPE remains the largest single class, the DESIRE category represents nearly one-third of the dataset, followed by BELIEF at just over one-fifth. The COUNTERFACTUAL category is notably underrepresented, comprising less than 5% of all texts. These proportions remain consistent across

Category	HopeDrone	PolyHope	WISH Corpus	HopeEDI	Total
COUNTERFACTUAL	15 (0.34%)	112 (2.56%)	62 (1.42%)	14 (0.32%)	**203 (4.65%)**
DESIRE	224 (5.13%)	682 (15.60%)	360 (8.24%)	113 (2.59%)	**1,379 (31.56%)**
BELIEF	390 (8.92%)	202 (4.62%)	226 (5.17%)	103 (2.36%)	**921 (21.08%)**
NOT-HOPE	649 (14.85%)	279 (6.39%)	569 (13.02%)	370 (8.47%)	**1,867 (42.72%)**
Total	**1,278 (29.24%)**	**1,275 (29.18%)**	**1,217 (27.85%)**	**600 (13.73%)**	**4,370 (100%)**

Table 1: The CDB dataset, following Ferreira Leite da Silva et al. (2025). The "Total" column aggregates the instance counts across all four datasets—HopeDrone, PolyHope, WISH Corpus, and HopeEDI—for each class in the CDB taxonomy. The bottom row summarizes the total number and relative size of each dataset.

Binary	Train	Dev	Total
NOT HOPE	2,245 (49.44%)	816 (49.45%)	3,061 (49.44%)
HOPE	2,296 (50.56%)	834 (50.55%)	3,130 (50.56%)
Multiclass	**Train**	**Dev**	**Total**
NOT HOPE	2,245 (49.44%)	816 (49.45%)	3,061 (49.44%)
GENERALIZED HOPE	1,284 (28.28%)	467 (28.30%)	1,751 (28.28%)
REALISTIC HOPE	540 (11.89%)	196 (11.88%)	736 (11.89%)
UNREALISTIC HOPE	472 (10.39%)	171 (10.36%)	643 (10.39%)

Table 2: Distribution of classes for binary and multiclass settings (PolyHope) by Split. The "Total" column presents the aggregate number of instances for each class, obtained by summing the respective values from the training and development splits.

the training and test splits.

Importantly, we verified that **1,020 texts** in the CDB dataset were originally drawn from the PolyHope corpus used in the shared task—**543** from the training set, **203** from the development set, and **274** from the test set. This overlap is explicitly reported to ensure transparency. During fine-tuning, the Mistral model was trained jointly on the PolyHope and CDB training sets. Crucially, the PolyHope test set remained unlabeled and was never used during training. Although some texts may have been seen with alternative annotations from the CDB taxonomy, their original PolyHope labels were hidden throughout. Rather than constituting test contamination, this setup enables robust cross-taxonomy learning and allows the model to internalize divergent labeling schemes over shared inputs.

Binary Label	Train	Test	Total
NOT-HOPE	1,599 (43.03%)	268 (40.98%)	1,867 (42.72%)
HOPE	2,117 (56.97%)	386 (59.02%)	2,503 (57.28%)
Multiclass Label	**Train**	**Test**	**Total**
NOT-HOPE	1,599 (43.03%)	268 (40.98%)	1,867 (42.72%)
DESIRE	1,149 (30.92%)	230 (35.17%)	1379 (31.56%)
BELIEF	801 (21.56%)	120 (18.35%)	921 (21.08%)
COUNTERFACTUAL	167 (4.49%)	36 (5.50%)	203 (4.65%)

Table 3: Distributions of classes for binary and multiclass settings (CDB model) by Split.

3.3 Cross-Taxonomy Comparison

Figures 2 and 3 illustrate the cross-taxonomy relationship between PolyHope and CDB, as reported in (Ferreira Leite da Silva et al., 2025). Each figure presents a correlation matrix based on randomly selected 1,022 PolyHope instances that were re-annotated using the CDB schema.

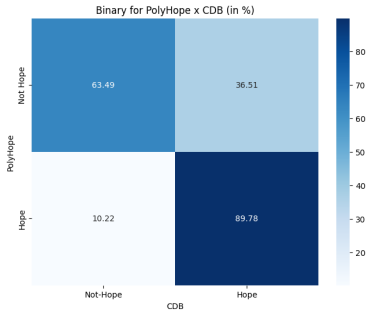

Figure 2: CDB vs. PolyHope binary annotations.

As previously stated, these datasets are not competing but complementary. Despite what Figure 2 may initially suggest—many instances labeled as NOT HOPE in PolyHope are reclassified as HOPE in CDB—the multiclass perspective reveals their compatibility. In Figure 3, we observe that the CDB category DESIRE serves as a good approximation for all three hope categories in PolyHope.

The fact that most PolyHope instances labeled as GENERALIZED, REALISTIC, or UNREALISTIC hope are mapped to the DESIRE category in CDB highlights a key difference between the taxonomies: CDB places greater emphasis on temporal and modal structure, while PolyHope focuses on plausibility and affective nuance.

This also reflects, among other factors, a structural divergence—PolyHope excludes past-oriented hope from its schema, whereas CDB explicitly encodes it through the COUNTERFACTUAL category. The divergence becomes even more apparent in the multiclass comparison (cf. Figure 3),

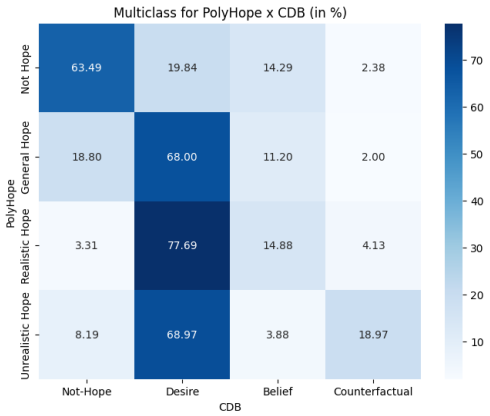

Figure 3: CDB vs. PolyHope multiclass annotations.

where there is no one-to-one mapping between categories across the two taxonomies.

We provide below some examples of PolyHope and CDB annotations. Instances (1) and (2) show cases where utterances labeled as NOT HOPE in PolyHope were annotated as BELIEF in CDB:

(1) "**Don't expect Chris Rock to talk** about 'the slap' when he performs Friday at the AVA at Casino del Sol."

(2) "Getting on your knees to pray **should stay off** the football field and stay in church settings."

Similarly, utterances (3) and (4) labeled as NOT HOPE in PolyHope, were classified as COUNTERFACTUAL in CDB, due to the presence of past-tense constructions:

(3) "Thank you for being brave and speaking up—your work is so beautiful! **Wish we had met** in NFT NYC."

(4) "I **wish it was** Speak Now, but the signs seem to point to 1989."

This structural mismatch has direct implications for training strategies. While the multiclass setup benefits from the complementarity of both taxonomies, the binary classification reveals overlapping boundaries that risk making the label space more concorrente than complementary. We therefore did not adopt the multi-task learning in the binary setting, as merging categories could introduce conflicting signals during training. Conversely, in the multiclass scenario, the alignment between affective subtypes and temporal-motivational roles supports more synergistic learning.

4 Methodology

4.1 Models

We designed models based on transformers and Large Language Models (LLMs). The reported scores are averaged over 3 runs on the PolyHope development set, as the test set has not been released. The hyper-parameters used for fine-tuning both the LLMs and transformer based models are available in the supplementary material associated to this submission.

4.1.1 Transformers

We use BERT (Bidirectional Encoder Representations from Transformers) as a baseline transformer model for hope speech classification. Known for its strong performance across a variety of NLP tasks, BERT was fine-tuned separately for both binary and multiclass classification.

4.1.2 Large Language Models

We make use of two models:

–**GPT-4**, the Generative Pretrained Transformer 4 developed by OpenAI, was also explored in our experiments. Although its performance on the development dataset was comparatively lower, it was still used to generate predictions under a zero-shot prompting strategy for the test set and included in the final submission.

–**Mistral FT**. It is the Mistral-7B-Instruct-v0.3, a 7-billion parameter open-weight LLM developed by Mistral AI, optimized for instruction-following tasks. We fine-tune Mistral using a parameter-efficient strategy based on QLoRA, a lightweight variant of Low-Rank Adaptation (LoRA) that enables scalable tuning of large language models with limited compute. This setup allows us to fine-tune Mistral-7B on structurally complementary datasets (PolyHope and CDB) while preserving efficiency and reducing overfitting. Unlike prior work comparing multiple adaptation methods, our goal is not to benchmark fine-tuning techniques, but rather to validate the value of combining taxonomically divergent supervision sources. During training, each instance included three fields: a dataset-specific prompt, the corresponding input text, and the gold classification label.

The set of prompting strategies used to fine-tune Mistral are as follows:

1. **Zero-Shot Prompting:** The prompt includes a description of the task, definitions for each label, and the expected output format, without providing any training examples.

2. **Few-Shot Prompting:** In addition to the task description and label definitions, this prompt includes two randomly chosen examples per class. There are 4 examples for binary classification (2 - HOPE and NOT HOPE), and 8 examples for multiclass classification (2 per class resp.). Each example is paraphrased before including it in the prompts to prevent overfitting and ensure the model learns from the examples and does not rely on the specific phrasing. We use a tool called quillbot[2] for paraphrasing the chosen examples after which these examples are included in the prompt for classification.

3. **Decision Tree Prompting:** This strategy implements a logical flowchart in prompt form, requiring the model to follow a step-by-step decision process. We pose several sequential queries to the model, inquiring about the presence of hope, any specific goal mentioned, and the feasibility of the goal by asking whether it involves a particular result and whether they are attainable or not. This prompt was elaborated based on the concept of *Decision Tree Reasoning for Prompting*, proposed as a structured decomposition strategy within the Tree-of-Thought framework by Yao et al. (2023), and categorized under Logical and Sequential Processing in (Fagbohun et al., 2023). It expands on classic Chain-of-Thought prompting by introducing a branching logic format for inference.

4. **Confidence-Structured Output Prompting:** This technique generates both a classification label and a graded confidence score for each category, based on observable features of the input. The prompt structure guides the model through identifying hope-related language, assessing the specificity of the desired outcome, and evaluating its feasibility. These components are followed by a confidence estimate for each label and the final classification. The method is inspired by uncertainty-aware reasoning and structured prediction techniques in LLMs. This prompt was elaborated based on the concept of *Uncertainty-Routed Chain-of-Thought (CoT)* prompting, proposed in (Schulhoff et al., 2025), and classified under Thought Generation and Self-Criticism strategies. It leverages confidence estimation techniques to refine final predictions.

5. **Multiple Reasoning Path Prompting:** This method encourages LLMs to perform a multi-perspective analysis of the text by decomposing the reasoning process into three steps: linguistic cues, goal assessment, and contextual framing. Each perspective contributes to the final classification. Such an approach is related to multi-perspective CoT prompting. This prompt was elaborated based on the concept of *Multi-View or Multi-Faceted Reasoning*, explicitly discussed in (Schulhoff et al., 2025) under Contrastive CoT and Meta-CoT, and structurally aligned with the Cognitive Verifier pattern in (White et al., 2023), which decomposes reasoning into modular sub-analyses to enhance robustness and explanatory power.

The prompts used for LLMs and the hyperparameters used for fine-tuning both the LLMs and transformer-based models, are provided in the Appendix.

4.2 Submitted Systems

A total of 9 systems have been submitted for the shared task:

1. **Binary Classification:** GPT-4, Mistral FT (Zero-shot$_P$, Few-shot$_P$, Confidence Score$_P$, Multiple Reasoning$_P$).

2. **Multiclass Classification:** BERT$_P$, GPT-4, Mistral FT (Zero-shot$_P$, Zero-shot$_{P+CDB}$)

Where $Prompt_d$ indicates **Mistral FT** model fine-tuned with one of the previous 5 $Prompt$ on the dataset $d \in \{P,\ P + CDB\}$. **BERT**$_P$ has only been fine-tuned on PolyHope, **GPT4**, being prompted in a zero-shot fashion. These configurations were selected based on their superior performance on the development set (see next Section).

5 Results

Table 4 presents the results for the binary classification, best scores are in bold font. We observe that most Mistral FT variants achieve relatively stable

[2]https://quillbot.com/paraphrasing-tool

	Development Set				Test Set			
Model	**P**	**R**	**F1**	**Acc**	**P**	**R**	**F1**	**Acc**
GPT-4	77.78	76.83	76.55	76.73	53.00	52.02	51.87	77.00
Mistral FT Variants								
Zero-shot$_P$	**84.19**	**84.17**	**84.18**	84.18	85.06	84.85	84.97	84.98
Few-shot$_P$	83.08	83.01	83.01	83.03	**85.44**	**85.34**	**85.35**	85.37
Confidence Score$_P$	83.66	83.62	83.63	83.64	85.30	85.25	85.26	85.27
Multiple Reasoning$_P$	84.03	83.98	83.99	84.00	85.01	84.91	84.92	84.93

Table 4: Performances of binary classification in development vs. test sets in terms of macro P, R and F1 scores.

	Development Set				Test Set			
Model	**P**	**R**	**F1**	**Acc**	**P**	**R**	**F1**	**Acc**
BERT$_P$	**71.01**	66.58	68.24	74.30	**73.31**	69.28	70.78	77.14
GPT-4	53.55	47.62	42.55	56.67	55.80	49.54	44.87	57.86
Mistral FT Variants								
Zero-shot$_P$	68.12	68.49	68.07	74.00	68.66	69.97	69.12	75.06
Zero-shot$_{P+CDB}$	70.40	**69.98**	**70.12**	75.58	71.19	**71.09**	**71.11**	76.80

Table 5: Multiclass classification results in the development vs. test sets in terms of macro P, R, and F1 scores.

	Binary				Multiclass		
Label	**P**	**R**	**F1**	**Label**	**P**	**R**	**F1**
BERT$_P$							
HOPE	80.00	**86.81**	83.27	GEN. HOPE	64.86	76.66	70.26
NOT HOPE	**85.23**	77.82	81.36	REAL. HOPE	69.33	53.06	60.12
				UNREAL. HOPE	66.91	54.39	60.0
				NOT HOPE	82.94	82.23	82.58
GPT-4$_P$							
HOPE	83.48	67.27	74.50	GEN. HOPE	76.81	11.35	19.78
NOT HOPE	72.09	86.40	78.60	REAL. HOPE	31.27	64.29	42.07
				UNREAL. HOPE	39.02	28.07	32.65
				NOT HOPE	67.11	**86.76**	75.68
Mistral FT							
Zero-shot$_P$				Zero-shot$_P$			
HOPE	83.75	85.25	**84.49**	GEN. HOPE	72.25	61.88	66.67
NOT HOPE	84.64	83.09	83.86	REAL. HOPE	53.62	64.29	58.47
				UNREAL. HOPE	64.24	61.99	63.10
				NOT HOPE	82.35	85.78	84.03
Few-shot$_P$				Zero-shot$_{P+CDB}$			
HOPE	81.99	85.13	83.53	GEN. HOPE	70.45	66.38	68.36
NOT HOPE	84.18	80.88	82.50	REAL. HOPE	59.62	64.80	62.10
				UNREAL. HOPE	67.72	62.57	65.05
				NOT HOPE	**83.79**	86.15	**84.95**
Confidence-score$_P$				Confidence-score$_P$			
HOPE	82.94	85.13	84.02	GEN. HOPE	52.38	30.62	38.65
NOT HOPE	84.38	82.11	83.23	REAL. HOPE	19.45	54.59	28.69
				UNREAL. HOPE	24.64	39.77	30.43
				NOT HOPE	76.89	19.98	31.71
Multiple Reasoning$_P$				Multiple Reasoning$_P$			
HOPE	83.29	85.49	84.38	Gen. Hope	37.87	63.81	47.53
NOT HOPE	84.76	82.48	83.60	REAL. HOPE	24.41	31.63	27.56
				UNREAL. HOPE	33.33	1.75	3.33
				NOT HOPE	70.28	51.59	59.51

Table 6: Performances per class in the development set in both binary and multiclass settings.

performance across all prompting strategies, with only minor variations, outperforming GPT4. The highest performance is observed using Few-shot prompting trained on the PolyHope dataset.

Table 5 shows the results for the multiclass classification task, where Mistral has been fine-tuned either on PolyHope or PolyHope+CDB. The $Prompt_{P+CDB}$ setups consistently outperform $Prompt_P$ by approximately 2%, indicating that incorporating the additional CDB data enhances the model's capacity for fine-grained classification.

We finally provide in Table 6 per-class performance results for both the binary and multiclass classification tasks on the development set, as the testset has not been released. It offers a comprehensive overview of all evaluated strategies.

Overall, our system achieved a **6th** place ranking in the Binary Classification task and a **3rd** place in the Multiclass Classification task on the English dataset. These results confirm the effectiveness of our tailored prompting and fine-tuning strategies, particularly for multiclass scenarios.

6 Error Analysis

Here we analyze the most frequent misclassifications—**377 instances**—as predicted by our best model Mistral FT using Zero-Shot$_{P+CDB}$ on the development set. These errors can be grouped into two main categories:

– Generalized Hope (Gold) vs. Not Hope (Prediction)

1. GENERALIZED HOPE (Gold) → NOT HOPE (Prediction) = 90 instances, as in "This is awful. **Please pray for these poor people**. No one should have died that way, but will this administration do anything? Nope, they have a clown tribunal to attend to, and a constitution to ignore".

2. NOT HOPE (Gold) → GENERALIZED HOPE (Prediction) = 81 instances, as in "All task done [...] thank you and **wish** me luck".

These confusions suggest that the model struggles with vague or subtle expressions of hope (highlighted in bold in the examples). In (a), for instance, short hopeful spans are embedded in longer neutral or non-hopeful content, which may dominate the model's representation. Conversely, in (b), lexical cues like *hope* or *wish* lead to overgeneralization.

– REALISTIC HOPE (Gold) vs. GENERALIZED HOPE (Prediction)

1. GENERALIZED HOPE (Gold) → REALISTIC HOPE (Prediction) = 53 instances, like in "I just hope my 3 years of Spanish lessons and streak are still there".

2. REALISTIC HOPE (Gold) → GENERALIZED HOPE (Prediction) = 34 instances, e.g., "Well I hope we're singing Turn Out the Lights the Party's Over, when this hearing is done."

In these last two cases, the model may struggle to distinguish between grounded, outcome-oriented hopes and more diffuse or emotive expressions, suggesting limited sensitivity to contextual or pragmatic features that signal speaker intent.

7 Conclusion

We proposed novel prompting strategies that achieved top-tier performance in the shared task. In addition, our fine-tuning methodology demonstrates the feasibility of combining structurally distinct datasets—each with its own label taxonomy—for multiclass classification using large language models and transformer architectures. This cross-taxonomy approach enables richer supervision and improved generalization.

In the future, we plan to consider the idea of unifying hope speech taxonomies via latent label modeling or joint annotation projection. This could offer a principled way to formalize cross-taxonomy alignment.

Acknowledgment

This work has been supported by DesCartes: the National Research Foundation, Prime Minister's Office, Singapore under its Campus for Research Excellence and Technological Enterprise (CREATE) program. Alda Mari gratefully thanks ANR-17-EURE-0017 FrontCog. Tulio Ferreira Leite da Silva is thankful to Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) 23/1010-1 for their support.

Limitations

The prompting strategies we explored such as decision tree prompting, confidence-structured output

prompting, and multiple reasoning path prompting-were selected precisely for their cross-domain generalization, as highlighted in recent work (e.g., (Schulhoff et al., 2025; White et al., 2023)). However, we acknowledge that task-specific adaptation is often necessary to fully leverage their benefits.

For instance, the structural logic of decision tree prompting can be transferred across tasks, but the branching criteria must be adapted to the domain's ontology. Similarly, while the general idea behind confidence-structured output prompting is domain-agnostic (e.g., eliciting outputs with associated self-assesses certainty), the format and calibration of confidence levels might require tuning. In multiple reasoning path prompting, the principle of diverse inference paths remain reusable, but the types of reasoning paths must reflect the target task's cognitive demands. In short, while the strategies are reusable at a conceptual level, they often require lightweight, task-aware instantiations to reach optimal performance.

References

Fazlourrahman Balouchzahi, Sabur Butt, Grigori Sidorov, and Alexander Gelbukh. 2023a. ReDDIT: Regret detection and domain identification from text. *Expert Systems with Applications*, 225:120099.

Fazlourrahman Balouchzahi, Grigori Sidorov, and Alexander Gelbukh. 2023b. PolyHope: Two-level hope speech detection from tweets. *Expert Systems with Applications*, 225:120078.

Meni Brief, Oded Ovadia, Gil Shenderovitz, Noga Ben Yoash, Rachel Lemberg, and Eitam Sheetrit. 2024. Mixing It Up: The Cocktail Effect of Multi-Task Fine-Tuning on LLM Performance – A Case Study in Finance. ArXiv:2410.01109 [cs].

Son Bui Hong, Quan Le Minh, and Van Thin Dang. 2024. ABCD team at HOPE 2024: Hope detection with BERTology models and data augmentation. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2024), co-located with the 40th Conference of the Spanish Society for Natural Language Processing (SEPLN 2024)*, CEUR Workshop Proceedings. CEUR-WS.org.

Bharathi Raja Chakravarthi. 2020. HopeEDI: A Multilingual Hope Speech Detection Dataset for Equality, Diversity, and Inclusion. In *Proceedings of the Third Workshop on Computational Modeling of People's Opinions, Personality, and Emotion's in Social Media*, pages 41–53, Barcelona, Spain (Online). Association for Computational Linguistics.

Bharathi Raja Chakravarthi, Vigneshwaran Muralidaran, Ruba Priyadharshini, Subalalitha Cn, John McCrae, Miguel Ángel García, Salud María Jiménez-Zafra, Rafael Valencia-García, Prasanna Kumaresan, Rahul Ponnusamy, Daniel García-Baena, and José García-Díaz. 2022. Overview of the Shared Task on Hope Speech Detection for Equality, Diversity, and Inclusion. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 378–388, Dublin, Ireland. Association for Computational Linguistics.

Oluwole Fagbohun, Rachel Harrison, and Anton Dereventsov. 2023. An Empirical Categorization of Prompting Techniques for Large Language Models: A Practitioner's Guide. *Journal of Artificial Intelligence, Machine Learning and Data Science*, 1:1–11.

Daniel García-Baena, Fazlourrahman Balouchzahi, Sabur Butt, Miguel Ángel García-Cumbreras, Atnafu Lambebo Tonja, José Antonio García-Díaz, Selen Bozkurt, Bharathi Raja Chakravarthi, Hector G. Ceballos, Rafael Valencia-García, Grigori Sidorov, L. Alfonso Ureña-López, Alexander Gelbukh, and Salud María Jiménez-Zafra. 2024. Overview of HOPE at IberLEF 2024: Approaching Hope Speech Detection in Social Media from Two Perspectives, for Equality, Diversity and Inclusion and as Expectations. *Procesamiento del Lenguaje Natural*, 73(0):407–419. Number: 0.

Anastasia Giannakidou and Alda Mari. 2021. *Truth and veridicality in grammar and thought: Mood, modality, and propositional attitudes*. University of Chicago Press.

Anastasia Giannakidou and Alda Mari. 2026. *Modal Sentences*. Cambridge University Press.

Andrew B. Goldberg, Nathanael Fillmore, David Andrzejewski, Zhiting Xu, Bryan Gibson, and Xiaojin Zhu. 2009. May All Your Wishes Come True: A Study of Wishes and How to Recognize Them. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 263–271, Boulder, Colorado. Association for Computational Linguistics.

Hamish Ivison, Noah A. Smith, Hannaneh Hajishirzi, and Pradeep Dasigi. 2023. Data-efficient finetuning using cross-task nearest neighbors. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 9036–9061, Toronto, Canada. Association for Computational Linguistics.

Salud María Jiménez-Zafra, Miguel Ángel Garcia-Cumbreras, Daniel García-Baena, José Antonio Garcia-Díaz, Bharathi Raja Chakravarthi, Rafael Valencia-García, and Luis Alfonso Ureña-López. 2023. Overview of HOPE at IberLEF 2023: Multilingual Hope Speech Detection. *Procesamiento del Lenguaje Natural*, pages 371–381.

Angelika Kratzer. 1991. Modality. *In von Stechow, A. and Wunderlich, D., editors, Semantics: An international handbook of contemporary research.*

Prasanna Kumar Kumaresan, Bharathi Raja Chakravarthi, Subalalitha Cn, Miguel Ángel García-Cumbreras, Salud María Jiménez Zafra, José Antonio García-Díaz, Rafael Valencia-García,

Momchil Hardalov, Ivan Koychev, Preslav Nakov, Daniel García-Baena, and Kishore Kumar Ponnusamy. 2023. Overview of the Shared Task on Hope Speech Detection for Equality, Diversity, and Inclusion. In *Proceedings of the Third Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 47–53, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Wenna Lai, Haoran Xie, Guandong Xu, and Qing Li. 2024. Multi-task learning with llms for implicit sentiment analysis: Data-level and task-level automatic weight learning.

Shriphani Palakodety, Ashiqur R. KhudaBukhsh, and Jaime G. Carbonell. 2020. Hope Speech Detection: A Computational Analysis of the Voice of Peace. ArXiv:1909.12940 [cs].

Paul Portner. 2009. *Modality*. OUP Oxford.

Sander Schulhoff, Michael Ilie, Nishant Balepur, Konstantine Kahadze, Amanda Liu, Chenglei Si, Yinheng Li, Aayush Gupta, HyoJung Han, Sevien Schulhoff, Pranav Sandeep Dulepet, Saurav Vidyadhara, Dayeon Ki, Sweta Agrawal, Chau Pham, Gerson Kroiz, Feileen Li, Hudson Tao, Ashay Srivastava, Hevander Da Costa, Saloni Gupta, Megan L. Rogers, Inna Goncearenco, Giuseppe Sarli, Igor Galynker, Denis Peskoff, Marine Carpuat, Jules White, Shyamal Anadkat, Alexander Hoyle, and Philip Resnik. 2025. The prompt report: A systematic survey of prompt engineering techniques.

Tulio Ferreira Leite da Silva, Gonzalo Freijedo Aduna, Farah Benamara, Alda Mari, Zongmin Li, Li Yue, and Jian Su. 2025. CDB: A Unified Framework for Hope Speech Detection Through Counterfactual, Desire and Belief. In *The 2025 Annual Conference of the Nations of the Americas Chapter of the ACL*.

Nguyen Thi Thuy and Dang Van Thin. 2024. An empirical study of prompt engineering with large language models for hope detection in english and spanish. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2024), co-located with the 40th Conference of the Spanish Society for Natural Language Processing (SEPLN 2024)*, CEUR Workshop Proceedings. CEUR-WS.org.

Jules White, Quchen Fu, Sam Hays, Michael Sandborn, Carlos Olea, Henry Gilbert, Ashraf Elnashar, Jesse Spencer-Smith, and Douglas C. Schmidt. 2023. A prompt pattern catalog to enhance prompt engineering with chatgpt.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models.

A Model Hyper-parameters

Tables 7 and 8 present the hyperparameter used to fine-tune the transformer-based models and large language models in our experiments.

Parameter	Value
Pre-trained Model	BERT-base-uncased
Max Sequence Length	256
Batch Size	16
Learning Rate	2×10^{-5}
Optimizer	AdamW
Number of Epochs	10

Table 7: Hyperparameter for BERT fine-tuning.

Parameter	Value
Sequence Length	2048
Gradient Accumulation	2
Learning Rate	2×10^{-5}
Scheduler	Cosine
Number of Epochs	3
Lora Rank (r)	8
Save Checkpoints	Every 1000 steps

Table 8: Hyperparameter for fine-tuning Mistral.

B Prompt Design and Examples

B.1 Prompt Structure

Figures 4, 5, 6, 7, and 8 illustrate the various prompting strategies applied in assessing the large language model.

Your Job is to classify user sentences using only **ONE** of the following categories. Please generate ONLY ONE WORD classification:
Categories:
Generalized Hope: A broad sense of optimism not tied to specific outcomes.
Realistic Hope: An expectation of optimism not tied to specific outcomes.
Unrealistic Hope: A desire for outcomes that are unlikely or impossible
Not Hope: Texts that do not express hope

Text: "{text}"
CLASSIFICATION:

Figure 4: Zero-Shot (Multiclass).

B.2 Dataset Prompt Examples

Figures 9 and 10 illustrate the prompt example instances from the PolyHope and CDB datasets, including the prompt, input text, and corresponding gold classification labels.

Your Job is to classify user sentences using only **ONE** of the following categories. Please generate ONLY ONE WORD classification:
Categories:
Hope: Any expression of optimism, positive expectation, or belief in a desired outcome. This includes:
a) A broad sense of optimism not tied to specific outcomes.
b) An expectation that is grounded in achievable goals.
c) A desire for outcomes that are unlikely or impossible.
Not Hope: Texts that do not express hope.
Examples …..
Text: "{text}"
CLASSIFICATION:

Figure 5: Few-Shot (Binary).

You are performing a hope speech classification task using a decision tree approach. For each given text, follow this decision process:

1) Does the text express any form of hope, optimism, or positive expectation?
- If NO → Classify as "Not Hope"
2) If YES, does the text express a broad sense of optimism not tied to specific outcomes?
- If YES → Classify as "Generalized Hope"
- If NO, proceed to the next step.
3) If YES, is the text grounded in achievable goals and expectations?
- If YES → Classify as "Realistic Hope"
- If NO, proceed to the next step.
4) If YES, does the text express a desire for outcomes that are unlikely or impossible?
- If YES → Classify as "Unrealistic Hope"

Analyze the input text and provide only the classification.

Examples….

Output Format:
Text: "{text}"
CLASSIFICATION:

Figure 6: Decision Tree (Multiclass).

You are performing a hope speech classification task based on multiple reasoning. Analyze this text for expressions and classify user sentences using only **ONE** of the following categories.
* Linguistic Features: What hope-related words or phrases appear in the text? What is the general sentiment? (not more than 50 words)
* Goal Assessment: Are specific outcomes mentioned? How feasible are these outcomes? (not more than 50 words)
* Context Evaluation: What is the broader context of the hope expression? Are there conditional elements to the hope? (not more than 50 words)

Categories:
* "Generalized Hope": A broad sense of optimism not tied to specific outcomes.
* "Realistic Hope": An expectation grounded in achievable goals.
* "Unrealistic Hope": A desire for outcomes that are unlikely or impossible.
* "Not Hope": Texts that do not express hope.

Examples….

Output Format:
Text: "{text}"
Linguistic Features:
Goal Assessment:
Context Evaluation:
CLASSIFICATION:

Figure 7: Multiple Reasoning (Multiclass).

You are performing a hope speech classification task based on confidence structured output. Analyze this text for expressions and classify user sentences using only **ONE** of the following categories.

Output Format:
* Hope-related elements: List any words, phrases, or sentiments related to hope
* Outcome Specificity: Are specific outcomes mentioned? What are they?
* Feasibility Assessment: If outcomes are mentioned, how feasible are they?
* Classification Confidence: Generalized Hope[0-100%], Realistic Hope[0-100%], Unrealistic Hope[0-100%], Not Hope: [0-100%]

Categories:
* "Generalized Hope": A broad sense of optimism not tied to specific outcomes.
* "Realistic Hope": An expectation grounded in achievable goals.
* "Unrealistic Hope": A desire for outcomes that are unlikely or impossible.
* "Not Hope": Texts that do not express hope.

Examples….

Output Format:
Text: "{text}"
Hope-related elements:
Outcome Specificity:
Feasibility Assessment:
Classification Confidence:
CLASSIFICATION:

Figure 8: Confidence Score (Multiclass).

You will classify user-provided sentences into ONE of the following categories. Output Format: Provide ONLY ONE WORD for the classification. Do not include any explanations or introductory phrases.

Categories:
- Generalized Hope: A broad sense of optimism not tied to specific outcomes.
- Realistic Hope: An expectation grounded in achievable goals.
- Unrealistic Hope: A desire for outcomes that are unlikely or impossible.
- Not Hope: Texts that do not express hope.

Text: "Well the club season is over. We ended up tied for 44th place out of 130 in the Aspire division at the AAU National Championships. I had a great time! Played well. On to Utica High School volleyball bright and early tomorrow. Go U!"
CLASSIFICATION: Generalized Hope

Figure 9: Prompt, input instance, and gold classification in the PolyHope dataset.

You will be presented with user sentences, and you must classify each sentence into one of the following categories:

- Desire: The expression is future-oriented and indicates a wish, intention, or command.
- Belief: The expression is future-oriented and indicates a belief or prediction, often epistemic or deontic.
- Counterfactual: The expression is past-oriented and implies a desire for an alternative outcome.
- Not Hope: Any sentence that does not match the above categories.

Text: "Democrats would literally say anything for more votes."
CLASSIFICATION: Belief

Figure 10: Prompt, input instance, and gold classification in the CDB dataset.