BioNLP at #SMM4H-HeaRD 2026 Task 3 Estimating Flu Vaccine Effectiveness: A Temporal-Aware Fine-Tuning and Similarity-Based Few-Shot Prompting Approach

Ioana Irina Patularu
University of Bucharest, Faculty of Mathematics and Computer Science
ioana-irina.patularu@s.unibuc.ro

Abstract

This paper presents our systems for the SMM4H 2026 shared task on flu-related tweet classification across two subtasks: flu vaccination status and flu test outcome classification. For each subtask, we evaluate two approaches: fine-tuning BERTweet-large with a temporal-aware architecture, cross-validation ensembling, and regularization techniques, and a GPT-4o few-shot prompting system with similarity-based dynamic example retrieval, chain-of-thought reasoning and contrastive label ranking. Fine-tuning proves superior for the flu vaccination subtask (micro-F1: 87.90%), where sufficient and relatively balanced training data is available, while few-shot prompting performs better for the flu test subtask (micro-F1: 95.74%), where limited and heavily imbalanced training data renders fine-tuning less effective.

1 Introduction

Seasonal influenza remains a significant public health burden, with the 2024-2025 season classified as the most severe since 2017-2018. Traditional influenza surveillance methods suffer from delayed reporting and limited geographic coverage (Xu et al., 2025). In contrast, social media provides a continuously updated, diverse stream of health data. While prior NLP research has utilized these platforms for general flu detection (Weissenbacher et al., 2019), using automated methods to extract fine-grained clinical signals - such as vaccination status and test outcomes - from tweets still remains to be explored.

To address this, the Social Media Mining for Health (SMM4H) 2026 shared task (Lopez-Garcia et al., 2026) introduces two tweet classification subtasks focused on the 2020-2021 influenza season: Subtask 1, which requires classifying tweets according to the author's flu vaccination status (Currently-Vaccinated, Currently-Unvaccinated, Previously-Vaccinated, Possibly-Vaccinated, or Other), and Subtask 2, which targets flu test outcome classification (Currently-Positive, Currently-Negative, Previously-Positive, Previously-Negative, or Other). Both subtasks present notable challenges, including informal language, temporal ambiguity, flu and COVID-19 mixed references, and significant class imbalance (particularly in Subtask 2 where the Other class accounts for over 70% of the training data).

In this paper, we present two complementary systems developed for both subtasks. For the first one we fine-tune BERTweet-large, a RoBERTa-large model pre-trained on Twitter data, augmented with a temporal-aware architecture that encodes flu season context both as a text prefix and as explicit numerical features, along with cross-validation ensembling and regularization techniques. For the second one, well-suited where limited and imbalanced training data makes fine-tuning less effective, we adopt a few-shot prompting approach powered by GPT-4o, combining similarity-based dynamic example retrieval, chain-of-thought reasoning, and contrastive label ranking to guide the model toward accurate classification. Both systems are compared against the baselines established by (Xu et al., 2025), as detailed starting with Section 2.2.

2 System description

2.1 Dataset

The dataset consists of 4,216 tweets from the 2020–2021 flu season that contain information about vaccination and flu testing. For the flu vaccination subtask, we have 2809 tweets (1977 for train, 270 for validation and 562 for test). For the flu test subtask we have 1407 tweets (990 for train, 135 for validation and 282 for test). Table 1 presents the distribution of annotated labels. Table 4 and Table 5 from the supplementary material provide examples of samples for all classes, for both sub-

Proceedings of the 11th Social Media Mining for Health Research and Applications (SMM4H-HeaRD 2026) Workshop and Shared Tasks, pages 252–259

Flu Vaccine			Flu Test		
Label	**Counts**	**Prop (%)**	**Label**	**Counts**	**Prop (%)**
Currently-Vaccinated	581	20.68%	Currently-Positive	87	6.18%
Currently-Unvaccinated	709	25.24%	Currently-Negative	116	8.24%
Previously-Vaccinated	186	6.62%	Previously-Positive	41	2.91%
Possibly-Vaccinated	324	11.53%	Previously-Negative	161	11.44%
Other	1009	35.92%	Other	1002	71.22%

Table 1: Distribution of annotated labels for flu vaccine and flu test outcome in tweets collected from the 2020–2021 influenza season. It provides counts and proportions (%) of each label category (Xu et al., 2025).

tasks.

2.2 Baseline

The baseline provided by the organizers that obtained the best results (Xu et al., 2025) is based on a prompt optimization approach using Large Language Models. In order to create the prompt, in context learning (ICL) (Dong et al., 2024) and chain of thought (CoT) (Wei et al., 2022) strategies are used: the prompt contains specific instructions for the subtask type along with a few examples. Two examples per label were randomly selected where each example consisted of the tweet content, date posted, label and the generated rationale. For the flu test subtask, 2 prompts are proposed to be used sequentially, the first one for determining the test outcome and the second one to identify if the test occurred during the current flu season. The prompts were used with LLaMA-3-70B-Instruct. This approach is compared to traditional supervised baselines, specifically BERT-large (Devlin et al., 2019), RoBERTa-large (Liu et al., 2019) and BERTweet-large (Nguyen et al., 2020), that are shown to perform worse on the dataset.

2.3 Prompt optimization approach

For the prompt-based solution we implemented a few-shot prompting system using Chain-of-Thought (CoT) reasoning. Following (Liu et al., 2022), we retrieve in-context examples based on semantic similarity to the test query. To further ensure label coverage and avoid topical bias in the selected examples, we adopt a stratified retrieval strategy, selecting the most semantically similar example for each valid label class.

For generating the rationale for every example in the dataset, we used Llama-3.3-70B to reverse-engineer a 2-step reasoning rationale for each tweet, explaining why the correct label is correct. The rationale follows a fixed structure, focusing on the timing (whether the event took place during the current flu season) and on the outcome (why the provided label is the correct one). Given the size of the dataset, we manually verified that the generated rationales adhered to the imposed structure and that the model did not exhibit hallucinations. The enriched training data (tweets + rationales) is embedded using `mixedbread-ai/mxbai-embed-large-v1` (a sentence transformer) and stored in a persistent ChromaDB vector database that uses cosine similarity as its default distance metric.

Using the prompts from the baseline paper as a starting point, we explored several manually crafted prompt formulations and selected the best-performing configuration on the validation dataset for each subtask. For the flu vaccination subtask, the prompt comprises task-specific classification rules, a disambiguation directive to prevent the model from conflating flu and COVID-19 references, and five dynamically retrieved few-shot examples, one per label class. For the flu test subtask, the optimal prompt includes explicit label definitions, a timing rule to correctly anchor events relative to the flu season window, the same COVID-19 disambiguation directive, and five retrieved examples following the same stratified retrieval strategy. In both cases, rather than asking the model to predict a single label directly, the prompt elicits a full ranking of all five candidate labels, encouraging the model to reason contrastively across classes before committing to a final answer. The complete prompts are provided in Figure 1 and Figure 2 of the supplementary material. At inference time, all predictions were generated using GPT-4o (gpt-4o-2024-08-06) with deterministic decoding settings (temperature=0.0, top_p=1.0, seed=42) and a maximum output length of 500 tokens.

2.4 Fine tuning approach

For this solution, we fine-tuned BERTweet-large, a RoBERTa-large model pre-trained on Twitter data, making it a well-suited choice for tweet classification. Rather than using the tweet text alone, we

	Flu Vaccination			Flu Test		
	Cur-Vac	**Cur-Unvac**	**Overall**	**Cur-Pos**	**Cur-Neg**	**Overall**
Prompt optimization	80.00	93.71	81.85	**94.12**	**88.00**	**97.04**
BERTweet-large	**85.47**	**94.03**	**85.93**	77.78	80.00	88.89

Table 2: F1 scores achieved with the prompt optimization approach and fine-tuning BERTweet-large on the validation dataset. Best scores are highlighted in bold.

enriched each input with a human-readable temporal prefix prepended to the tweet. In parallel, we extracted from the tweet's posting date four temporal features (calendar month, flu-season week, season progress, and a binary peak-season indicator) and passed them through a dedicated two-layer MLP projection head, expanding them to a 64-dimensional representation. The resulting 64-dimensional temporal representation was concatenated with the `[CLS]` token embedding (1024-dim) from the BERTweet encoder. The main reason behind this expansion was to ensure the temporal signal carries sufficient weight when fused with the much larger `[CLS]` representation. We applied a LayerNorm to normalize the fused representation and used a Multi-Sample Dropout classification head (Inoue, 2019) (k=5, p=0.3), producing the final logits over the 5 label classes.

To address class imbalance, we computed inverse-frequency class weights from each fold's training split and incorporated them into a label-smoothed (ε=0.1) cross-entropy loss. We additionally applied R-Drop regularization (Liang et al., 2021) with α=0.5, encouraging the model to produce consistent predictions under dropout noise.

Training followed a 5-fold stratified cross-validation scheme repeated over 2 random seeds, yielding 10 checkpoints in total. We used AdamW (Loshchilov and Hutter, 2019) with layer-wise learning rate decay (LLRD) (Howard and Ruder, 2018; Sun et al., 2019), where the base learning rate (10^{-5}) was decayed by a factor of 0.9 per encoder layer from top to bottom, while the classification head and temporal projection branch were trained at 5$\times$ the base rate. A cosine schedule with 10% linear warmup was applied. Early stopping with patience of 5 epochs (monitored on micro-F1) prevented overfitting, with a maximum of 15 epochs per fold. We used input sequences of 128 tokens, with an effective batch size of 16 (batch size 8 with 2 gradient accumulation steps). At inference time, soft-voting ensemble averaging was applied across all 10 checkpoints by averaging the softmax probability distributions and taking the argmax.

For the flu test subtask we used the same architecture; given the smaller size of the training set, we adjusted the training configuration accordingly: maximum epochs were increased to 20, early stopping patience to 7, and the R-Drop regularization coefficient was reduced to α=0.3 to avoid over-constraining the model during learning.

3 Results and interpretation

3.1 Validation dataset

The validation results are presented in Table 2. For the flu vaccination subtask, fine-tuning BERTweet-large yielded the best performance whereas for the flu test subtask, the prompt-based solution proved superior. For the flu vaccination subtask, we obtained an F1 score for Currently-Vaccinated of 85.47 and 94.03 for Currently-Unvaccinated with BERTweet-large while the prompt solution got only F1 scores of 80.00 for Currently-Vaccinated and 93.71 for Currently-Unvaccinated. For the flu test subtask, the prompt approach obtained F1 scores of 94.12 for Currently-Positive and 88.00 for Currently-Negative, while BERTweet-large scored only 77.78 for Currently-Positive and 80.00 for Currently-Negative.

We attribute this difference primarily to the smaller training set size and more pronounced class imbalance in the flu test subtask, conditions under which fine-tuning struggles to generalize, while few-shot prompting remains effective by relying on only a handful of retrieved examples rather than learning from the full training distribution.

3.2 Test dataset

Table 3 presents the results of our two systems alongside the baseline configurations. For the flu vaccination subtask, our fine-tuned BERTweet-large model achieves a micro-F1 of 87.90%, outperforming all baseline systems, including the strongest LLM-based baseline (Few-shot + CoT, 84.88%) and the fine-tuned baselines (BERTweet: 84.36%, RoBERTa: 84.48%). Notably, compared

		Flu Vaccination			Flu Test		
		Cur-Vac	**Cur-Unvac**	**Overall**	**Cur-Pos**	**Cur-Neg**	**Overall**
Baseline fine tuning	BERT-large	82.76	89.90	83.80	64.71	73.47	90.07
	RoBERTa-large	86.09	91.70	84.48	77.27	81.82	91.84
	BERTweet-large	84.62	90.46	84.36	55.17	75.56	88.56
Baseline LLM	Zero-shot	75.34	83.71	75.09	82.05	69.70	87.59
	Zero-shot + CoT	82.21	84.78	78.83	81.08	71.70	90.43
	Few-shot	80.43	87.63	80.78	83.72	86.96	93.26
	Few-shot + CoT	87.35	**92.14**	84.88	88.89	89.36	95.03
Our systems	Prompt optimization	85.06	91.84	83.99	**91.43**	**91.67**	**95.74**
	BERTweet-large	**90.21**	90.65	**87.90**	60.00	68.29	88.30

Table 3: F1 scores for baseline models from (Xu et al., 2025) (for fine tuning BERT-large, RoBERTa-large and BERTweet-large and prompt optimization with LLM) and our systems (the prompt optimization and fine tuning BERTweet-large approaches) on the test dataset. Best scores are highlighted in bold while the second best scores are underlined.

to the fine-tuned baselines, our model achieves the highest F1 scores on both primary labels, reaching 90.21% on Currently-Vaccinated and 90.65% on Currently-Unvaccinated. We attribute this improvement over the BERTweet baseline largely to our temporal-aware architecture (the prepended date prefix and the dedicated temporal MLP branch) which allowed the model to anchor flu-related events within the correct season window, a signal that plain text encoding would otherwise underweight. Additionally, the ensemble of 10 checkpoints with soft-voting, combined with R-Drop regularization and Multi-Sample Dropout, contributed to a more robust and stable classifier, reducing the variance that typically affects fine-tuning on moderately sized, imbalanced datasets. Our prompt-based solution for this subtask, however, performs below the baselines with a micro-F1 of 83.99%, suggesting that the larger and more balanced training set of flu vaccination subtask favors a fine-tuning approach over few-shot prompting.

For the flu test subtask, the pattern reverses. Our prompt-based solution achieves a micro-F1 of 95.74%, surpassing the strongest baseline (Few-shot + CoT, 95.03%) and yielding particularly strong results on the two primary labels (Currently-Positive with 91.43% and Currently-Negative with 91.67%). We attribute this gain primarily to the dynamic, similarity-based retrieval strategy: by selecting the most semantically similar example for each label class from the training set, the prompt exposes GPT-4o to highly relevant context at inference time. The explicit timing rule included in the prompt further helped the model correctly interpret the flu season window, addressing the same temporal ambiguity that motivated our temporal features in the fine-tuning solution. Asking the model to rank all five labels rather than predict a single one also likely contributed by encouraging more deliberate reasoning across classes before committing to a final answer. In contrast, our BERTweet-large fine-tuned model achieves only 88.30% micro-F1, performing comparably to the BERTweet baseline (88.56%) and falling well short of the prompt solution. The considerably lower F1 scores on the primary labels (60.00% Currently-Positive, 68.29% Currently-Negative) highlight BERTweet's difficulty in learning from this subtask's limited and highly imbalanced training data, even with the regularization and ensemble strategies applied.

4 Conclusions

In this paper, we presented two systems for flu-related tweet classification in SMM4H 2026, showing that dataset characteristics dictate the optimal approach. With abundant, relatively balanced data (vaccination subtask), full fine-tuning with ensembling and regularization performs best. Conversely, for scarce, imbalanced data (flu test subtask), similarity-based few-shot prompting with chain-of-thought reasoning is more robust. A consistent finding across both sub-tasks was the importance of temporal awareness, whether encoded as explicit features in the fine-tuning architecture or as disambiguation rules in the prompt.

References

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.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Baobao Chang, Xu Sun, Lei Li, and Zhifang Sui. 2024. A Survey on In-context Learning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1107–1128, Miami, Florida, USA. Association for Computational Linguistics.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 328–339, Melbourne, Australia. Association for Computational Linguistics.

Hiroshi Inoue. 2019. Multi-sample dropout for accelerated training and better generalization. *arXiv preprint arXiv:1905.09788*.

Xuancheng Liang, Lijun Wu, Juntao Li, Yue Wang, Qi Meng, Tao Qin, Wei Chen, Min Zhang, and Tie-Yan Liu. 2021. R-drop: Regularized dropout for neural networks. In *Advances in Neural Information Processing Systems*, volume 34, pages 10961–10973. Curran Associates, Inc.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2022. What makes good in-context examples for GPT-3? In *Proceedings of Deep Learning Inside Out: The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures (DeeLIO 2022)*, pages 84–92. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.

Guillermo Lopez-Garcia, Jose Miguel Acitores Cortina, Jacob Berkowitz, Joey Chan, Ganesh Chandrasekar, Sumon Kanti Dey, Ivan Flores Amaro, Fernando Gallego, Lauren Gryboski, Ari Z Klein, Martin Krallinger, Salvador Lima-López, Tomohiro Nishiyama, Lisa Raithel, Ahmad Rezaie Mianroodi, Roland Roller, Judith Rosell, Frank Rudzicz, Abeed Sarker, and 8 others. 2026. Overview of the 11th Social Media Mining for Health (#SMM4H) and Health Real-World Data (HeaRD) Shared Tasks at ACL 2026. In *Proceedings of the 11th Social Media Mining for Health (#SMM4H) and Health Real-World Data (HeaRD) Workshop and Shared Tasks*. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. 2020. BERTweet: A pre-trained language model for English tweets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 9–14, Online. Association for Computational Linguistics.

Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. 2019. How to fine-tune BERT for text classification? *arXiv preprint arXiv:1905.05583*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc.

Davy Weissenbacher, Abeed Sarker, Arjun Magge, Ashlynn Daughton, Karen O'Connor, Michael J. Paul, and Graciela Gonzalez-Hernandez. 2019. Overview of the fourth social media mining for health (SMM4H) shared tasks at ACL 2019. In *Proceedings of the Fourth Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task*, pages 21–30, Florence, Italy. Association for Computational Linguistics.

Dongfang Xu, Guillermo López García, Karen O'Connor, Haily Holston, Ari Z Klein, Ivan Flores Amaro, Matthew Scotch, and Graciela Gonzalez-Hernandez. 2025. Mining Social Media Data for Influenza Vaccine Effectiveness Using a Large Language Model and Chain-of-Thought Prompting. In *AMIA Annual Symposium Proceedings*, pages 1404–1413. American Medical Informatics Association. PMC12919567.

A Examples of samples from the dataset

Label	Tweet	Date
Currently-Vaccinated	*My left arm is hurting so bad from the flu shot*	2021-01-05 19:56:00
Currently-Unvaccinated	*Never had a flu shot nor will I. I've also never had the flu and I'm 55!*	2020-11-23 02:49:19
Previously-Vaccinated	*I did nearly black out in the shower one time after getting the flu shot. I went blind for about 10 seconds. Scary as hell*	2020-12-18 02:00:55
Possibly-Vaccinated	*I truly feel like I'm being forced to take the flu shot. It's either get the shot or be jobless :(*	2020-11-30 12:21:35
Other	*Y'all really need to get ya flu shot.*	2020-11-21 17:11:55

Table 4: One example per label class from the flu vaccination subtask training set.

Label	Tweet	Date
Currently-Positive	*I got a Covid test today. They said I was negative but positive for the flu. Well this stinks.*	2020-12-31
Currently-Negative	*Negative flu test, covid test i gotta wait on. Overall think i got a virus of some sort*	2020-11-24 02:40:30
Previously-Positive	*I had been really ill during a spirit night at work once a few years ago and my boss couldn't have cared less... I tested positive for flu next day.*	2021-01-17
Previously-Negative	*Last January I wa SO sick! Tested negative for the flu but had all of the symptoms. I was in bed for 3 weeks!*	2020-12-01 16:10:46
Other	*Good News! Flu Cases Disappear in US - Number of Positive Flu Tests at All-Time Low for Some Reason? #News*	2020-12-30 00:41:13

Table 5: One example per label class from the flu test outcome subtask training set.

B Prompts used in our solution

Few-Shot Prompt Template for Flu Vaccination Classification

INSTRUCTIONS
You are an expert public health annotator for the SMM4H competition (Subtask 1: Flu Vaccination). Classify the tweet based on the user's SELF-REPORTED vaccination status.

VALID LABELS
Currently-Vaccinated, Currently-Unvaccinated, Previously-Vaccinated, Possibly-Vaccinated, Other

FOLLOW THESE RULES TO CLASSIFY IN THIS ORDER

1. If user doesn't reference anything about flu vaccination OR he references flu vaccination of other people, not a personal statement, classify it as Other
2. If user expresses intention or consideration to get a flu shot but provides no evidence of actual vaccination, classify it as Possibly-Vaccinated
3. If user mentions receiving a flu shot before the current flu season which is (Sept 2020 – Aug 2021), with no indication of any recent vaccination, classify it as Previously-Vaccinated
4. If user explicitly states receiving a flu shot THIS season (Sept 2020 – Aug 2021), classify it as Currently-Vaccinated
5. If user explicitly states they have NOT received or plan NOT to receive a flu shot. Past refusal patterns implying a continuous unvaccinated status are also included.

COVID DISAMBIGUATION
After Dec 1, 2020, COVID vaccines were being rolled out alongside flu vaccines.

- Generic phrases like 'got my shot', 'got vaccinated', 'got my jab' WITHOUT a flu qualifier (flu shot, flu vaccine, flu jab, influenza vaccine, flumist, fluzone) $\rightarrow$ label as Other.
- Explicit flu vaccine signals always override: 'flu shot', 'flu jab', 'fluzone' $\rightarrow$ Currently-Vaccinated.
- Explicit COVID vaccine signals (pfizer, moderna, booster, j&j) with no flu mention $\rightarrow$ Other.

EXAMPLES
[5 stratified examples are inserted here in the following format:]
Date: {date}
Tweet: {text}
Rationale: {rationale}
...

RANKING LOGIC
Your ranking must represent a decrease in confidence from left to right:

- Label1: Highest confidence match.
- Label2: The 'near-miss' - the label that would be correct if the most ambiguous part of the tweet was interpreted differently.
- Label5: The least likely label that the tweet definitely does NOT represent.

TASK
Analyze the tweet below. Respond in exactly this format:

Rationale: [Step-by-step reasoning: check date, tense, flu vs COVID vaccine, self-report]
Ranking: Label1 $>$ Label2 $>$ Label3 $>$ Label4 $>$ Label5

Rules: Use each of the 5 valid labels exactly once. No brackets, no numbers.

Date: {tweet_date}
Tweet: {test_tweet}
Rationale:

Figure 1: The full prompt template used for the flu vaccination classification. The *EXAMPLES* section dynamically inserts five few-shot demonstrations prior to the target inference tweet.

Few-Shot Prompt Template for Flu Test Classification

INSTRUCTIONS
You are an expert public health annotator for the SMM4H competition (Subtask 2: Flu Test). Classify the tweet based on the user's SELF-REPORTED flu test outcome.

VALID LABELS

- **Currently-Positive:** User explicitly reports a recent positive flu test or diagnosis.
- **Currently-Negative:** User explicitly reports a recent negative flu test or a diagnosis excluding flu.
- **Previously-Positive:** User mentions a positive flu test/diagnosis from a PREVIOUS season only.
- **Previously-Negative:** User mentions a negative flu test/non-diagnosis from a PREVIOUS season only.
- **Other:** Tweet references flu testing WITHOUT the user's personal test status. Use Other for: another person's result, encouraging others to test, symptoms only (no test), unconfirmed suspicions, general flu discussions, or truly ambiguous statements.

TIMING RULE
Current Flu Season: Sept 1, 2020 - Aug 31, 2021.

- Tests/diagnoses WITHIN this window $\rightarrow$ 'Currently-'
- Tests/diagnoses BEFORE Sept 1, 2020 $\rightarrow$ 'Previously-'

COVID DISAMBIGUATION
After Dec 1, 2020, 'tested positive/negative' WITHOUT an explicit flu qualifier (flu test, influenza, rapid flu, tamiflu, H1N1, H3N2) likely refers to COVID $\rightarrow$ label as Other. Only use Currently/Previously-Positive or -Negative if influenza is explicitly mentioned.

EXAMPLES
[5 stratified examples are inserted here in the following format:]
Date: {date}
Tweet: {text}
Rationale: {rationale}
...

RANKING LOGIC
Your ranking must represent a decrease in confidence from left to right:

- Label1: Highest confidence match.
- Label2: The 'near-miss' - the label that would be correct if the most ambiguous part of the tweet were interpreted differently.
- Label5: The least likely label that the tweet definitely does NOT represent.

TASK
Analyze the tweet. Respond in exactly this format, nothing else:

Rationale: [Step-by-step reasoning: check date, tense, flu specificity, self-report, COVID overlap]
Ranking: Label1 $>$ Label2 $>$ Label3 $>$ Label4 $>$ Label5

Use each of the 5 labels exactly once. No brackets, numbers, or extra text.

Date: {tweet_date}
Tweet: {test_tweet}
Rationale:

Figure 2: The full prompt template used for the flu test classification. The *EXAMPLES* section dynamically inserts few-shot demonstrations prior to the target inference tweet.