PushingBoundaries at NakbaVirality Shared Task: Recursive Prompt Improvement for Multimodal Virality Classification

Ashhadul Islam[1], **Md Rafiul Biswas**[2], **Samir Brahim Belhaouari**[2], **Wajdi Zaghouani**[3]
[1]KTH Royal Institute of Technology, Stockholm, Sweden
[2]Hamad Bin Khalifa University, Doha, Qatar
[3]Northwestern University in Qatar, Doha, Qatar
aisla@kth.se, {mbiswas,sbelhaouari}@hbku.edu.qa, wajdi.zaghouani@northwestern.edu

Abstract

This paper describes our participation in the NakbaVirality shared task at the NakbaNLP Workshop (LREC–COLING 2026). We investigate Recursive Prompt Improvement (RPI), an instruction-level optimization strategy for virality classification in high-stakes geopolitical discourse. In this work, we propose a self-supervised approach to iteratively improve the classification prompt without human intervention. We begin with a basic prompt that guides the LLM to perform multi-class classification, incorporating contextual information about the tweets. After obtaining predictions, we identify misclassified tweets and feed them back to the model with an instruction to refine and improve the original classification prompt. This process is repeated over multiple iterations to assess whether performance improves over time. Our results show a remarkable improvement in F1 score from the first iteration to the final one. Although the proposed method does not reach the accuracy of models fine-tuned directly on task-specific data, it demonstrates that iterative, self-supervised prompt refinement can serve as a viable proxy for fine-tuning. By leveraging the model's own errors as feedback, this approach reduces reliance on computationally expensive training procedures and heavy GPU usage, while preserving much of the adaptability typically associated with fine-tuned models. This paradigm opens promising avenues for resource-efficient model adaptation and suggests new directions for scalable, low-cost performance improvement without traditional fine-tuning. The code has been made publicly available at GitHub repository.

Keywords: virality, social media, multimodal, LLM

1. Introduction

Social media platforms such as X, Facebook, and Instagram have become central arenas for information diffusion and narrative contestation during geopolitical conflicts (Hussain and Howard, 2013; Castells, 2015). In these contexts, virality is shaped by emotional intensity, polarization, and algorithmic amplification, and can directly influence public opinion and mobilization (Vosoughi et al., 2018; Stray et al., 2023; Kamin, 2019; Tufekci, 2015).

The Nakba and post–October 7th Gaza war discourse exemplify highly polarized, emotionally charged online communication, where moral-emotional language and identity signaling strongly drive engagement (Garimella et al., 2018; Barberá, 2015; Brady et al., 2017). Predicting virality in such settings therefore requires models that go beyond surface lexical features to capture emotional framing and stance.

Conflict discourse is also inherently multimodal, with images, symbols, and memes playing a key role in shaping interpretation and engagement (Highfield and Leaver, 2016; Alam et al., 2024). Prior work shows that integrating textual and visual signals improves social media understanding tasks (Zaghouani et al., 2025; Hasanain et al., 2024), yet multimodal virality prediction in sensitive geopolitical domains remains underexplored.

The NakbaVirality shared task (Ezzini et al., 2026) addresses this gap by challenging participants to predict reach and engagement for both textual and multimodal posts related to the Nakba and the post–October 7th war, fostering research on context-rich and polarizing data.

We participated in the NakbaVirality Shared Task, which focuses on predicting virality levels (low, medium, high) for posts in a high-stakes geopolitical discourse setting. We have seen in previous works LMMs and LLMs can be converted into classifiers by suitable prompting (Islam et al., 2023). However, to improve performance the prompt needs to be changed and tested multiple times manually. Our primary objective in this task was to examine whether *Recursive Prompt Improvement (RPI)*—an iterative system prompt refinement strategy—can serve as a viable alternative to conventional model fine-tuning. Rather than updating model parameters, we explored whether structured prompt evolution driven by misclassified examples could achieve competitive performance in a data-efficient and cost-effective manner.

- **Recursive Prompt Improvement (RPI):** We introduce an iterative prompt refinement framework that leverages misclassified examples to progressively improve task-specific system prompts without parameter updates.
- **Alternative to Fine-Tuning:** We empirically demonstrate that recursive prompt optimiza-

Proceedings of the 2nd International Workshop on Nakba Narratives as Language Resources (Nakba-NLP 2026) @ LREC 2026, pages 201–205

tion can substantially improve virality classification performance, highlighting its effectiveness as a lightweight alternative to traditional fine-tuning.

- **Cross-Model Prompt Transferability:** We show that prompts optimized using a smaller open-source model generalize effectively to larger and closed-source LLMs, suggesting that prompts can act as portable task adapters.
- **Cost-Efficient Optimization Pipeline:** We propose a practical workflow where a cheaper model is used for iterative prompt refinement, and the resulting optimized prompt is deployed on more powerful models, reducing computational and financial overhead.

2. Background

The NakbaVirality shared task (Ezzini et al., 2026) formulates virality prediction as a three-class classification problem, requiring systems to label posts as *Low*, *Medium*, or *High* based on a normalized engagement score derived from likes, shares/retweets, and comments. The dataset contains approximately 2,600 curated post–image pairs collected from X and Reddit after October 7, 2023, filtered using conflict-related keywords (e.g., "Gaza," "Nakba," "Palestine," "Israel"). The content reflects highly polarized, multilingual discourse, with the majority in English and Arabic.

Although the task emphasizes multimodal modeling, our work focuses primarily on a text-only setting to isolate the effect of **Recursive Prompt Improvement (RPI)**. We conducted experiments on roughly 1,600 tweet–image pairs, using only 100 samples for iterative prompt optimization and the remainder for validation and evaluation in a data-efficient setup.

Unlike prior approaches centered on parameter fine-tuning or supervised multimodal architectures (Ezzini et al., 2026), we propose RPI as a lightweight, model-agnostic alternative. By iteratively refining the system prompt using misclassified examples, we improve Macro-F1 without updating model weights and demonstrate cross-model transferability. This positions RPI as a computationally efficient strategy for adapting large language models to sensitive, high-stakes virality classification tasks.

3. System Overview

We are given 1,600 tweet–image pairs annotated with three virality classes: *Low (0)*, *Medium (1)*, and *High (2)*. Although the shared task supports multimodal inputs, our primary system focuses on the **text-only setting** in order to evaluate prompt optimization independently from visual reasoning.

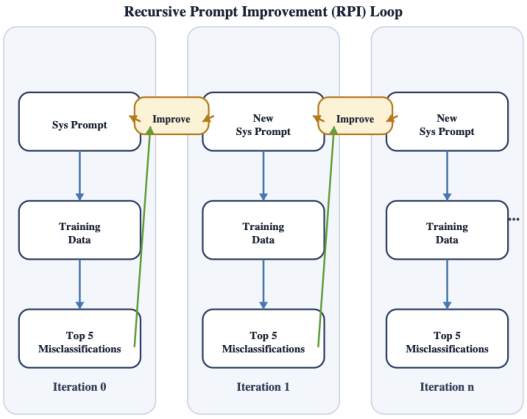

Figure 1: Recursive prompt flow Diagram

3.1. System Configurations

We evaluate three configurations:

- **System A (Baseline Prompt, Epoch 0):** A naïve fixed system prompt defining virality rules.
- **System B (Recursive Prompt Improvement):** Iteratively refined prompt using misclassification feedback.
- **System C (Cross-Model Transfer):** Applying the optimized prompt to different LLM architectures.

3.2. Recursive Prompt Improvement (RPI)

Instead of optimizing model parameters θ, we optimize the **system prompt** P.

Let:

$$D = \{(x_i, y_i)\}_{i=1}^{N}$$

where x_i is tweet text and $y_i \in \{0, 1, 2\}$ is the virality label.

Given model M and prompt P_t at iteration t:

$$\hat{y}_i = M(x_i; P_t)$$

We identify misclassified examples:

$$E_t = \{x_i \mid \hat{y}_i \neq y_i\}$$

The prompt update rule is:

$$P_{t+1} = \text{Improve}(P_t, \text{TopK}(E_t))$$

where the `Improve` function is implemented via a meta-prompt asking the model to revise classification instructions based on observed errors.

Figure 1 illustrates the Recursive Prompt Improvement (RPI) workflow as an iterative refinement process. In each iteration, the current system prompt generates virality predictions on the training subset, which are then compared with ground-truth labels to identify classification errors. Rather than updating model parameters, refinement occurs at the instruction level.

We select the top-5 misclassified tweets as diagnostic signals, highlighting weaknesses in the prompt formulation, such as ambiguous boundaries between Medium and High virality or insufficient emphasis on urgency cues. These examples are incorporated into a meta-prompt that revises the classification guidelines. The resulting updated system prompt is used in the next iteration.

This closed-loop process can continue for multiple rounds, progressively clarifying decision rules and improving separation between Low, Medium, and High virality classes without gradient-based fine-tuning.

Each iteration therefore consists of three structured stages:

1. **Prediction Stage:** Apply the current system prompt to generate virality labels.
2. **Error Analysis Stage:** Compare predictions with ground truth and extract the top-5 most informative misclassifications.
3. **Prompt Refinement Stage:** Update the system prompt using a meta-instruction informed by these errors.

Through this structured cycle, prompt engineering becomes a systematic optimization procedure rather than a one-shot manual design process.

4. Experimental Setup

We did not use any additional labeled datasets, external annotations, or handcrafted features beyond the data provided by the shared task. Our methodology is intentionally lightweight and relies primarily on large pre-trained language models and structured prompt design rather than parameter fine-tuning.

Our approach leverages:

- **Pre-trained LLMs:** DeepSeek-chat, OpenAI `gpt5.2`, and Qwen2.5-VL-7B-Instruct, accessed via the OpenRouter API. We rely on the models' pre-trained world knowledge and contextual reasoning capabilities for interpreting geopolitical and emotionally charged content.
- **Data partitioning strategy:** The dataset is divided into train/dev/test splits, with 100 samples used for iterative prompt optimization and the remaining data used for validation and testing.
- **Iterative refinement setup:** The recursive loop was executed for up to 30 epochs, with the best-performing prompt observed at epoch 7.
- **Computational environment:** Experiments were conducted using Google Colab Pro with an NVIDIA A100 GPU (40GB GPU RAM) and 80GB system RAM.
- **API usage and cost:** Running the closed-source model (`openai/gpt5.2`) for approximately 3,000 prompts across two epochs incurred an estimated cost of $15 USD.

5. Results

Figure 2 shows the Macro-F1 scores across iterations for both training and validation sets. The validation trend closely follows the training curve, indicating stable generalization during prompt refinement. Performance improves in the early iterations and reaches its maximum validation score at Epoch 7, after which gains plateau. This suggests that the recursive prompt optimization converges within a small number of iterations without evident overfitting.

It is instructive to compare the initial prompt with the final prompt obtained after iterative self-supervised refinement.

Initial Prompt

```
You are a tweet virality scor-
ing model.
Your job:  read the tweet
text and return a virality
score.
Virality score rules:
0 = low viral (ordinary,
niche, no hook, low engage-
ment potential)
1 = medium viral (interesting,
relatable, decent hook, moder-
ate engagement potential)
2 = high viral (strong hook,
highly shareable, controver-
sial, emotional, breaking
news, meme-worthy, strong
CTA)
```

Final Self-Supervised Prompt

```
You are a tweet virality scor-
ing model.  Your task:  read
a tweet and return a viral-
ity score (0=low, 1=medium,
```

```
2=high). Consider emotional
impact, public interest, and
potential for shares or dis-
cussions. Assign low scores
(0) to tweets that are short,
lack details, report routine
developments, or contain ex-
treme claims without evidence
of broad impact. Reserve
high scores (2) only for
tweets with urgent, highly
contentious elements likely
to spark widespread debate or
action. Be cautious with
tweets about ongoing con-
flicts or political issues;
default to lower scores un-
less they contain urgent
breaking news, major policy
shifts, or highly provocative
accusations with clear poten-
tial for mass discussion.
```

The refined prompt demonstrates a clearer operationalization of the scoring criteria. While the initial prompt provides broad qualitative descriptors (e.g., "strong hook" or "highly shareable"), the final prompt introduces more structured decision boundaries and explicit cautionary rules. In particular, it distinguishes between emotional intensity and verifiable public impact, discourages over-scoring routine or unsubstantiated claims, and provides nuanced handling of political or conflict-related content. Such refinements enable the model to better capture subtle contextual cues that would otherwise require extensive manual inspection. By encoding these nuanced distinctions directly into the prompt, the self-supervised process effectively internalizes criteria that a human annotator might take considerable time to apply consistently across large volumes of tweets.

5.1. Cross-Model Settings

The recursive prompt refinement procedure was first conducted using text-only LLMs: DeepSeek-chat (open-source) and OpenAI `gpt5.2` (closed-source) to assess cross-model generalization. This setup allows us to examine whether improvements achieved through recursive refinement on one model transfer effectively to other architectures. To further investigate multimodal adaptability, we conducted additional experiments using Qwen2.5-VL-7B-Instruct, a vision-language model. For the multimodal model, the prompt was minimally adapted to explicitly reference both textual and visual inputs (e.g., "You are a tweet and image virality scoring model"), while preserving the refined classification instructions learned during recursive optimization.

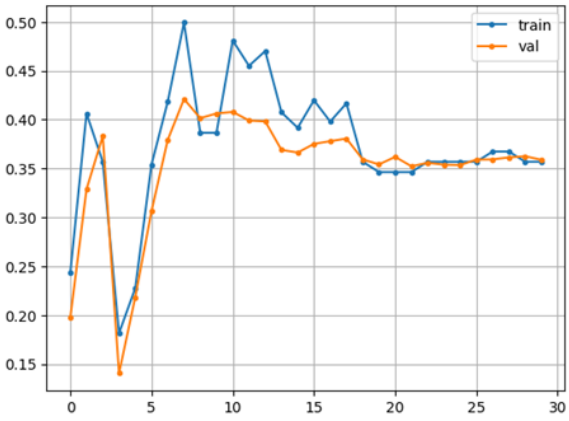

Figure 2: Measurement of F1-score over number of epoch

Model	Baseline	RPI	Gain
Deepseek-chat	0.20	**0.44**	**+120%**
Openai/gpt5.2	0.25	**0.42**	+68%
Qwen2.5-VL-7B-Instruct	0.20	**0.30**	+50%

Table 1: Relative improvement after Recursive Prompt Improvement (RPI).

Table 1 shows that the optimized prompt (Epoch 7) consistently outperforms the baseline configuration. DeepSeek exhibits the largest improvement (+0.24 absolute F1), indicating strong alignment between prompt refinement and model behavior. OpenAI `gpt5.2` also benefits substantially (+0.17), demonstrating that prompt-level optimization is transferable even to larger, closed-source systems.

5.2. Multimodal Settings

In the multimodal setting, Qwen2.5-VL-7B-Instruct shows a measurable improvement (+0.10), suggesting that the refined decision boundaries encoded in the optimized prompt remain beneficial when image information is introduced. Although the gain is smaller compared to text-only LLMs, the improvement confirms that Recursive Prompt Improvement produces architecture-agnostic refinements that generalize beyond the model used during optimization.

6. Conclusion

We investigated Recursive Prompt Improvement (RPI) as a parameter-free alternative to fine-tuning for virality classification. Iterative, error-driven prompt refinement consistently improved performance across multiple language models and demonstrated cross-model transferability. Several multimodal models behaved unstably: `mbzuai/ain` achieved low performance (F1 =

0.14), while Molmo and `idefics2` collapsed to single-class predictions. In contrast, text-focused LLMs were more stable and achieved stronger results, suggesting the task relies heavily on textual reasoning. Prompt optimization performed with a smaller open-source model transferred effectively to others. Notably, DeepSeek outperformed the larger closed-source `gpt5.2`, showing that scale alone does not ensure better performance under prompt-based refinement.

Acknowledgments

This work was funded by the NPRP Grant No. 14C-0916-210015 from the Qatar National Research Fund, part of the Qatar Research Development and Innovation Council (QRDI). The findings and conclusions expressed in this paper are solely those of the authors.

7. References

Firoj Alam, Md Rafiul Biswas, Uzair Shah, Wajdi Zaghouani, and Georgios Mikros. 2024. Propaganda to hate: A multimodal analysis of arabic memes with multi-agent llms. In *International Conference on Web Information Systems Engineering*, pages 380–390. Springer.

Pablo Barberá. 2015. Birds of the same feather tweet together: Bayesian ideal point estimation using twitter data. *Political analysis*, 23(1):76–91.

William J Brady, Julian A Wills, John T Jost, Joshua A Tucker, and Jay J Van Bavel. 2017. Emotion shapes the diffusion of moralized content in social networks. *Proceedings of the National Academy of Sciences*, 114(28):7313–7318.

M Castells. 2015. Networks of outrage and hope: Social movements in the internet age. john wiley & sons.

Saad Ezzini, Salima Lamsiyah, Shadi Abudalfa, Samir El-Amrany, and Walid Alsafadi. 2026. The nakbavirality shared task on multimodalvirality prediction in high-stakes discourse. In *Proceedings of the 2nd International Workshop on Nakba Narratives as Language Resources (Nakba-NLP 2026), co-located with the Language Resources and Evaluation Conference (LREC 2026)*, Palma, Mallorca, Spain.

Kiran Garimella, Gianmarco De Francisci Morales, Aristides Gionis, and Michael Mathioudakis. 2018. Political discourse on social media: Echo chambers, gatekeepers, and the price of bipartisanship. In *Proceedings of the 2018 world wide web conference*, pages 913–922.

Maram Hasanain, Md Arid Hasan, Fatema Ahmad, Reem Suwaileh, Md Rafiul Biswas, Wajdi Zaghouani, and Firoj Alam. 2024. Araieval shared task: propagandistic techniques detection in unimodal and multimodal arabic content. In *Proceedings of The Second Arabic Natural Language Processing Conference*, pages 456–466.

Tim Highfield and Tama Leaver. 2016. Instagrammatics and digital methods: Studying visual social media, from selfies and gifs to memes and emoji. *Communication research and practice*, 2(1):47–62.

Muzammil M Hussain and Philip N Howard. 2013. What best explains successful protest cascades? icts and the fuzzy causes of the arab spring. *International studies review*, 15(1):48–66.

Ashhadul Islam, Md Rafiul Biswas, Wajdi Zaghouani, Samir Brahim Belhaouari, and Zubair Shah. 2023. Pushing boundaries: Exploring zero shot object classification with large multimodal models. In *2023 Tenth International Conference on Social Networks Analysis, Management and Security (SNAMS)*, pages 1–5. IEEE.

Julia Kamin. 2019. *Social Media and Information Polarization: Amplifying Echoes or Extremes?* Ph.D. thesis, University of Michigan.

Jonathan Stray, Ravi Iyer, and Helena Puig Larrauri. 2023. The algorithmic management of polarization and violence on social media. Technical report, Knight First Amendment Institute, Columbia University. KnightColumbia.org, forthcoming.

Zeynep Tufekci. 2015. Algorithmic harms beyond facebook and google: Emergent challenges of computational agency. *Colo. Tech. LJ*, 13:203.

Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *science*, 359(6380):1146–1151.

Wajdi Zaghouani, Md Rafiul Biswas, Mabrouka Bessghaier, Shimaa Ibrahim, George Mikros, Abul Hasnat, and Firoj Alam. 2025. Mahed shared task: Multimodal detection of hope and hate emotions in arabic content. In *Proceedings of the third arabic natural language processing conference: shared tasks*, pages 560–574.