Reddit-V: A Virality Prediction Dataset and Zero-Shot Evaluation with Large Language Models

Samir El-amrany, Matthias R. Brust, Salima lamsiyah, Pascal Bouvry

Faculty of Science, Technology and Medicine (FSTM), University of Luxembourg

{samir.el-amrany, matthias.brust, salima.lamsiyah, pascal.bouvry}@uni.lu

Abstract

We present Reddit-V, a new dataset designed to advance research on social media virality prediction in natural language processing. The dataset consists of over 27,000 Reddit posts, each enriched with images, textual content, and pre-engagement metadata such as post titles, categories, sentiment scores, and posting times. As an initial benchmark, we evaluate several instruction-tuned large language models (LLMs) in a zero-shot setting, prompting them with post titles and metadata to predict post virality. We then fine-tune two multimodal models, CLIP and IDEFICS, to assess whether incorporating visual context enhances predictive performance. Our results show that zero-shot LLMs perform poorly, whereas the fine-tuned multimodal models achieve better performance. Specifically, CLIP outperforms the best-performing zero-shot LLM (CodeLLaMA) by 3%, while IDEFICS achieves an 7% improvement over the same baseline, highlighting the importance of visual features in virality prediction. Despite these improvements, the task remains challenging, with none of the models surpassing 50% accuracy. We release the Reddit-V dataset and our evaluation results to facilitate further research on multimodal and text-based virality prediction. Our dataset and code will be made publicly available on Github[1].

1 Introduction

Social media platforms like Reddit, Twitter, and Instagram can propel a simple post to global visibility in a matter of hours. This rapid spread, known as virality shapes public discourse, drives marketing campaigns, and influences the design of recommendation systems (Berger and Milkman, 2012; Weng et al., 2013). Despite its importance, predicting virality before a post gains traction remains an open challenge, since most successful methods rely on early engagement signals or detailed network information that are unavailable at publish time (Gao et al., 2021).

Previous research has primarily used retrospective features such as like counts, comment volumes or diffusion graphs to forecast which content will catch on (Doerr et al., 2012; Gao et al., 2021). Although effective for analysis after the fact, these approaches offer little guidance to creators and platforms who need to decide what to publish next. A handful of studies have experimented with combining text and image data to improve prediction (Singhal et al., 2019; El-Amrany et al., 2024; El-amrany et al., 2025; Xu and Qian, 2023), but they depend on large, specialized datasets, often proprietary or unreleased and on complex model pipelines that require resource-intensive components (e.g., custom visual encoders, graph construction frameworks) and extensive preprocessing. Implementation details of these methods such as data-collection procedures, feature engineering scripts, and hyperparameter configurations are frequently omitted or under-specified, making exact replication infeasible.

To address the lack of resources designed specifically for pre-engagement prediction, we present Reddit-V, a new dataset of over 27,000 Reddit posts labeled viral or non-viral based on top-percentile thresholds for upvotes and comments. Unlike existing collections, Reddit-V focuses on text and multimodal classification: each example provides only the post title, image and basic metadata available shortly after posting (within 1 hour), such as subreddit category, posting time and subscriber counts.

Using Reddit-V, we evaluated two complementary strategies. First, we test state-of-the-art instruction-tuned language models without any task-specific training, by crafting natural-language prompts that ask the models to predict whether a post will go viral (Brown et al., 2020; Kojima et al.,

[1] https://github.com/EL-Amrany/Reddit-V

Proceedings of Recent Advances in Natural Language Processing,pages 334–341
Varna, Sep 8–10, 2025
https://doi.org/10.26615/978-954-452-098-4-041

2022). Second, we fine-tune vision-language architectures (CLIP; (Radford et al., 2021), 2021 and IDEFICS; [2]) on both the title text and any associated image, to see if adding visual context can boost accuracy. Our results demonstrate that integrating images through multimodal fine-tuning increases F1 scores by up to 10% compared to zero-shot, text-only prompting, proof that visual signals provide meaningful gains in early virality prediction.

We make three contributions. First, we introduce Reddit-V, the first public dataset for pre-engagement virality prediction in both text-only and multimodal settings. Second, we define zero-shot benchmarks on modern language models, revealing their current blind spots in anticipating viral reach. Third, we show that fine-tuning vision-language architectures with paired text and image data yields substantial performance improvements, illustrating the practical value of visual features.

The remainder of the paper is organised as follows. Section 2 reviews related work; Section 3 describes the dataset creation process; Section 4 details the evaluation methods, presents results, and discusses their implications; Section 5 outlines the limitations of this work; and Section 6 concludes the paper.

2 Related Work

Research on social media virality prediction has traditionally focused on leveraging retrospective engagement signals and network structure to forecast the eventual spread of content. Early studies demonstrated that temporal patterns of initial interactions and the topology of user communities strongly influence whether a meme or post will "go viral" (Weng et al., 2013) and that diffusion graphs capturing shares, likes, and comment volumes can be effective predictors of cascade formation (Doerr et al., 2012; Gao et al., 2021). However, these approaches depend on post-hoc features such as early engagement counts or detailed social graphs that are inherently unavailable at the moment of publication. Consequently, they offer limited utility for content creators and recommendation systems that must decide what to publish or promote before any user feedback has accrued.

A parallel line of work has explored multimodal fusion, combining textual and visual features to enhance predictive accuracy. For instance, recent frameworks integrate image embeddings with metadata to detect misleading or "fake" content, reporting performance gains over text-only baselines (El-Amrany et al., 2024; Xu and Qian, 2023). Yet these methods rely on large, proprietary datasets and complex architectures that hinder reproducibility and generalization. Moreover, they typically frame virality as a by-product of detection tasks rather than as a primary prediction objective, and they seldom isolate pre-engagement features in a controlled experimental setting.

Meanwhile, the rise of instruction-tuned LLMs has raised the possibility of zero-shot classification across a wide array of tasks. Research showed that generative models like GPT-3 can perform translation, summarization, and sentiment analysis purely through carefully crafted prompts (Brown et al., 2020), and that encouraging chain-of-thought reasoning further refines zero-shot performance (Kojima et al., 2022). Despite these advances, virality prediction remains underexplored in a zero-shot context, and it is unclear whether structured metadata (e.g. sentiment scores, posting times, or subreddit subscriber counts) can be meaningfully interpreted by general-purpose LLMs to anticipate viral outcomes.

A third stream leverages vision language pre-training to ground textual and visual representations in large multimodal datasets. CLIP demonstrated that aligning image and text embeddings yields strong zero-shot transfer on classification benchmarks (Radford et al., 2021) while IDEFICS extended this paradigm with a transformer-based fusion architecture optimised for diverse multimodal tasks (Lee et al., 2022). However, neither model has been systematically fine-tuned for the specific challenge of predicting social media virality using pre-publication features, leaving open the question of how well visual context can compensate for the absence of early engagement signals.

Taken together, these strands of prior research reveal a clear gap: no publicly available resource enables rigorous, reproducible evaluation of both text-only and multimodal approaches to pre-engagement virality prediction. Existing datasets either lack comprehensive metadata and multimodal content or constrain models to retrospective features and proprietary data. Our work addresses this need by introducing Reddit-V, a dataset of more than 27,000 Reddit posts annotated with only those attributes accessible immediately after posting, and by establishing both zero-shot LLM benchmarks and fine-tuned vision-language base-

[2] https://huggingface.co/blog/idefics

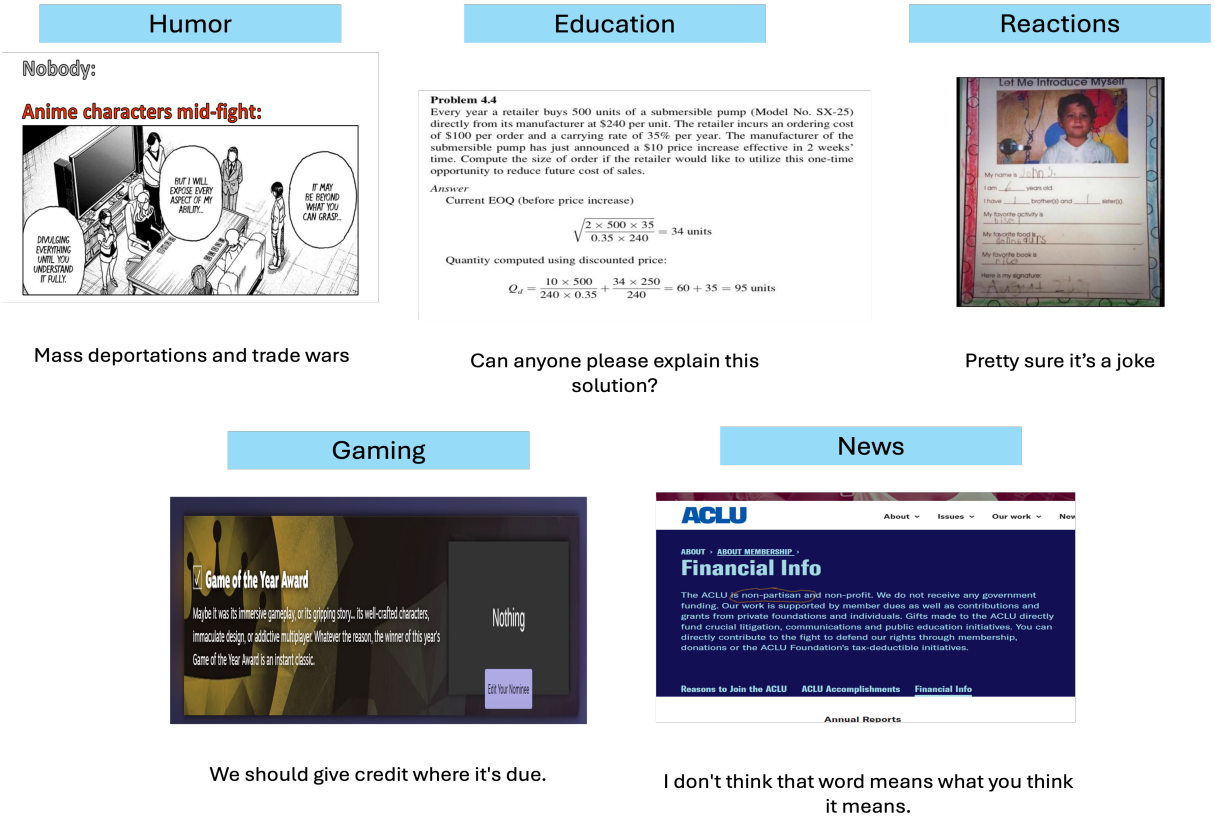

Figure 1: Examples of posts from each category.

lines. In doing so, we provide the first unified framework for assessing how textual and visual pre-publication signals contribute to the emergence of viral content.

3 Dataset

To evaluate virality before any user feedback, we introduce Reddit-V, a dataset of Reddit submissions comprising each post's image, text, and key metadata. Posts are drawn from five categories including humor, reaction, educational, gaming, and news, to cover the main types of entertainment content on the platform. By bringing together each post's visual and textual characteristics, Reddit-V enables the study of how these features alone influence early engagement on Reddit.

3.1 Data Collection

We retrieved 27,587 posts published between January 2018 and December 2023 from five representative subreddits, selected to cover diverse entertainment modalities:

- **Humor** (e.g. r/funny)
- **Educational** (e.g. r/todayilearned)
- **Reaction** (e.g. r/reactiongifs)
- **Gaming** (e.g. r/gaming)
- **News** (e.g. r/worldnews)

The above communities were chosen to capture a mix of playful content (Humor, Reaction), informative snippets (Educational, News), and community-driven gameplay discussion (Gaming). Examples from each category are shown in Figure 1. Using the Python Reddit API Wrapper (PRAW) [3], we fetched up to 10,000 of the latest posts from each subreddit, then kept only those with at least one image file (JPEG, PNG, or GIF). For every post, we recorded data that would be available immediately after it went live, namely, the title, the subreddit's subscriber count, the timestamp, and basic image details (format, dimensions, file size).

3.2 Data Preprocessing

Consistent with prior work that operationalizes virality via top-percentile engagement (Berger and Milkman, 2012; Weng et al., 2013), we label a post as *viral* if it simultaneously falls within the top

[3] https://praw.readthedocs.io/en/stable/

20% of upvote counts and the top 20% of comment counts relative to all posts. We chose the 20% cutoff for three reasons:

- **Literature alignment**: Focusing on the upper quintile aligns with established methodologies for isolating highly engaging content while excluding moderate-performance outliers (Berger and Milkman, 2012).
- **Statistical power**: A 20% threshold yields a sufficient volume of viral examples to support robust training and evaluation, without diluting the label with marginally engaging posts.
- **Empirical robustness**: Sensitivity analyses using 15% and 25% cutoffs produced qualitatively similar class balances and model performance trends, indicating that our findings are not idiosyncratic to the exact threshold choice.

All remaining posts are labeled as *non-viral*. The dataset statisitics are summarized in Table 1.

Table 1: Summary of dataset statistics.

Statistic	Value
Total Posts	27,587
Number of Subreddits	112
Number of Tokens	205,529
Number of Non Viral Posts	24,090
Number of Viral Posts	3,497
Average Title Length (characters)	72
Average Comments per Post	15
Average Upvotes per Post	120

4 Experimental Results

This section consolidates our zero-shot evaluation of LLMs and the fine-tuning results for multimodal baselines into a unified experimental framework. We first detail the overall evaluation protocol, then present results for the LLM benchmarks and the vision-language models, and conclude with a concise discussion of comparative performance and implications.

4.1 Experimental Setup

All experiments use the Reddit-V dataset described in Section 3. For zero-shot LLM evaluation, models receive exactly one inference pass per example, without any parameter updates. By contrast, the multimodal models (CLIP and IDEFICS) are fine-tuned using a train-validation-test split of 70%-10%-20%, stratified by subreddit and virality label to preserve class balance. We report results on the held-out 20% test partition.

Zero-Shot Prompt Template We convert each post into a plain-text prompt that concatenates the title and pre-publication metadata fields (cross-post count, comment sentiment, subreddit name, title sentiment, first-hour upvotes, subscriber count, posting date). A representative prompt appears below:

> **Instruction:** Predict whether the following Reddit post will go *viral* or *non-viral*.
>
> **Title:** "<post title>"
> **Crossposts:** <value>
> **Comment Sentiment:** <value>
> **Subreddit:** <name>
> **Title Sentiment:** <value>
> **Upvotes (1h):** <value>
> **Subscribers:** <value>
> **Date:** <YYYY-MM-DD>
>
> **Answer:** "viral" or "non-viral."

4.2 LLM Baseline Models

We evaluate the following publicly available language models via the Ollama interface. Each model is used in a true zero-shot setting with no examples or fine-tuning beyond the prompt template described in Section 4.1.

- **Gemma3** (Team et al., 2025): A family of transformer-based models (1 B-12 B parameters) released by the Gemma team. We test the base checkpoint (architecture only) and two parameter-specific variants, none of which have been instruction-tuned.
- **LLaMA 3.2** (Grattafiori et al., 2024): An instruction-tuned update of Meta's LLaMA 3 series, estimated at 7-13 B parameters. Designed for broad NLP tasks, it includes conversational fine-tuning.
- **Phi4-mini** (Phi-2) [4]: A compact, instruction-tuned model (< 1.3 B parameters) from Microsoft Research, optimised for efficient zero-shot performance on classification and reasoning tasks.
- **Mistral 7B** (Jiang et al., 2023): A 7 B-parameter instruction-tuned model known for strong performance on standard language benchmarks without additional fine-tuning.

[4] https://www.microsoft.com/en-us/research/blog/phi-2-the-surprising-power-of-small-language-models/

- **Starling-LM** (Zhu et al., 2024): A 7 B-parameter, LLaMA-based model with instruction-tuning aimed at dialogue and question-answering applications.
- **CodeLLaMA 7B** (Roziere et al., 2023): A variant of LLaMA fine-tuned on code corpora. Although not instruction-tuned for conversational tasks, its text understanding capabilities make it an informative baseline.
- **LLaVA 7B** (Liu et al., 2023): A vision-language model built on LLaMA and CLIP, instruction-tuned to accept image inputs alongside text prompts.

4.3 LLM Zero-Shot Performance

We interpret any model response beginning with "viral" (case-insensitive) as the positive class; all others are "non-viral." We compute accuracy, precision, recall, and the balanced F1 score to account for class imbalance. Table 2 reports the balanced F1 for each LLM.

Table 2: Zero-Shot LLM Balanced F1 Scores

Model	Balanced F1
Gemma3	0.335
Gemma3-12B	0.339
LLaMA 3.2	0.369
Phi4-mini	0.339
Mistral 7B	0.398
Starling-LM	0.333
CodeLLaMA 7B	0.399
LLaVA 7B	0.333

The zero-shot experiments reveal that off-the-shelf language models struggle to use pre-publication metadata and post titles to predict virality. The best-performing model, CodeLLaMA 7B, achieves a F1-score of 0.399. All other models cluster between 0.33 and 0.40 in F1-score, indicating consistent shortcomings across architectures and parameter scales.

Three factors help to explain this outcome. First, instruction-tuned LLMs are optimised for general natural language understanding and following user instructions, not for extracting statistical signals from structured metadata. Features like "first-hour upvotes" or "subscriber count" lack the rich contextual patterns (word usage, syntax, narrative elements) at which these models excel. As a result, LLMs default to heuristic reasoning, often over-relying on title sentiment or subreddit name rather than learning reliable associations between numeric fields and eventual engagement.

Second, the flat prompt format limits the model's ability to compare values systematically. Presenting metadata as a list of text lines places the burden on the model to interpret relative magnitudes ("Is 1,200 subscribers high or low?") without any reference distribution. In a few-shot or fine-tuned setting, the model could calibrate its internal scale, but in zero-shot mode it lacks the grounding necessary to distinguish, say, whether 50 upvotes in the first hour signifies likely virality.

Third, virality is driven by complex interactions among content, timing, and community dynamics, factors that are not fully captured by static metadata. For example, a humorous image posted at peak hours in a highly active subreddit can outperform a similar post in a smaller community, even with identical features. Zero-shot LLMs have no mechanism for weighting these contextual differences, so they treat all metadata fields with equal or arbitrary importance.

Taken together, these limitations explain why generic instruction tuning does not translate into effective virality prediction from pre-publication signals alone. Improving performance will require either (1) prompt strategies that embed reference distributions or comparative examples, (2) few-shot calibration runs to teach the model how to interpret numeric ranges, or (3) fine-tuning on a labelled virality dataset. Until such targeted interventions are applied, zero-shot LLMs should be viewed as coarse filters rather than reliable predictors in content-recommendation pipelines.

4.4 Multimodal Fine-Tuned Models

We fine-tune CLIP (Radford et al., 2021) and IDEFICS [5] on our train split, using cross-entropy loss and early stopping on validation F1. Both models ingest paired image embed-

[5] https://huggingface.co/blog/idefics

dings and textual metadata. Table 3 presents test-set performance.

Table 3: Fine-Tuned Multimodal Model F1 Scores

Model	Balanced F1
CLIP	0.475
IDEFICS	0.436

Both CLIP and IDEFICS, when fine-tuned on Reddit-V, achieve significantly higher F1 scores than any zero-shot LLM (paired t-test, $p < 0.01$). This gap highlights two key factors:

First, **visual context** carries non-redundant information that pure text prompts cannot convey. Images capture tone, layout, and subject matter qualities that often trigger rapid engagement but are invisible to text-only models. By incorporating image embeddings, our fine-tuned baselines learn to associate visual patterns (e.g., bright colors, human faces, or recognisable memes) with higher likelihoods of virality.

Second, **task-specific training** calibrates the model's decision boundary around the precise characteristics of our dataset. Zero-shot LLMs apply general language understanding and may misinterpret metadata fields or fail to weigh them appropriately. In contrast, fine-tuning optimises feature representations and classifier weights jointly, yielding a model that is sensitive to the combination of image cues and textual metadata most predictive of high engagement.

4.5 Discussion

The comparative evaluation shows a clear performance gap between zero-shot LLMs and fine-tuned multimodal models in predicting Reddit post virality prior to any engagement. Zero-shot LLMs achieve F1 scores around 0.33-0.40, indicating that general-purpose language understanding alone cannot reliably infer which posts will attract disproportionate attention. In contrast, CLIP and IDEFICS fine-tuned on Reddit-V exceed an F1 of 0.43, confirming that visual information and task-specific adaptation materially strengthen predictive power.

The results reveal two practical implications. First, incorporating image embeddings is essential: visual features carry signals such as color contrast, facial expressions, or meme templates, these signals often drive user interaction yet are inaccessible to text-only systems. Second, task-specific training aligns model representations with the target distribution, enabling the classifier to weight metadata fields and visual information according to their true relevance for virality. Together, these factors produce a model that operates meaningfully in a pre-publication setting.

From an application standpoint, the findings suggest that content recommendation or creation tools should integrate lightweight fine-tuning on representative examples rather than relying exclusively on prompt engineering. It is noteworthy, that prompt-based zero-shot methods may still serve as rapid baselines or for out-of-distribution checks, but they cannot match the accuracy demanded by real-time decision systems.

5 Limitations

This work explores a narrow, well-defined setup, and there are a few limitations worth keeping in mind. The models we evaluated were not specifically trained for virality prediction. Since they were used in a zero-shot setting, their understanding of the task depended entirely on how they interpreted the prompt, without any exposure to examples or fine-tuning.

The prompt used in this work was kept fairly structured: it included the post title and a small set of metadata features. That design made the evaluation process more consistent across models, but it may have also constrained their ability to reason about factors that do not fit neatly into structured inputs, for example, whether the post touches on a trending topic, reflects community sentiment, or appears at a moment when the audience is particularly active.

The way in which we evaluated the performance of the model is also something to keep in perspective. To obtain a comprehensive comparison, we utilized F1 scores; however, this metric does not always capture the entire

picture, particularly when one class presents greater prediction difficulties or when certain predictions are slightly inaccurate.

It is important to note that the data set was constructed exclusively using Reddit content, which implies that our results are related to Reddit structure, user behavior, and engagement signals. Other platforms, such as Twitter or TikTok, function quite differently. Even the way we defined virality (i.e. based on upvotes and comments) captures just one version of what 'going viral' might look like, and there is room to explore alternative approaches in future work.

6 Conclusion

This paper introduces Reddit-V, a new dataset designed specifically for research on natural language processing (NLP)-based research into social media virality prediction. Reddit-V includes over 27,000 Reddit posts annotated with titles, categories, sentiment, posting times, and clearly defined virality labels. This structured metadata allows for realistic evaluations, where predictions must rely exclusively on information available before users engage with the posts.

We conducted initial evaluations using two approaches. First, we evaluated several instruction-tuned LLMs in a zero-shot scenario, where the models predicted virality based on textual content and pre-engagement metadata only. Second, we fine-tuned two multimodal models (i.e. CLIP and IDEFICS) using both images and associated textual information. These models performed notably better, demonstrating that visual context can improve virality prediction compared to text-only zero-shot methods.

Reddit-V, along with these baseline results, provides a practical resource for researchers aiming to improve virality prediction models. Future studies can build upon this work by exploring multimodal fusion strategies, richer contextual embeddings, or improved prompting techniques. By openly releasing both the dataset and our evaluation methods, our objective is to support reproducible advances in understanding and predicting content virality on social networks.

Ethical Considerations

All data in this study were collected from publicly available Reddit posts. No personally identifying information (such as usernames or email addresses) was retained in the dataset, and analyses focused on aggregate metrics (e.g., upvotes, comments). This approach minimizes potential risks to user privacy while still enabling examination of engagement patterns.

References

Jonah Berger and Katherine L Milkman. 2012. What makes online content viral? *Journal of marketing research*, 49(2):192–205.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Benjamin Doerr, Mahmoud Fouz, and Tobias Friedrich. 2012. Why rumors spread so quickly in social networks. *Communications of the ACM*, 55(6):70–75.

Samir El-Amrany, Matthias R Brust, Johnatan E Pecero, and Pascal Bouvry. 2024. Tri-fusiondet: Leveraging user engagement, textual, and visual features for enhanced fake news detection. In *2024 28th International Computer Science and Engineering Conference (ICSEC)*, pages 1–6. IEEE.

Samir El-amrany, Salima Lamsiyah, Matthias R Brust, and Pascal Bouvry. 2025. Guardharmem and harmdetect: a multimodal dataset and benchmlark model for fine-grained harmful meme classification. *Social Network Analysis and Mining*, 15(1):63.

Liqun Gao, Yujia Liu, Hongwu Zhuang, Haiyang Wang, Bin Zhou, and Aiping Li. 2021. Public opinion early warning agent model: A deep learning cascade virality prediction model based on multi-feature fusion. *Frontiers in Neurorobotics*, 15:674322.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Albert Q Jiang, A Sablayrolles, A Mensch, C Bamford, D Singh Chaplot, Ddl Casas, F Bressand, G Lengyel, G Lample, L Saulnier, et al. 2023. Mistral 7b. arxiv. *arXiv preprint arXiv:2310.06825*, 10.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

Doyup Lee, Chiheon Kim, Saehoon Kim, Minsu Cho, and Wook-Shin Han. 2022. Autoregressive image generation using residual quantization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11523–11532.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. arXiv preprint arXiv:2103.00020.

Baptiste Roziere, Jonas Gehring, Fabian Gloeckle, Sten Sootla, Itai Gat, Xiaoqing Ellen Tan, Yossi Adi, Jingyu Liu, Romain Sauvestre, Tal Remez, et al. 2023. Code llama: Open foundation models for code. *arXiv preprint arXiv:2308.12950*.

Shivangi Singhal, Rajiv Ratn Shah, Tanmoy Chakraborty, Ponnurangam Kumaraguru, and Shin'ichi Satoh. 2019. Spotfake: A multi-modal framework for fake news detection. In *2019 IEEE fifth international conference on multimedia big data (BigMM)*, pages 39–47. IEEE.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

Lilian Weng, Filippo Menczer, and Yong-Yeol Ahn. 2013. Virality prediction and community structure in social networks. *Scientific reports*, 3(1):1–6.

Zhixuan Xu and Minghui Qian. 2023. Predicting popularity of viral content in social media through a temporal-spatial cascade convolutional learning framework. *Mathematics*, 11(14):3059.

Banghua Zhu, Evan Frick, Tianhao Wu, Hanlin Zhu, Karthik Ganesan, Wei-Lin Chiang, Jian Zhang, and Jiantao Jiao. 2024. Starling-7b: Improving helpfulness and harmlessness with rlaif. In *First Conference on Language Modeling*.