LARGE LANGUAGE MODELS ARE NATURAL VIDEO POPULARITY PREDICTORS

Anonymous authors
Paper under double-blind review

ABSTRACT

Predicting video popularity is typically formalized as a supervised learning problem, where models classify videos as popular or unpopular. Traditional approaches rely heavily on meta-information and aggregated user engagement data, but video popularity is highly context-dependent, influenced by cultural and social factors that such approaches fail to capture. We argue that Large Language Models (LLMs), with their deep contextual awareness, are well-suited to address these challenges. However, bridging the modality gap between pixel-based video data and token-based LLMs is a key challenge. To address this, we transform frame-level visual data into sequential text representations using Vision-Language Models (VLMs), enabling LLMs to process multimodal video content—titles, frame-based descriptions, and captions—capturing both engagement intensity (view count) and geographic spread (the number of countries where a video trends). Evaluating on 13,639 videos, we show that while a supervised neural network using content embeddings achieved 80% accuracy, our LLM-based method reached 82% without fine-tuning. A combined approach, integrating the neural network's predictions into the LLM, further improved accuracy to 85.5%. Additionally, the LLM generates interpretable hypotheses explaining its predictions based on theoretically sound attributes. Manual validations confirm the quality of these hypotheses and address concerns about hallucinations in the video-to-text conversion process. Our findings highlight that LLMs, equipped with textually transformed multimodal representations, offer a powerful, interpretable, and data-efficient solution to tasks, requiring rich contextual and cultural insights, such as video popularity prediction.

1 INTRODUCTION

Video consumption now accounts for the majority of internet traffic and continues to grow rapidly, making video popularity prediction a critical challenge for content creators, social media platforms, and advertisers (Cisco, 2021). Accurately predicting which videos will become popular is not only important for these stakeholders but also provides valuable insights for researchers studying information diffusion (Park et al., 2017; Rezvanian et al., 2023), social influence (Park et al., 2016; Lin et al., 2023b), cultural dynamics (Park et al., 2017; Haldar et al., 2023), and potential misuse in online networks (Beni et al., 2023). Despite its significance, predicting video popularity remains a complex and challenging task due to the wide range of influencing factors, such as historical context (Ng & Taneja, 2023), cultural trends (Park et al., 2017), and even emotional engagement with viewers (Guadagno et al., 2013; Park et al., 2016). The vast diversity and volume of video content online further complicate these challenges.

Despite the growing interest in this area, most existing research has employed statistical and supervised learning approaches that primarily rely on meta-information and aggregated user metrics, such as uploader statistics, view/comment/like counts, and external factors like social network size and early engagement in other platforms (Zhou et al., 2010; Shamma et al., 2011; Borghol et al., 2012; Park et al., 2016). While these factors provide useful signals, they mainly reflect the uploader's reputation or early user reactions and often fail to capture the deeper contextual and cultural significance encoded within the video content. These intrinsic qualities of video content may play a critical role in how a video resonates with both local and global audiences. However, traditional approaches struggle to process and leverage such rich information due to limited technological capacity in processing complex multimodal data.

In this paper, we propose a novel approach to video popularity prediction that shifts the focus to the intrinsic qualities of a video's textual, verbal, *and* visual content, excluding after-the-fact user engagement data such as early view counts and social network signals. Our method leverages the power of Vision-Language Models (VLMs) and Large Language Models (LLMs) to extract and interpret these intrinsic qualities, complemented by conventional descriptors such as titles and descriptions. To address the challenge of integrating pixel-based video data with the token-based architecture of LLMs, we use VLMs to transform frame-level visual data into sequential textual representations. These representations are then combined with conventional video descriptors such as titles, descriptions, and captions (extracted from the video's audio) to create a comprehensive multimodal textual representation. This transformation allows LLMs to perform the video popularity prediction task, effectively capturing both vertical aspects of popularity, such as view counts, and horizontal aspects, such as the global reach of the video across different countries, by utilizing the deep contextual understanding embedded in LLMs.

Empirically, we introduce a prompting strategy that incorporates a novel approach of integrating supervised learning signals into LLM-based predictions. Our method not only outperforms traditional deep learning models based on content embeddings but also provides human-interpretable predictions in the form of attribute-based hypotheses. Additionally, we introduce the *Global Popular Video Dataset (GPVD)*, a large-scale dataset of 1.3M unique popular YouTube videos, enriched with titles, descriptions, and detailed metadata. The dataset uniquely includes three key popularity metrics: view counts, the number of countries where the video trended, and the number of days it remained on trending lists. The latter two metrics, which reflect the video's global reach and sustained popularity, have been overlooked in prior research but are crucial for a comprehensive understanding of video virality. For our experiments, we subsampled this dataset to create a balanced subset that captures different popularity classes, reflecting both engagement intensity and geographic spread, as detailed in the Methods section. Furthermore, we address concerns related to hallucinations in video-to-text conversion and validate the quality of attribute-based hypotheses through survey experiments.

The key contributions of this paper are as follows:

1. We formulate a video popularity prediction task that not only accounts for engagement intensity (e.g., view counts), commonly used in prior work, but also incorporates geographic spread as a critical dimension of popularity.
2. We introduce the *GPVD*, comprising 1.3M representative popular YouTube videos from 109 countries, supplemented with various metadata to support future studies in the field.
3. We propose a framework combining VLMs and LLMs to predict video popularity by transforming multimodal video content into textual representations. This pipeline innovatively integrates supervised learning signals into LLM-based predictions, enhancing performance.
4. We systematically explore and evaluate the impact of different prompting techniques, such as hypothesis generation, KNN-based example retrieval, and supervised signals while also providing human-interpretable predictions validated through systematic human evaluations.

2 Related Work

Video Popularity Prediction Video popularity prediction has traditionally been treated as a supervised learning task, focusing primarily on predicting view counts based on factors like title length, runtime, and user engagement (Zhou et al., 2010; West, 2011; Borghol et al., 2012; Wang et al., 2012). Later work introduced more sophisticated methods, such as time-series analysis and user behavior modeling, to track how popularity evolves (Broxton et al., 2013; Pinto et al., 2013; Vallet et al., 2015; Park et al., 2016; Jog et al., 2021). Despite these advancements, video popularity prediction remains challenging, as it demands capturing cultural and social dynamics that traditional approaches, which often focus narrowly on aggregated platform metrics, fail to account for.

Moreover, these studies typically treat video popularity as a unidimensional problem, focusing exclusively on view counts. However, high view counts alone do not fully capture a video's popularity—especially in the era of streaming, where a video with many views may not necessarily be a global hit; it could be localized to specific regions. To address this limitation, we redefine the prediction task by incorporating both engagement intensity (view counts) and geographic spread (global-local reach),

making the task more nuanced and realistic. We argue that LLMs, with their extensive contextual awareness, are uniquely positioned to capture these complexities, encoding cultural and contextual subtleties that previous studies have been unable to leverage.

Multimodal Learning and Modality Gaps Integrating multimodal data for tasks like video understanding has long been a challenge, particularly in bridging the modality gap between pixel-based visual data and token-based language models. Existing approaches typically employ modular architectures that use pre-trained visual encoders (e.g., ViT) and language models (e.g., BERT) to process visual and textual data independently, before concatenating their embeddings (Zeng et al., 2022; Alayrac et al., 2022; Li et al., 2023a;b; Lu et al., 2022). While these methods improve performance, they fail to fully capture the complex interactions between visual and textual modalities, notably the rich temporal and contextual relationships in videos (Chen et al., 2023; Qin et al., 2023).

Our approach tackles this problem by introducing a *frame-to-text transformation* that converts video frames into sequential textual descriptions using VLMs. This enables LLMs to process the visual information as richly contextualized text, allowing them to reason across modalities in a unified format. By treating video frames as text, we bridge the modality gap and leverage the LLMs' powerful reasoning capabilities, achieving deeper integration of multimodal data and capturing the complex interactions between visual and textual content more effectively than modular architectures that simply concatenate embeddings.

Natural Language Explanation and Prompt Engineering Generating explanations alongside predictions has been shown to enhance both model understanding and performance in complex tasks. Techniques like Chain-of-Thought prompting (Wei et al., 2022) and self-consistency sampling (Wang et al., 2022) have demonstrated how reasoning chains can improve model accuracy while maintaining interpretability. Two-stage approaches, such as hypothesis generation followed by task solving (Wang et al., 2023), suggest that explanations contribute to better performance, but these approaches often add complexity to the prediction pipeline.

Building on the work of Hanu et al. (2023), who demonstrated the use of textual descriptions for multimodal classification, we extend this approach by integrating hypothesis generation directly into the prediction process, reducing the two-stage approach into a more efficient one-step process. Additionally, we incorporate the supervised learning prediction outcomes into the prompt as additional signals to further improve performance. By combining prompt engineering with contextualized inputs from our frame-to-text transformation, we enhance both the prediction accuracy and the generation of interpretable hypotheses, creating a unified and efficient approach to video popularity prediction.

3 METHODS

The idea behind our approach is to transform video content into a sequence of textual descriptions in order to enable video popularity prediction through LLMs (see Appendix A.1 for an overview of the pipeline and further details). This approach involves a multi-step process. Let $\mathbf{V} = \{f_1, f_2, ..., f_M\}$ represent the set of key frames for a video, where M is the number of frames. We define $\mathbf{C} = \{C_1, C_2, ..., C_C\}$ as the set of captions, T as the title, and D as the description provided by the user. We first perform frame-to-text transformations using *VideoLLava* model (Lin et al., 2023a), which effectively captures the essence of the video content. The output of this model is represented as $S_{\text{VideoLLava}}(\mathbf{V}) = S_1, S_2, ..., S_S$, where $S_i = \text{VideoLLava}(f_i^W)$ and f_i^W represents a frame window around f_i. To further enhence the textual representation, we integrate time-matched captions to provide additional context. This integration ensures that each segment of frames is accurately described in both visual and verbal terms, improving the overall representation with extended contextual details. We then establish baseline performance using supervised multimodal models. Finally, we introduce our novel LLM-based method, which leverages the reasoning capabilities of LLMs to predict video popularity. In the following subsections, we provide a detailed step-by-step explanation of each component of our methodology.

3.1 TASK DEFINITION AND DATASET

In this work, we redefine the video popularity prediction task by focusing on two key dimensions: *engagement intensity* and *geographic spread*. Traditional approaches have typically focused on

predicting view counts, but this often provides an incomplete picture of a video's overall success. A video may garner high views without achieving significant global reach, especially in the era of streaming, where content can be popular in localized regions without becoming a global hit. Our approach captures both:

- **Engagement Intensity**: The total number of views a video receives, reflecting its overall reach and audience engagement.
- **Geographic Spread**: The number of countries where the video trends, indicating its global reach and resonance.

By incorporating these dimensions, we offer a more nuanced understanding of video popularity. Formally, given a video v and its content features (e.g., frames, audio, captions), we aim to predict a popularity score $p(v) \in [0, 1]$, representing the likelihood of the video being classified as either a 'local hit' or a 'global big hit.'

To support this prediction task, we introduce the *Global Popular Video Dataset (GPVD)*, which comprises the top 50 trending videos for each of 109 countries for 589 days between February 13, 2021 and March 17, 2023 (approximately 5,450 observations per day), resulting in a total of 3,210,050 observations and 1,302,698 unique videos. Each observation includes the unique ID of the video, the countries where it was trending, its category, title, tags, and popularity metrics, including views, likes, and dislikes. [1] Given that the majority of videos neither go viral nor achieve significant consumption levels, gathering a representative sample of globally popular videos is inherently challenging. Previous studies collected similar datasets but over much shorter periods (1-5 months) (Park et al., 2017; Ng & Taneja, 2023), whereas our dataset spans more than two years with larger coverage of countries.

This dataset captures both short-term trends and long-term patterns across diverse geographic regions, allowing for cross-cultural comparisons of popularity. Based on the two key dimensions—*Engagement Intensity* and *Geographic Spread*—we classify videos into 16 categories using 4×4 quantiles. Two primary classes emerge from this classification:

- **Global Big Hit**: Videos in the top 25% for both views and geographic spread.
- **Local Hit**: Videos in the bottom 25% for both dimensions, indicating localized popularity.

This classification enables us to analyze the characteristics that distinguish globally highly popular videos from those with regional appeal. For our experiments, we selected a random sample from the dataset, which contains 6,279 videos classified as 'Local Hit' and 7,360 as 'Global Big Hit,' to ensure balanced experimentation.

3.2 Transforming Video Content into Text

Our approach for transforming video content into text for LLM-based video popularity prediction follows a structured multi-step process:

1. **Frame Extraction**: We define $\mathbf{V} = f_1, f_2, ..., f_M$ as the set of equally spaced frames extracted from the video, with 10 frames per minute selected to ensure comprehensive coverage of the video content.
2. **Frames to Text Conversion**: The *VideoLLava* model (Lin et al., 2023a) is then applied to these frames to generate descriptive textual representations. The model, pretrained to capture the essence of visual content, produces contextually relevant descriptions for each frame. The output is represented as $S_{\text{VideoLLava}}(\mathbf{V}) = \{S_1, S_2, ..., S_S\}$, where each S_i is a textual description generated from a window of frames around f_i.
3. **Caption Matching and Data Integration**: The sequential frame-based textual descriptions are aligned with the video's existing captions, ensuring that visual and verbal elements are

[1] We will release the dataset, including video IDs, metadata, and Python code for video downloading, in a publicly accessible repository upon the paper's acceptance. This will ensure reproducibility and usability for future research.

synchronized. This step improves contextual accuracy by combining frame descriptions with corresponding timestamps. The frame descriptions and captions are then integrated into a cohesive textual narrative that comprehensively represents the video content, incorporating both visual and temporal aspects.

4. **Summarization**: The integrated textual data is summarized using an LLM ($\Phi_{\texttt{LLM}}$), which refines the content into a polished final text. This step eliminates redundancies and ensures a coherent narrative. The LLM processes the integrated summaries $\mathcal{I}$, along with the title T and description D, to generate a final text $\mathcal{F}$ that accurately reflects the video's content while maintaining narrative consistency.

Note that the *VideoLLava* model was chosen for its effectiveness in translating visual content into high-quality descriptive text, forming a strong foundation for subsequent steps such as caption matching and summarization. This approach improves the overall accuracy of video content representation, enhancing the effectiveness of our video popularity prediction model.

An excerpt of the summarising prompt is shown below:

```
You're an expert in YouTube videos with extensive experience in analyzing
video content and trends.
...
<output>
<scratchpad>
Step 1:  Identify key elements -...  Step 2:  Analyze Segment Transitions -
...
...
</scratchpad>
<complete_summary>
<segment_summaries>
<introduction> Provide an overview of the video's theme and initial setting.
</introduction>
...
</segment_summaries>
</output>
```

3.3 Video Popularity Prediction through Prompting

Our method for video popularity prediction leverages LLMs through a series of carefully designed prompts. These prompting techniques are conceptually categorized into three main sets: *context*, *reasoning*, and *transfer*, each playing a distinct role in improving prediction performance. While these sets provide a logical framework, our experiments proceeded sequentially, progressively incorporating each set to refine performance. The experimentation began with a simple vanilla LLM setup and systematically added features such as reasoning, few-shot learning, near examples, hypothesis generation, and supervised signals.

3.3.1 Overview of Prompt Components

1. **Context Set**: This set establishes the foundation for the task by including instructions, defining the task, and specifying the expected output format. It ensures the LLM understands what to predict and how to deliver predictions.
2. **Reasoning Set**: This set encourages the LLM to generate intermediate reasoning steps before making predictions, building on chain-of-thought and hypothesis generation literature. It includes prompts like "think before evaluating" and "hypothesis generation," which help the model process more complex information and provide explanations for its predictions.
3. **Transfer Set**: This set focuses on improving predictions by transferring knowledge from other examples. Techniques such as few-shot learning, near-example selection, and even supervised signals are used to provide the LLM with external examples and guidance to refine its predictions.

3.3.2 Sequential Prompting Approach

We frame the video popularity prediction task as a four-class classification problem, where classes $c \in 1, 2, 3, 4$ represent increasing levels of popularity, ranging from local to global hits. Specifically, classes 1 and 2 correspond to varying levels of local popularity (local hits), while classes 3 and

4 indicate broader, global popularity (global big hits). This granularity allows the LLM to better capture nuances in the video's geographical spread and engagement intensity, with classes 1 through 4 reflecting progressively broader and higher audience reach and engagement. For final predictions, we consolidate these into two categories: classes 1 and 2 as 'local hits,' and classes 3 and 4 as 'global big hits.' The input includes integrated summaries, titles, and descriptions of the video, denoted as $\mathcal{F}(\mathcal{I}, T, D)$.

The actual experimentation followed a progressive structure where each new component was added incrementally to refine performance. The sequence of prompt addition is described below:

Vanilla LLM Prompt (Context Set) The initial prompt, denoted as $\mathcal{P}_{\text{vanilla}}$, included the basic components necessary to perform the video popularity prediction task:

$$\mathcal{P}_{\text{vanilla}} = \mathcal{P}_{\text{instructions}} + \mathcal{P}_{\text{task}} + \mathcal{P}_{\text{output}} \tag{1}$$

where $\mathcal{P}_{\text{instructions}}$ provides general instructions to the LLM; $\mathcal{P}_{\text{task}}$ defines the video popularity prediction task; and $\mathcal{P}_{\text{output}}$ specifies the expected output format (i.e., the predicted popularity class). This vanilla prompt serves as the foundation for subsequent enhancements and this structure ensures that each component is seamlessly integrated into the final prompt.

Thinking (Context + Reasoning Set) To improve the prediction process, we added a reasoning step based on the Chain-of-Thought approach (Wei et al., 2022). This was done through the "thinking before evaluating" prompt, denoted as $\mathcal{P}_{\text{think}}$, which encouraged the LLM to process the data thoroughly before making a decision. This intermediate reasoning improved the model's capacity to interpret the complex relationships within the video content.

Few-shot Learning (Context + Reasoning + Transfer Set) Building on the reasoning step, we introduced few-shot learning (Brown et al., 2020) by providing the LLM with a small set of labeled examples $\mathcal{E} = \{(T_i, \mathcal{F}(\mathcal{I}_i, T_i, D_i), y_i)\}_{i=1}^{N}$, where $y_i \in \{1, 2, 3, 4\}$ is the popularity class of the i-th example video. This few-shot learning step helped the LLM make more informed predictions by allowing it to learn from analogous examples.

Near Examples (Context + Reasoning + Transfer Set) Next, we introduce a subset of examples $\mathcal{E}_{\text{near}} \subseteq \mathcal{E}$ that are semantically similar to the given video based on a similarity function $\Phi_{\text{sim}}(T, T_i)$, which computes the cosine similarity between the embeddings of the given video title T and each example video title T_i (Liu et al., 2021).

The textual modalities (title T) are processed using the MPNet base v2 encoder Song et al. (2020) $\mathcal{E}_{\texttt{MPNet}} : \mathcal{T} \rightarrow \mathcal{Z}$ to generate embeddings. We select the top-k examples with the highest similarity scores to form $\mathcal{E}_{\text{near}}$. The selected near examples are then incorporated into the prompt as follows: $\mathcal{P}_{\text{near}} = \mathcal{P}_{\text{vanilla}} + \sum_{(T_i, \mathcal{F}(\mathcal{I}i, T_i, D_i), y_i) \in \mathcal{E}\text{near}}$ where T_i, $\mathcal{F}(\mathcal{I}_i, T_i, D_i)$, and y_i represent title, full integrated description, and popularity, respectively.

This ensures that the most relevant examples are provided to the LLM as additional context to aid in making more accurate predictions.

Hypothesis Generation (Full Context + Reasoning + Transfer Set) We then added hypothesis generation, prompting the LLM to create a set of hypotheses $\mathcal{H} = \{h_j\}_{j=1}^{M}$ based on the near examples $\mathcal{E}_{\text{near}}$ using a hypothesis generation function $\Phi_{\text{hypothesis}}$. In essence, the hypothesis generation function is the specific prompt we design and use in conjunction with the LLM to produce hypotheses (see Figure A2 for the final prompt). While Wang et al. (2023) adopts a two-stage approach—first generating hypotheses and then solving the task—for inductive reasoning tasks, we streamline the process by integrating hypothesis generation directly into the 'thinking' process. This adjustment enables the LLM to process and synthesize information more effectively, leading to both more accurate and interpretable predictions by embedding reasoning within the task-solving step.

Supervised Signal (Final Prompt) The final enhancement involved incorporating a supervised signal from the baseline classifier $\mathcal{F}_{\texttt{classifier}}$. Specifically, we append to the prompt information stating, "*A*

supervised model (x% accurate) predicts a popularity rating of {prediction} with {confidence}." This signal, though noisy, provided the LLM with an external estimate of the video's potential popularity, thereby encouraging the LLM to weigh its predictions against an additional model's insights as well as to consider the inherent uncertainty in such predictions.

The final prompt, $\mathcal{P}_{\text{Final}}$, was constructed through a straightforward concatenation:

$$\mathcal{P}_{\text{Final}} = \mathcal{P}_{\text{vanilla}} + \mathcal{P}_{\text{think}} + \mathcal{P}_{\text{few-shot}} + \mathcal{P}_{\text{near}} + \mathcal{P}_{\text{hypothesis}} + \mathcal{P}_{\text{supervised}} \quad (2)$$

This prompt guided the LLM to combine all previous reasoning, examples, and external signals to make an informed, final popularity prediction.

3.4 Manual Validations of Hallucinations and Hypothesis Quality

To evaluate the video-to-text conversion process and the quality of the LLM-generated hypotheses, we conducted two surveys with human evaluators in the US, recruited through Mechanical Turk (MTurk). All participants held at least a Master's degree and were compensated at an hourly rate equivalent to USD 15 for tasks taking approximately 10-15 minutes each. The survey instructions and questions are fully provided in Section A.2 of the Appendix.

For the evaluation, we selected 5 videos from each popularity category (i.e., 5 local hits and 5 global big hits; 10 videos in total), with each video reviewed by 30 independent evaluators. This setup resulted in a total of 300 evaluations for each task: assessing video-to-text conversion and hypothesis and analysis quality.

We implemented screening questions at the end of the survey to ensure high-quality feedback. These questions required detailed attention to video content, focusing on the video's title, activity, and evaluation metrics. Participants who answered all questions correctly were classified as having "passed." Notably, the results were consistent across both groups—those who passed and those who did not—demonstrating the robustness and reliability of the outputs of our pipeline.

3.4.1 Validation 1: Video-to-Text Conversion Quality

The first survey evaluated the accuracy and reliability of the video-to-text conversion process, focusing on potential hallucinations. Participants were tasked with reviewing short video clips and their corresponding model-generated text descriptions. They rated the descriptions on four criteria—accuracy, adherence, consistency, and coverage of the main topic—using a 1-5 scale. The mean ratings from participants are as follows (mean $\pm$ std):

Metric	All Participants (N = 30)	Passed Screening (N = 12)
Accuracy	4.35 ± 0.30	4.32 ± 0.28
Adherence	4.28 ± 0.40	4.22 ± 0.10
Consistency	4.40 ± 0.25	4.36 ± 0.22
Main Topic	4.56 ± 0.24	4.55 ± 0.30

The high ratings, all above 4.22, indicate that the text descriptions accurately reflect video content without significant hallucinations.

3.4.2 Validation 2: Hypothesis Quality and LLM Analysis

The second survey assessed the LLM's hypotheses explaining video popularity predictions. Participants rated the quality of the hypotheses and overall analysis on a 1-5 scale (1: Strongly Disagree, 5: Strongly Agree). The mean ratings from participants are as follows:

Metric	All Participants (N = 30)	Passed Screening (N = 13)
Hypothesis Quality	4.44 ± 0.26	4.45 ± 0.30
LLM Analysis Quality	4.15 ± 0.25	4.24 ± 0.42

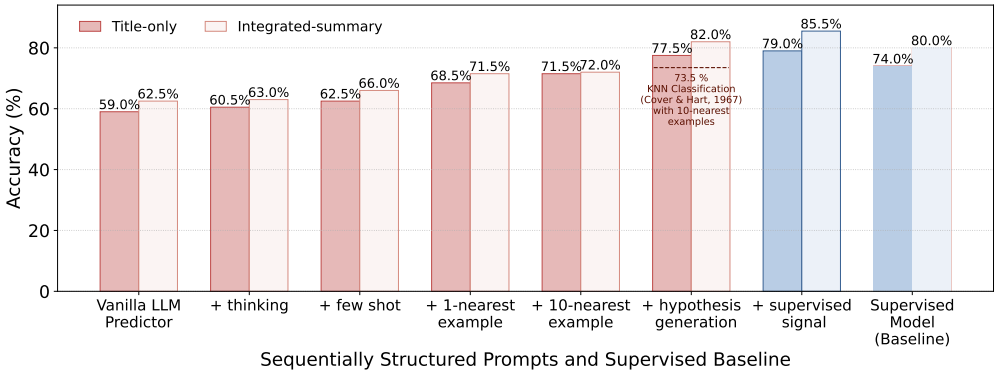

Figure 1: Comparison of accuracy for video-title-only and full-integrated-description-based predictions across models using different prompt sets. The plot shows performance improvements with each model enhancement, beginning with the baseline 'Vanilla' model and culminating in the final configuration incorporating supervised signals. The KNN (Cover & Hart, 1967) and supervised models are included as baselines.

These results confirm that the LLM-generated hypotheses are meaningful and its overall analysis of video popularity is high-quality, with ratings consistently above 4.15. Overall, both surveys validate the accuracy of the video-to-text conversion and the interpretability of the LLM's predictions.

4 Experiments and Results

4.1 Supervised Multimodal Approach (Baseline)

As a baseline for video popularity prediction, we implemented a supervised deep learning model that integrates multimodal embeddings from various video features. Textual features such as the title T, description D, and captions $\mathbf{C}$ are encoded using the MPNet base v2 encoder (Song et al., 2020, $\mathcal{E}_{\texttt{MPNet}} : \mathcal{T} \rightarrow \mathcal{Z}$), generating embeddings $\mathbf{e}_T$, $\mathbf{e}_D$, and $\mathbf{e}C$. Visual features, including video frames $\mathbf{V}$ and thumbnails, are encoded using the CLIP model (Radford et al., 2021; Mendelevitch & Aguynamed, 2023, $\mathcal{E}_{\texttt{CLIP}} : \mathcal{I} \rightarrow \mathcal{Z}$) and its video counterpart ($\mathcal{E}_{\texttt{CLIP-Video}} : \mathcal{V} \rightarrow \mathcal{Z}$) generates frame-level embeddings $\mathbf{e}_{\mathbf{I}_i}$ and a video-level embedding $\mathbf{e}_{\mathbf{V}}$. These multimodal embeddings are concatenated into a unified video representation $\mathbf{v}$, which is fed into a deep neural network for the binary classification ($\mathcal{F}_{\texttt{classifier}} : \mathcal{Z} \rightarrow \{0, 1\}$), predicting whether a video is a local or global hit, $y \in \{0, 1\}$. This baseline model provides a strong foundation by effectively combining multiple modalities.

Implementation Details We used the official Claude Sonnet 3.5 (Anthropic AI, 2024), a commercial model, for our experiments. For the zero-shot and in-context learning experiments, we utilized the LLaMa 3 model (Touvron et al., 2023). Additional implementation details, including specific configurations for LLaMa, are provided in Appendix A.3. For training the supervised baseline model, we utilized a learning rate of 0.001 with the Adam optimizer and early stopping (patience=6). See Section A.8 for more details.

Evaluation Metrics The model's performance is evaluated using accuracy as the primary metric, with additional tracking of precision and recall to assess classification quality.

4.2 Impact of Sequential Prompts on Prediction Performance

The performance of the LLM-based models was evaluated on both title-only and full-integrated-description-based prediction tasks. Figure 1 shows the incremental improvements in accuracy. Starting with the vanilla prompt ($\mathcal{P}_{\text{vanilla}}$), the model achieved 59.3% accuracy for title-based prediction and 62.5% for description-based prediction. The richer information provided in the description resulted in a noticeable improvement in accuracy.

Next, we incorporated additional prompting techniques: thinking prompts ($\mathcal{P}_{\text{think}}$), few-shot examples ($\mathcal{P}_{\text{few-shot}}$), and one-nearest-example retrieval ($\mathcal{P}_{\text{near}}$), which gradually improved performance, enhancing title-based prediction accuracy from 60.6% to 68.5% and description-based prediction from 63.2% to 71.4%.

The most substantial gains were observed with the integration of 10-nearest examples ($\mathcal{P}_{\text{near}}$) and hypothesis generation ($\mathcal{P}_{\text{hypothesis}}$), boosting accuracy by 9.0 and 10.6 percentage points, reaching 77.5% for title-based prediction and 82.0% for integrated description-based prediction. This significant improvement highlights the strength of hypothesis generation, allowing the model to generate and test multiple hypotheses, leading to more accurate and robust predictions.

Finally, the model incorporating all components, including the supervised signal ($\mathcal{P}_{\text{supervised}}$), achieved the highest accuracies of 79.2% for title-based prediction and 85.5% for description-based prediction. Compared to the traditional supervised baseline, this represents a substantial improvement of 5.2 and 4.3 percentage points, respectively.

The performance improvement from the baseline to the final configuration is particularly noteworthy, with increases of 19.9 and 23.0 percentage points for title- and description-based predictions, respectively. These results highlight the significant potential of our approach to enhance LLM performance in video popularity prediction tasks. Notably, the hypothesis generation stage contributed the most substantial improvement, underscoring the value of integrating reasoning and learning into the prediction process.

The consistent outperformance of our model, particularly with the supervised signals, over traditional supervised methods and the near examples baselines, also underscores the advantages of our hybrid approach. By combining the strengths of LLMs with supervised learning techniques, our approach delivers superior performance across both prediction tasks, demonstrating the potential for further improvements with targeted enhancements and supervised fine-tuning.

Ablation Study We conducted extensive ablation studies to evaluate the robustness of our model. Specifically, we analyzed the effects of temperature settings, the number of near-examples, and different embedding type choices on performance. Detailed results are presented in Appendix A.6, providing insights into the model's hyperparameter sensitivity and stability across configurations. Furthermore, we extended our analysis to include experiments with Gemini 1.5 Pro, a state-of-the-art Vision-Language Large Model (VLLM). The results, detailed in Appendix A.9, demonstrate consistent performance, further validating the robustness and generalizability of our approach and prompting strategies across diverse architectures.

4.3 Qualitative Evaluations

Figure 2 presents two successful predictions alongside one incorrect prediction, offering insights into both the strengths and limitations of the approach. For videos like "Mexico vs. Brazil Highlights" and "Minecraft Survivor VS 3 Hitmen," the framework correctly identified key elements that drive popularity. In the case of the football highlights, the model accurately attributed the video's success to the involvement of Brazil's national team, a globally recognized entity, and star players such as Vinícius and Richarlison. Similarly, for the Minecraft video, the model successfully captured the unique mechanics of the speedrun challenge and the personalities involved, which helped the video achieve high engagement. These examples demonstrate the model's ability to handle diverse types of content and identify relevant attributes, from popular personalities to novel content features, that drive video success.

However, the framework made an erroneous prediction for the "Dad Jokes" video, which underscores some of the model's limitations. Despite correctly identifying the universal appeal of humor and the presence of well-known creators, the model misjudged the overall reach and impact. This error highlights the challenge of predicting the success of content that resonates on an emotional or cultural aspect, where the usual popularity markers, such as star power or novelty, may not fully apply. This emphasizes the need for further refinement of the model to capture subtler factors like humor, sentiment, and cultural resonance, which can transcend traditional indicators of popularity.

Figure 2: The figure illustrates the LLM-based framework's explainable video popularity prediction process, including text summaries, similar video retrieval, hypothesis generation, and popularity score prediction. Two successful predictions (bordered in black) and one erroneous prediction (bordered in red) are highlighted.

5 Discussion and Future Work

The results of our experiments demonstrate the effectiveness of our **hybrid approach** to video popularity prediction, where LLMs are progressively enhanced through structured prompting techniques and supervised fine-tuning. The improvements seen in the model's performance—from the vanilla prompt to the final model incorporating hypothesis generation and supervised signals—underscore the potential of combining LLM reasoning capabilities with supervised learning methods. This hybrid approach consistently outperformed traditional supervised models, showing 5.2 percentage points improvement for title prediction and 5.5 percentage points improvement for description prediction over the supervised baseline. Additionally, the final model achieved 79.2% accuracy for title-based prediction and 85.5% for integrated description-based prediction, reflecting a substantial increase of 19.9 and 23.0 percentage points, respectively, compared to the vanilla LLM model.

One of the most significant contributors to these gains is the **hypothesis generation** stage, which alone boosted prediction accuracy by 9.0 percentage points for title prediction and 10.6 percentage points

for description prediction. This feature enhances not only the performance but also its explainability—validated through survey experiments—making the predictions more transparent and providing useful insights into the factors driving video popularity.

Moreover, the **integration of supervised signals** into the prompting framework contributed to further improvements, particularly in refining prediction accuracy when combined with the LLM's reasoning capabilities. By leveraging LLMs' contextual understanding with external supervision, we demonstrated that this approach is more effective than traditional supervised methods alone, particularly in complex tasks like video popularity prediction.

An important finding from our ablation study, detailed in the Appendix, is that the framework demonstrates robustness across various temperature settings, indicating its potential for stable performance under different conditions. This robustness strengthens the generalizability of our approach, making it applicable across a wide range of contexts and ensuring reliable predictions even when model parameters fluctuate.

The results of this study suggest that the hybrid LLM-based framework, which integrates sequential prompting techniques and supervised signals, is a highly effective solution for multimodal prediction tasks. It significantly outperforms not only traditional supervised models but also simpler models that rely on example-based guidance, such as few-shot and near examples, establishing a new benchmark for video popularity prediction. Additionally, by focusing on two key dimensions of popularity—**engagement intensity and geographic spread**—our framework offers a more nuanced and holistic understanding of what drives video success, compared to the one-dimensional focus on view counts in previous research.

Beyond video popularity prediction, the techniques developed in this study have broader implications and applications across various domains. For instance, our VLM-to-LLM pipeline and hypothesis generation methods could generalize to social media analysis, enabling trend, sentiment, or engagement prediction on multimodal platforms. In healthcare, the interpretability of hypotheses generated by LLMs could enhance transparency in medical imaging and diagnosis, where explainable AI is critical for trust and adoption. Similarly, the approach could support education by offering explainable feedback for student assessments or personalized content. Finally, in computational social science, the high-quality hypothesis generation demonstrated by our framework could transform theoretical exploration by offering nuanced explanations, shifting the focus beyond simple statistical coefficients to a richer understanding of sociological and cultural phenomena. These broader applications underscore the versatility and transformative potential of our approach.

There are several promising directions for future research. While our model already shows substantial improvements, further refinement of the hypothesis generation process could enhance accuracy, particularly by incorporating more advanced reinforcement learning techniques. More specifically, the LLM acts as an agent generating hypotheses about video popularity factors, where each hypothesis represents an action within the state space of possible predictions. The prediction accuracy serves as a reward signal, guiding the system to learn which types of hypotheses are most effective. Additionally, improving the model's ability to account for cultural and emotional factors could boost prediction quality, especially for content like humor or emotionally resonant videos, where traditional popularity metrics (such as view counts) may fall short. A deeper understanding of these factors could also enable the model to work more effectively in conjunction with specific user contexts, gauging the micro-level appeal of a video based on user-provided intent, situational factors, and video characteristics. Furthermore, integrating more diverse multimodal data, such as audio analysis or more granular sentiment analysis of comments, could offer richer insights into the drivers of video popularity. Future work could also explore real-time prediction capabilities and the adaptation of this model to predict popularity trends across platforms beyond YouTube.

REFERENCES

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a Visual Language Model for Few-Shot Learning. In *Advances in Neural Information Processing Systems*, volume 35, pp. 29463–29478, 2022.

Anthropic AI. The claude 3 model family: Opus, sonnet, haiku. *Claude-3 Model Card*, 2024. Accessed: 2024-09-05.

Nima Ghorbani Beni, Ahlem Bouyer, and Alireza Aghajani Rezaei. An improved independent cascade model for influence maximization in social networks. *IEEE Transactions on Network Science and Engineering*, 2023. doi: 10.1109/TNSE.2023.3270422. Early Access.

Yassine Borghol, Sacha Ardon, Niklas Carlsson, Derek Eager, and Anirban Mahanti. Content-agnostic factors affecting YouTube video popularity. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1186–1194, 2012.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language Models are Few-Shot Learners. *arXiv preprint arXiv:2005.14165*, 2020.

Tim Broxton, Yannet Interian, Jonathan Vaver, and Markus Wattenhofer. Catching a viral video. In *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pp. 296–303, 2013.

Yiwei Chen, Yunjin Choi, Suk-Jae Hwang, and Jungseock Kim. Enhancing social media post popularity prediction with visual content. *arXiv preprint arXiv:2405.02367*, 2023.

Cisco. Cisco Visual Networking Index: Forecast and Trends, 2017–2022. `https://cloud.report/Resources/Whitepapers/eea79d9b-9fe3-4018-86c6-3d1df813d3b8_white-paper-c11-741490.pdf`, 2021. Accessed: 2024-09-29.

Thomas Cover and Peter Hart. Nearest neighbor pattern classification. *IEEE transactions on information theory*, 13(1):21–27, 1967.

Rosanna E Guadagno, Daniel M Rempala, Shannon Murphy, and Bradley M Okdie. What makes a video go viral? an analysis of emotional contagion and internet memes. *Computers in human behavior*, 29(6):2312–2319, 2013.

Rajarshi Haldar, Don Towsley, and Tauhid Zaman. A temporal cascade model for understanding information spreading dynamics in evolving networks. *IEEE Transactions on Network Science and Engineering*, 2023.

Laura Hanu, Anita L Verő, and James Thewlis. Language as the medium: Multimodal video classification through text only. *arXiv preprint arXiv:2309.10783*, 2023.

Shivam Jog, Bhargav Siras, Akash Fender, Parikshit Mandurkar, Yash Nikalje, and Sharda Chhabria. Video popularity prediction using machine learning. *International Research Journal of Modernization in Engineering Technology and Science*, 3(5), 2021.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023a.

KunChang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. Videochat: Chat-centric video understanding. *arXiv preprint arXiv:2305.06355*, 2023b.

Bin Lin, Bin Zhu, Yang Ye, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. *arXiv preprint arXiv:2311.10122*, 2023a.

Yiming Lin, Lixin Gao, Yike Guo, and Yingying Zhu. A tensor-based independent cascade model for influence maximization in social networks. *IEEE Transactions on Knowledge and Data Engineering*, 2023b.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. What makes good in-context examples for gpt-3? *arXiv preprint arXiv:2101.06804*, 2021.

Jiasen Lu, Christopher Clark, Rowan Zellers, Roozbeh Mottaghi, and Aniruddha Kembhavi. Unified-io: A unified model for vision, language, and multi-modal tasks. In *The Eleventh International Conference on Learning Representations*, 2022.

Daniel Mendelevitch and Rich Aguynamed. *iejMac/clip-video-encode*. github, 10 2023. URL https://github.com/iejMac/clip-video-encode.

Yee Man Margaret Ng and Harsh Taneja. Web use remains highly regional even in the age of global platform monopolies. *PloS one*, 18(1):e0278594, 2023.

Minsu Park, Mor Naaman, and Jonah Berger. A data-driven study of view duration on youtube. In *Proceedings of the international AAAI conference on web and social media*, volume 10, pp. 651–654, 2016.

Minsu Park, Jaram Park, Young Min Baek, and Michael Macy. Cultural values and cross-cultural video consumption on youtube. *PLoS one*, 12(5):e0177865, 2017.

Henrique Pinto, Jussara M Almeida, and Marcos A Gonçalves. A novel time series based approach to predict popularity of online content. *Journal of Information and Data Management*, 4(3):347, 2013.

Jie Qin, Yifan Ding, Yiwei Chen, Jiawei Jiang, and Yong Xu. Predicting video popularity based on video covers and titles using a multimodal large-scale model and pipeline parallelism. *ResearchGate*, 2023.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

Alireza Rezvanian, S Mehdi Vahidipour, and Mohammad Reza Meybodi. A new stochastic diffusion model for influence maximization in social networks. *Scientific Reports*, 13(1):6122, 2023.

David Shamma, Jude Yew, Lyndon Kennedy, and Elizabeth Churchill. Viral actions: Predicting video view counts using synchronous sharing behaviors. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 5, pp. 618–621, 2011.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. Mpnet: Masked and permuted pre-training for language understanding. *arXiv preprint arXiv:2004.09297*, 2020.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

David Vallet, Miriam Fernandez, and Pablo Castells. Social media and information retrieval: A survey. In *Proceedings of the 2015 International Conference on Social Media and Society*, pp. 1–10, 2015.

Ruocheng Wang, Eric Zelikman, Gabriel Poesia, Yewen Pu, Nick Haber, and Noah D Goodman. Hypothesis search: Inductive reasoning with language models. *arXiv preprint arXiv:2309.05660*, 2023.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2022.

Zhi Wang, Shaojie Ye, Bernardo A Huberman, Chuang Li, and Dongmei Sun. Characterizing and modeling the dynamics of online popularity. In *Proceedings of the 21st international conference on World Wide Web*, pp. 623–632, 2012.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

Tyler West. Going viral: Factors that lead videos to become internet phenomena. *Public Relations Review*, 37(1):101–104, 2011.

Andy Zeng, Maria Attarian, Brian Ichter, Krzysztof Choromanski, Adrian Wong, Stefan Welker, Federico Tombari, Aveek Purohit, Michael Ryoo, Vikas Sindhwani, Johnny Lee, Vincent Vanhoucke, and Pete Florence. Socratic models: Composing zero-shot multimodal reasoning with language. *arXiv*, 2022.

Ruoming Zhou, Sartaj Khemmarat, and Lixin Gao. The impact of youtube recommendation system on video views. In *Proceedings of the 10th ACM SIGCOMM conference on Internet measurement*, pp. 404–410. ACM, 2010.

A APPENDIX

The appendix provides an in-depth exploration of our video popularity prediction framework, offering detailed analyses and insights to supplement the main text. Its primary objectives are to demonstrate the robustness of our approach, provide a comprehensive understanding of the dataset, and highlight the performance improvements achieved over baseline models. Additionally, we present key ablation studies to examine the impact of various hyperparameters on our model's performance.

A.1 OVERVIEW OF PIPELINE

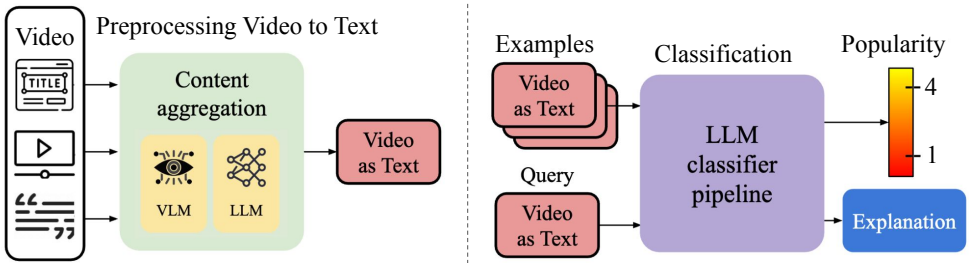

Figure A1: A training-free framework for video popularity prediction utilizing modality-aligned vision-language models (VLMs) and large language models (LLMs). The **left** shows the video preprocessing and content aggregation stages, where video content is transformed into sequential text representations through VLMs. This transformation generates *Video as Text* summaries, combining visual and textual information. The **right** illustrates the classification and prediction stages, where the LLM processes the *Video as Text* summaries to predict a popularity score and provides an explanation based on the identified patterns.

Figure A1 presents an overview of our approach. This high-level view depicts our training-free framework for video popularity prediction, which leverages modality-aligned Vision-Language Models (VLMs) and Large Language Models (LLMs) to generate *Video as Text* summaries. In the preprocessing stage (**left**), video content is transformed into sequential text representations using VLMs. During the content aggregation stage, visual and textual information is aligned and combined. The LLM then processes these text summaries to predict a video's popularity score (ranging from 1 to 4) and generates explanations based on identified patterns (**right**). These explanations take the form of hypotheses grounded in theoretically sound attributes. For example, if the video is about a national football game organized by FIFA, the model may highlight its global appeal due to the prominence of the organization and the attention drawn by specific teams, such as Brazil.

A.2 SURVEYS FOR HUMAN EVALUATION

To ensure data quality and participant reliability in our human evaluation studies, we conducted two separate surveys with distinct screening protocols. The first survey focused on video-to-text conversion quality, while the second evaluated hypothesis quality and LLM analysis. Each survey included comprehensive instructions, evaluation criteria, and targeted screening questions to verify participant attention and understanding. Participants were recruited through Amazon Mechanical Turk (MTurk Masters) and compensated with a base payment of USD 1.25, with potential bonuses up to USD 5 for high-quality responses.

A.2.1 Survey 1: Video-to-Text Conversion Quality

Video-to-Text Conversion Quality: Initial Instructions

Welcome to our research study on video-to-text transcription quality. We are academic researchers from ****, investigating the accuracy of automated transcription systems.

Survey Overview

In this survey, you will:

- Watch short video clips
- Read the corresponding automated transcriptions
- Answer questions about the accuracy and quality of the transcriptions

The survey should take approximately 8-12 minutes to complete. You will receive:

☐ Base compensation: USD 1.25

☐ Potential bonus: Up to USD 5 total for high-quality responses

Your Role as an Evaluator

Describing video content accurately is an incredibly complex task for AI. It requires understanding context, nuance, and implied information—skills that come naturally to humans but are extremely challenging for machines. Your task will involve:

- Watching diverse video clips
- Reviewing AI-generated content descriptions
- Providing detailed feedback on accuracy and quality

Key Evaluation Areas

When assessing the AI-generated descriptions, please consider:

- Overall accuracy in capturing key themes and concepts
- AI's ability to understand context and implied information
- Areas where the AI shows particular understanding
- Opportunities for improvement

Important Considerations

Please note:

- The AI system provides a comprehensive overview, not word-for-word transcription
- Focus on overall meaning and key points rather than exact phrasing
- The AI may make contextual inferences

Ready to Begin?

☐ I understand all the above instructions thoroughly

Video-to-Text Conversion Quality: Evaluation Criteria

1. Overall Accuracy

How accurately does the text description match the content of the video?

☐ 1: Completely inaccurate

☐ 2: Mostly inaccurate

☐ 3: Somewhat accurate

☐ 4: Mostly accurate

☐ 5: Highly accurate

2. CONTENT ACCURACY

How closely does the text description stick to the content presented in the video?

☐ 1: Mostly unrelated to video content

☐ 2: Significant deviations from video content

☐ 3: Moderate adherence to video content

☐ 4: Close adherence to video content

☐ 5: Perfectly matches video content

3. MAIN TOPIC CAPTURE

How well does the text description capture the main topic(s) discussed in the video?

☐ 1: Misses all main topics

☐ 2: Captures few main topics

☐ 3: Captures some main topics

☐ 4: Captures most main topics

☐ 5: Accurately captures all main topics

4. KEY POINT COVERAGE

To what extent are key points from the video included in the text description?

☐ 1: Misses all key points

☐ 2: Includes few key points

☐ 3: Includes some key points

☐ 4: Covers most key points

☐ 5: Covers all key points

Video-to-Text Conversion Quality: Task: Video Transcription Evaluation

Watch this Video, you will be given the task of evaluating a short transcript about this video next:
Video title: Cristiano Ronaldo Hat-Trick! | Manchester United 3-2 Norwich | Highlights

[Youtube Video]

Important Note

Your careful attention to this video is essential for accurately understanding the issues related to AI behavior and the quality of AI-generated video transcriptions. The more accurately you understand and remember the video's content, the more accurate and valuable your evaluation will be. We encourage you to watch the entire video attentively, as your insights will directly impact the assessment of AI performance! Participants who demonstrate a thorough understanding of the video content will be eligible for bonus compensation. Thank you for your dedication to this task!

TRANSCRIPT

INTRODUCTION

This video is a recording of a football match between two teams, featuring commentary and analysis throughout. The initial setting is a stadium with players from both teams on the field.

SEGMENT DETAILS

- **Segment 1:** The first segment shows a man walking on the field while another man walks in the background. The commentator describes the scene, mentioning the ball and a goal scored by Adrian Luna.
- **Segment 2:** In this segment, a man is seen walking towards the camera, wearing a yellow shirt, while another man sits on the ground, wearing a red shirt. The commentators discuss the game, highlighting great deliveries and goals scored.
- **Segment 3:** This segment shows more gameplay, with the commentators analyzing the players' moves and discussing the score.

OVERALL

The overall impact of this video is an immersive and engaging experience for football fans. The combination of exciting commentary, intense gameplay, and skilled players creates a thrilling narrative that will likely appeal to viewers who enjoy sports content.

EVALUATION	1 Not at all	2 Slightly	3 Somewhat	4 Mostly	5 Completely
How accurately does the text description match the content of the video?	☐	☐	☐	☐	☐
How closely does the text description stick to the content presented in the video?	☐	☐	☐	☐	☐
How consistent is the information in the text description with the facts presented in the video?	☐	☐	☐	☐	☐
How well does the text description capture the main topic(s) discussed in the video?	☐	☐	☐	☐	☐

Video-to-Text Conversion Quality: Screening Questions

Q1. Which videos did you evaluate in this survey?

- ☐ SUPAHOTFIRE vs BLUEFACE
- ☐ Cristiano Ronaldo Hat-Trick
- ☐ Kerala Blasters FC vs Jamshedpur FC Highlights
- ☐ Where I'm Travelling Next- Solo Trip?

Q2. In the videos you evaluated, which of the following activities was *not* mentioned?

- ☐ Playing rock-paper-scissors
- ☐ Eating cake
- ☐ A football match
- ☐ Skydiving

Q3. In the evaluation process, what were you asked to rate about the video transcriptions?

- ☐ Overall Accuracy
- ☐ Content Accuracy
- ☐ Main Topic Capture
- ☐ Video Production Quality
- ☐ Key Point Coverage

Q4. Thank you for participating in our study. Your insights will help us improve our AI model for predicting video popularity. Do you have any additional comments or feedback about the model's predictions or this survey?

A.2.2 Survey 2: Hypothesis Quality and LLM Analysis

Hypothesis Quality and LLM Analysis: Initial Instructions

Welcome

Welcome to our research study on video popularity prediction using AI models. We are academic researchers from $\cdots$, investigating how Large Language Models (LLMs) can predict video popularity.

Survey Overview

In this survey, you will:

- Read video descriptions and LLM-generated hypotheses about video popularity
- Rate the accuracy and relevance of these hypotheses
- Assess the LLM's ability on critical analysis and judgment

The survey should take approximately 8-12 minutes to complete. You will receive:

☐ Base compensation: USD 1.25
☐ Potential bonus: Up to USD 5 total for high-quality responses

Study Overview

We are evaluating a Language Learning Model (LLM) designed to predict video popularity based on content analysis. The LLM analyzes videos from YouTube's trending page and generates hypotheses about what makes videos popular, as well as providing a detailed analysis of each video's content. Your role is to:

- Rate the hypotheses generated by the LLM
- Assess the LLM's critical analysis and judgment for TWO separate videos

Video Popularity Rating Scale

The LLM rates videos on a 4-point scale:

- **Popular**: Likely to have general appeal and be popular for a short while
- **Moderately Popular**: Has several appealing elements for more than basic popularity
- **Highly Popular**: Likely to be popular among a broad audience but may not reach ultra popularity
- **Ultra Popular**: Strong potential to become ultra popular, featuring unique, engaging, and broadly appealing content

Note: While the LLM uses this 4-point scale to rate video popularity, your task will be to rate your agreement with the LLM's hypotheses and analysis using a different 4-point scale.

Hypothesis Quality and LLM Analysis: Instructions - Part 2

Your Tasks

Task 1: Rate the LLM's Hypotheses

You will be presented with video descriptions and the LLM's hypotheses about what makes them popular. You will rate your agreement with 4-5 specific hypotheses generated by the LLM about what makes the video popular. For example, a hypothesis looks like '*Sport highlights, especially from important matches, tend to be ultra popular*', to which you can Strongly agree, Agree, Disagree or Strongly Disagree. Your job is to rate how much you generally agree or disagree with given hypothesis based on the same information provided to the LLM.

Task 2: Assess the LLM's Critical Analysis

You will evaluate the LLM's critical analysis of the video, including its assessment of factors influencing popularity and its final popularity prediction.

Rating Scale for Your Responses

You will use a 4-point scale to rate both the LLM's hypotheses and its critical analysis. This scale is designed to encourage you to form a definitive opinion based on your knowledge and the information provided. For both tasks, use the following scale:

- **1 - Strongly Disagree**: The hypothesis or analysis is clearly incorrect or irrelevant
- **2 - Disagree**: The hypothesis or analysis has major flaws or inaccuracies
- **3 - Agree**: The hypothesis or analysis is mostly accurate and relevant
- **4 - Strongly Agree**: The hypothesis or analysis is highly accurate and insightful

Tips for Completing the Tasks

- Read each video description and LLM hypothesis carefully before rating
- Consider each hypothesis and analysis point carefully. Draw on your own knowledge of popular online content, but focus primarily on the information provided in the video description
- For instance, if you think the LLM's hypothesis about sports highlights is accurate based on the video description and your knowledge, you might select 'Agree' or 'Strongly Agree'
- Try to be consistent in your ratings across similar types of content

Your thoughtful evaluations will help us improve the LLM's ability to predict video popularity, ultimately contributing to a better understanding of content trends on platforms like YouTube.

Hypothesis Quality and LLM Analysis: Video: Minecraft but there's Cartoon Hearts

Video summary

Introduction: The video opens with a mysterious green figure walking around a dark room, setting the tone for a fantastical and humorous adventure. Segment Details - Segment 1: Introduces the green figure, referencing Shrek and showcasing magic powers $\cdots$ Conclusion: The video's impact is significant, as it showcases the creators imagination and ability to blend disparate elements into a cohesive narrative. The humor, entertainment value, and references to popular franchises will likely appeal to viewers who enjoy fantasy, scifi, and comedy.

In the next two pages, you will be shown 2 tasks:

- Task 1: Rate the hypotheses generated by the LLM
- Task 2: Assess the LLM's critical analysis and judgment

Task 1: Rate the LLM's Hypotheses

Video Information

Important Note

Your careful attention to this video description is essential for accurately understanding the quality of AI-generated hypothesis. The more accurately you understand the video's content, the more accurate and valuable your evaluation will be. We encourage you to read the entire video description attentively, as your insights will directly impact the assessment of AI performance!
Participants who demonstrate a thorough understanding of the video content will be eligible for bonus compensation. Thank you for your dedication to this task!

MODEL'S HYPOTHESES

	Strongly Disagree	Disagree	Agree	Strongly Agree
H1: Videos with unique Minecraft concepts tend to be ultra-popular	☐	☐	☐	☐
H2: Content that blends multiple franchises or pop culture elements has broader appeal	☐	☐	☐	☐
H3: Videos with humorous and imaginative content encourage sharing and discussion	☐	☐	☐	☐
H4: Fast-paced content with diverse visual elements keeps viewers more engaged	☐	☐	☐	☐

If you Disagree/Strongly Disagree, do you have any better hypotheses or suggestions for improving the provided hypotheses?

TASK 2: ASSESS THE LLM'S CRITICAL ANALYSIS

FACTORS IN THE GIVEN VIDEO THAT COULD INFLUENCE POPULARITY:

- F1: Creative concept: "Minecraft but there's Cartoon Hearts" (very positive)
- F2: Blending of multiple franchises (Shrek, Teen Titans, Star Wars, Scooby-Doo) (positive)
- F3: Humorous and playful tone (positive)
- F4: Imaginative scenarios (magic powers, cyber crystals, outer space) (positive)
- F5: Alignment with geek culture trends (positive)
- F6: Potential for viewer engagement and discussion (positive)

LLM's Final Analysis: Considering all factors, especially the similarity to other ultra popular videos and the supervised model prediction, this video is likely to be Ultra Popular. It has all the elements of highly engaging content that tends to perform exceptionally well, particularly in the gaming and geek culture niches.

	Strongly Disagree	Disagree	Agree	Strongly Agree
F1: Creative concept	☐	☐	☐	☐
F2: Blending of multiple franchises	☐	☐	☐	☐
F3: Humorous and playful tone	☐	☐	☐	☐
F4: Imaginative scenarios	☐	☐	☐	☐
F5: Alignment with geek culture trends	☐	☐	☐	☐
F6: Potential for viewer engagement	☐	☐	☐	☐

Hypothesis Quality and LLM Analysis: Screening Questions

Q1. What was the primary task you were asked to perform in this survey?

☐ Predict which videos would become viral
☐ Assess the LLM's hypotheses and analysis about video popularity
☐ Provide your own theories about what makes videos popular
☐ Compare different AI models' performance in analyzing videos

These video-specific screening questions were randomly assigned based on the viewed video:

Q2. Which of the following elements were present in the video you analyzed? (select all that apply)

☐ Commentary and analysis
☐ Great deliveries and goals scored
☐ Interviews with team managers

- ☐ Slow-motion replays
- ☐ Adrian Luna scoring a goal
- ☐ Penalty shootout

Q3. Which of the following elements were present in the video you analyzed? (select all that apply)

- ☐ References to Shrek
- ☐ Mining of cyber crystals
- ☐ Outer space scenes
- ☐ Pokémon battles
- ☐ Teen Titans-inspired content
- ☐ Underwater exploration

Q4. Which of the following elements were present in the video you analyzed? (select all that apply)

- ☐ Internal struggle of the protagonist
- ☐ Car chase scenes
- ☐ Self-harm depicted
- ☐ Comedic dialogue
- ☐ Apology and plea for forgiveness
- ☐ Transformation sequences

Q5. Which of the following elements were present in the video you analyzed? (select all that apply)

- ☐ Two men playing video games
- ☐ Reaction to a music video
- ☐ Wearing headphones
- ☐ Dancing performances
- ☐ Occasional singing into microphones
- ☐ Cooking demonstrations

A.3 Implementation Details

The video popularity prediction pipeline was implemented using PyTorch 2.1.0 and the transformers 4.35.0 library. The content aggregation was performed using a custom module that combined textual information from titles, descriptions, and generated captions. We employed Claude 3.5 Sonnet (accessed via Anthropic's API) (Anthropic AI, 2024) as our primary LLM for classification and hypothesis generation, while also using LLaMa 3 70B (Touvron et al., 2023) Instruct offline on two NVIDIA RTX A4000 GPUs using tensor parallelism and quantization (2.8 bits per weight) for efficient video-to-text generations. The VLM component (of VideoLLava) used a fine-tuned CLIP model to align visual and textual features. The entire pipeline was orchestrated using a custom Python script that handled data flow between components, with batching implemented to optimize throughput. For the offline LLaMa 3 70B setup, we achieved approximately 2 tokens per second inference speed. All experiments maintained consistent hyperparameters (Temperature: 0.5, Max Tokens: 4096, Top-p: 0.95) to ensure reproducibility.

A.4 Final prompt

The final prompt is also described in Fig A2. The Large Language Model processes the final prompt and generates a structured output containing the thinking process, hypotheses, judgments, and the predicted popularity class.

Final Prompt for Video Popularity Prediction

<Instructions>
You will be predicting the potential popularity of a video based on its title and a description of its content.
Note: All videos in this dataset are from YouTube's trending page, meaning they have already achieved a significant level of popularity.
Your task is to provide a 'popularity rating' indicating how likely the video is to become popular among viewers, using the following 1-4 scale:
1 - Locally Moderately Popular: The video is likely to appeal to be popular, and has elements of general appeal and is probable to be popular for shorter while.
2 - Locally Popular: The video has several appealing elements for more than basic popularity.
3 - Globally Highly Popular: The video is likely to be popular among a broad audience but may not have elements that lead to ultra popularity status.
4 - Globally Ultra Popular: The video has strong potential to become ultra popular, featuring unique, engaging, and broadly appealing content.
</Instructions>

<output>
<scratchpad>
Think step by step inside <scratchpad>Your analysis here</scratchpad>.
Step 1: Look at the given <video_description>{description}</video_description>, and First, answer the question: "comparing this video with videos in <similar_examples>, are videos similar to this video in the popular or ultra popular category?".
Step 1.5: A supervised model (80% accurate) predicts a popularity rating of {supervised_prediction} with {supervised_confidence:.2f} confidence. Factor this into your analysis.
Step 2: Then create 4 hypothesis about why videos in the <similar_examples>are ultra popular (Evaluation=4) and some popular (Evaluation=1). Try to catch patterns from these example videos, try to generalise patterns that make a video reach high popularity, and why some stay in basic popularity.
...
Step 4: Now expand on that reasoning think about whether the given and <video_description>are going to be Locally Moderately Popular, Locally Popular, Globally Highly Popular, or Globally Ultra Popular, give more weight to answer of step 1: 1 is for "Locally Moderately Popular" and 4 is for "Globally Ultra Popular."
</scratchpad>
<Evaluation>rating</Evaluation>
</output>

Now predict the popularity for the given video title and description, using the proper output format. Your insights could help shape the future of video content creation.

Figure A2: The final prompt structure for video popularity prediction using Large Language Models. The prompt incorporates instructions, a structured output format, and a step-by-step analysis process.

A.5 Dataset Analysis

Our dataset analysis uncovers several key insights that help contextualize the dynamics of video popularity and inform our prediction task design. We highlight the importance of considering both *geographical reach* and *view count* as critical factors in assessing a video's popularity. Our analysis shows a positive correlation between a video's international presence and its view count, with notable clusters of videos distributed across different quadrants of this relationship.

Additionally, a three-dimensional analysis that includes the duration a video remains on trending lists reveals a more nuanced relationship between trending duration, geographical reach, and view counts. We observe an optimal range of 100-200 units of trending duration and a reach of 15-35 countries as indicators of peak performance. However, outliers in this analysis suggest that content quality and other unquantified factors play important roles in determining a video's success.

A.5.1 Dataset features and video categorization

We present the features used for each video (Figure A3, left) and a heatmap categorizing videos based on engagement intensity and geographical reach (Figure A3, right). This information provides crucial context for understanding the nature of our dataset and how we distinguish globally viral videos from those with more localized popularity.

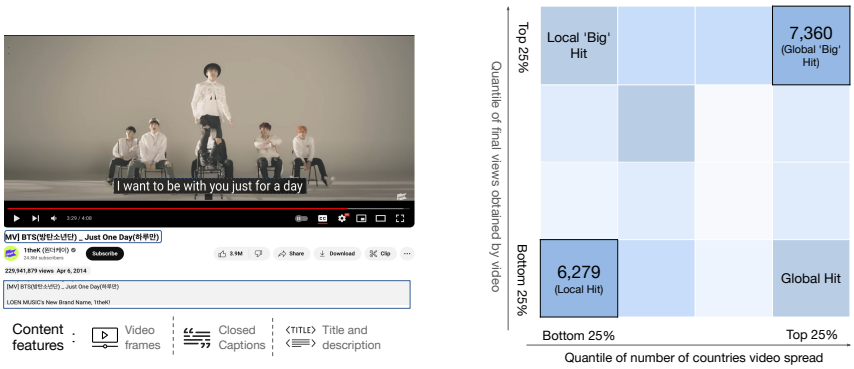

Figure A3: **Left:** The list of features for a youtube video. **Right:** Heatmap categorization of YouTube videos into 16 quantiles based on two key dimensions: the number of views and the number of countries in which the video trended. Videos are classified into 'Global Big Hit' (top 25% in both dimensions) and 'Local Hit' (bottom 25% in both dimensions), with cell colors indicating the relative density of each class.

A.5.2 GEOGRAPHICAL REACH VS. VIEW COUNT

To better understand how a video's international presence relates to its popularity, we analyzed our dataset, focusing on the connection between the number of countries a video reaches and its total views. Figure A4 visualizes this relationship.

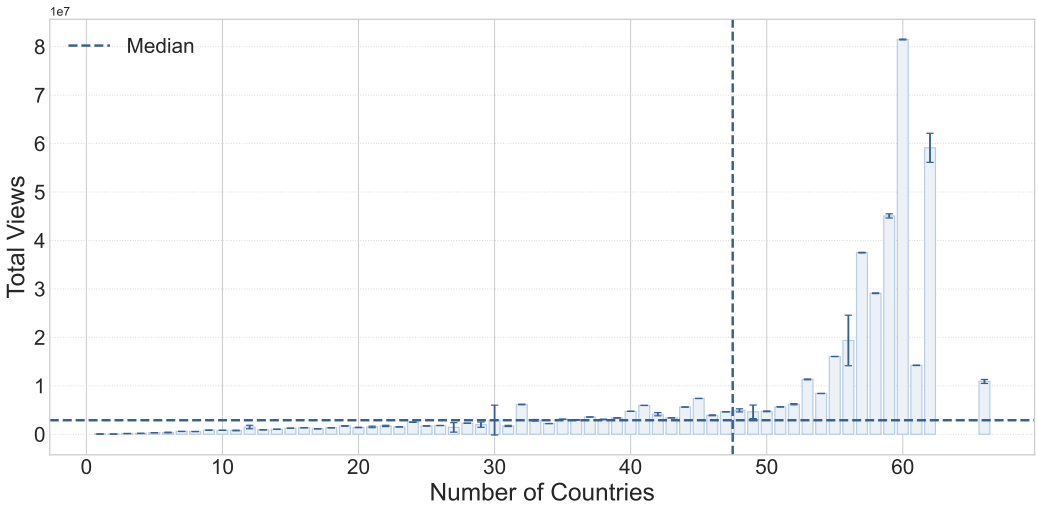

Figure A4: The plot depicts the relationship between the number of countries a video reaches and its total view count. Red lines represent the median values for each dimension, dividing the plot into quadrants.

The plot is bisected by two red lines representing the median values for each dimension, effectively partitioning the data into four distinct quadrants. The median number of countries reached by a video is 47.5, while the median total view count is approximately 2.9 million (2,896,886 views). The visualization shows that videos reaching more countries tend to get more views, but not all videos are spread out evenly. We observe a general positive correlation between a video's geographical reach and its view count, suggesting that videos with broader international appeal tend to accumulate more views. Also, a significant cluster of videos is concentrated in the lower-left quadrant, indicating a substantial number of videos with both limited geographical reach and relatively low view counts (less than 2.9 million views). Conversely, the upper-right quadrant, while less densely populated, contains

videos that have achieved both high view counts and extensive geographical reach, representing the most globally popular content in our dataset.

A.5.3 Three-dimensional analysis: Trending lists, reach, and view counts

We conducted a three-dimensional analysis that examines the relationship between the duration a video remains on trending lists, its geographical reach, and its total view count. This analysis, visualized through a heatmap (Figure A5), provides a more comprehensive understanding of the complex dynamics of spread.

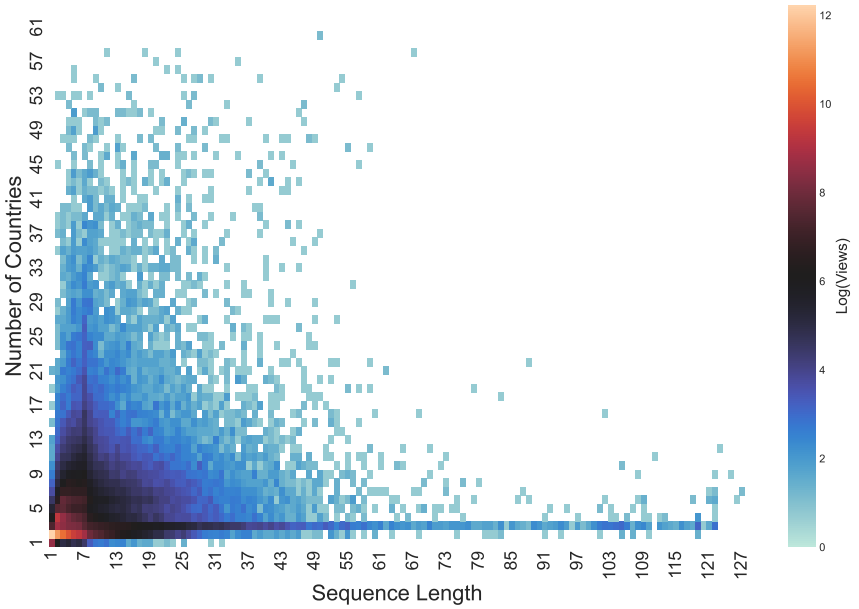

Figure A5: Heatmap depicting the relationship between trending duration (x-axis), number of countries reached (y-axis), and the logarithm of total views (color intensity) (say that image is truncated to 127 sequence length).

The heatmap reveals that videos with longer trending durations and wider international presence generally attract more views, though the relationship isn't straightforward. Peak viewership, represented by the darkest areas on the heatmap, is concentrated in the upper-right quadrant. This indicates that videos with longer sequences (approximately 150-200 units) and broader international appeal (reaching 30-40 countries) tend to accumulate the highest number of views. The single highest point, with a log(views) value of approximately 12.5, corresponds to a sequence length of 175 and reaches 35 countries.

Examining sequence length patterns, we observe a notable increase in viewership as sequence length increases from 0 to about 100 units. Beyond this point, the relationship becomes more complex, with videos between 100-200 units performing particularly well, especially when they reach a moderate to high number of countries. Interestingly, there's a slight decline in viewership for extremely long sequences (200+), suggesting an optimal range for sequence length.

The impact of country reach on viewership is evident, with videos reaching more countries generally receiving more views. However, this relationship varies across different sequence lengths. For shorter sequences (0-50), the impact of reaching more countries is less pronounced, while it becomes more significant for medium to long sequences.

Several notable patterns emerge from this analysis:

- A distinct area of low viewership in the bottom-left corner, representing short sequences with limited country reach.
- A "hot zone" in the middle-right area of the heatmap (sequence lengths of 100-175 and country counts of 15-35), consistently showing high viewership.

- Potential diminishing returns for very high sequence lengths (200+) and country counts (40+).
- Isolated "hot spots" throughout the heatmap, representing outlier videos with unexpectedly high views.

A.5.4 VIDEO CATEGORIES

YouTube's content ecosystem is diverse, encompassing a wide range of video categories. Our dataset provides a unique opportunity to analyze popularity trends across these categories, offering insights that are typically challenging to obtain. In this section, we present an analysis based on 11 heatmaps, each representing a distinct YouTube category (Figure A6). These heat maps visualize the complex interaction between the length of the sequence, the number of countries reached, and the total views for each category. This multi-dimensional analysis reveals both overarching trends and category-specific patterns in video popularity.

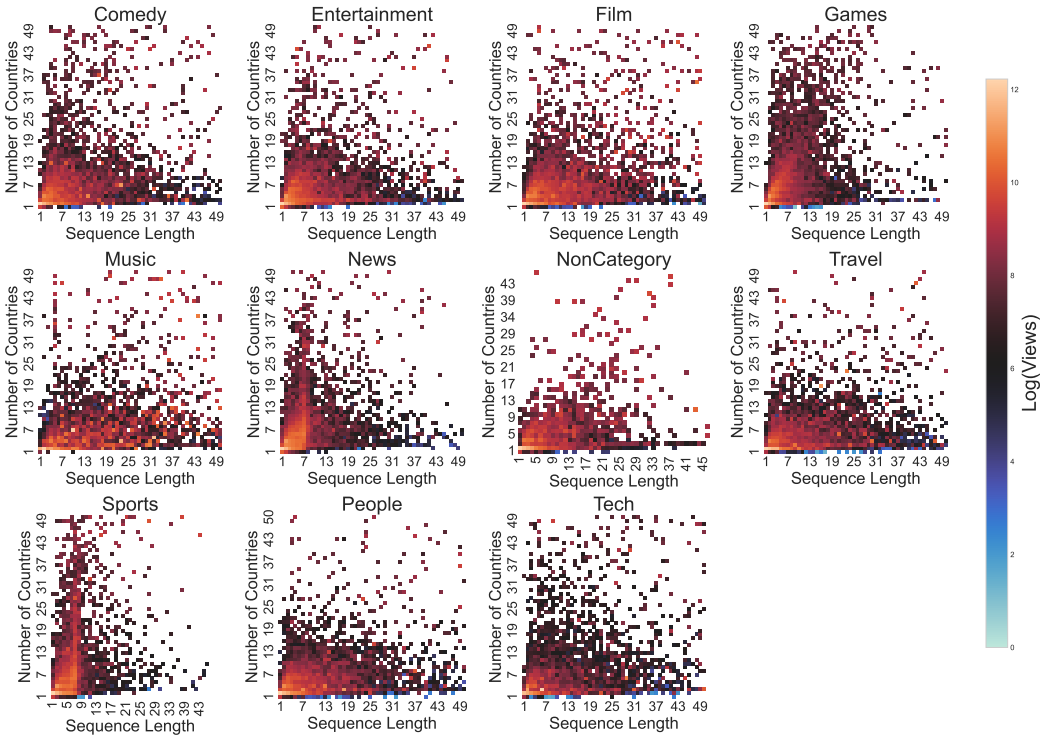

Figure A6: Heatmaps depicting the relationship between sequence length, number of countries reached, and logarithm of total views for 11 YouTube categories. Each heatmap represents a different category, with color intensity indicating log_{10}(Views).

Our analysis reveals several consistent patterns across categories. The maximum number of views ranges from 182 million to 1.48 billion views. The average number of views for most categories falls between 63,000 and 251,000 views. Interestingly, for almost all categories, maximum views occur at very short sequence lengths (mostly 1) and low number of countries (2), suggesting that brief, targeted content can achieve high viewership.

Each category exhibits unique characteristics. The Games category demonstrates the highest maximum views (about 1.48 billion views) and one of the highest average views, indicating high engagement. In contrast, the News category contains the largest number of videos but shows a lower average view count, suggesting a high volume of content with more moderate individual performance. The Music category, despite having the fewest videos, maintains a competitive average view count, indicating that music videos tend to perform well relative to their number. The NonCategory exhibits the lowest average views, which might be expected for content that doesn't fit into standard categories.

The relationship between sequence length, country reach, and views varies across categories. Across most categories, shorter sequence lengths (0-20) tend to have higher average view counts. The Games category shows particularly high performance for short sequences (about 31,000 views on average). Regarding country reach, videos reaching 6-10 countries often have the highest average views across categories. There's a consistent decline in average views as the number of countries increases beyond 15, suggesting that very broad international appeal is rare.

Studying Popularity and Reach We looked at how different video categories do across popularity and international reach using a grid of pie charts (Figure A7). Each pie chart represents a different quantile combination of video popularity (views) and international reach (number of countries), providing a comprehensive view of category distribution across various levels of success.

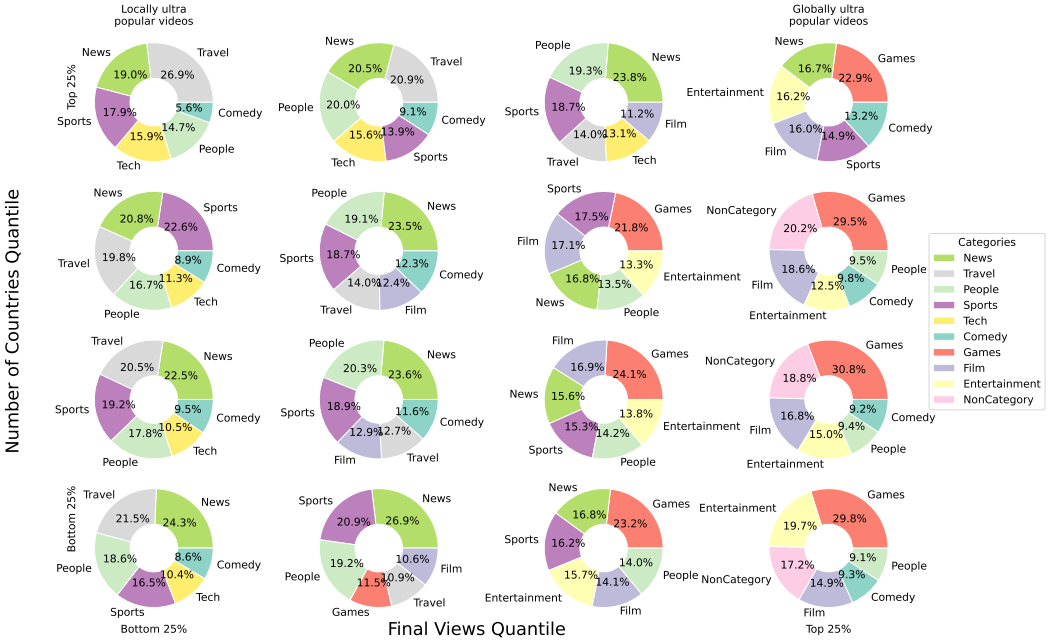

Figure A7: 4x4 grid of pie charts showing the distribution of video categories across different quantiles of popularity (views) and international reach (number of countries).

Our analysis uncovers the universal appeal of certain video categories across different levels of popularity and geographical spread. Notably, "People," "News," and "Sports" stand out, appearing in the majority of quantiles, indicating their widespread popularity. "Comedy" and "Film" also show strong presence, suggesting their content resonates across various levels of success and international reach.

A.6 Ablation Study

A detailed ablation study was conducted to examine the effects of key model hyperparameters on prediction accuracy. This investigation focused on the interaction between the embedding type used for near example retrieval, the count of these examples, and the temperature parameter of the Large Language Model (LLM). By methodically adjusting these parameters, the study aimed to reveal how variations in these elements influence the framework's predictive capabilities. The analysis closely examines the impact of embedding types, whether sourced from video descriptions or titles, the strategic selection of near example quantities, and the temperature settings within the LLM, providing a thorough examination of their combined effects on performance. This study is crucial as it illuminates the framework's operational nuances and informs potential adjustments, enhancing its effectiveness in the intricate task of video content analysis and prediction. Through this rigorous analysis, we aim to explore the framework's responsiveness to different hyperparameters, contributing to a nuanced understanding of its predictive mechanisms. This endeavor is not about optimizing the

model per se but about uncovering how the framework behaves under varied conditions, offering valuable insights into its structure and function.

A.6.1 IMPACT OF EMBEDDING TYPE ON MODEL PERFORMANCE

In our study of our proposed framework, we focused on understanding how the choice of embedding type, whether video descriptions or titles, used to find similar videos affects the accuracy of predicting video popularity. It's important to note that while both methods use full video descriptions, the key difference lies in how these similar videos are identified. This study looks into how choosing between video descriptions or titles to find similar videos affects prediction accuracy, showing that using titles to retrieve examples, surprisingly make our pipeline's predictions better.

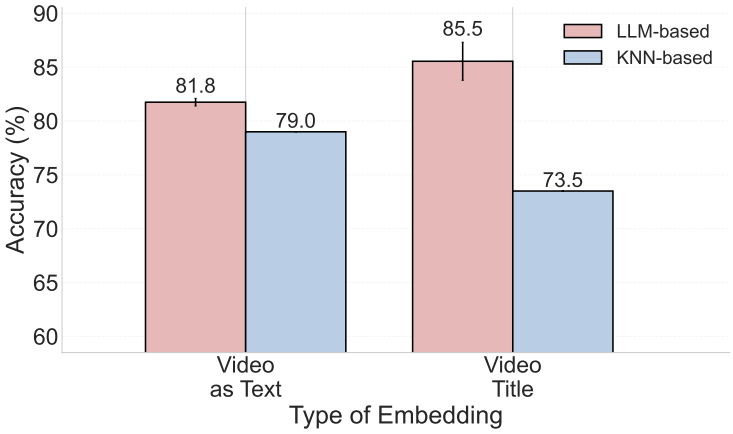

Figure A8: This graph compares how well our pipeling and near-examples models predict video popularity using 10 examples each, showing that using titles to find similar videos works better, even though both methods use examples containing full video descriptions as input. The only difference lies in how we find these examples: using titles or descriptions.

We report the findings in Figure A8, which show a nuanced resut of different kinds retrieved examples. Video-to-text achieved a mean accuracy of 81.75%, showcasing a consistent prediction capability with a standard deviation of 0.35%. This suggests that descriptions provide a reliable basis for similarity matching, albeit with a marginally lower accuracy compared to titles. Conversely, title embeddings yielded a higher mean accuracy of 85.5%, indicating their effectiveness in accurately identifying highly popular videos, albeit with increased variability, as evidenced by a standard deviation of 1.77%. This discrepancy may stem from the complexity and length of generated video descriptions, which could introduce extraneous information, diluting the core elements necessary for precise similarity matching. Titles, being more succinct, appear to offer a more focused approach for example retrieval, likely due to their ability to encapsulate the video's essence more directly. In contrast, the KNN model exhibited a mean accuracy of 79% with video descriptions and 73.5% with titles, highlighting a different pattern of performance that underscores the importance of model choice in leveraging embedding types effectively. This means that the way we select similar videos is key, and using titles, which are shorter, might predict popularity better by focusing on the main points of the video. The study shows that how we choose similar videos can make a big difference in how well our pipeline works, suggesting that we should think carefully about how we find these videos to improve predictions.

A.6.2 IMPACT OF NUMBER OF NEAR EXAMPLES ON MODEL PERFORMANCE

In this part of our study, we looked at how changing the number of similar videos (near examples) used by the our pipeline affects its ability to predict video popularity. Near examples are similar videos retrieved from the database that the model uses to identify patterns and make predictions. This analysis aims to determine the relationship of near examples, prediction accuracy and computational efficiency. We kept everything else the same and only changed the number of these examples to see how it impacts accuracy and how efficiently the model works.

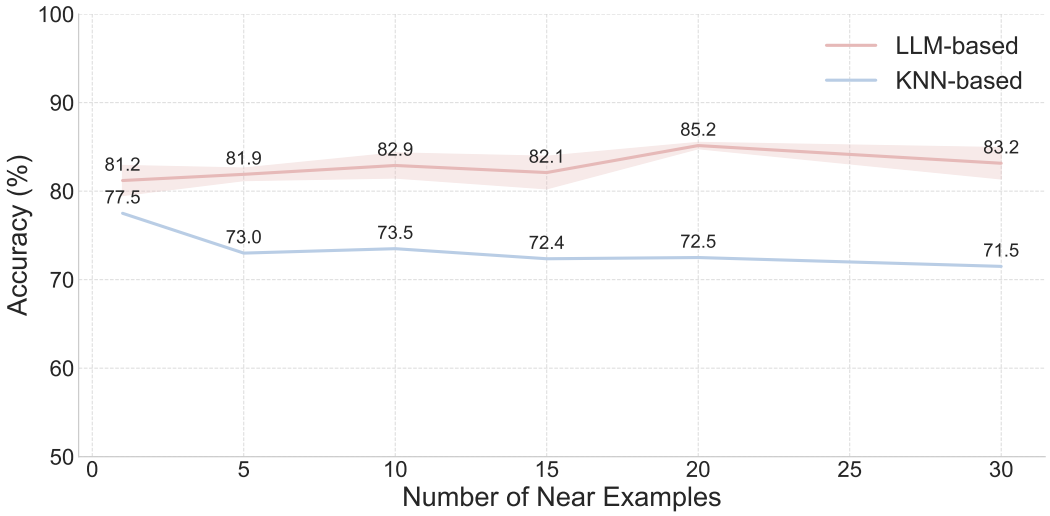

Figure A9: Impact of the number of near examples on accuracy. The plot shows mean accuracy and standard deviation for 1, 5, 10, 15, 20, and 30 near examples.

We tested using different numbers of near-examples, from 1 to 30, and found that more examples can actually help the model predict better, even though they might not seem as accurate when used in simpler models. Figure A9 shows this. The model's accuracy changes as we use more examples, with the best accuracy at 20 examples, suggesting this number might be just right for our model.

Interestingly, as we add more examples, the model's predictions become a bit less consistent, but it gets better at predicting overall. This means that even if more examples don't always lead to better results in simpler models, the LLM can use them to understand videos better and make more accurate predictions. However, using too many examples can actually make predictions a bit worse, showing there's a sweet spot at 20 examples where the model is both accurate and consistent. This finding is important because it shows that the LLM can use more information to improve, but there's a point where adding more doesn't help as much. It also reminds us that while more examples can help, we need to consider how much work the model has to do. Looking at how different types of videos respond to more examples could help us fine-tune the model even more, making it better at predicting video popularity.

A.6.3 Impact of temperature on model performance

In language models, temperature is a hyperparameter that controls the randomness of the model's output. A lower temperature makes the model more deterministic in its predictions, while a higher temperature increases randomness and creativity. In the context of LLM based video popularity prediction, temperature plays a crucial role in balancing between making consistent, safe predictions and exploring more diverse, potentially insightful outcomes. Finding the optimal temperature is essential for maximizing the model's predictive accuracy while maintaining its ability to generalize across various video types and popularity patterns.

We looked at how different temperature settings affect the model's accuracy to find the best balance. Figure A10 shows that the model works well across a range of temperatures, with 0.3 and 0.6 being optimal, both hitting 85.4% accuracy with little variation. This means the model can handle different temperatures well, but there's a slight dip at 0.8 that needs more study. Future work could look closer at temperatures between 0.6 and 1.0 to understand why and improve predictions.

A.7 Additional Ablation Study - Supervised Model

This section presents a detailed ablation study of our supervised model for video popularity prediction. By systematically analyzing the performance of different feature combinations, we aim to identify

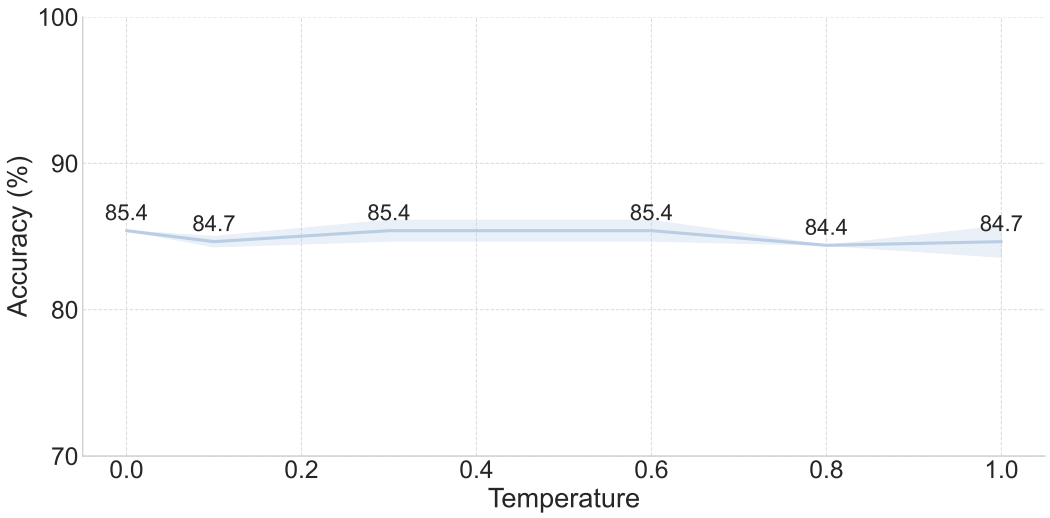

Figure A10: Impact of temperature settings. The plot shows mean accuracy for temperatures ranging from 0.0 to 1.0.

the most effective features and understand the trade-offs between model complexity and prediction accuracy. To investigate the effectiveness of different feature combinations in predicting video popularity, we conducted a series of experiments using various feature sets. We analyzed the performance of individual features, pairwise combinations, triple combinations, and complex feature sets. This comprehensive analysis aims to identify the most effective feature combinations and understand the trade-offs between model complexity and prediction accuracy.

A.8 Additional Ablation Study - Baseline Multimodal Model

Feature	Embedding model	Embedding Size
Text (Title, Description, Captions)	MPNet	768
Image (Thumbnail)	CLIP	512
Video (Key Frame Aggregation)	VideoCLIP	512

Table A1: Baseline Embeddings

To establish a strong foundation for comparison, we implement a baseline multimodal model that leverages deep learning techniques to predict video popularity. The architecture of this baseline model is described in Table A2. The model consists of several key components, including a preprocessing layer to handle variable input sizes, main layers to learn complex representations, dimension matching layers to facilitate residual connections, an attention layer to weigh the importance of different parts of the input data, and a final classification layer to produce the output.

The baseline model utilizes various embeddings to represent the different features of the video data, as detailed in Table A.8. For textual features, such as the title, description, and captions, the MPNet model is employed to generate embeddings of size 768. Visual features, including the thumbnail, are processed using the CLIP model, resulting in embeddings of size 512. Finally, the video content is represented using key frame aggregation and the VideoCLIP model, producing embeddings of size 512. The baseline multimodal model serves as a robust point of comparison for our proposed framework, allowing us to assess the performance improvements achieved through the integration of VLMs and LLMs in the video popularity prediction task.

Component	Details
Preprocessing Layer	Uses a linear transformation to map input data to a fixed dimensionality (processed_dim).
Main Layers	Composed of fully connected layers with batch normalization and ReLU activation. Includes dropout for regularization. Layers are sequentially connected with increasing reduction in dimensionality ($1024 \rightarrow 512 \rightarrow 256 \rightarrow 128$).
Dimension Matching Layers	Linear layers that adjust dimensions to enable addition of residual connections at each main layer stage.
Attention Layer	Consists of a linear transformation, a tanh activation, and a softmax output to produce attention weights.
Final Classification Layer	A fully connected layer that takes the attended features and outputs the final classification results.
Overall Model Architecture	Input data is processed through layers that include preprocessing, main processing with residuals, attention application, and final classification.

Table A2: Baseline Multimodal Model Description

A.8.1 INDIVIDUAL FEATURE PERFORMANCE

We begin by examining the predictive power of each feature type in isolation. Table A3 presents the performance scores for individual features.

Feature	Score
Thumbnail	0.77
Video	**0.79**
Title	0.75
Description	**0.79**
Caption	0.76

Table A3: Performance Scores for Individual Features

As shown in Table A3, video features and description features achieved the highest individual performance with a score of 0.79, followed closely by thumbnail features (0.77) and caption features (0.76). Title features showed the lowest individual performance at 0.75, suggesting that while titles contribute to prediction, they may not be as informative as other features when used alone.

A.8.2 FEATURE COMBINATION PERFORMANCE

Next, we explore the synergistic effects of combining different feature types. Table A4 illustrates the performance scores for various feature combinations.

The combination of title features with other modalities consistently improved performance. Notably, the combination of thumbnail, description, and caption features achieved the highest score of 0.83, demonstrating the complementary nature of these modalities in predicting video popularity.

A.8.3 COMPLEX FEATURE COMBINATIONS

Finally, we investigated the impact of combining four or five feature types. Table A5 shows the performance scores for these complex feature combinations.

The combination of videotext, thumbnail, video, description, and caption features achieved the highest score of 0.83. However, it's important to note that this score is not significantly higher than some of the triple feature combinations, suggesting a point of diminishing returns in terms of prediction accuracy as we increase feature complexity.

These results highlight the importance of considering multiple modalities in video popularity prediction. While individual features provide valuable information, the combination of complementary

Combination	<TITLE>			
-	**75**	**75**	73	74
<TITLE>	-	**76**	**76**	72
<TITLE> &	-	-	**77**	76
& & <TITLE>	-	-	-	**78**

Table A4: Performance analysis of multimodal feature combinations for video popularity prediction. Icons represent title and description (<TITLE>), video content ( ), thumbnail ( ), and caption ( ). Bold numbers indicate the highest score in each row.

Feature 1	Feature 2	Feature 3	Feature 4	Feature 5	Score
Videotext	Thumbnail	Video	Title	Description	0.81
Videotext	Thumbnail	Video	Title	Caption	0.81
Videotext	**Thumbnail**	**Video**	**Description**	**Caption**	**0.83**
Videotext	Thumbnail	Title	Description	Caption	0.82
Videotext	Video	Title	Description	Caption	0.80
Thumbnail	Video	Title	Description	Caption	0.82

Table A5: Performance Scores for Complex Feature Combinations

features leads to improved prediction accuracy. However, the experiments also reveal a trade-off between model complexity and performance gains, as the most complex feature combinations do not necessarily yield significantly better results than some simpler combinations. The analysis suggests that careful feature selection and combination can lead to efficient and effective video popularity prediction models. In particular, the combination of thumbnail, description, and caption features appears to be a strong predictor of video popularity, achieving the highest score of 0.83. This combination likely captures a diverse range of information about the video content, including visual appeal, textual context, and spoken content.

These findings suggest that while incorporating multiple modalities can improve prediction accuracy, there is a point of diminishing returns. Future model development should focus on optimizing the balance between feature complexity and performance gains, potentially prioritizing the most informative feature combinations identified in this study.

A.9 Comparison with Vision-Language Large Model

To further strengthen our contribution, we conducted additional experiments using an advanced Vision-Language Large Model (VLLM), Gemini. Specifically, we used the Gemini 1.5 Pro model, accessed as API in Vertex library. The input to the model included a combination of the summariser prompt and the final prediction prompt. The sequential frames were aligned with the video's existing captions to ensure that visual and verbal elements were synchronized before being fed into the LLM to generate the video-to-frame summary for the entire video. This process closely follows the steps described in Section 3.2 where "Frame Extraction" step (extracting 5 frames per minute) and the "Caption Matching and Data Integration" step were employed to align captions and frames for LLM processing. Subsequently, "Frames to Text Conversion and Summarization" step generated the final video summary using an LLM call. For prediction, we used the same final prompt detailed in Figure A2, ensuring consistency with our primary method.

These experiments confirmed that our strategy—incorporating sequential prompting, hypothesis generation, and supervised signals—consistently improves prediction performance, even with this

state-of-the-art model (see Figure A11. This underscores the generalizability and robustness of our approach across different model architectures.

Notably, this result demonstrates that more advanced models do not inherently outperform others across all aspects; rather, each model tends to excel in specific areas. This highlights the importance of strategically combining models based on their unique strengths to achieve optimal results. Additionally, the prompting strategies developed in this work offer practical guidance for designing effective multimodal solutions. These insights not only inform future research on multimodal learning but also emphasize the value of integrating pre-trained models tailored to specific task requirements.

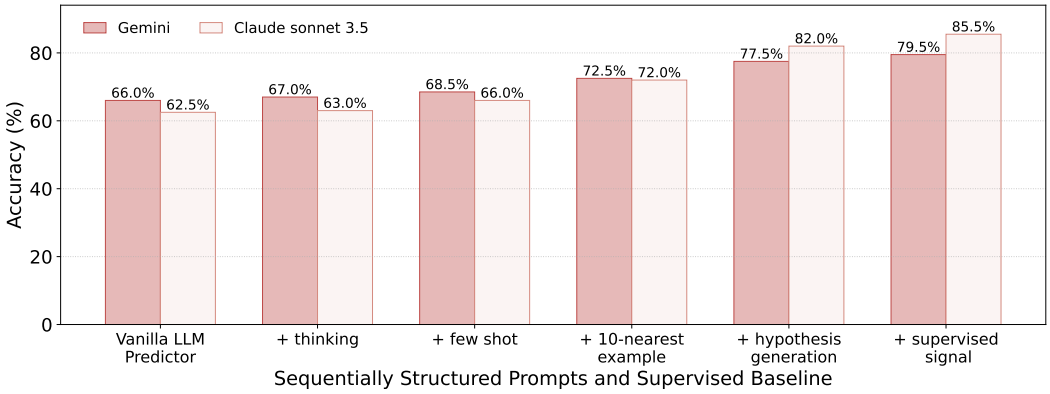

Figure A11: Performance evaluation of the Gemini 1.5 Pro model for video popularity prediction, demonstrating the impact of sequential prompting, hypothesis generation, and supervised signals.