Multi-Criteria Evaluation Framework of Selecting Response-worthy Chats in Live Streaming

Zhantao Lai
Bandai Namco Research Inc.
z-lai@bandainamco-mirai.com

Kosuke Sato
Bandai Namco Research Inc.
k18-sato@bandainamco-mirai.com

Abstract

Live streaming, a dynamic medium that merges real-time audiovisual content with interactive text-based chat, presents unique challenges for maintaining viewer engagement and ensuring streamers' well-being. This study introduces a multi-criteria evaluation framework designed to identify response-worthy chats during live streaming. We proposed a system that evaluates chats based on sentiment polarity and intensity, contextual relevance, and topic uniqueness. We also constructed a dataset annotated by human reviewers who validates the framework, demonstrating a closer alignment with human preferences compared to single-criterion baselines. This framework not only supports the development of more responsive and engaging live streaming environments but also contributes to the broader field of dialog systems by highlighting the distinct needs of real-time, large-scale conversational contexts.

1 Introduction

Live streaming, which merges real-time audiovisual content with simple text-based chat, has seen a surge in popularity and is now influential in various sectors (Haimson and Tang, 2017; Hamilton et al., 2014). Live streaming is transforming how streamers and viewers interact online, creating a novel type of dialog system that can either facilitate human interaction or autonomously host the live streaming conversation (Lu et al., 2017). The challenge for these live streaming dialogue systems lies in boosting user engagement, prolonging viewing duration, and improving viewer satisfaction. (Cai and YvetteWohn, 2019).

The main goal of introducing a system to selecting chats in live streaming is to address the challenges that human streamers face due to their limited time and capabilities. For instance, when dealing with large audiences, it's not feasible for streamers to sift through and reply to every chat during live interactions with potentially thousands of viewers. Automation can support streamers by helping them identify important chats and craft responses. Additionally, the demands of streaming for extended periods and frequently can take a toll on streamers' health, both physically and mentally. Through the implementation of automation in live streaming, we can reduce the burden on streamers and contribute to their overall well-being (Lu et al., 2019).

Figure 1: Go Round Game (GoRanGe) is an experimental AI YouTuber project from Bandai Namco Entertainment. The proposed dataset in this study comprises a selection of chats obtained from this project.

Research into dialogue systems, both traditional and those tailored for live streaming, reveals distinct differences in their design and functionality. Traditional dialogue systems are built for one-on-one interactions, whereas those for live streaming must handle simultaneous real-time conversations with numerous users. This demands that the system quickly processes inputs from potentially thousands of participants (DeVito et al., 2017). While traditional dialogue systems strive to offer a personalized experience, those on live streaming also need to personalize but prioritize delivering responses that are relevant to a wide audience (BWalther, 1996). Content moderation is a feature of traditional dialogue systems, but it is not as critical as it is for live streaming. Here, dialogue systems

Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue, pages 186–191
September 18–20, 2024. ©2024 Association for Computational Linguistics

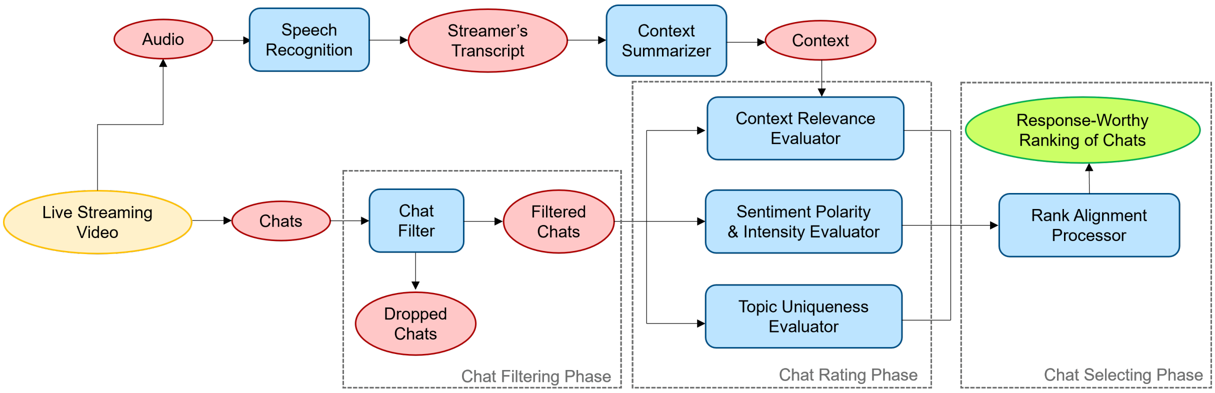

Figure 2: The architecture of the framework used to evaluate chats is illustrated. Live streaming video (input) is shown in yellow, processed data in red, pipeline components in blue and evaluation results (output) in green. The diagram is divided into three sections by dashed lines, with each section corresponding to one of the three phases in the evaluation pipeline.

require advanced monitoring and filtering tools to immediately address and eliminate any inappropriate content (Seering et al., 2017).

This study proposed a novel pipeline in capturing the most interactively significant chats from the real-time interactions in live streaming. The key contributions of this study include:

- We proposed a framework for evaluating chats in live streaming with multiple assessment criteria.
- We constructed a dataset annotated by humans to validate our framework, demonstrating its closer alignment with human preferences when compared to the baseline.

2 Related Work

Automated dialog systems for live streaming systems can be categorized into two types: those that partly assist human interaction and those that are fully automated, with an AI streamer taking the place of a human host. An example of the former is NightBot[1], a tool used on platforms such as Twitch, YouTube and Trovo. It helps manage live chats by filtering out spam and facilitating custom chat commands. The framework in our study incorporated a module for filtering that draws on strategies similar to NightBot. However, these assisted systems rely on a predefined set of keywords to filter or respond, which can limit their ability to adapt to the dynamic context of live streams.

On the other hand, fully automated live streaming systems are often performed as VTubers, or virtual YouTubers (Lu et al., 2021). These are streamers who utilize animated avatars. AI-hosted VTubers generate replies and animate their avatar's expressions and movements by feeding chats into a large language model. For instance, Neuro-sama[2] is recognized for engaging in smooth dialogue with viewers. However, it was temporarily banned from Twitch for generating hateful speech and has shown difficulty in grasping the context of conversations (Seiji, 2023). AI streamers are also expected to not only chitchatting but also handling multimodal information. The open-source framework Luna AI[3] equips AI streamers with tools for voice and singing synthesis, as well as image generation. Meanwhile, GoRoundGame[4] presents an AI streaming project tailored for gaming broadcasts. AI streamers in GoRoundGame streams while playing mahjong against another AI streamer but struggles to strike a balance between commenting on the game and interacting with chats. We gathered chat data from a segment of the GoRoundGame live stream replays and included it in the evaluation dataset.

3 Framework

Figure 2 presents the proposed framework for evaluating chat from viewers in this study. The framework is designed to filter, evaluate, and finally identify the response-worthy chats. This process is structured into three distinct phases: chat filtering, chat rating, and chat selection. In chat rating phase,

[1] https://nightbot.tv/

[2] https://www.twitch.tv/vedal987

[3] https://github.com/0x648/luna-ai

[4] https://virtualyoutuber.fandom.com/wiki/Go_Round_Game

three criteria are employed: sentiment polarity and intensity, contextual relevance, and topic uniqueness.

3.1 Chat Filtering

The objective of this phase is to review viewers' chats and identify those that are unsuitable for interaction. This includes chats that are too brief to convey meaningful content, those that include personal attacks or violate social norms, and chats that are off-topic such as advertisements. The filtering process is achieved through four methods: removing chats that do not meet the established character count threshold, excluding chats with symbols like "http" or "@", which are often associated with promotional content, eliminating chats that contain predefined banned words, and using a language model to evaluate the potential harm of chat content, discarding any chats that surpass a harmfulness score threshold. In this study, We utilized OpenAI's Content Moderation[5] for harmful chats detection.

3.2 Chat Rating

The aim of the chat rating phase is to evaluate chats using various criteria. Since these criteria are measured on different scales, we use the relative positions of the chats in a ranked order rather than their absolute numerical scores. These rankings are then applied in the chat selecting phase. The criteria for ranking are as follows:

Sentiment Polarity and Intensity This criterion assesses the emotional tone and strength in the viewers' chats. We predict the sentiment polarity and intensity for each chat by applying a BERT model that has been finetuned on the WRIME dataset (Tomoyuki et al., 2021). Chats that express a positive tone and exhibit a higher intensity are assigned better rankings.

Contextual Relevance This criterion evaluates how closely the chats align with the ongoing discussion in the live stream. For this purpose, we transcribe the steamer's speech from YouTube videos into transcript by Whisper-v3[6] and periodically summarize the transcript by OpenAI's GPT-4[7] to capture the essence of the live topic. We then encode the summary of the current topic and the chats into vector by utilizing OpenAI's text-embedding-ada-002 and measure the cosine similarity between them. Chats that show a closer vector alignment with the topic summary, indicating greater relevance, receive higher rankings.

Topic Uniqueness This criterion is designed to gauge the informational richness and specificity of the viewers' chats in relation to the live stream's subject. In our approach, we create a matrix that identifies co-occurring keywords within each chat using Rapid Automatic Keyword Extraction (RAKE) (Stuart et al., 2010), and assign a score to each word based on its frequency within the chat's keywords compared to its overall frequency across all chats. The aggregate of these scores for the words in a chat reflects its uniqueness. Consequently, chats that include phrases with higher aggregate scores are deemed to have greater uniqueness and are ranked accordingly.

3.3 Chat Selecting

The objective of this phase is to identify the response-worthy chats by utilizing the rankings derived from previous phase. We employ the Reciprocal Rank Fusion (RRF) (Cormack et al., 2009), a prevalent algorithm in search systems, to amalgamate the three distinct sets of rankings into a unified ranking. From this ranking, we select the highest-ranked viewer chats for interaction as results.

4 Evaluation

This chapter discusses the evaluation of the proposed multi-criteria framework for selecting response-worthy chats in live streaming. It involves the creation of a dataset from YouTube live streams, annotated by human reviewers to reflect preferences. The framework's accuracy is compared to single-criterion baselines, showing improved alignment with human selections, and highlights differences between AI-hosted and human-hosted streams.

4.1 Dataset

To evaluate our proposed framework, we created a dataset from YouTube live streaming replays by following steps:

Replays Selection We selected 28 replays, with 12 hosted by human VTubers and 16 by AI, to account for potential differences in viewer interaction and content. We used the YouTube Data API to

[5] https://platform.openai.com/docs/guides/moderation

[6] https://huggingface.co/openai/whisper-large-v3

[7] https://platform.openai.com/docs/models

collect all chat messages and their corresponding timestamps, ensuring that any personal information, except for the text and the posting time, was excluded.

Periodic Extraction of Chats For the chat evaluation phase, we converted the video replays into audio to enable speech recognition. We then grouped the viewer chats into 5-second intervals based on when they were posted. Each group, containing all messages sent during that interval, was considered as a single input batch. We excluded any batch with no chat or only one chat. Consequently, we gathered 20,514 batches of chats, with an average of 11.91 chats per batch.

Human Annotation To gather labels that match human preferences, we recruited crowd-workers to take on the role of streamers and review YouTube live stream replays. Crowd-workers were between 20 and 40 years old, regularly viewed VTuber live streams. Their task was to identify the most response-worthy chat from a batch and note its id. If no chat in the batch was appropriate for a response, they could label it as 'no reply'. Any batch labeled 'no reply' was removed from the final dataset. 10 crowd-workers were involved in this task. Each replay was annotated by a single crowd-worker, who handled all of the chat batches. After the labeling task, we interviewed each crowd-worker to understand their perspectives for choosing the most response-worthy chat.

4.2 Result

We employed the proposed framework to process each batch of the evaluation dataset. The chat id with the highest rank in each batch was designated as the predicted id. We assessed the accuracy by comparing the pipeline's predictions with human labels. Additionally, we contrasted these results with a baseline that utilized only a single criterion in the chat rating phase.

Table 1: Accuracy (%) of the evaluation dataset. Baseline are categorized as follows: (a) utilizes only sentiment polarity and intensity, (b) utilizes only contextual relevance, and (c) utilizes only topic uniqueness. Hybrid w/voting refers to the combination of the three rankings based on a majority vote to determine the final ranking. Hybrid w/RRF indicates the amalgamation of rankings with RRF (our method)

Method	Accuracy (%)		
	Overall	AI-hosted	Human-hosted
Baseline (a)	39.40	47.57	34.50
Baseline (b)	31.17	48.84	20.59
Baseline (c)	32.76	42.10	27.16
Hybrid w/voting	43.84	51.16	39.45
Hybrid w/RRF	55.46	63.39	50.71

The data in Table 1 indicates that using a combination of criteria aligns more closely with human preferences than relying on a single criterion. Additionally, there are noticeable differences between human and AI streamers. For AI streamers, the accuracy of the proposed evaluation method is relatively high, with the uniqueness of the chat topics standing out as the most significant criterion. This may be due to the AI's limited range in generating diverse dialogues, prompting a need to introduce new topics more frequently.

In contrast, the accuracy of the proposed method for human-hosted live streams is lower than that for AI-hosted streams. It has been noted that in streams hosted by humans, viewer emotions tend to vary more, making the sentiment expressed in viewer chats a more critical factor for interaction.

Our survey indicates that when the audience knows the streamer is an AI, their expectations for interaction quality are generally lower than for human streamers. This reduced expectation is often due to the audience for AI streamers being more sensitive to and tolerant of AI technology. For future research, we recommend using live streaming data from human streamers as the evaluation benchmark.

4.3 Perspectives from Crowd-workers

We have collected the perspectives for selecting the most response-worthy chat from crowd-workers and compared those three criteria proposed in this study.

Opinions consistent with our framework's criteria include: steering clear of negative chats, choosing chats pertinent to the ongoing discussion, favoring chat contributions that stem from the streamer's remarks and have the potential to spark a new conversation.

Conversely, aspects not reflected in our criteria include: giving priority to replies to greetings, particularly for newcomers to the live stream, which can significantly boost viewer loyalty for future sessions. We have also received recommendations to focus more on picking out questions or suggestions, as these often originate from the most engaging viewers.

4.4 Latency

In this study, we compare the outcomes of our evaluation framework with those of human annotators. A key consideration in implementing this framework is its real-time processing capability. The system's latency is influenced by two main factors:

External Factor These include the time required to fetch chats content via the streaming API. This encompasses the frequency of API requests, live broadcast delay settings, and the time it takes for comments to appear on the streaming platform after submission. These response times are largely dictated by the limitations of the live streaming platform and the API's quota restrictions, typically ranging from a few seconds to several tens of seconds, depending on the configuration.

Internal Factor These pertain to the inference time of modules within the framework. Most of these modules complete their inference in under one second. The component with the highest latency is the summarization of chat contexts using GPT-4, which averages several tens to hundreds of milliseconds per token for inference. However, since summarization does not require the most current chat input, it can be processed asynchronously during the latency from external factor. In future research, we also plan to explore the use of local specialized summarization models, such as T5(Raffel et al., 2019),, to replace modules using commercial LLM services, thereby reducing the overall inference time.

5 Conclusion

In this study, we proposed a framework based on various criteria, including sentiment polarity and intensity, contextual relevance, and topic uniqueness—to evaluate view chats in live streaming. We also constructed a dataset reflecting human preferences to assess the performance of above framework. Our findings suggested that a composite criteria better reflects human preferences than a single approach, and identified differences in interaction preferences between human-hosted and AI-hosted live streams.

Moving forward, we plan to improve our method by incorporating feedback from crowd-workers and train a chat-scoring model directly from the labels of human feedback. Additionally, we intend to make this dataset publicly available to support further research in enhancing automated dialog systems for live streaming.

References

Joseph BWalther. 1996. Computer-mediated communication: Impersonal, interpersonal, and hyperpersonal interaction. *Communication research*, 23(1):3–43.

Jie Cai and Donghee YvetteWohn. 2019. Live streaming commerce: Uses and gratifications approach to understanding consumers' motivations. In *Proceedings of the 52nd Hawaii International Conference on System Sciences*.

Gordon V. Cormack, Charles L A Clarke, and Stefan Buettcher. 2009. Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, page 758–759. Association for Computing Machinery.

Michael A DeVito, Jeremy Birnholtz, and Jeffery T Hancock. 2017. Platforms, people, and perception: Using affordances to understand self-presentation on social media. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*.

Oliver L Haimson and John C Tang. 2017. What makes live events engaging on facebook live, periscope, and snapchat. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing System*, pages 48–60.

William A Hamilton, Oliver Garretson, and Andruid Kerne. 2014. Streaming on twitch: fostering participatory communities of play within live mixed media. In *Proceedings of the SIGCHI Conference on Human Factors in Computing System*, pages 1315–1324.

Zhicong Lu, Michelle Annett, and Daniel Wigdor. 2019. "i feel it is my responsibility to stream" streaming and engaging with intangible cultural heritage through livestreaming. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*.

Zhicong Lu, Chenxinran Shen, Jiannan Li, Hong Shen, and Daniel Wigdor. 2021. More kawaii than a real-person live streamer: Understanding how the otaku community engages with and perceives virtual youtubers. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*.

Zhicong Lu, Haijun Xia, Seongkook Heo, and Daniel Wigdor. 2017. You watch, you give, and you engage: a study of live streaming practices in china. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing System*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *CoRR*.

Joseph Seering, Robert Kraut, and Laura Dabbish. 2017. Shaping pro and anti-social behavior on twitch through moderation and example-setting. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*.

Narita Seiji. 2023. Ai vtuber neuro-sama is back from its twitch ban and acting as strange as ever. *Automation. Active Gaming Media*.

Rose Stuart, Engel Dave, Cramer Nick, and Cowley Wendy. 2010. *Automatic Keyword Extraction from Individual Documents*. John Wiley & Sons Inc.

Kajiwara Tomoyuki, Chu Chenhui, Takemura Noriko, Yuta Nakashima, and Hajime Nagahara. 2021. WRIME: A new dataset for emotional intensity estimation with subjective and objective annotations. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2095–2104. Association for Computational Linguistics.

A Example of Batch in Evaluation Dataset

Table 2: Example of Batch in Evaluation Dataset. The original texts are in Japanese, and the examples provided in the table are translated into English.

Video ID	DtAFgs_gAzE
Video Title	[First Broadcasting] The Debut of AITuber Popuri!
Batch ID	31
Batch Context	Hello everyone, my name is Popuri Miyako. Nice to meet you!
Batch Chats	1: Hello Popuri-chan, it's nice to meet you! 2: Hello♪ 3: Congratulations on Popuri-chan's debut!! 4: Popuri-chan! 5: :clapping_hands::clapping_hands: 6: This BGM is pleasant 7: LoL
Response Flag	True
Response Chat ID	3
Response Chat	Congratulations on Popuri-chan's debut!!