Aggregating Impressions on Celebrities and their Reasons from Microblog Posts and Web Search Pages by LLMs

Hibiki Yokoyama[1]**, Rikuto Tsuchida**[1]**, Kosei Buma**[1]**, Sho Miyakawa**[1]**,**
Takehito Utsuro[1]**, Masaharu Yoshioka**[2]**,**

[1]University of Tsukuba, [2]Hokkaido University,
s2320808@_u.tsukuba.ac.jp, s2110466@_u.tsukuba.ac.jp, s2313594@_u.tsukuba.ac.jp,
s2320794@_u.tsukuba.ac.jp, utsuro@_iit.tsukuba.ac.jp, yoshioka@_ist.hokudai.ac.jp,

Abstract

This paper aims to augment fans' ability to critique and explore information related to celebrities of interest. First, we collect posts from X (formerly Twitter) that discuss matters related to specific celebrities. For the collection of major impressions from these posts, we employ ChatGPT as a large language model (LLM) to analyze and summarize key sentiments. Next, based on collected impressions, we search for Web pages and collect the content of the top 30 ranked pages as the source for exploring the reasons behind those impressions. Once the Web page content collection is complete, we collect and aggregate detailed reasons for the impressions on the celebrities from the content of each page. For this part, we continue to use ChatGPT, enhanced by the retrieval augmented generation (RAG) framework, to ensure the reliability of the collected results compared to relying solely on the prior knowledge of the LLM. Evaluation results by comparing a reference that is manually collected and aggregated reasons with those predicted by ChatGPT revealed that ChatGPT achieves high accuracy in reason collection and aggregation. Furthermore, we compared the performance of ChatGPT with an existing model of mT5 in reason collection and confirmed that ChatGPT exhibits superior performance.

1 Introduction

In recent years, social networking services (SNS) such as X (formerly Twitter) have become platforms where various opinions are expressed. As shown in Figure 1, a significant number of posts on these platforms contain impressions and critiques of celebrities, often triggered by events such as TV drama broadcasts, commercials, or news reports. Among celebrity fans, there are individuals who have a strong interest in this type of information. For example, when an event or an incident that is related to a popular celebrity occurs, people express their own thoughts regarding those events and incidents in SNS such as microblog (e.g., X) posts. Since a number of those posts are distributed through SNS, this makes it unexpectedly difficult to correctly identify what people actually intend to express in their posts. The reasoning behind these impressions is often implicit and can be influenced by various factors such as the stance of the writers of the posts, recently occurring related events, and the contexts provided by external sources like news articles and ads. However, such background information is not always detailed in the posts themselves. Therefore, it is necessary to utilize not only the information within the posts but also external information to gain comprehensive understanding.

Considering those situations, this paper aims to augment fans' ability to critique and explore information related to celebrities of interest. To achieve this overall goal, we first collect posts from X that discuss matters related to specific celebrities and gather major impressions on those celebrities.

Our ChatGPT-based approach overcomes the limitations of a previous research (Sugawara and Utsuro, 2022), allowing for a more flexible and comprehensive collection of aspects and impressions about celebrities. The details of our ChatGPT-based method for collecting and aggregating impressions are to be explained in Section 4. This approach allows us to more effectively identify and aggregate the major impressions on celebrities' aspects from the vast amount of information available in X posts, while taking into account the context of the posts. This enables a more comprehensive and nuanced understanding of the public's perceptions of celebrities, going beyond the limitations of the previous method. We then use the corresponding pair of a celebrity's aspect and an impression as a keyword for collecting detailed information and their reasons from Web pages.

After selecting the keyword, we search for Web

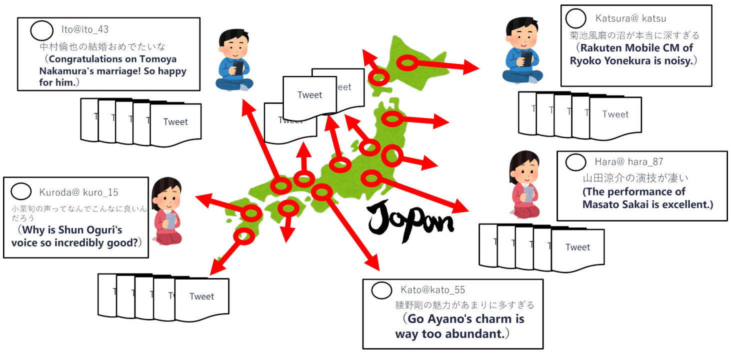

Ito@ito_43
中村倫也の結婚おめでたいな
（Congratulations on Tomoya Nakamura's marriage! So happy for him.）
Tweet
Kuroda@ kuro_15
小栗旬の声ってなんでこんなに良いんだろう
（Why is Shun Oguri's voice so incredibly good?）
Tweet
Tweet
Tweet
Katsura@ katsu
菊池風磨の沼が本当に深すぎる
（Rakuten Mobile CM of Ryoko Yonekura is noisy.）
Tweet
Hara@ hara_87
山田涼介の演技が凄い
(The performance of Masato Sakai is excellent.)
Tweet
Japan
Kato@kato_55
綾野剛の魅力があまりに多すぎる
（Go Ayano's charm is way too abundant.）

Figure 1: Numerous Posts on Celebrities triggered by Various Celebrities related Events

pages using the keyword as a query and collect the content of the top 30 ranked pages as the source for exploring the reasons behind the impressions.

Once the Web page content related to the keyword is collected, we explore detailed reasons for the impressions within the content of each page. For this part, we utilize ChatGPT as a large language model (LLM). A crucial aspect of our research is the use of RAG (Lewis et al., 2020) in this reason collection process, which plays a significant role in enhancing the reliability of LLM outputs. The RAG framework allows LLMs to refer to information retrieved from external databases, thereby improving the reliability of the generated content. In this paper, we aim to enhance the reliability of the collected results by leveraging the RAG framework to collect reasons for impressions based on the content of Web pages, compared to relying solely on the prior knowledge of LLMs. We also show that ChatGPT outperforms an existing model of mT5 (Xue et al., 2021) in reason collection.

The reasons for impressions obtained through this method, however, are highly duplicated and hence redundant, making it difficult for users to recognize the critiques and related information about celebrities at a glance. Therefore, we categorize and rank the multiple reasons for impressions obtained for each keyword, considering the frequency of the reasons. This allows users to easily understand the reasons behind the impressions on celebrities' aspects in an aggregated ranked format, enabling the exploration of critiques and relevant information on celebrity-related topics. We employ ChatGPT also for this part.

The followings give the contribution of this paper:

1. We proposed a novel approach using ChatGPT, a large language model, to effectively collect and aggregate impressions on celebrities' aspects from X posts.
2. In the RAG framework, we showed that ChatGPT is highly effective in collecting and aggregating reasons for the impressions on celebrities from Web pages.
3. In collecting reasons for impressions on celebrities, we demonstrated that ChatGPT outperforms mT5, highlighting the effectiveness of ChatGPT in extracting relevant information from Web pages.

2 Related Work

Previous work on assisting information access regarding celebrities includes studies on constructing large-scale celebrity profile datasets by combining Twitter and Wikidata (Wiegmann et al., 2019) and analyzing persuasion strategies in celebrities' language use on social media to predict their influence (Chang et al., 2021). Regarding assisting fans of celebrities, previous work includes studies on determining the relationship between celebrities and impressions in microblog posts (Nozaki et al., 2022) and those on mining impressions on celebrities' aspects in microblog posts (Sugawara and Utsuro, 2022). This paper differs from those previous work in that we search Web pages for reasons behind impressions on celebrities' aspects mined from microblog posts. This paper also differs from the previous work in that we employ ChatGPT, a large language model, to extract impressions on celebrities' aspects from microblog posts, while the previous studies relied on other methods such as co-occurrence frequency statistics.

Furthermore, one of the key characteristics of this paper is the use of RAG (Lewis et al., 2020), which improves the reliability of LLM-generated output by allowing reference to external information. RAG allows LLMs to refer to information

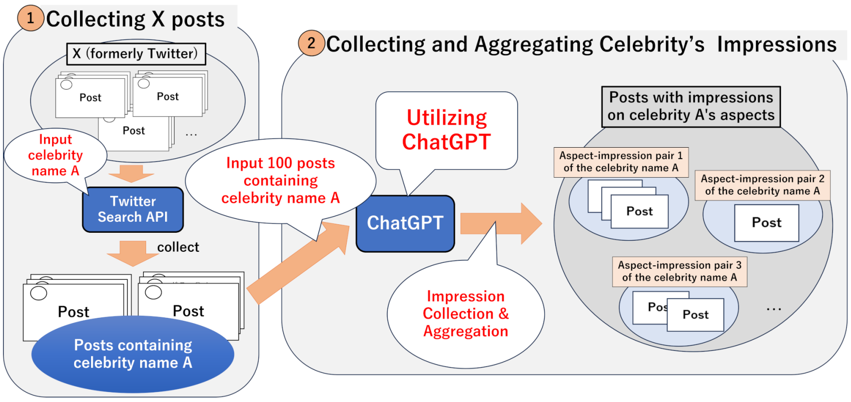

①Collecting X posts
X (formerly Twitter)
Post
Post
Post
…
Input celebrity name A
Twitter Search API
collect
Post
Post
Posts containing celebrity name A
Input 100 posts containing celebrity name A
②Collecting and Aggregating Celebrity's Impressions
Utilizing ChatGPT
ChatGPT
Impression Collection & Aggregation
Posts with impressions on celebrity A's aspects
Aspect-impression pair 1 of the celebrity name A
Post
Aspect-impression pair 2 of the celebrity name A
Post
Aspect-impression pair 3 of the celebrity name A
Pos
Post
…

Figure 2: Overview of Collecting and Aggregating Impressions on Aspects of Celebrities from X Posts

retrieved from external databases, enhancing the accuracy and trustworthiness of the generated content. In this paper, we leverage RAG to collect reasons for impressions based on the content of Web pages, aiming to significantly improve the reliability of the collected results compared to traditional methods that rely solely on the knowledge stored within the LLMs. Recent studies have explored various RAG applications and improvements, such as context tuning for tool retrieval and plan generation (Anantha and Vodianik, 2024) improving open-domain table question answering with late interaction models and joint training (Lin et al., 2023), few-shot multilingual image captioning without requiring supervised training (Ramos et al., 2023), and incorporating additional components for more powerful question answering systems (Tan et al., 2023). Other works have focused on improving zero-shot performance on low-resource languages using prompts from high-resource languages (Nie et al., 2023), leveraging retrieval for non-knowledge-intensive tasks with a two-stage framework (Guo et al., 2023), and incorporating rich answer encoding for better generation quality in knowledge-intensive tasks (Huang et al., 2023).

ChatGPT-related research also includes entity linking (Peeters and Bizer, 2023), and dialogue analysis (Finch et al., 2023), and text summarization (Zhang et al., 2023b; Pu and Demberg, 2023; Zhang et al., 2023a). This paper differs in that we utilize ChatGPT for both collecting and aggregating impressions on celebrities' aspects, as well as collecting and aggregating the reasons for these impressions.

3 Aspect, Impression, and Reason

In this study, we define "aspect", "impression", and "reason" as follows:

aspect: a specific attribute, characteristics, or topic related to a celebrity. This can include physical features, skills or talents, specific works or performances, interactions or relationships, behaviors, or other notable elements of their public persona.

impression: a subjective opinion, evaluation, or feeling about a celebrity's aspect, often expressed through adjectives, descriptive phrases, or statements of recognition.

reason: the underlying explanations, justifications, or evidences that support a particular impression about a celebrity's aspect. Reasons are typically more detailed and context-rich than impressions, often found in longer-form content such as Web articles or detailed social media posts.

4 Collecting Impressions from X Posts using ChatGPT

This section describes the procedure of collecting posts containing celebrity names from X, identifying posts that mention impressions on specific aspects of celebrities, and aggregating them into aspect-impression pairs using ChatGPT. An overview is illustrated in Figure 2.

4.1 Collecting X posts

In this paper, we selected 10 celebrities who are frequently discussed on X and collected posts using their names as search queries from September 7,

celebrity name	number of posts	number of non-repost posts
Ryosuke Yamada	938,882	213,886
Kazunari Ninomiya	851,579	164,130
Fuma Kikuchi	1,131,863	185,823
Shun Oguri	425,188	120,583
Go Ayano	272,232	95,890
Kentaro Sakaguchi	284,622	64,472
Ryoma Takeuchi	83,368	31,670
Kasumi Arimura	370,956	105,084
Tomoya Nakamura	702,807	206,632
Mei Nagano	246,489	58,636
Total	5,307,986	1,246,806

Table 1: Numbers of Collected Posts for Each Celebrity Name

celebrity name	aggregated aspect	impression
Ryosuke Yamada	beauty	outstanding
	quality of dance	high
	interaction with Daiki Shigeoka	touching
	Karubi harassment	funny
	kidnapping of Jr.	cute
	eye contact with camera	charming
Kazunari Ninomiya	movie “Ragelee yori Ai wo Komete”	masterpiece and moving
	acting skills	recognized as a good actor
	activities during year-end and New Year	enjoyment for fans
	personality	loved and respected by fans
	radio program	enjoyment for fans

Table 2: Examples of Aspect-Impression Pairs aggregated by ChatGPT

2022 to April 9, 2023. This process is depicted in the “Collecting X posts” part of Figure 2. The Twitter Search API[1] was used for post collection. The numbers of posts and non-repost posts collected for each celebrity name are shown in Table 1. In this paper, we only use non-repost posts.

4.2 Collecting/Aggregating Impressions from Posts

Next, we perform two main tasks on the X posts containing a specific celebrity name collected in the previous section. First, we collect posts that mention impressions on specific aspects of that celebrity. Second, we aggregate the collected information into aspect-impression pairs. These tasks are illustrated in the “Collecting and Aggregating Celebrity’s Impressions” part of Figure 2. As the framework for these tasks, we utilize ChatGPT[2] model, specifically `gpt-4-turbo-2024-04-09`. The specific prompts given to ChatGPT are shown in Figure 5 of Appendix A. Here, we show an example of a prompt targeting the celebrity “Ryosuke Yamada”. The prompts begin by providing posts, and instruct to first collect posts that mention what aspects (impression targets) of Ryosuke Yamada and what kind of impressions are associated with those aspects. Next, it instructs to aggregate the collected posts based on the impression targets and their corresponding impressions. The desired output format is then specified, indicating to output the pairs of the impression target and the corresponding impression, along with the specific relevant posts. Due to the limitation of input token numbers, 100 posts are provided as an example. The prompts also instruct not to include the celebrity’s name “<Ryosuke Yamada>” in the impression targets, and to be careful not to make the impression targets and their corresponding impressions redundant. Finally, before outputting, it instructs to double-check if there exist any remaining posts that were not collected nor aggregated.

To evaluate the performance of the proposed approach, we manually annotate a subset of the collected posts to create a reference dataset. The evaluation is conducted for both the collection and the aggregation tasks by comparing the outputs generated by ChatGPT with the reference. The results are summarized in Table 3. For the collecting task,

---

[1] https://developer.twitter.com/en/docs/tweets/search/api-reference/get-search-tweets

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

celebrity name	total posts	collecting celebrity's impressions		aggregating celebrity's impressions	
		recall $\left[\frac{\#\,(\text{ref} \cap \text{collected})}{\#\,\text{ref}}\right]$	precision $\left[\frac{\#\,(\text{ref} \cap \text{collected})}{\#\,\text{collected}}\right]$	recall $\left[\frac{\#\,(\text{ref} \cap \text{aggregated})}{\#\,\text{ref}}\right]$	precision $\left[\frac{\#\,(\text{ref} \cap \text{aggregated})}{\#\,\text{aggregated}}\right]$
Ryosuke Yamada	100	0.42 (=10/24)	0.91 (=10/11)	0.82 (=9/11)	1.00 (=9/9)
Kazunari Ninomiya	100	0.73 (=8/11)	0.80 (=8/10)	0.38 (=3/8)	0.60 (=3/5)
Fuma Kikuchi	100	0.63 (=5/8)	0.71 (=5/7)	0.60 (=3/5)	0.60 (=3/5)
Shun Oguri	100	0.67 (=2/3)	0.40 (=2/5)	0.67 (=2/3)	0.50 (=2/4)
Go Ayano	100	0.69 (=9/13)	0.69 (=9/13)	0.50 (=3/6)	0.60 (=3/5)
Total/Micro Average	500	0.58 (=34/59)	0.74 (=34/46)	0.61 (=20/33)	0.71 (=20/28)

Table 3: Manual Evaluation Results of Collecting/Aggregating Impressions on Aspects of Celebrities

the evaluation results are shown in the "collecting celebrity's impressions" section of Table 3. The table presents the total number of posts used for evaluation in the "total posts" column. The "recall" and "precision" columns display the recall and precision of ChatGPT's performance for the collection task, respectively, where recall is calculated as [# (ref ∩ collected) / # ref] and precision as [# (ref ∩ collected) / # collected]. Similarly, for the aggregation task, the evaluation results are presented in the "aggregating celebrity's impressions" section of Table 3. The evaluation is conducted using the posts that were identified as containing impressions on aspects of celebrities by ChatGPT in the collection task. The "recall" and "precision" columns show the recall and precision of ChatGPT's performance for the aggregating task, respectively, where recall is calculated as [# (ref ∩ aggregated) / # ref] and precision as [# (ref ∩ aggregated) / # aggregated].

Table 2 shows examples of the aspect-impression pairs aggregated by ChatGPT for the celebrities Ryosuke Yamada and Kazunari Ninomiya. The table presents the aggregated aspects and their corresponding impressions for each celebrity. As can be seen from the examples in the table, ChatGPT is capable of capturing and aggregating a wide range of aspects and impressions for both celebrities. For Ryosuke Yamada, this includes physical appearance, performance skills, interactions with others, behavior on variety shows, roles in dramas, and even eye contact with the camera. For Kazunari Ninomiya, ChatGPT aggregates aspects such as his highly reputed movie, acting skills, activities during year-end and New Year, personality, and radio program. These examples demonstrate that our proposed method using ChatGPT can effectively address the limitations of the previous research (Sugawara and Utsuro, 2022), which considered the aspects of celebrities to be in the form of "A (celebrity name)'s B (noun)" and used a language model to determine whether an sentiment relation exists between the celebrity's aspect and the impression. By leveraging the advanced natural language understanding capabilities of ChatGPT, our approach allows for a more flexible and comprehensive analysis of celebrity aspects and impressions. Our method can identify and analyze aspects that may not fit the "A's B" format, capture impressions expressed in various parts of speech, not just adjectives, and consider the broader context of posts, leading to more accurate interpretation of sentiments. Our approach can better identify and aggregate the major impressions on celebrities' aspects from the vast amount of information available in X posts, while taking into account the context of the posts. This allows for a more comprehensive and nuanced understanding of the public's perceptions of celebrities, going beyond the limitations of the previous method.

5 Reason Collection/Aggregation

This section describes the procedure of collecting and aggregating reasons for impressions from Web pages using ChatGPT. Section 5.1 discusses the process of selecting pairs of aspects and their corresponding impressions from those collected and aggregated in Section 4, which will be used as keywords for searching for Web pages. Section 5.2 describes the method for searching for Web pages using the selected keywords and collecting the content of the Web pages. Section 5.3 explains the procedure for collecting reasons for impressions from the collected Web page contents and presents the results of manual evaluation. Section 5.4 discusses the procedure for aggregating the collected reasons for impressions and presents the results of manual evaluation.

5.1 Selecting Aspect-Impression Pairs

In this section, we describe the process of selecting pairs of aspects and their corresponding impressions from those collected and aggregated in

Section 4, which will be used as keywords for searching for Web pages. As will be explained in detail in Section 5.2, we use the Google search engine[3] for searching for Web pages in this study. Among the aspect-impression pairs collected and aggregated in Section 4, some may not yield sufficient number of Web pages when used directly as search keywords on the Google search engine. Examples of such pairs collected for the celebrity "Ryosuke Yamada" include "interaction with Daiki Shigeoka - touching" and "kidnapping of Jr. - cute" as shown in Table 2. Therefore, instead of simple Google searches, it is necessary to make significant efforts in the search process, such as collecting many Web pages related to the celebrity in advance and performing Semantic search or Embedding search (Reimers and Gurevych, 2019; Cer et al., 2018; Karpukhin et al., 2020) within those pages. This is a challenge that should be addressed in the future. Considering this, in this study, we use 10 aspect-impression pairs that are judged to be directly usable as search keywords on the Google search engine for the subsequent processes. These 10 pairs are listed in the "aggregated aspect" and "impression" columns of Table 4 of Appendix B. The aim of the following sections is to clarify whether it is possible to collect and aggregate reasons for impressions using these selected aspect-impression pairs as search keywords.

5.2 Web Page Search

First, we search for Web pages using the Google search engine with the keywords selected in the previous section. Next, we manually collect the content of the top 30 Web pages in the search results. This series of operations are performed for all the keywords. For example, in the case of "Ryosuke Yamada's acting performance - amazing", we first search for Web pages using "Ryosuke Yamada's acting performance - amazing" as the query and collect the content of the top 30 Web pages. Those Web pages are expected to contain reasons for the impressions expressed in the keywords such as "reasons why Ryosuke Yamada's acting performance is amazing".

5.3 Reason Collection

5.3.1 The Procedure

Next, we use the content of the Web pages collected in the previous section to collect reasons for impressions for each Web page. We use the ChatGPT model `gpt-4-0613` as the framework for collecting reasons for impressions. The entire prompt given to ChatGPT is shown in Figure 6 of Appendix B[4]. Here, we show an example of a prompt targeting the keyword "Ryosuke Yamada's acting performance - amazing". First, we use the prompt in Figure 6 of Appendix B to instruct ChatGPT to search for reasons for impressions based on the collected Web pages without using prior knowledge of ChatGPT itself but referring to the content of the retrieved Web pages as added as the context. If the added context information does not contain reasons for impressions, ChatGPT is instructed to output only "not included". By having ChatGPT search for reasons for impressions based on the content of Web pages rather than the prior knowledge of ChatGPT itself, we expect to suppress the output of information that differs from or does not exist in the Web search results at that moment, a phenomenon known as hallucination.

5.3.2 Manual Evaluation

Here, we evaluate the reasons for impressions collected by ChatGPT by comparing them with manually collected reference reasons for impressions, where the evaluation is performed with 10 sets of keywords.

Based on the content of Web pages obtained for each keyword in Section 5.2, the first author manually collected reasons for impressions. For example, for the keyword "Ryosuke Yamada's acting performance - amazing", the first author manually examined each collected Web page and extracted statements that correspond to reasons why "Ryosuke Yamada's acting is said to be amazing". These extracted reasons were compiled into a list for each Web page, serving as our reference data. We then assess whether ChatGPT can output corresponding reasons, allowing for variations in wording.

Based on the multiset[5] of reasons $S(d)$ output by ChatGPT for a given Web page d and the multiset of reference reasons $R(d)$ manually prepared for the Web page d[6], the recall and precision are defined as follows:

---

[3] `https://www.google.co.jp/`

[4] We confirmed through experimental ablation studies that, although all the prompts in Figure 6 of Appendix B and Figure 7 of Appendix B can be replaced with similar sentences, the performance of ChatGPT is severely damaged if any of them is removed.

[5] Note here that it can happen that ChatGPT redundantly outputs a single reason several times from a single Web page d. Similarly, it is allowed that reference reasons manually collected from a single Web page d may include a single reason several times, resulting in a multiset.

$$\text{Recall} = \sum_d |R(d) \cap S(d)| / \sum_d |R(d)|,$$

$$\text{Precision} = \sum_d |R(d) \cap S(d)| / \sum_d |S(d)|$$

The evaluation is performed for each collected Web page, and the micro-average is used as the evaluation result for each keyword. The overall evaluation results are measured as the macro-average of the evaluation results for the total 10 keywords for evaluation. The overall evaluation results for reason collection are shown in Figure 3(a). As a result, in reason collection, high performance around 0.9 are achieved for recall, precision, and F1-score. As will be presented in Table 4 in section 6.2 and in Appendix B, half of the retrieved Web pages are without reasons. Thus, high performance of reason collection by ChatGPT reveals that ChatGPT is highly tolerant of noisy context such as those Web page retrieval errors, where ChatGPT does not collect incorrect reasons even from those noisy Web pages.

5.4 Reason Aggregation

5.4.1 The Procedure

Next, we aggregate the reasons for impressions collected in the previous section. Here, we use the ChatGPT model `gpt-4-1106-preview`. The entire prompt given to ChatGPT is shown in Figure 7 of Appendix B. Again, we show an example of a prompt targeting the keyword "Ryosuke Yamada's acting performance - amazing". First, the prompt in Figure 7 of Appendix B indicates that, the series of instructions are followed by summaries of Web pages related to the specified keyword. Furthermore, we instruct ChatGPT to perform reason aggregation by categorizing reasons given as the content mentioned in each Web page summary. Those instructions represent how ChatGPT aggregates reasons for impressions. In the example of Figure 7 of Appendix B, we begin by stating that we will provide ChatGPT with summaries of Web pages related to the keyword "Ryosuke Yamada's acting performance - amazing". Next, we present examples of categories and the corresponding sentences that are regarded as examples of reasons, instructing ChatGPT to categorize reasons of impressions following these examples. We also instruct ChatGPT to create new categories if the Web page summary includes categories that do not correspond to the provided examples.

By providing category examples in advance, we aim to stabilize ChatGPT's output. The actual category examples and corresponding sentences provided here are totally unrelated to the specified keyword "Ryosuke Yamada's acting performance - amazing". The subsequent instructions are further given with examples to simply guide the output format to obtain results in a format that is easy to automatically interpret. Details on the output format can be found in Appendix D.

5.4.2 Manual Evaluation Procedure

Here, we evaluate the reasons for impressions aggregated using ChatGPT by comparing them with manually aggregated reference reasons, where the evaluation is performed with 10 keywords. The manually aggregated reference reasons were created solely by the first author, who grouped similar reasons from the reference data used in the reason collection step. This process involved carefully examining the collected reasons for each keyword and combining those that expressed similar concepts or ideas, ensuring a concise yet comprehensive set of aggregated reasons. At this point, we define the following seven multisets/sets. Specifically, first, we define S' as the multiset of reasons aggregated by ChatGPT based on the collected reasons, and R as the set of distinct reference reasons prepared manually after aggregation[7]. Next, we define S'_r as the multiset of elements of S', where, for each of their elements, a corresponding reason exists in R. In contrast, we define $S'_{\neg r}$ as the multiset of the elements of S', where, for each of their elements, no corresponding reason exists in R. Then, we obtain S_r as the set of elements of S'_r by aggregating multiple reasons corresponding to a single reason in R into one reason. In contrast, we also obtain $S_{\neg r}$ as the set of elements of $S'_{\neg r}$ by aggregating multiple reasons into one reason. Finally, we define S as the union of S_r and $S_{\neg r}$.

Based on these multisets/sets, recall, precision, and redundancy are defined as follows. Here, redundancy measures how well ChatGPT can avoid redundancy in the aggregated reasons by comparing the number of redundant reasons before and after aggregation. The overall evaluation results are calculated as the macro-average of the evaluation results for the total 10 keywords.

---

[6] See Appendix C.1 for details on the inter-annotator agreement in reason collection.

[7] See Appendix C.2 for details on the inter-annotator agreement in reason aggregation.

$$\text{Recall} = |S_r|/|R|, \ \ \text{Precision} = |S_r|/|S|,$$

$$\text{Redundancy} = |S'_r|/|S_r|$$

5.4.3 Manual Evaluation Results

We conducted an experiment to investigate the impact of providing category examples within the prompts on the manual evaluation results. The overall evaluation results for reason aggregation are shown in "w/ examples" of the "Reason Aggregation" section in Figure 3. As a result, in reason aggregation with examples, recall was 0.77, precision was 0.88, F1-score was 0.81, and redundancy was 1.40, where, overall, reason aggregation with examples outperforms that without examples in terms of recall and F1-score, while it was more redundant than that without examples, simply because it outputs more reasons than that without examples. These results suggest that by excluding category examples, ChatGPT can perform more concise and accurate categorization of reasons. However, it also tends to fail in detecting several reasons as illustrated in the damage in recall. In other words, providing examples allows ChatGPT to generate results closer to human annotations, but at the cost of potentially performing more redundant categorization.

6 Automatic Evaluation of Reason Detection/Collection

6.1 Task Definition: Reason Detection/Collection

In this section, we focus on two tasks for automatic evaluation: reason detection and reason collection. Given a keyword consisting of a celebrity name, an aspect and an impression such as "Ryosuke Yamada's acting performance - amazing" and a retrieved Web page in relation to the keyword, the reason detection task outputs a binary judgment whether or not there exist one or more reasons in the retrieved Web page for the question composed from the keyword as in "Why is Ryosuke Yamada's acting performance said to be amazing?". The output of the reason detection task is "YES" or "not included". The evaluation metrics for this task are recall, precision, and F1-score based on the reference judgment result.

The reason collection task aims to generate reasons for impressions from retrieved Web pages, where the inputs to the task are the same as the reason detection task. In the automatic evaluation, ROUGE-L is used as an evaluation metric for the reason collection results by ChatGPT and mT5, which measures the longest common subsequence between the generated reasons and the manually created reference reasons[8] . The sentences corresponding to reasons are rarely concentrated in one location but often span multiple parts within the text. Therefore, it is more appropriate to apply the procedure of generating reasons based on context rather than extracting reason chunks from the context. Here, we apply mT5 (Xue et al., 2021) as a comparison to ChatGPT for both tasks.

6.2 Evaluation

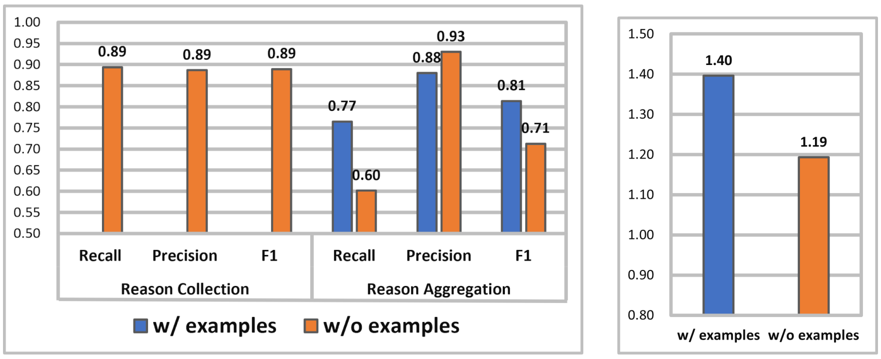
This section describes the automatic evaluation procedure for reason detection and reason collection by ChatGPT and mT5. The dataset for fine-tuning mT5[9] was created using the Web pages collected in section 5.2. Specifically, first, question sentences for fine-tuning of mT5 are set based on the keywords used for collecting Web pages. For example, for a Web page collected in relation to the keyword "Ryosuke Yamada's acting performance - amazing", the question sentence is set as "Why is Ryosuke Yamada's acting performance said to be amazing?" Next, the context for answering the question is set. Here, the collected Web pages are first split into sentences by periods, and then split sentences are concatenated as a chunk under the restriction of satisfying the input token length upper bound of mT5. After that, for each chunk, if it contains sentences that are the reasons for impressions, all the relevant sentences are manually extracted and combined to form the reference answer. If no chunk contains a sentence that is regarded as the reason for impressions, the Web page is judged as unanswerable to the question and "" (blank) is set as the reference answer. The statistics of the numbers of Web pages are shown in Table 4 of Appendix B.

For ChatGPT, as in Figure 6 of Appendix B, it is instructed to output "not included" if there is no reason. Thus, if only "not included" is output, it is treated as no reason is observed. For mT5, if the output is "" (blank), it is treated as no reason is observed. Moreover, there could be cases where

---

[8] While ROUGE-L relies on exact string matching, future work will explore metrics that better capture embedding based semantic similarity beyond string matching, such as BERTScore, BARTScore, and SentenceBERT, for a more comprehensive evaluation.

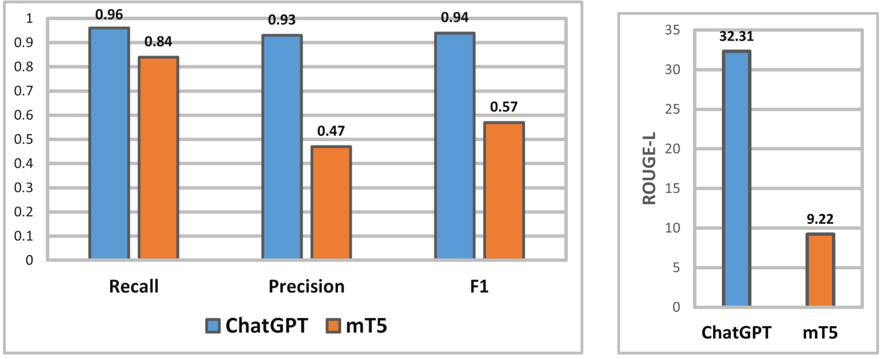
[9] `https://huggingface.co/google/mt5-base`



(a) Recall, Precision, and F1 Score

(b) Redundancy of Reason Aggregation

Figure 3: Manual Evaluation Results of Reason Collection/Aggregation by ChatGPT



(a) reason detection

(b) reason collection

Figure 4: Automatic Evaluation Results of Detecting/Collecting Reasons for Impressions

mT5 outputs only symbols such as "." or "?", which are also treated as no reason is observed.

For the training and evaluation of mT5, 5-fold cross validation was performed based on the dataset shown in Table 4 of Appendix B, where each line is counted as the unit of 5-fold cross validation. In each fold of 5-fold cross validation, the dataset for 8 out of the 10 keywords shown in Table 4 was used as the training data[10], and the dataset for the remaining 2 keywords was used as the evaluation data. When generating answers to the evaluation data, answer generation is first performed on each chunk. Then, for each Web page, the generated answers are concatenated and then further used as the context when generating an answer for the whole Web page span again. Similar to reason collection by ChatGPT in section 5.3, this procedure allows for generating an answer for each Web page span[11].

ChatGPT outperformed mT5 for all evaluation results. In the evaluation of reason detection, as shown in Figure 4(a), ChatGPT outperformed mT5 in all the metrics, with a particularly large difference in precision. This means that mT5 tends to erroneously output reasons, corresponding to over detection of reasons. In contrast, ChatGPT achieved over 90% recall and precision. From the ROUGE-L evaluation results shown in Figure 4(b), on the other hand, ChatGPT is able to generate reasons much closer to the reference compared to mT5.

7 Conclusion

In this paper, we proposed a method to augment fans of celebrities to critique and explore information concerning celebrities. We conducted evaluation on the results obtained by the proposed method by comparing them with manually collected and aggregated reasons for impressions. We also evaluated the methods for reason collection with ChatGPT and mT5, confirming that ChatGPT shows higher performance. Beyond mT5, we plan to compare ChatGPT with larger models that have a comparable number of parameters, such as Mistral Large[12] or LLaMA 3 70B[13] , to provide a more meaningful evaluation of the proposed method.

8 Limitations

While our proposed method demonstrates promising results, it is important to acknowledge its limitations. LLMs, including ChatGPT, are prone to hallucinations and may generate plausible but incorrect information. While our use of the RAG framework mitigates this risk, it does not eliminate it entirely. The manual aspects of our data collection and processing methods may pose challenges for exact reproducibility, despite our efforts to provide detailed descriptions.

---

[10] The number of training epochs is set as five.

[11] See Appendix C.3 for details on the inter-annotator agreement in reason detection.

[12] `https://mistral.ai/news/mistral-large/`

[13] `https://ai.meta.com/blog/meta-llama-3/`

Our approach relies on aspect-impression pairs that can be directly used as search keywords. For pairs that do not yield sufficient Web pages, more sophisticated information retrieval techniques, such as semantic or embedding search (Reimers and Gurevych, 2019; Cer et al., 2018; Karpukhin et al., 2020), may be necessary in future research.

9 Ethical Statements

The use of AI to analyze and aggregate information about celebrities raises several ethical concerns that we must address. While we use publicly available information, the aggregation and analysis of this data may have unintended consequences for the individuals involved. We emphasize the importance of using this information responsibly and respectfully. The potential for generating or amplifying false information is a significant concern. We acknowledge that our method, despite safeguards, could inadvertently contribute to the spread of misinformation if not used cautiously. There is also a risk that our system could reinforce existing biases or create echo chambers. We encourage users to seek diverse sources and perspectives beyond what our system provides.

We stress that the intent of this research is not to facilitate unwarranted criticism or invasion of privacy, but to promote more informed and nuanced understanding of public figures and media representation. We recognize the broader implications of developing tools that aggregate and analyze public sentiment. We call for ongoing dialogue about the ethical use of such technologies and their impact on public discourse.

In light of these considerations, we recommend that users of our system approach the generated information critically, cross-reference with reliable sources, and use the tool as a starting point for further exploration rather than as a definitive source of information. As researchers, we commit to continuing to refine our methods to address these limitations and ethical concerns, and to contribute to the responsible development of AI technologies in media analysis.

References

R. Anantha and D. Vodianik. 2024. Context tuning for retrieval augmented generation. In *Proc. UncertaiNLP*, pages 15–22.

D. Cer, Y. Yang, S. Kong, N. Hua, N. Limtiaco, R. S. John, N. Constant, M. Guajardo-Cespedes, S. Yuan, C. Tar, Y. Sung, B. Strope, and R. Kurzweil. 2018. Universal Sentence Encoder. *arXiv preprint arXiv:1803.11175*.

Y. Chang, P. A. Wang, H. Hung, K. Khóo, and S. Hsieh. 2021. Examine persuasion strategies in Chinese on social media. In *Proc. 35th PACLIC*, page 108–118.

S. E. Finch, E. S. Paek, and J. D. Choi. 2023. Leveraging large language models for automated dialogue analysis. In *Proc. 24th SIGDIAL*, pages 202–215.

Z. Guo, S. Cheng, Y. Wang, P. Li, and Y. Liu. 2023. Prompt-Guided Retrieval Augmentation for Non-Knowledge-Intensive Tasks. In *Findings ACL*, pages 10896—10912.

W. Huang, M. Lapata, P. Vougiouklis, N. Papasarantopoulos, and J. Pan. 2023. Retrieval Augmented Generation with Rich Answer Encoding. In *Proc. 13th IJCNLP and 3rd AACL*, pages 1012—1025.

V. Karpukhin, B. Oguz, S. Min, P. Lewis, L. Wu, S. Edunov, D. Chen, and W. Yih. 2020. Dense Passage Retrieval for Open-Domain Question Answering. In *EMNLP*, pages 6769–6781.

P. Lewis, E. Perez, et al. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Proc. 34th NeurIPS*, pages 483–498.

W. Lin, R. Blloshmi, B. Byrne, A. de Gispert, and G. Iglesias. 2023. LI-RAGE: Late Interaction Retrieval Augmented Generation with Explicit Signals for Open-Domain Table Question Answering. In *Proc. 61st ACL*, pages 1557—-1566.

E. Nie, S. Liang, H. Schmid, and H. Schütze. 2023. Cross-Lingual Retrieval Augmented Prompt for Low-Resource Languages. In *Findings ACL*, pages 8320—-8340.

Y. Nozaki, K. Sugawara, Y. Zenimoto, and T. Utsuro. 2022. Tweet review mining focusing on celebrities by MRC based on BERT. In *Proc. 36th PACLIC*, pages 757–766.

R. Peeters and C. Bizer. 2023. Using ChatGPT for entity matching. *arXiv preprint arXiv:2305.03423*.

D. Pu and V. Demberg. 2023. ChatGPT vs Human-authored Text: Insights into Controllable Text Summarization and Sentence Style Transfer. In *Proc. 61st ACL-SRW*, pages 1–18.

R. Ramos, B. Martins, and D. Elliott. 2023. LMCap: Few-shot multilingual image captioning by retrieval augmented language model prompting. In *Findings of ACL*, pages 1635—-1651.

N. Reimers and I. Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *EMNLP-IJCNLP*, pages 3982–3992.

K. Sugawara and T. Utsuro. 2022. Developing a dataset for mining reviews in tweets focusing on celebrities' aspects. In *Proc. 7th ABCSS*, pages 466–472.

W. Tan, Y. Li, et al. 2023. Reimagining retrieval augmented language models for answering queries. In *Findings of ACL*, pages 6131–6146.

M. Wiegmann, B. Stein, and M. Potthast. 2019. Celebrity Profiling. In *Proc. 57th ACL*, pages 2611–2618.

L. Xue, N. Constant, et al. 2021. mT5: A massively multilingual pre-trained text-to-text Transformer. In *Proc. NAACL*, pages 483–498.

H. Zhang, X. Liu, and J. Zhang. 2023a. Extractive Summarization via ChatGPT for Faithful Summary Generation. In *Findings of EMNLP*, pages 3270–3278.

H. Zhang, X. Liu, and J. Zhang. 2023b. SummIt: Iterative text summarization via ChatGPT. In *Findings of EMNLP*, pages 10644–10657.

A Collecting/Aggregating Impressions from X Posts

Figure 5 presents the specific prompts given to ChatGPT for collecting and aggregating impressions on celebrities' aspects from X posts as described in Section 4.2.

B Reason Collection/Aggregation

Table 4 shows the selected aspect-impression pairs for Web page search and the numbers of Web pages in the dataset for reason detection and collection as described in Section 5.1 and Section 6.2, respectively.

Figure 6 shows the prompts given to ChatGPT for collecting reasons for impressions from Web pages, as described in Section 5.3.

Figure 7 presents the prompts given to ChatGPT for aggregating the collected reasons for impressions, as described in Section 5.4.

C Inter-annotator Agreement for Reason Collection, Aggregation and Detection

C.1 Reason Collection

The multiset $R(d)$ of reference reasons is manually prepared by the first author following exactly the same procedure as presented in the previous section for ChatGPT. Another annotator ID=SK also manually prepared $R(d)$ for 6 keywords out of the overall 10, where, out of all the 5,625 sentences within the retrieved Web pages, 242 agreed to be collected as specifying reasons, 5,200 agreed not to be collected as specifying reasons, while 183 not agreed (collected by only one of the two annotators), resulting in 97% agreement rate and Cohen's kappa coefficient as 0.71, which is sufficiently high agreement.

C.2 Reason Aggregation

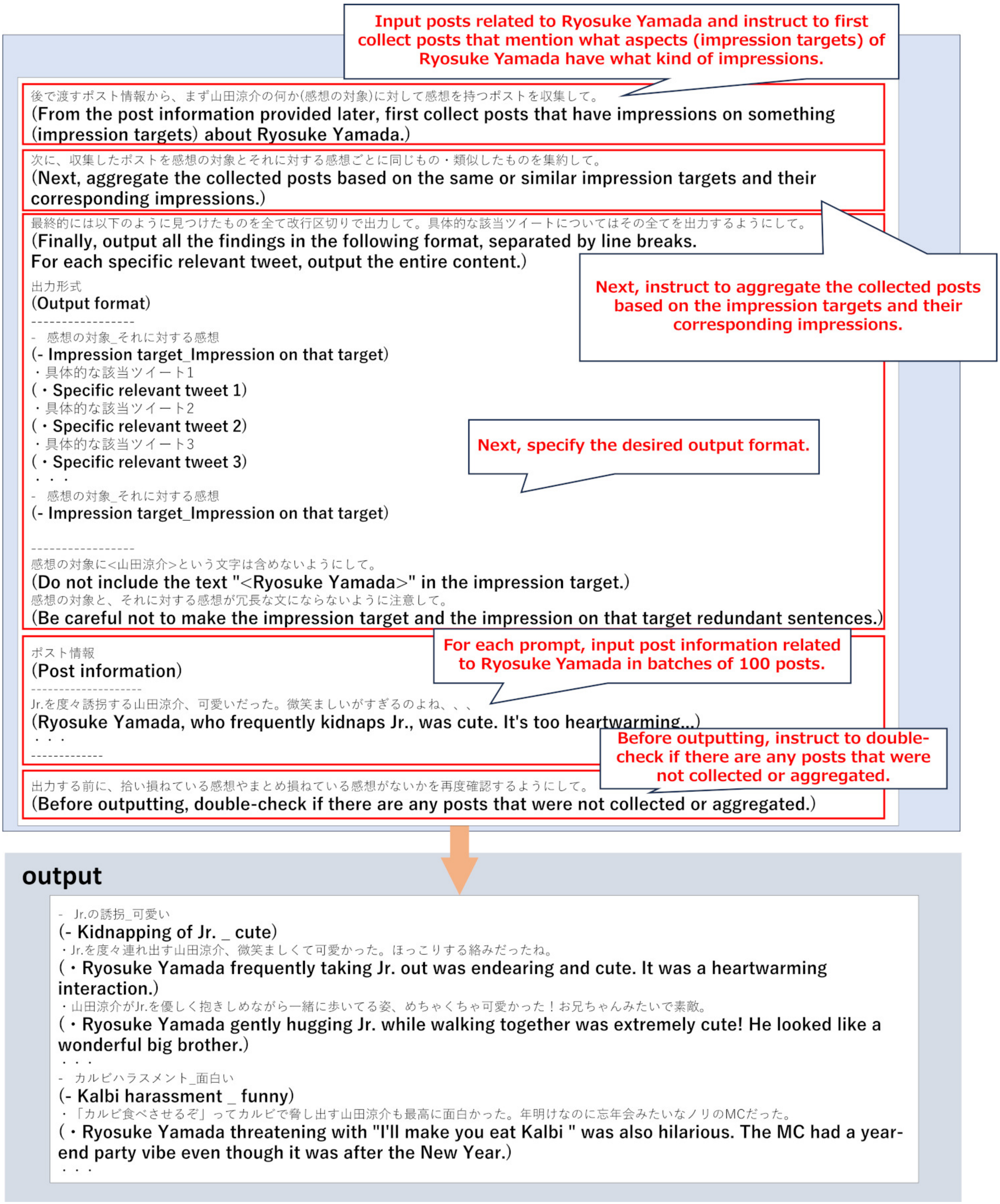
R was prepared by the first author in the overall evaluation. Here, for 6 keywords out of the overall 10, R was prepared independently by the annotator ID=SK from one's own result of collecting reasons in the previous section, where the agreement rate between the first author and the annotator ID=SK was 71%.

C.3 Reason Detection

The reference data for reason detection is directly constructed from the multiset $R(d)$ of reference reasons for the Web page d prepared by the first author. The agreement rate between the first author and the annotator ID=SK was 98% and Cohen's kappa coefficient was 0.95. The reference text for reason collection is composed by concatenating reasons manually prepared by the first author for each Web page.

D ChatGPT Output Format for Reason Aggregation

Specifically, we instruct ChatGPT to output the estimated category names and the corresponding Web page IDs in a ranked format. Each category name represents a reason for an impression accompanied with Web paged IDs, where the corresponding reason is collected from each of those Web pages. Those categories each representing a reason are ranked in descending order of the frequencies of their observation, expecting users to more easily understand the reasons for impressions. The output format is specified to be in a JSON format.



Input posts related to Ryosuke Yamada and instruct to first collect posts that mention what aspects (impression targets) of Ryosuke Yamada have what kind of impressions.

```
後で渡すポスト情報から、まず山田涼介の何か(感想の対象)に対して感想を持つポストを収集して。
(From the post information provided later, first collect posts that have impressions on something
(impression targets) about Ryosuke Yamada.)
```

```
次に、収集したポストを感想の対象とそれに対する感想ごとに同じもの・類似したものを集約して。
(Next, aggregate the collected posts based on the same or similar impression targets and their
corresponding impressions.)
```

Next, instruct to aggregate the collected posts based on the impression targets and their corresponding impressions.

```
最終的には以下のように見つけたものを全て改行区切りで出力して。具体的な該当ツイートについてはその全てを出力するようにして。
(Finally, output all the findings in the following format, separated by line breaks.
For each specific relevant tweet, output the entire content.)
出力形式
(Output format)
-----------------
-  感想の対象_それに対する感想
(- Impression target_Impression on that target)
・具体的な該当ツイート1
(・Specific relevant tweet 1)
・具体的な該当ツイート2
(・Specific relevant tweet 2)
・具体的な該当ツイート3
(・Specific relevant tweet 3)
・・・
-  感想の対象_それに対する感想
(- Impression target_Impression on that target)

-----------------
感想の対象に<山田涼介>という文字は含めないようにして。
(Do not include the text "<Ryosuke Yamada>" in the impression target.)
感想の対象と、それに対する感想が冗長な文にならないように注意して。
(Be careful not to make the impression target and the impression on that target redundant sentences.)
```

Next, specify the desired output format.

```
ポスト情報
(Post information)
--------------------
Jr.を度々誘拐する山田涼介、可愛いだった。微笑ましいがすぎるのよね、、、
(Ryosuke Yamada, who frequently kidnaps Jr., was cute. It's too heartwarming...)
・・・
-------------
```

For each prompt, input post information related to Ryosuke Yamada in batches of 100 posts.

```
出力する前に、拾い損ねている感想やまとめ損ねている感想がないかを再度確認するようにして。
(Before outputting, double-check if there are any posts that were not collected or aggregated.)
```

Before outputting, instruct to double-check if there are any posts that were not collected or aggregated.

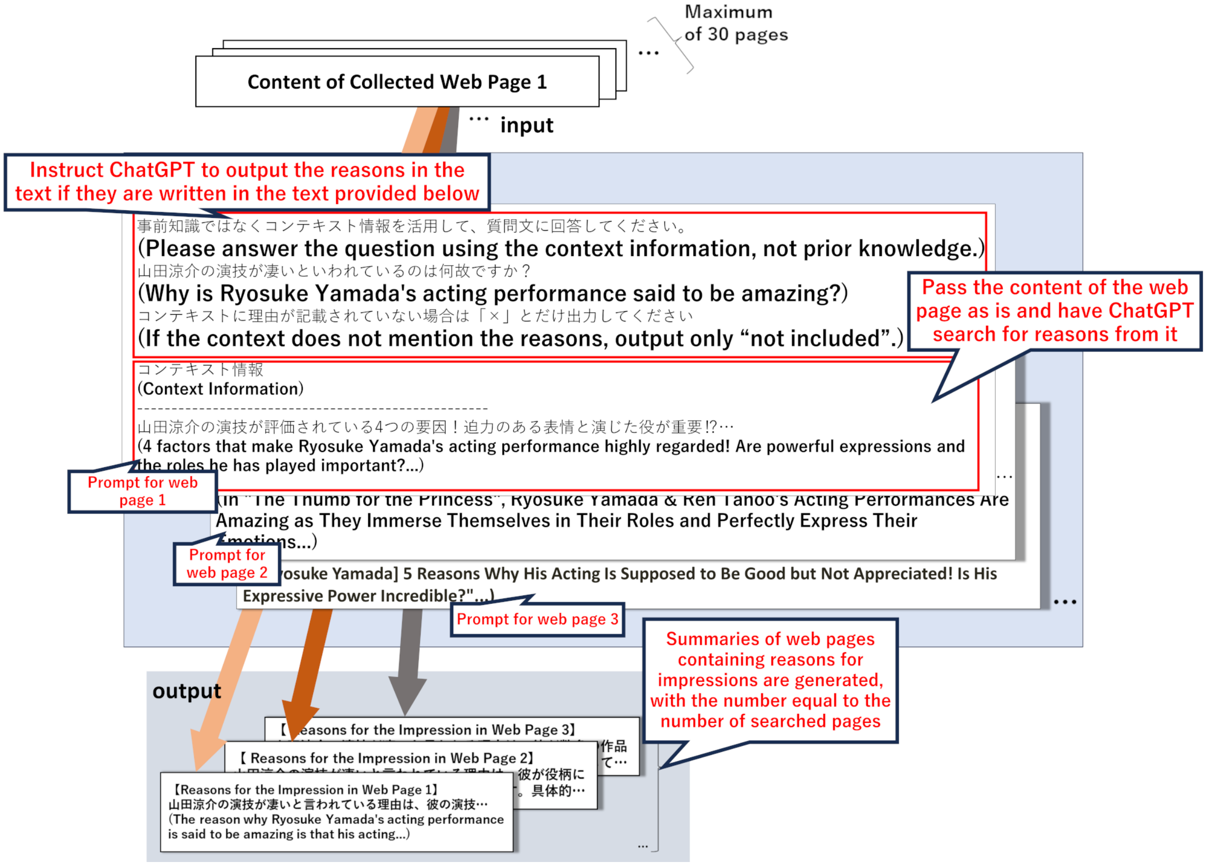
output

```
-  Jr.の誘拐_可愛い
(- Kidnapping of Jr. _ cute)
・Jr.を度々連れ出す山田涼介、微笑ましくて可愛かった。ほっこりする絡みだったね。
(・Ryosuke Yamada frequently taking Jr. out was endearing and cute. It was a heartwarming
interaction.)
・山田涼介がJr.を優しく抱きしめながら一緒に歩いてる姿、めちゃくちゃ可愛かった！お兄ちゃんみたいで素敵。
(・Ryosuke Yamada gently hugging Jr. while walking together was extremely cute! He looked like a
wonderful big brother.)
・・・
-  カルビハラスメント_面白い
(- Kalbi harassment _ funny)
・「カルビ食べさせるぞ」ってカルビで脅し出す山田涼介も最高に面白かった。年明けなのに忘年会みたいなノリのMCだった。
(・Ryosuke Yamada threatening with "I'll make you eat Kalbi " was also hilarious. The MC had a year-
end party vibe even though it was after the New Year.)
・・・
```

Figure 5: Prompts for Collecting/Aggregating Impressions from X Posts

celebrity name	aspect	impression	# Web pages		
			w/ reason	w/o reason	total
Ryosuke Yamada	acting performance	amazing	56	56	112
	drama	scary	22	22	44
		interesting	36	36	72
	face	good	20	20	40
Fuma Kikuchi	acting performance	amazing	33	33	66
	swamp	deep	8	8	16
Shun Oguri	acting performance	amazing	42	42	84
		bad	15	15	30
	face	good	12	12	24
	voice	good	13	13	26
total	—	—	**257**	**257**	**514**

Table 4: Selected Aspect-Impression Pairs and Numbers of Web Pages in the Dataset for Reason Detection/Collection


Maximum of 30 pages
Content of Collected Web Page 1
…
input
Instruct ChatGPT to output the reasons in the text if they are written in the text provided below
Pass the content of the web page as is and have ChatGPT search for reasons from it
Prompt for web page 1
Prompt for web page 2
Prompt for web page 3
Summaries of web pages containing reasons for impressions are generated, with the number equal to the number of searched pages
output
…

事前知識ではなくコンテキスト情報を活用して、質問文に回答してください。
(Please answer the question using the context information, not prior knowledge.)
山田涼介の演技が凄いといわれているのは何故ですか？
(Why is Ryosuke Yamada's acting performance said to be amazing?)
コンテキストに理由が記載されていない場合は「×」とだけ出力してください
(If the context does not mention the reasons, output only “not included”.)

コンテキスト情報
(Context Information)

山田涼介の演技が評価されている4つの要因！迫力のある表情と演じた役が重要!?…
(4 factors that make Ryosuke Yamada's acting performance highly regarded! Are powerful expressions and the roles he has played important?...)

(In "The Thumb for the Princess", Ryosuke Yamada & Ren Tanoo's Acting Performances Are Amazing as They Immerse Themselves in Their Roles and Perfectly Express Their Emotions...)

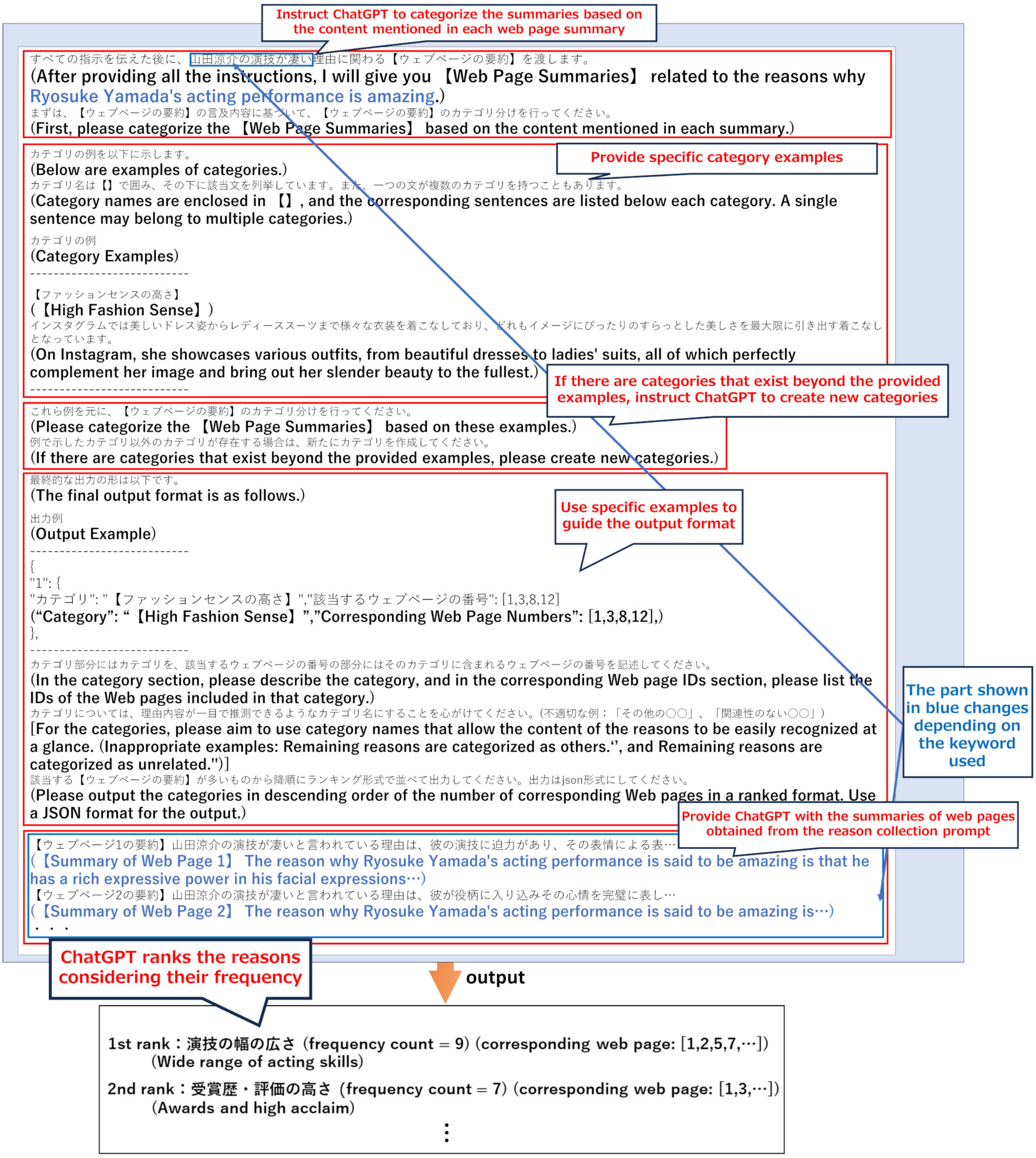
osuke Yamada] 5 Reasons Why His Acting Is Supposed to Be Good but Not Appreciated! Is His Expressive Power Incredible?"...)

【 easons for the Impression in Web Page 3】
作品
て…

【 Reasons for the Impression in Web Page 2】
彼が役柄に
。具体的…

【Reasons for the Impression in Web Page 1】
山田涼介の演技が凄いと言われている理由は、彼の演技…
(The reason why Ryosuke Yamada's acting performance is said to be amazing is that his acting...)

Figure 6: Prompts for Reason Collection by a Large Language Model (ChatGPT)



Instruct ChatGPT to categorize the summaries based on the content mentioned in each web page summary

すべての指示を伝えた後に、山田涼介の演技が凄い理由に関わる【ウェブページの要約】を渡します。

(After providing all the instructions, I will give you 【Web Page Summaries】 related to the reasons why Ryosuke Yamada's acting performance is amazing.)

まずは、【ウェブページの要約】の言及内容に基づいて、【ウェブページの要約】のカテゴリ分けを行ってください。

(First, please categorize the 【Web Page Summaries】 based on the content mentioned in each summary.)

Provide specific category examples

カテゴリの例を以下に示します。

(Below are examples of categories.)

カテゴリ名は【】で囲み、その下に該当文を列挙しています。また、一つの文が複数のカテゴリを持つこともあります。

(Category names are enclosed in 【】, and the corresponding sentences are listed below each category. A single sentence may belong to multiple categories.)

カテゴリの例

(Category Examples)

【ファッションセンスの高さ】

(【High Fashion Sense】)

インスタグラムでは美しいドレス姿からレディーススーツまで様々な衣装を着こなしており、どれもイメージにぴったりのすらっとした美しさを最大限に引き出す着こなしとなっています。

(On Instagram, she showcases various outfits, from beautiful dresses to ladies' suits, all of which perfectly complement her image and bring out her slender beauty to the fullest.)

If there are categories that exist beyond the provided examples, instruct ChatGPT to create new categories

これら例を元に、【ウェブページの要約】のカテゴリ分けを行ってください。

(Please categorize the 【Web Page Summaries】 based on these examples.)

例で示したカテゴリ以外のカテゴリが存在する場合は、新たにカテゴリを作成してください。

(If there are categories that exist beyond the provided examples, please create new categories.)

Use specific examples to guide the output format

最終的な出力の形は以下です。

(The final output format is as follows.)

出力例

(Output Example)

{
"1": {
"カテゴリ": "【ファッションセンスの高さ】","該当するウェブページの番号": [1,3,8,12]
(“Category”: “【High Fashion Sense】”,”Corresponding Web Page Numbers”: [1,3,8,12],)
},

カテゴリ部分にはカテゴリを、該当するウェブページの番号の部分にはそのカテゴリに含まれるウェブページの番号を記述してください。

(In the category section, please describe the category, and in the corresponding Web page IDs section, please list the IDs of the Web pages included in that category.)

カテゴリについては、理由内容が一目で推測できるようなカテゴリ名にすることを心がけてください。(不適切な例：「その他の〇〇」、「関連性のない〇〇」)

[For the categories, please aim to use category names that allow the content of the reasons to be easily recognized at a glance. (Inappropriate examples: Remaining reasons are categorized as others.‘', and Remaining reasons are categorized as unrelated.")]

該当する【ウェブページの要約】が多いものから降順にランキング形式で並べて出力してください。出力はjson形式にしてください。

(Please output the categories in descending order of the number of corresponding Web pages in a ranked format. Use a JSON format for the output.)

The part shown in blue changes depending on the keyword used

Provide ChatGPT with the summaries of web pages obtained from the reason collection prompt

【ウェブページ1の要約】山田涼介の演技が凄いと言われている理由は、彼の演技に迫力があり、その表情による表…

(【Summary of Web Page 1】 The reason why Ryosuke Yamada's acting performance is said to be amazing is that he has a rich expressive power in his facial expressions…)

【ウェブページ2の要約】山田涼介の演技が凄いと言われている理由は、彼が役柄に入り込みその心情を完璧に表し…

(【Summary of Web Page 2】 The reason why Ryosuke Yamada's acting performance is said to be amazing is…)

・・・

ChatGPT ranks the reasons considering their frequency

output

1st rank：演技の幅の広さ (frequency count = 9) (corresponding web page: [1,2,5,7,…])
(Wide range of acting skills)

2nd rank：受賞歴・評価の高さ (frequency count = 7) (corresponding web page: [1,3,…])
(Awards and high acclaim)

⋮

Figure 7: Prompts for Reason Aggregation by a Large Language Model (ChatGPT)