ECE1786 Final Report

Briefly

Word Count: 1934 (excluding captions/figures/tables but including text in tables)

Penalty: 0%

Ambrose Man

Anson Lau

Jihoon Chung

Joe Longpre

Permissions

Ambrose Man

- Permission to post video: yes
- Permission to post final report: yes
- Permission to post source code: yes

Anson Lau

- Permission to post video: yes
- Permission to post final report: yes
- Permission to post source code: yes

Jihoon Chung

- Permission to post video: yes
- Permission to post final report: yes
- Permission to post source code: yes

Joe Longpre

- Permission to post video: yes
- Permission to post final report: yes
- Permission to post source code: yes

1. Introduction

Briefly is a news summarization application which generates personalized news summaries tailored to a user's interests. If users are familiar with the topic of an article, the summaries generated by Briefly should be specific and detailed. If users are less familiar with the topic, the summaries should provide more background context. This approach allows users to either focus on the specifics of an issue or develop a broader understanding of an issue. By improving the quality and delivery of news summaries, users can stay up-to-date with less effort and time.

2. Illustration/Figure

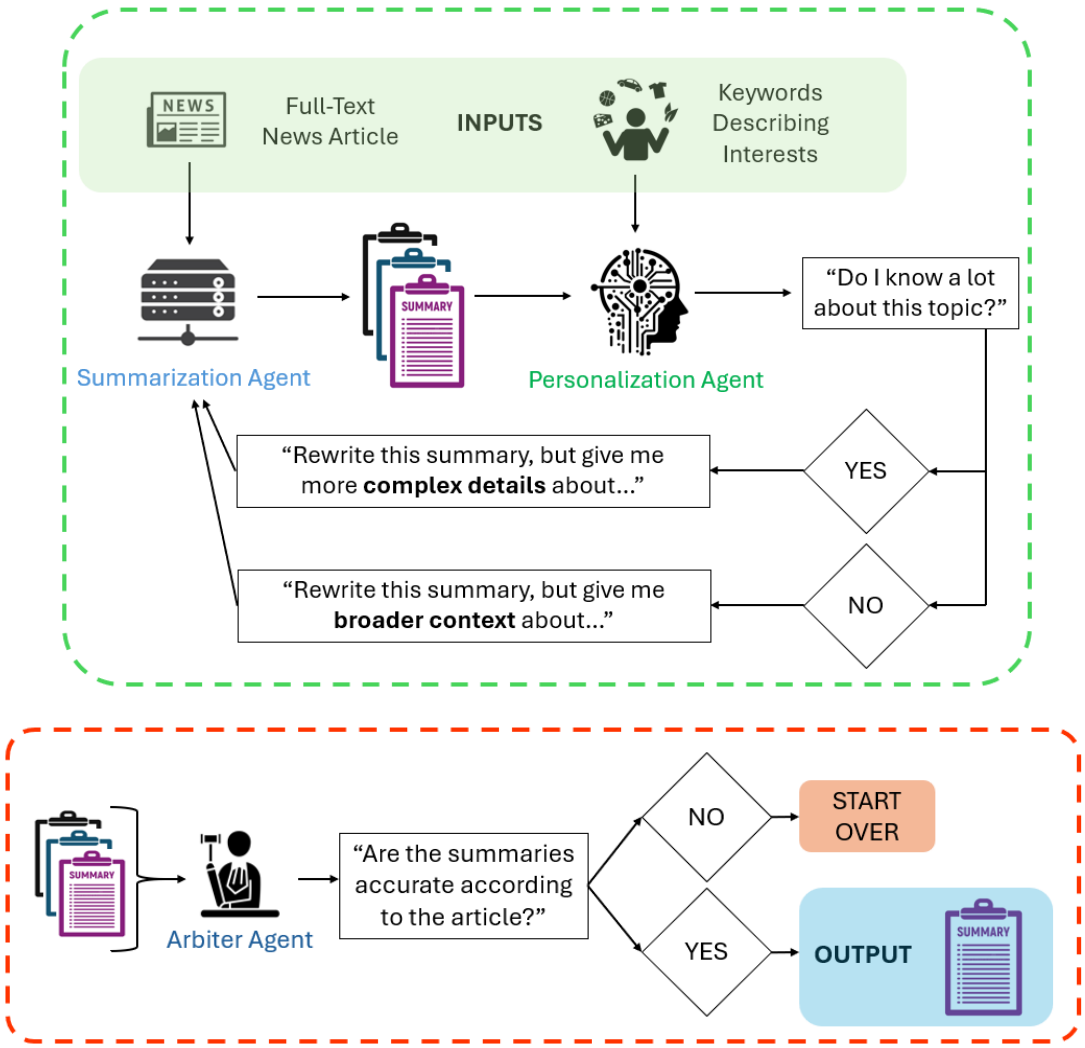

Figure 1. (Top - Green) The main feedback loop whereby the summarization and personalization agents work together to generate a list of three refined summaries tailored to the user's interests. (Bottom - Red) The arbiter agent ensures that the list of summaries remains accurate according to the article.

3. Background & Related Work

Personalization in an LLM-context has the potential to create more relevant and useful outputs across various applications. In medicine, a proposed agentic LLM framework translates medical reports from a medical professional to a general public persona in a zero-shot approach and then corrects the produced simplified summary with another LLM agent [1]. A review paper aimed to categorize the use of personas in LLMs, focusing on role-playing and personalization, where LLMs adapt to user personas in fields like recommendation, search, and healthcare [2]. It introduces LLM personality evaluation, whether the personality of an LLM agent accurately reflects the intended persona after adapting to the user. They bring forward challenges with personalization that are relevant to this project, such as long-context personas, bias, and lack of benchmarks.

4. Data and Data Processing

To create a personalized summary, two pieces of information are required: the full-text article to be summarized and the keyphrases describing user interests. These two inputs are processed to generate customized summaries tailored to each persona.

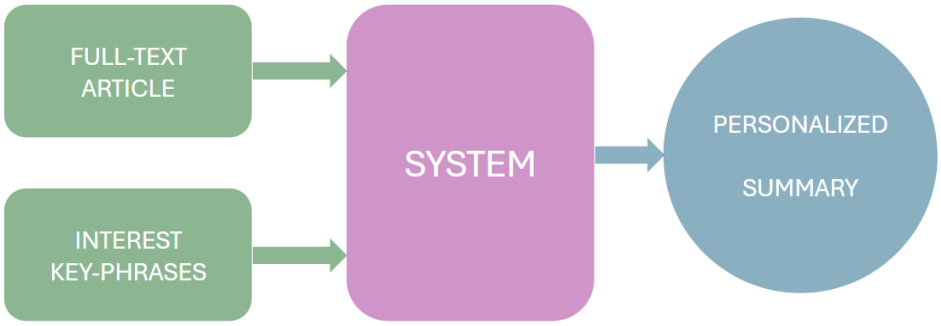

Figure 2. Inputs and output of the system

The system was tested using news articles belonging to four categories which match the interests of four generated personas: sports, entertainment, politics, and technology. An additional set of 10 general-interest articles was collected, selected to ensure that none aligned with the identified persona interests. The articles were scraped from news websites using *selenium* and the relevant text was extracted with *newspaper3k*.

	General	Sports	Entertainment	Politics	Technology	Total
Total Examples	10	10	10	10	10	50

Table 1: Breakdown of collected news articles used to test *Briefly*.

5. Architecture and Software

The system shown in Figure 1 contains three GPT-4o agents which interact throughout the article summarization process. These agents work together to provide a summary of an input article targeted to the background knowledge of a given user. The three main agents and their responsibilities are:

1. **Summarization Agent**: This agent receives a full text article and produces an initial zero-shot summary. This agent then receives feedback from the Personalization Agent and produces new summaries while taking into account requests in the feedback.

2. **Personalization Agent**: This agent first generates a fictional person based on a list of keywords passed as input; this person has a name, an age, a location, and in-depth knowledge about the keyword topics. The Personalization Agent receives summaries from the Summarization Agent, then requests new summaries tailored to the persona's knowledge of the article's topic. The Personalization Agent might ask for more in-depth details if the user is an expert in the topic, or more background context if the user is unfamiliar with the topic.

3. **Arbiter Agent**: After three conversational turns between the Summarization Agent and the Personalization Agent, the Arbiter Agent receives a list of refined summaries plus the original article. The Arbiter Agent determines if the final summary is accurate according to the article. If not, the process restarts until agreement between a summary and the original article is found. Otherwise, the most refined summary is passed as the output.

The article summarization process can be experienced through a front-end powered by *Gradio*. A user can choose a set of interests and an article to summarize, then watch as the system generates a refined summary. A debug tab gives access to the entire conversation recorded between agents.

6. Comparison

To evaluate the architecture's functionality, the final output summary is compared with the zero-shot summary to determine whether it better reflects the persona's background knowledge. The evaluation process involves a manual comparison of the zero-shot summary and final output summary from an article based on the rubric in Table 2. If the output summary meets any of the criteria in a category, it will automatically be awarded the lower score. For example, if the output summary is generic but contains hallucinations, it would receive a score of "0," as hallucinations are deemed unacceptable. Comparing the final output with the zero-shot verifies the functionality of the Personalization Agent. In some cases, the system may not be able to produce a differentiable summary from the zero-shot; some points are awarded as this is still considered acceptable functionality. A normalized score was obtained by averaging the scores over all the test cases.

Criteria (final output vs zero-shot)	**Score Awarded**
Worse than Zero-shot • Does not summarize well • Hallucinates / adds information unrelated to article • Removes important information	0
Similar or No distinction to Zero-shot • Final output summary has no noticeable difference compared to the zero-shot summary	0.5
Better than Zero-shot • Summary is more generic to cater to a persona not familiar with the topic of the article • Summary is detailed and specific to cater to a persona familiar with the topic of the article	1.0

Table 2. Rubric for the evaluation of system outputs.

7. Quantitative Results

Four personalization agents were developed, focusing on broad areas like sports, entertainment, politics, and technology. Each persona was tested with some general-interest articles (about which the personas have little background knowledge) and specific-interest articles (about which they are experts). The results in Table 3 show that, in general, the final summaries produced by the system are more personalized than the zero-shot summaries initially produced by the Summarization Agent.

	Worse Than Zero-Shot		Similar to Zero-Shot		Better Than Zero-Shot		Score
	Off-Persona	On-Persona	Off-Persona	On-Persona	Off-Persona	On-Persona	
Sports	0/10	2/10	2/10	1/10	8/10	7/10	0.825
Entertainment	1/10	0/10	3/10	1/10	6/10	9/10	0.85
Technology	1/10	1/10	1/10	1/10	8/10	8/10	0.85
Politics	0/10	0/10	2/10	3/10	8/10	7/10	0.875
Total Score (Average of 4 Persona Scores): 0.85							

Table 3. Summary of evaluated system outputs using the rubric in Table 2.

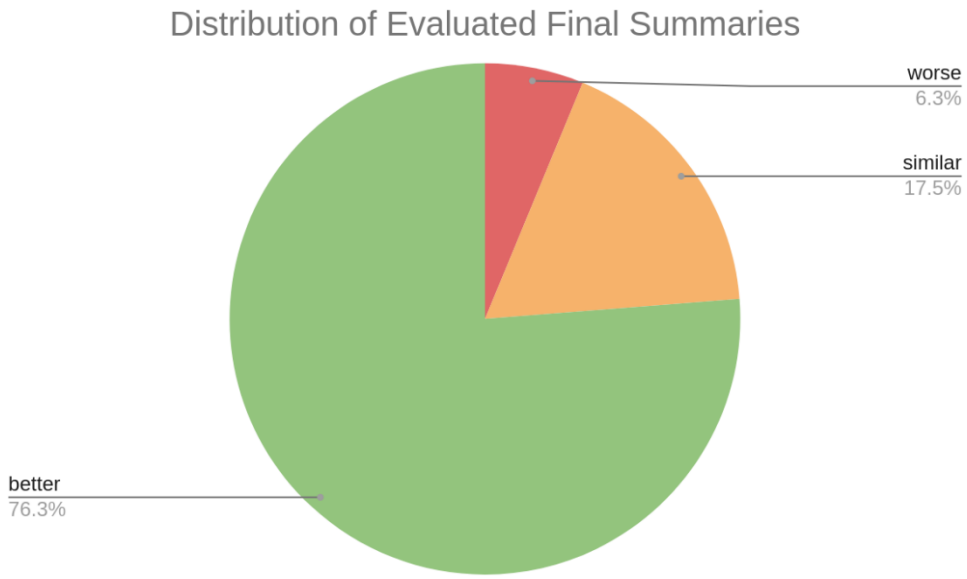

Figure 3. Distribution of evaluated system outputs (summaries).

After evaluating using the rubric, the final score of 85% was achieved, meaning that the system was able to create a useful summary for the user 85% of the time. 6.25% of the tests (2+3/80) were cases where the final output was worse than the zero-shot. Meanwhile 76.25% (30+31/80) of the tests resulted in a better than zero-shot. 17.25% (8+6/80) resulted in a final output similar to the zero-shot.

8. Qualitative Results

General Summary	Specific Summary
The Los Angeles Dodgers won their eighth World Series title by overcoming a significant deficit to defeat the New York Yankees 7-6 in a pivotal game. This marked their first opportunity for a championship parade since pandemic-related restrictions in 2020, as **key errors by Yankees players** like Aaron Judge and Anthony Volpe allowed the Dodgers to rally successfully. Freddie Freeman and Mookie Betts **were instrumental**, with Freeman tying a **World Series RBI record** and Betts later delivering the decisive run.	The Los Angeles Dodgers secured their eighth World Series title by defeating the New York Yankees 7-6 in Game 5, overcoming a **5-0 early deficit** through a combination of Yankee defensive errors and clutch performances. Freddie Freeman tied a World Series record with **12 RBIs**, contributing crucially during a five-run rally sparked by Yankees' mistakes, including Aaron Judge **dropping a routine catch** and Anthony Volpe **throwing wide of third base**. Mookie Betts brought home the winning run with a **sacrifice fly** in the eighth inning, showcasing the Dodgers' persistence and capitalization on New York's **bullpen fatigue**.

Table 4: Example outputs for a sports news article [3], comparing the output summary for a non-sports fan (general) and a sports fan (specific).

The summaries in Table 4 are generated from an article about the latest World Series Final, comparing the model's on-and-off persona behavior. In the green text, the first sentences of both summaries are similar because they are based on the same source article. However, the specific summary, highlighted in blue, includes additional match details that would appeal to a sports fan.

Furthermore, in the yellow and green-highlighted text of the specific summary in Table 4, errors and contributions are elaborated upon using baseball terminology. These detailed facts cater specifically to sports fans, while the general summary for non-sports fans focuses only on the overall result. This contrast demonstrates the model's ability to effectively tailor content to the intended persona.

Approximately 20% of the summaries show similar levels of personalization as zero-shot summaries. This occurs because the model struggles with simpler articles where the arbiter enforces strict alignment with facts from the source. The model performs best with detailed articles, as they provide a broader selection of facts to include in the summaries.

9. Discussions and Learnings

The quantitative results suggest that, in the majority of instances, *Briefly* produces news summaries which are better tailored to a user's background knowledge than the initial zero-shot summary. Indeed, the system produced summaries which were more appropriate for the user for 76.3% of articles tested. It is important to note that if the summaries generated by the system do not demonstrate improvements over the zero-shot summary, then the agentic approach offers no advantage over a simple single summarization agent. Thus, the zero-shot summary serves as the primary benchmark for comparison.

The benefits of the agentic approach are well illustrated when reviewing outputs from the arbiter. Interestingly, the summarization agent always positively confirms the accuracy of its summaries; however, the arbiter agent often disagrees. This likely occurs because the summarization agent has access to the entire conversation history between itself and the personalization agent, and thus has some "chain of thought" available to it. For example, initial tests showed the arbiter agent rejecting summaries which used a neutral tone when the article was originally written in a celebratory or sombre tone. Significant prompt engineering was required to adjust the arbiter agent to assess only factual claims.

Similarly, the personalization agent needed significant tuning. Beyond the prompt engineering required to generate the right *type* of feedback, the temperature parameter on the personalization agent was tuned up to 1.5 to increase the *creativity* of feedback. Increasing the temperature in this manner led to better feedback which produced summaries that deviated more favorably from the zero-shot summaries.

One area where the model struggles is when summarizing articles that are short or overly general. In these cases, the zero-shot summary must also be general. Attempts to further generalize produce summaries of similar quality, and requests to add detail or specifics cannot be met. In these cases, the system cannot earn the "better than zero-shot" grade according to the quantitative rubric.

Finally, there has been a modest attempt to address bias in the articles. Originally, the intention was to remove bias as a whole from all summaries. After initial investigation, it was found that GPT-4o already removes obvious, inflammatory bias from its responses while unconscious bias or bias by omission is more difficult to address. Since inflammatory bias is handled automatically by GPT-4o models, more subtle types of bias were deemed to be beyond the scope of this project. Instead, the arbiter agent provides a bias rating of a refined summary on a scale from 1-5 (with a lower grade being less biased). The user may interpret that rating as they see fit.

Individual Contributions

Each team member contributed equally to core aspects of the project, including initial idea generation, prompt engineering, testing, presenting, and document writing. Additionally, members had specific responsibilities, including:

Ambrose Man:

- Collecting 10 “Sports” articles for testing the “Sports” persona.
- Generating and testing the “Sports” persona and grading 20 summaries
- Rate bias functionality

Anson Lau:

- Collecting 10 “Politics” articles for testing the “Politics” persona.
- Generating and testing the “Politics” persona and grading 20 summaries.
- Building core implementation (agent.py + prototype.py)

Joseph Longpre:

- Collecting 10 “General Interest” articles for testing all personas.
- Collecting 10 “Entertainment” articles for testing the “Entertainment” persona.
- Generating and testing the “Entertainment” persona and grading 20 summaries.
- Connecting all agents and creating the gradio-powered front-end (frontend.py).

Jihoon Chung:

- Collecting 10 “Technology” articles for testing the “Technology” persona.
- Generating and testing the “Technology” persona and grading 20 summaries.
- Implemented knowledge extraction feature.

References

[1] M. Sudarshan *et al.*, "Agentic LLM Workflows for Generating Patient-Friendly Medical Reports," *arXiv Prepr. arXiv2408.01112*, 2024.

[2] Y.-M. Tseng *et al.*, "Two tales of persona in llms: A survey of role-playing and personalization," *arXiv Prepr. arXiv2406.01171*, 2024.

[3] F. Ardaya, C. Kirschner, B. Kuty, and T. Kepner, "Dodgers capture World Series by storming back in Game 5 to upend Yankees: Takeaways," *The New York Times*, Oct. 31, 2024. https://www.nytimes.com/athletic/5886965/2024/10/30/dodgers-yankees-world-series-game-5-takeaways/.