AdVize Final Report
Developing an LLM-Based Evaluation Tool for Advertisement Pushing Algorithms

Zhifei Dou, Qingyi Xia, Shirley Qin, Anthony Cui
Department of Electrical and Computer Engineering
University of Toronto

Word Count: 1733 (Penalty 0%)

December 10, 2024

Contents

1 Permissions

	Video	Final Report	Source Code
Zhifei Dou	Yes	Yes	Yes
Qingyi Xia	Yes	Yes	Yes
Sherly Qin	Yes	Yes	Yes
Anthony Cui	Yes	Yes	Yes

2 Introduction

Ad-pushing algorithms are machine learning-based systems designed to match user queries with relevant advertisements, optimizing metrics such as Click-Through Rates (Satisfaction Rate)[1]. These algorithms are vital for app developers like Instagram, as they allow these platforms to earn revenue from merchants like Sephora by effectively marketing their products to users through ads. However, traditional evaluation methods for these algorithms rely on human raters to assess advertisement quality and relevance, which incurs considerable recruitment costs and lengthy evaluation time[2]. To address this, the team proposes a Large Language Model (LLM) based multi-algorithm evaluation system that replaces human raters and provides a comprehensive, automative, and efficient evaluation methodology for various ad-pushing algorithms.

3 Illustration & Figure

Please refer to Figure 7 for the structure diagram of AdVize.

4 Background & Related Work

"Online Advertisements with LLMs: Opportunities and Challenges" by Feizi et al. has discussed the potential of leveraging LLMs in online advertising systems [3]. It addresses essential aspects such as privacy, latency, and reliability of LLMs, which provokes the team's thinking about implementing the LLM in evaluating advertisement-pushing algorithms.

Meanwhile, "PICLe: Eliciting Diverse Behaviors from Large Language Models with Persona In-Context Learning" proposed by Choi et al. presents a framework for persona elicitation, customizing LLM behaviors to align with target personas [4]. It gives the team guidelines about persona installation for engineering our primary model.

5 Data and Data Processing

5.1 The Persona Dataset

The Persona Dataset was collected from 20 real individuals through surveys. It contains "persona data" such as age, occupation, gender, topics of interest, and search history.

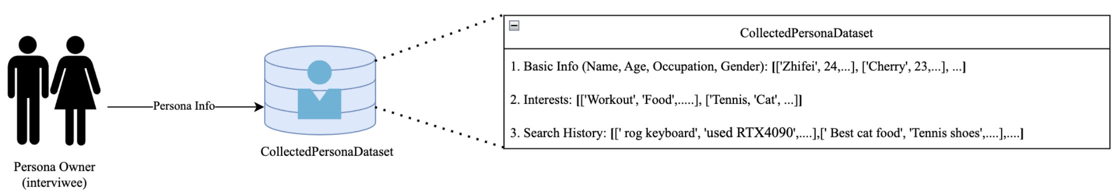


Figure 1: Persona Dataset

5.2 CommercialAds Dataset

The CommercialAds Dataset contains 250k search queries and the corresponding matched ads extracted from the original CommercialAds Dataset.

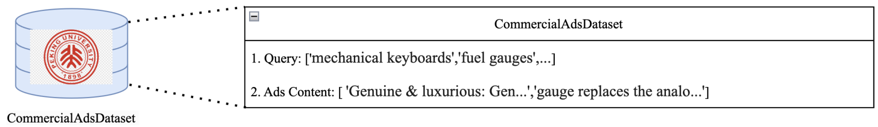


Figure 2: CommercialAds Dataset

The original dataset has 300k samples, where each sample includes a user search query, ad title, description, image, product manufacturer, seller information, and a binary label indicating ad relevance to the query. From this original dataset, the team kept the query and ad columns only while removing all the other columns, which constitute our CommercialAds Dataset.

6 Architecture and Software

6.1 Engineering Simulated Rater

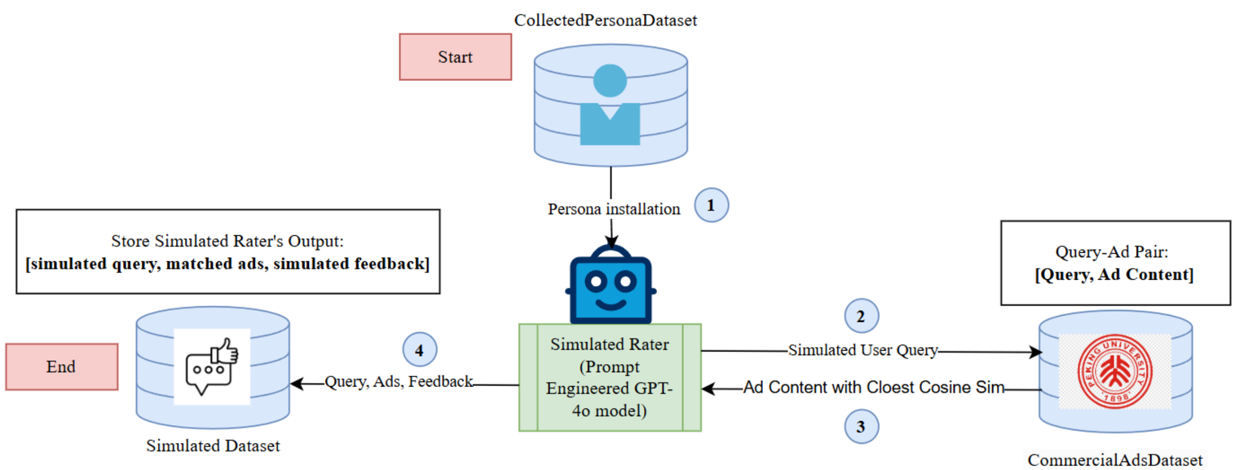


Figure 3: Module-level overview: query generation and feedback collection

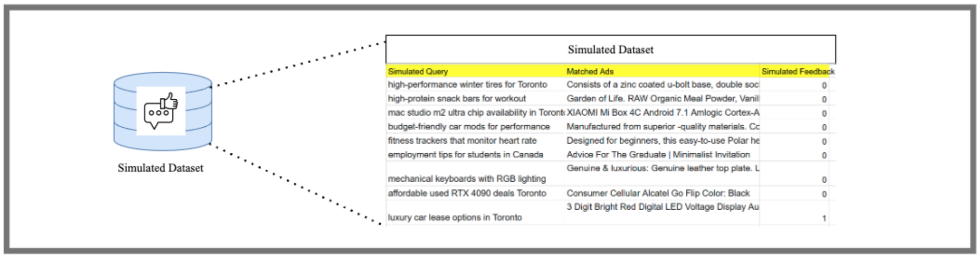
6.1.1 Persona-Based Initialization

As Figure 3 shows, the simulated rater is initialized using the CollectedPersonaDataset. The queries generated by the simulated rater should align with the persona's profile, reflecting the persona's search behaviours.

6.1.2 Ad Matching and Query Processing

After generating a persona-based simulated query, cosine similarity is used to match it with the most similar query in the Query-Ad Pair dataset. The ad linked to the most similar query is presented to the simulated rater for feedback assessment.

6.1.3 Simulated Rater's Feedback Collection



Simulated Query	Matched Ads	Simulated Feedback
high-performance winter tires for Toronto	Consists of a zinc coated u-bolt base, double soc	0
high-protein snack bars for workout	Garden of Life. RAW Organic Meal Powder, Vanil	0
mac studio m2 ultra chip availability in Toront	XIAOMI Mi Box 4C Android 7.1 Amlogic Cortex-A	0
budget-friendly car mods for performance	Manufactured from superior -quality materials. Cc	0
fitness trackers that monitor heart rate	Designed for beginners, this easy-to-use Polar he	0
employment tips for students in Canada	Advice For The Graduate | Minimalist Invitation	0
mechanical keyboards with RGB lighting	Genuine & luxurious: Genuine leather top plate. L	0
affordable used RTX 4090 deals Toronto	Consumer Cellular Alcatel Go Flip Color: Black	0
luxury car lease options in Toronto	3 Digit Bright Red Digital LED Voltage Display Au	1

Figure 4: Simulated Dataset

The simulated raters' binary feedback (1 for relevant, 0 for not relevant) will be stored along with the generated query and matched ads in the Simulated dataset. This dataset is used as a medium to demonstrate simulated queries and feedback to persona owners.

6.1.4 Persona Owners Inspection and Prompt-Engineering Loop

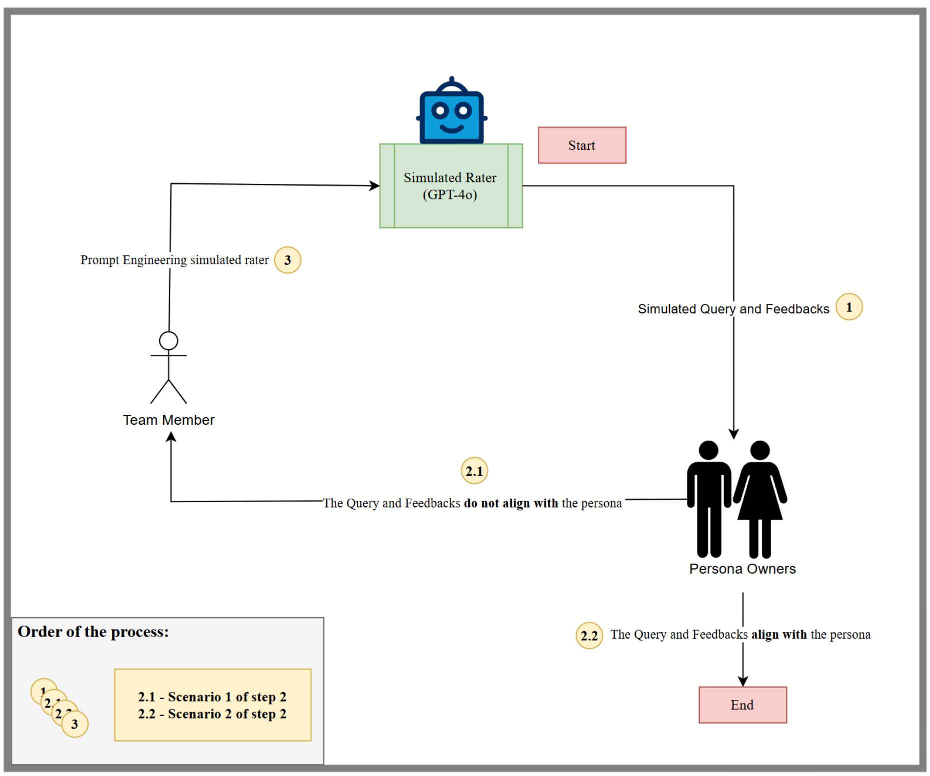


Figure 5: Prompt-engineering loop for Simulated Rater

Figure 5 demonstrates the prompt engineering loop of the simulated rater. Persona owners inspected the quality and correctness of each simulated query and feedback generated by the simulated rater. This inspection was used for team members to refine and optimize the simulated rater through iterative prompt engineering until all simulated queries and feedback have passed inspection.

During this process, one important modification involved addressing the simulated rater's initial tendency to over-focus on the details of matched ads, leading to feedback that often misaligned with the persona owner's general preferences. Hence, the team added prompts shown in Figure 6, to encourage the simulated rater adopting generalized perspectives which significantly boosted its performance.

```
- When you making the dicision of interested or not, think generally about the pushed ads don't get too hung up on the specific items in a particular ad,
  focus more on the category and features of the ad. For example, if a user searches for RTX4090, if the pushed advertisement is for an older graphics card such as GT260,
  the user you are playing may also be potentially interested.\

- When making the decision of whether an ad is interesting or not, please consider the persona's age and the typical preferences of their age group.
  Think about how someone in this age group might perceive the ad—would it appeal to their lifestyle, interests, or needs?
```

Figure 6: Generalization prompt for Simulated Rater

6.2 Evaluating Multiple Ad-Pushing Algorithms

After developing the simulated rater, the team has implemented it into the AdVize system which is demonstrated in Figure 7. Please find **Appendix A** for one comprehensive example used through the system.

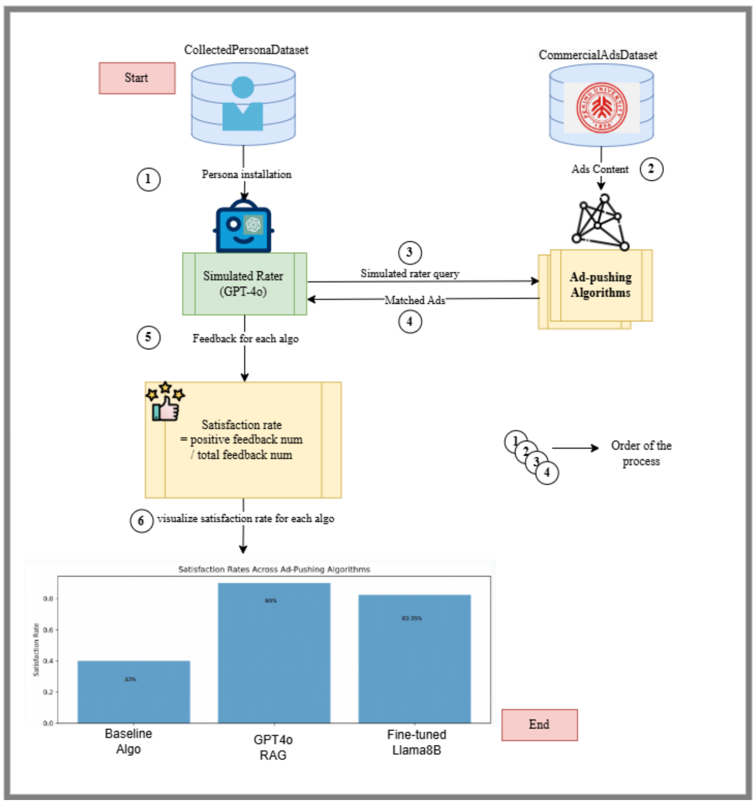


Figure 7: System architecture of AdVize

6.2.1 Persona Installation and Algorithm Building

The system begins by embedding the persona into the simulated rater. Subsequently, ad-pushing algorithms, including Cosine Similarity Matching (baseline), fine-tuned Llama 8B, and GPT-4o with Retrieval Augmented Generation, are developed utilizing ad content sourced from the CommercialAdsDataset.

6.2.2 Simulated Query Generation and Ads Evaluation

The simulated rater generates queries and sends them to the ad-pushing algorithms. These algorithms process the queries and return the corresponding ads, which are then evaluated by the simulated rater, providing feedback for each algorithm's pushed ads.

6.2.3 Feedback Aggregation

For each ad-pushing algorithm, the system analyzes the simulated rater's binary feedback and calculates satisfaction rates for each algorithm. The final output of our system presents the satisfaction rates for all evaluated algorithms and identifies the most effective algorithm.

6.3 Automated AdVize Application

6.3.1 Dockerized Algorithms

To streamline and automate the evaluation process, the simulated rater system incorporates a Dockerized architecture[5]. To encapsulate the necessary environment, dependencies, and code, each ad-pushing algorithm is containerized into a Docker image, as shown in Figure 8, ensuring compatibility and reproducibility across different systems.

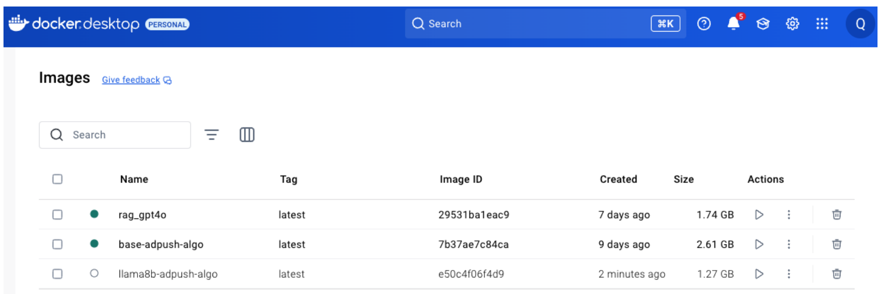


Figure 8: Docker images for ad-pushing algorithms

Figure 9 demonstrates our `AdVize` prototype application, we expect the user to provide their target ad-pushing algorithm by entering the name of the encapsulating Docker image, and after clicking the "Run" button, they will see a satisfaction rate showing up shortly in the "Result" section.

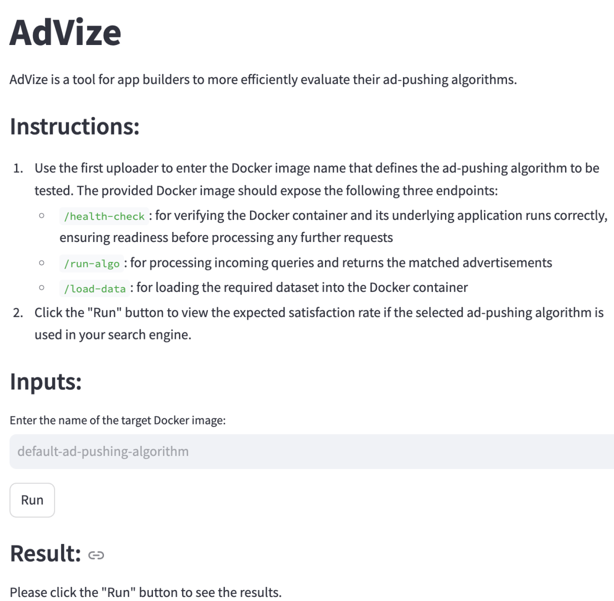


Figure 9: AdVize user interface

As demonstrated in Figure 9, the provided docker images should expose three essential endpoints (`/health-check`, `/load-data`, and `/run-algo`) to integrate seamlessly with the backend of the simulated rater system.

- **`/health-check`:** The endpoint for verifying the Docker container and its underlying application runs correctly, ensuring readiness before processing any further requests.
- **`/run-algo`:** The endpoint for processing incoming queries and returns the matched advertisements.
- **`/load-data`:** The endpoint for loading the required dataset into the Docker container. While some algorithms require a dataset of queries and advertisements to initialize or process matches, in cases where the dataset is pre-embedded in the model (e.g. the fine-tuned Llama 8B model), this endpoint should be used as a placeholder for compatibility without actually loading the dataset.

6.3.2 Automated Workflow

The `AdVize` system automates the testing and evaluation process, enabling a streamlined and reproducible workflow for the entire system, as shown in Figure 10 and 11.

```
Enter the Docker image name for the ad-pushing algorithm: base-adpush-algo
Starting the Docker container...

Started Docker container with ID: b5ebe96444f8569d5253918a86dca74fa6d0d9bc21c4fa8ea8973d55ed5988b3
Waiting for Flask to be ready...
Waiting for Flask to be ready...
Waiting for Flask to be ready...
Waiting for Flask to be ready...
Waiting for Flask to be ready...
Waiting for Flask to be ready...
Flask ready!
The docker URL is:  http://localhost:5001
Dataset loaded successfully into Docker container.
Starting user simulation...


Simulating user: Yuyang Zeng
Processing query: 1. Smart LED Desk Lamp with USB Charging Port
```

Figure 10: Docker container initialization, health check, data loading, and simulation

```
Stopping Docker container with ID: b5ebe96444f8569d5253918a86dca74fa6d0d9bc21c4fa8ea8973d55ed5988b3
b5ebe96444f8569d5253918a86dca74fa6d0d9bc21c4fa8ea8973d55ed5988b3
Stopped Docker container.
Feedback saved to user_feedback_base-adpush-algo_20241128_200137.csv
```

Figure 11: Docker container termination and result saving

1. **Starting the Docker Container and Health Check:** As shown in Figure 10, the system initializes the Docker container for the specified algorithm using `docker run`. The system repeatedly sends POST requests to the `/health-check` endpoint every 10 seconds for up to 20 attempts, and if the endpoint returns a status code of 200 within this 200 seconds, the system is deemed ready for further processing.
2. **Loading Data:** The system sends a POST request to the `/load-data` endpoint with the dataset path. This step prepares the algorithm to process queries, ensuring it has access to the required data.
3. **Running the Algorithm:** For each query, the system sends a POST request to the `/run-algo` endpoint. The query is processed, and the matched advertisement is returned. This interaction is the core functionality of the ad-pushing algorithms.
4. **Stopping the Docker Container:** Once the simulation is completed, the system stops the running Docker container to free up resources and ensure the environment is clean for future executions, as shown in Figure 11.

Integration and Compatibility: By standardizing the `/run-algo`, `/load-data`, and `/health-check` endpoints, all Dockerized ad-pushing algorithms are compatible with the main backend script. This ensures modularity, allowing algorithms to be swapped or added without disrupting the overall system architecture.

6.3.3 Data Pipeline and Feedback Generation

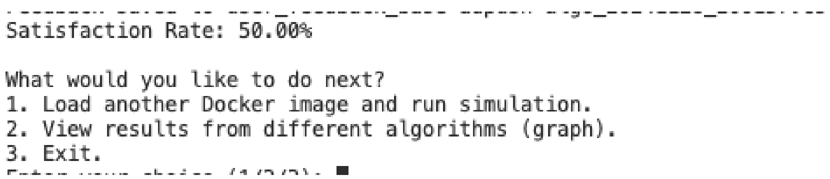


Figure 12: Interactive options after completing a simulation

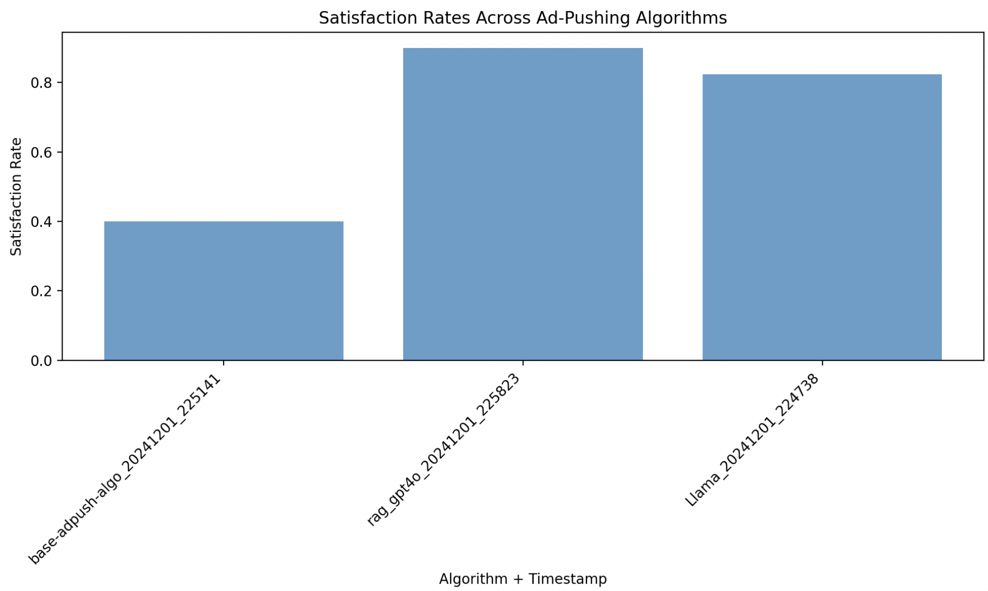


Figure 13: Satisfaction rates across different ad-pushing algorithms

Query Matching and Feedback Collection: The `AdVize` system evaluates the performance of each algorithm by processing a batch of persona-generated queries. For each query, the matched advertisement is retrieved from the `/run-algo` endpoint, and the simulated rater provides feedback. The system then records the feedback and justification for further analysis.

Output Generation: Once the Docker container is stopped and the simulation concludes, the results – including matched advertisements, feedback, justifications, and satisfaction rates – are saved to CSV files. The system also calculates the satisfaction rate for each ad-pushing algorithm based on the simulated feedback. The results are visualized through graphs, allowing users to compare the performance of different algorithms.

Figure 11 illustrates the interactive flow presented to the user after the simulation process. Users can choose to:

1. Test and run another ad-pushing algorithm by entering the corresponding Docker image name.

2. View the satisfaction rates and results graph generated from the CSV files.

3. Exit the simulation process.

The graph in Figure 13 displays satisfaction rates across different algorithms, enabling a clear comparison of their performance. This visualization helps identify the most effective ad-pushing algorithm based on quantitative metrics.

7 Baseline Model or Comparison

7.1 Where Evaluation is Needed

To measure the success of AdVize, we needed different evaluation metrics for different functionalities of our simulated rater model. There are two major functionalities of the model that we are interested in evaluating: query generation and feedback-on-ad generation.

7.2 Evaluating Query Generation

Every time the simulated rater generates a search query, we would like to know whether this query matches the assigned persona. For example, it is unlikely for a user whose interests lie in outdoor activities to search for products for indoor entertainment. Given a persona description and a search query (both in plain text), we use the following three evaluation methods to evaluate the query generation:

1. **BERT scores**: We use BERT F1 scores through the `bert_score` library in Python to compare the similarity between the persona description and the search query in their BERT contextual embeddings [6]. If the query does not match the persona, we would expect to see a low BERT F1 score due to low similarity.

2. **Human verification**: Human verification is done by asking a real person (the source of that persona) to rate on a scale of 1 to 3 whether they are likely to use that specific query themselves when searching. Scores of 2 and 3 here indicate 'Moderate' and 'High' likelihood.

3. **GPT-4o verification**: We could also let another GPT-4o agent to rate the query for that persona source. In this case, we will need to give GPT-4o the full persona description so that it thinks like the persona source when rating the queries.

For Method 2 and 3 above, a rubric was created as the guideline for query grading.

```
Evaluate the quality of how closely a search query is matched with the given persona on a scale of 1 to 3.
1 - The text is unlikely to be a search query from the given persona.
2 - The text might be a search query from the given persona.
3 - The text is very likely to be a search query from the given persona.
```

Figure 14: Evaluation Rubric for Search Queries

7.3 Evaluating Feedback-on-Ad Generation

Another functionality we need to measure the success on is the simulated rater generating feedback on an ad. Ideally, we want the simulated rater's feedback to match the human's feedback for as many ads as possible. Given one list of binary feedback (0 for interested, 1 for uninterested) from simulated rater and another one from human, we now compute the two metric values as follows for evaluation:

1. **Percentage of Matched Feedback**: For a given ad, when the feedback from simulated rater and human is the same (both being 0 or both being 1), we call it a "matched feedback". A high-performance Simulated User model is expected to have a high percentage of matched feedback.

2. **Cohen's Kappa**: We can also use Cohen's kappa statistics, which is used to evaluate the agreement between two annotators on a classification task. This value can range from -1 to 1 and should be above 0.8 to indicate good agreement between the raters [7].

8 Quantitative Results

The primary objective of AdVize is to replicate human rater performance. Accordingly, the evaluation metrics were specifically designed to assess the alignment and similarity between human raters and simulated raters, including BERT scores, manual assessments, and evaluations conducted by an LLM framework.

8.1 Query Evaluation

Manual Scoring:

- Human evaluators assessed the relevance of the generated queries using a team-curated rubric, assigning scores on a scale of 1 to 3. Among the 200 queries evaluated, 96% were rated as having "Moderate" to "High" relevance.

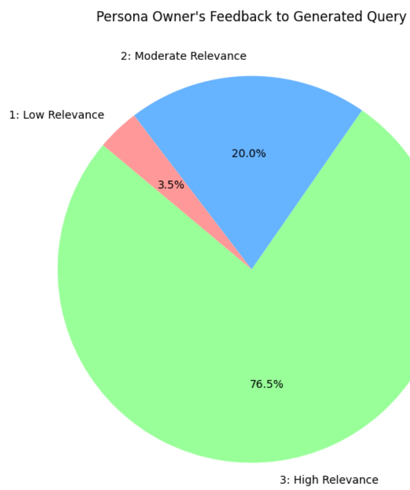


Figure 15: Manual evaluation result for query simulation

Automated Scoring - BERTScore F1:

- The contextual similarity between the persona prompt and the generated queries was quantified using the BERTScore metric. The model achieved an average F1 score of 86.9% across the 200 queries.

8.2 Ad-Query Matching Evaluation

Manual Scoring:

- The percentage of matched feedback reached 91.8% between the Simulated User and human raters.
- A Cohen's kappa of 0.8232 indicates that the Simulated User and human raters highly agree on their feedback on generated ads. The annotators' classification is visualized in Figure 16 below.

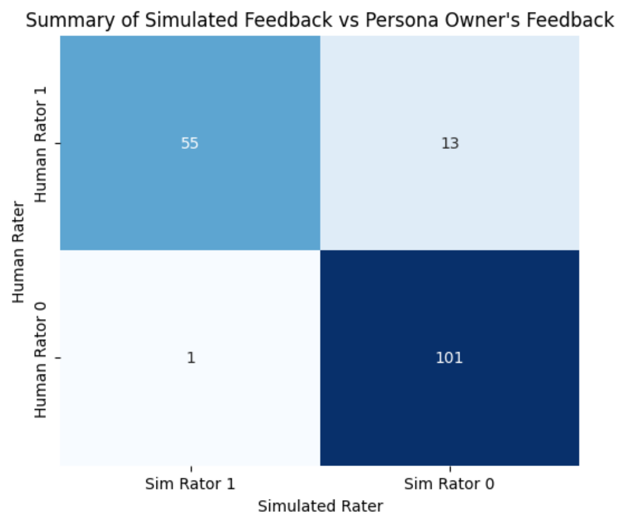


Figure 16: Agreement between raters

9 Qualitative Results

Simply providing digital relevance scores for generated queries and binary feedback for matched advertisements is insufficient to fully evaluate the effectiveness of the AdVize model. Therefore, an additional GPT-4o model was implemented to enhance the evaluation process. The following examples (Figure 17 and 18) illustrate how the GPT-4o model offers detailed justifications and analyses for its ratings.

- Query Evaluation:

Persona	**Generated Query**	**GPT-4o Score**	**GPT-4o Justification**
24-year-old Female Meng Student in UofT. Interested in Cats, Movies, Fashion, Tennis, Travel, Music	Stylish women's tennis outfits	3 (High Relevance)	The persona has a documented interest in fashion and tennis. This search query **directly aligns** with the user's interest in both categories, making it **very likely** to be a search query from this persona.

Figure 17: Qualitative result for query generation

- Ad-Query Matching Evaluation:

Query	Matched Ad	Simulated Rater Feedback	GPT-4o Analysis
Stylish women's tennis outfits	Nike Women's Pure 11.75 Inch Tennis Skort | The V-notch detail enhances mobility, while built-in shorts provide coverage and tennis ball storage. Dri-FIT technology ensures you stay cool and dry, with the iconic white Nike swoosh on the left hem.	1 (Interested)	The product combines functionality (Dri-FIT technology, built-in shorts for coverage) with fashion (sleek design, flattering fabric), **strongly matched** with the search query "Stylish women's tennis outfits", **likely appealing** to someone interested in both fashion and tennis.
Stylish women's tennis outfits	A stylish almost-knee-high lace-up shoe with a clear chunky heel, soft ribbons, and cushioned support from the Strut by J. Adams.	0 (Not Interested)	While I am interested in fashion and tennis, this advertisement **does not align** well with my query for stylish women's tennis outfits. The ad is focused on a specific type of shoe that features a chunky clear heel and soft lush ribbons, which **does not seem suitable** for playing tennis. Although I do appreciate fashion, this specific item **does not meet** the functional or stylistic criteria for tennis outfits as per my query.

Figure 18: Qualitative result for ad matching

10 Discussion and Learning

The quantitative and qualitative results presented in section 9 have demonstrated the simulated rater model's outstanding performance, highlighting its effectiveness in query generation and its ability to replicate human judgment in evaluating ad relevance.

The team was particularly surprised by the model's capability to achieve such high performance with minimal intervention, relying solely on prompt engineering with brief persona data that GPT-4o model had never encountered before. This outcome underscores the remarkable zero-shot ability of GPT-4o model to generalize across unseen data and domains.

During this project, an essential learning for the team was the persona installation technique, which is the process of embedding personas into the simulated rater model (special thanks to our TA Zafar). The success of the simulated rater in generating queries and evaluating ad relevance that reflects a real individual's behavior highlighted the effectiveness of persona installation. If the team implements this project again, we will further expand the geographic and demographic diversities of the persona dataset to enhance the generalizability of the simulated rater model. Expanding the dataset will introduce greater variability in personas, and also allow the simulated rater model to better capture cultural, linguistic, and behavioral nuances, ultimately improving its evaluation performance in real-world applications.

11 Individual Contributions

The following table highlights the individual contributions of each team member to the AdVize project, showcasing each member's responsibilities in the development process.

Name	**Contributions**
Zhifei	Project structure initialization, Collected persona data, Designed Cosine Similarity Baseline pushing algorithm, Designed Fine-tuned Llama8B pushing algorithm.
Anthony	Collection of persona data, Evaluator for Simulated Queries, AdVize app development.
Qingyi	Collection of persona data, RAG GPT-4o pushing algorithm, Simulation model draft.
Shirley	Collection of persona data, AdVize app backend development, Implementation of the automated evaluation workflow, Dockerized Architecture setup (Prepared Docker images for Cosine Similarity Baseline adpush-algo, rag-gpt4o, and llama8b-adpush-algo).

Table 1: Individual Contributions to the AdVize Project

References

[1] S. Yuan et al. Real-time bidding for online advertising: Measurement and analysis. https://arxiv.org/abs/1306.6542, 2013. Accessed: 10-Dec-2024.

[2] Google. Improving search with rigorous testing. https://www.google.com/search/howsearchworks/how-search-works/rigorous-testing/, 2024. Accessed: 10-Dec-2024.

[3] S. Feizi et al. Online advertisements with llms: Opportunities and challenges. https://arxiv.org/abs/2311.07601, 2023. Accessed: 10-Dec-2024.

[4] Choi et al. Picle: Eliciting diverse behaviors from large language models with persona in-context learning. PICLe:ElicitingDiverseBehaviorsfromLargeLanguageModelswithPersonaIn-ContextLearning, 2024. Accessed: 10-Dec-2024.

[5] Docker, Inc. *Docker: Empowering App Development for Developers*, 2023. Accessed: 2023-12-09.

[6] T. Zhang et al. BERTScore: Evaluating Text Generation with BERT. https://arxiv.org/pdf/1904.09675, 2020. Accessed: 08-Dec-2024.

[7] scikit-learn. Cohen's kappa. https://scikit-learn.org/dev/modules/model_evaluation.html#cohen-kappa. Accessed: 08-Dec-2024.

A Appendix A: One Example from the start of system to the end

This Appendix demonstrated an example from the start to the end of the automated AdVize system. **This example follows the structure and process of Figure 7, please refer to it for the visualization.**

A.1 Persona Installation and Query Generation

As introduced before, the AdVize system starts with embedding the persona into the simulated rater model through prompts. As demonstrated in Figure 19, the example persona of one of team members, Cherry, is input into the simulated rater model. Next, the model takes the persona prompt and outputs 10 simulated queries according to Cherry's persona.

Query Generation Task

You are tasked with generating 10 realistic and contextually appropriate product search queries for a user profile.

User Profile:
- **Name**: Cherry Xia
- **Age**: 24
- **Gender**: Female
- **Location**: Toronto, Canada
- **Occupation**: Uoft Meng Student
- **Interests**: Cats, Movies, Fashion, Tennis, Travel, Music
- **Recent Searches**: Huawei Pockets2, ON tennis shoes, Skims sale

Instructions:
- Generate 10 realistic and contextually appropriate product search queries for the user profile.
- Reflect on the user's interests, occupation, and recent searches.
- Generate general product search queries relevant to the user's profile.
- Do not include location-specific terms or time-sensitive events in the queries.
- Make sure the generated query is an object that is reasonable and realistic.

Output:
- A numbered list of 10 queries. You can only generate these 10 queries, don't output extra words

Output

1. Cat-themed phone case for Huawei Pockets2
2. Stylish women's tennis outfits
3. Affordable noise-cancelling headphones
4. Vintage movie poster wall art
5. Travel-sized skincare essentials
6. Skims loungewear reviews
7. Fashionable cat-themed jewelry
8. Ergonomic study desk chair
9. Latest album releases on vinyl
10. Trendy oversized sunglasses

Figure 19: Query Generation Example

A.2 Ad Pushing Algorithm and Pushing Ads

For demonstration, GPT-4o with RAG Ad pushing algorithm is shown as the example. In Figure 20, four simulated queries generated by the simulated rater are received as the first input. Then, ads Content from CommmercialAds Dataset is served as the second input into this algorithm. The algorithm will subsequently process the inputs and output the most relevant ad as shown in Figure 20. (Here only show three queries and ads for demo)

Example GPT-4o RAG Ads Pushing Prompt

Input - Simulated Queries

First input:

```
simulated_queries = [
    "Cat-themed phone case for Huawei Pockets2",
    "Stylish women's tennis outfits",
    "Fashionable cat-themed jewelry",
  ]
```

Second input:

<table>
<tr><th></th><th>Ads_Content</th></tr>
<tr><td>0</td><td>Klutch 15-Slot Universal Wrench Pouch | Sturdy...</td></tr>
<tr><td>1</td><td>TAG OFF Skin Natural Skin Tag Remover Take Ski...</td></tr>
<tr><td>2</td><td>Harley-Davidson Skull LED Fuel Gauge | This lo...</td></tr>
<tr><td>3</td><td>CHANEL Frt Pocket Handbag Quilted Patent Leath...</td></tr>
<tr><td>4</td><td>Dell WD15 Monitor Dock 4K with 180W Adapter, U...</td></tr>
<tr><td>5</td><td>Entwined Halo Diamond Engagement Ring - 14K Ro...</td></tr>
<tr><td>6</td><td>Emser Tile T06FONT0404UT Trav Fontane Tumbled ...</td></tr>
</table>

Output

Search Query: Cat-themed phone case for Huawei Pockets2

Ad 210144: Camouflage Cat Laptop case | Laptop sleeve. This cat is MADE OF LEAVES! Rhubarb, poplar and pine needle whiskers! Relaxing in his fave garden spot, hes just waiting for his next prey to stroll by cause this mighty hunter is really quite lazy. His camouflage serves him well for this hunting style as he is not seen so well until the last minute. And then we know, at that point, its entirely too late for someone!. Painting & Digital. By TeriMadeLeafArt. cat, cat lovers gift, nature lovers decor, domestic animal, cat in nature, colorful cat, cat t shirt, cat dress, cat pillow, cat clock, orange cat, cat leaves, cat outdoors

Search Query: Stylish women's tennis outfits

Ad 52929: Nike Women's Pure 11.75 Inch Tennis Skort | The Womens Pure Tennis Skort by Nike features a sleek design that allows you to bring power and finesse to the court The classic Nike looks are updated and improved with the Pure line bringing thicker and softer fabric that flatters all womens bodies The vnotch detail increases range of motion during play The builtin shorts provide coverage and security as well as an option for tennis ball storage DriFIT technology provides sweat management so you stay cool and dry on the court The exclusive recognizable Nike swoosh appears in White on the hem above the left knee

Search Query: Fashionable cat-themed jewelry

Ad 181834: Bestselling Classic Themed Charm Bracelets | Bestselling Classic Themed Charm Bracelets

---

Figure 20: Example of GPT-4o RAG Ad pushing Algo

A.3 Feedback Generation

After the ad pushing algorithm returns the ads, the simulated rater model takes in the ads as the input. As shown in Figure 21, it outputs the decision of if the simulated persona would find the ads appealing, along with the justification for it's decision.

Example Simulated Rater Feedback Generation

Input - Pushed Ads

Ad 210144: Camouflage Cat Laptop case | Laptop sleeve. This cat is MADE OF LEAVES! Rhubarb, poplar and pine needle whiskers! Relaxing in his fave garden spot, hes just waiting for his next prey to stroll by cause this mighty hunter is really quite lazy. His camouflage serves him well for this hunting style as he is not seen so well until the last minute. And then we know, at that point, its entirely too late for someone!. Painting & Digital. By TeriMadeLeafArt. cat, cat lovers gift, nature lovers decor, domestic animal, cat in nature, colorful cat, cat t shirt, cat dress, cat pillow, cat clock, orange cat, cat leaves, cat outdoors

Ad 52929: Nike Women's Pure 11.75 Inch Tennis Skort | The Womens Pure Tennis Skort by Nike features a sleek design that allows you to bring power and finesse to the court The classic Nike looks are updated and improved with the Pure line bringing thicker and softer fabric that flatters all womens bodies The vnotch detail increases range of motion during play The builtin shorts provide coverage and security as well as an option for tennis ball storage DriFIT technology provides sweat management so you stay cool and dry on the court The exclusive recognizable Nike swoosh appears in White on the hem above the left knee

Ad 181834: Bestselling Classic Themed Charm Bracelets | Bestselling Classic Themed Charm Bracelets

Output - Feedback for Pushed Ads

Interested: 1 (Yes)
Justification: I love cats and enjoy cat-themed items, so the unique and artistic design of this cat-themed laptop case does appeal to my general interests. Even though it is not specifically a phone case, the playful and creative description aligns with my love for cat-inspired fashion and accessories, making me interested in this ad.

Interested: 1 (Yes)
Justification: This ad perfectly aligns with my query for stylish women's tennis outfits. The Nike Women's Pure Tennis Skort is not only functional with features like built-in shorts and sweat management through Dri-FIT technology but also stylish with a sleek and flattering design. As someone who enjoys tennis and values fashionable yet practical sportswear, I find this ad highly relevant and appealing.

Interested: No
Justification: While I am interested in fashionable and cat-themed items, this ad lacks specificity about whether the charm bracelets are cat-themed. Without a clear indication that the jewelry aligns with my love for cats, it doesn't capture my interest. The vague description fails to convey relevance to my query.

Figure 21: Example of Feedback Generating for Simulated Rater

A.4 Satisfication Rate Calculation and Visualization

The stratification rate can be calculated with the following function:

$$\text{Satisfication Rate} = \frac{\sum \text{Positive Feedback Num}}{\sum \text{Total Feedback Num}}$$

In **this specific example**, the satisfaction rate is:

$$\text{Satisfication Rate} = \frac{\sum \text{Positive Feedback Num}}{\sum \text{Total Feedback Num}} = \frac{2}{3} = 66.6667\%$$

As we have 2 ”yes” feedbacks and in total 3 feedbacks. Then, this stratification rate is visualized as demonstrated in Figure 22

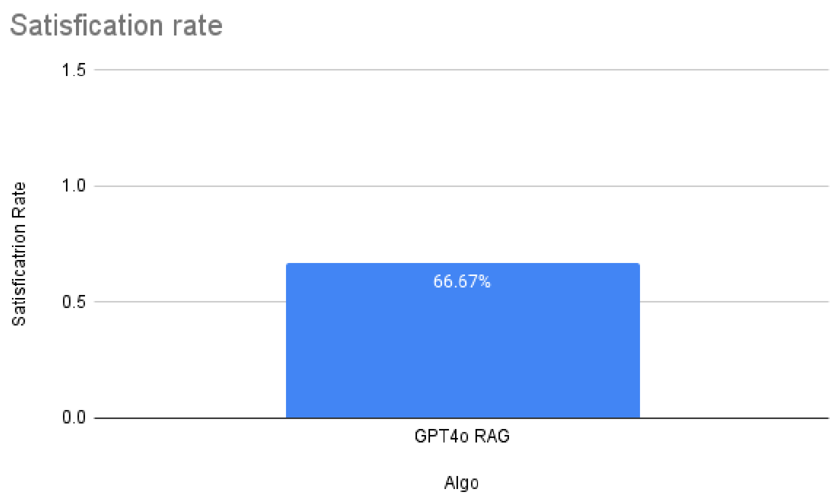


Figure 22: Example of GPT-4o RAG Ad pushing Algo