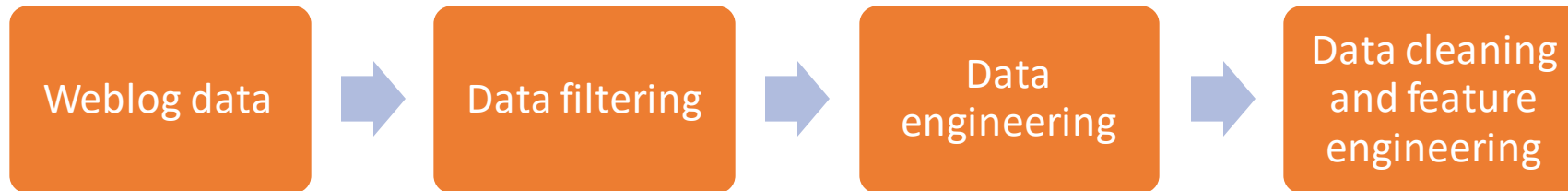


***InteraRec*: Interactive Recommendations using Multimodal Large Language Models**

Saketh Reddy Karra, **Theja Tulabandhula**
University of Illinois at Chicago

Current Recommendation Systems are Complex

- ❑ Recommendation systems leverage the data mined through **weblogs**.



- Interpreting the raw **weblogs** directly can be challenging for non-experts (e.g., product managers).
- Sophisticated data engineering techniques required to extract the necessary and relevant features.

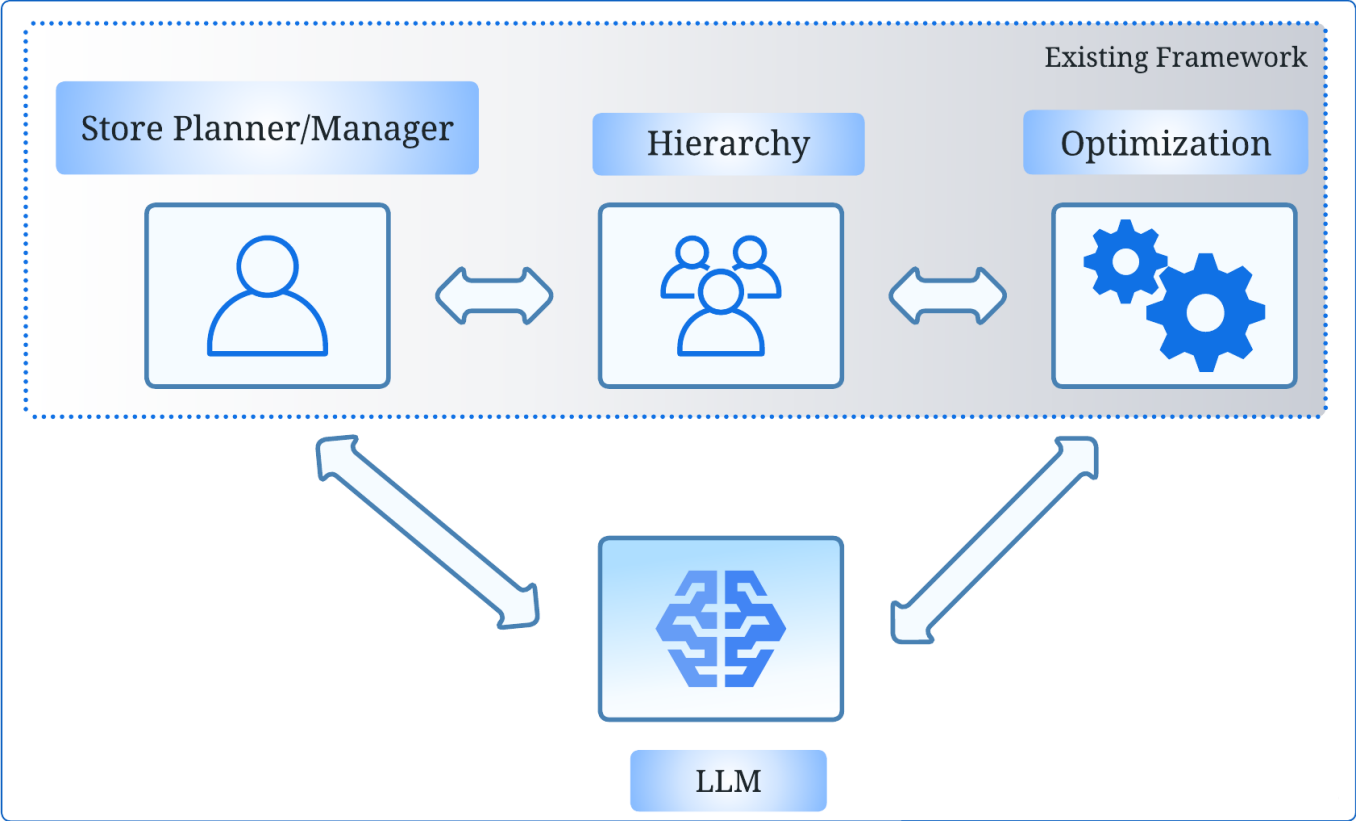
Our Goal: Venture Beyond Weblog-centric Systems using MLLMs

- ❑ MLLMs are proficient in vision-language tasks such as image captioning and reasoning (i.e., **visual question answering**).
- ❑ Screenshots as a sensing modality
 - Visual modality provides a lucid and transparent representation.

Potential advantages:

- Generate high quality real-time recommendations with lesser steps.
- Interpretability without the need for extensive engineering.
- Simplified inputs and outputs.

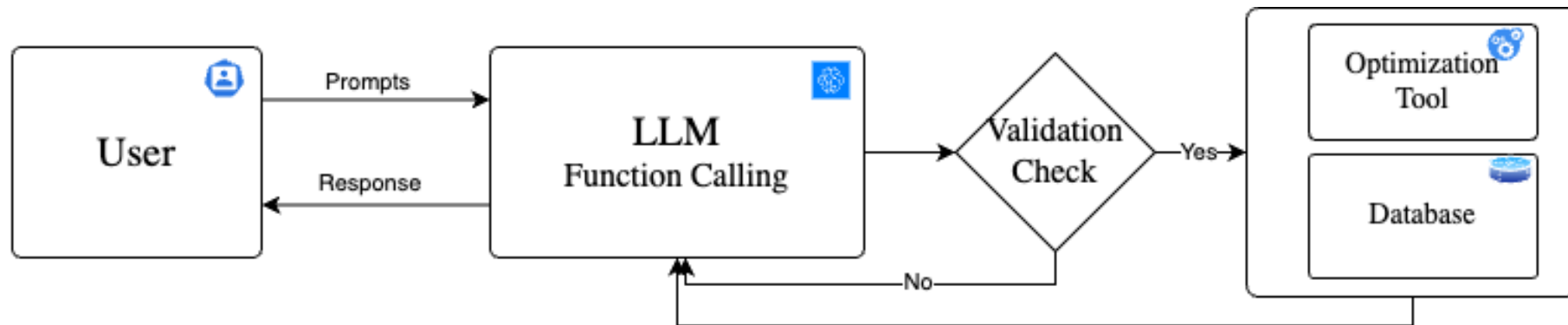
A Prelude: LLM as a Co-Pilot



Incorporating LLM as an intelligent assistant to an existing workflow.

Co-Pilot Design

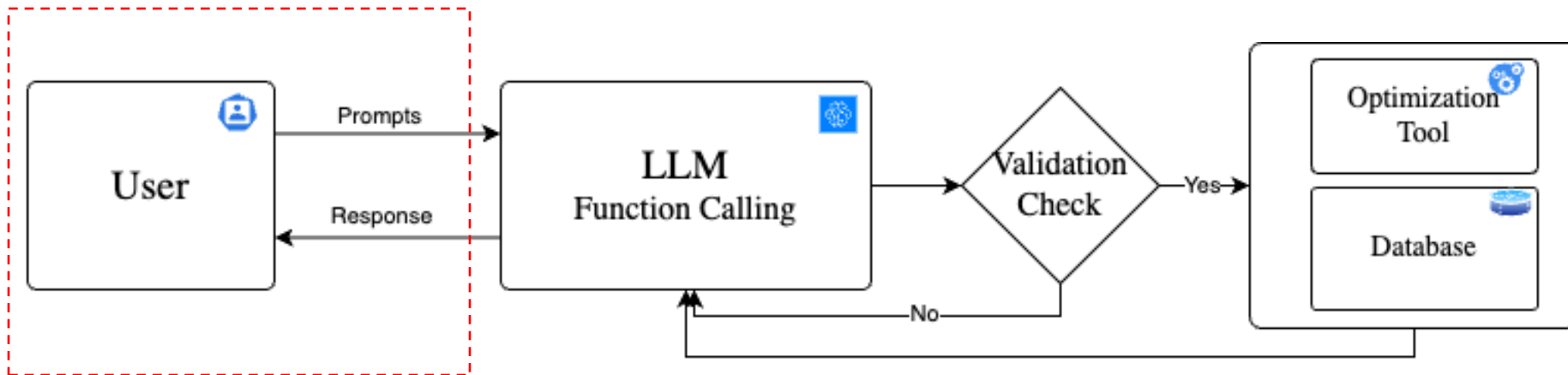
1. Prompt design
2. Prompt decomposition and function calling
3. Tool execution and response generation



Co-Pilot Design

□ Prompt design

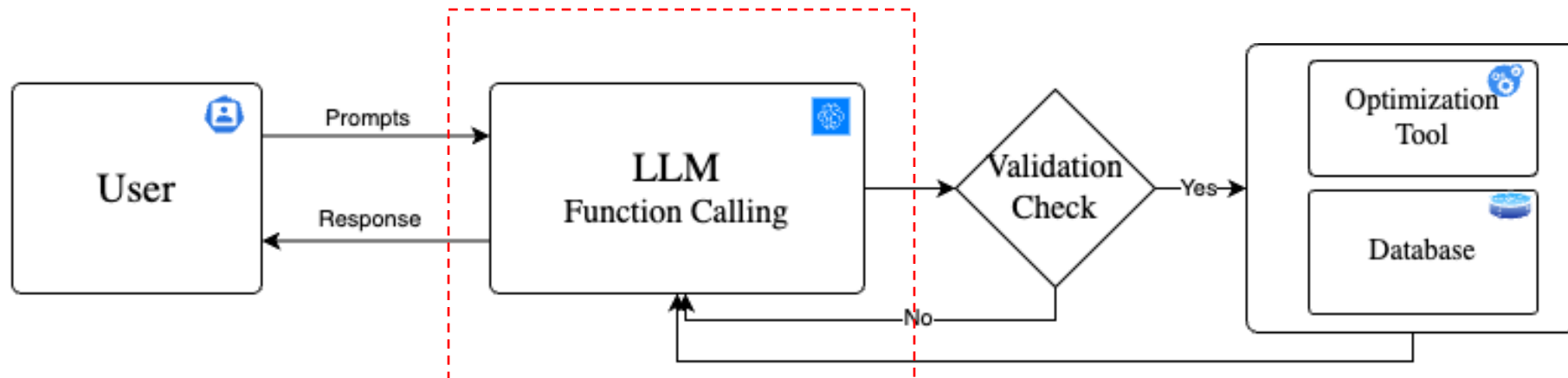
- Standardized template for input prompts to deduce information provided by the user for the task at hand.



Co-Pilot Design

□ Prompt decomposition and function calling

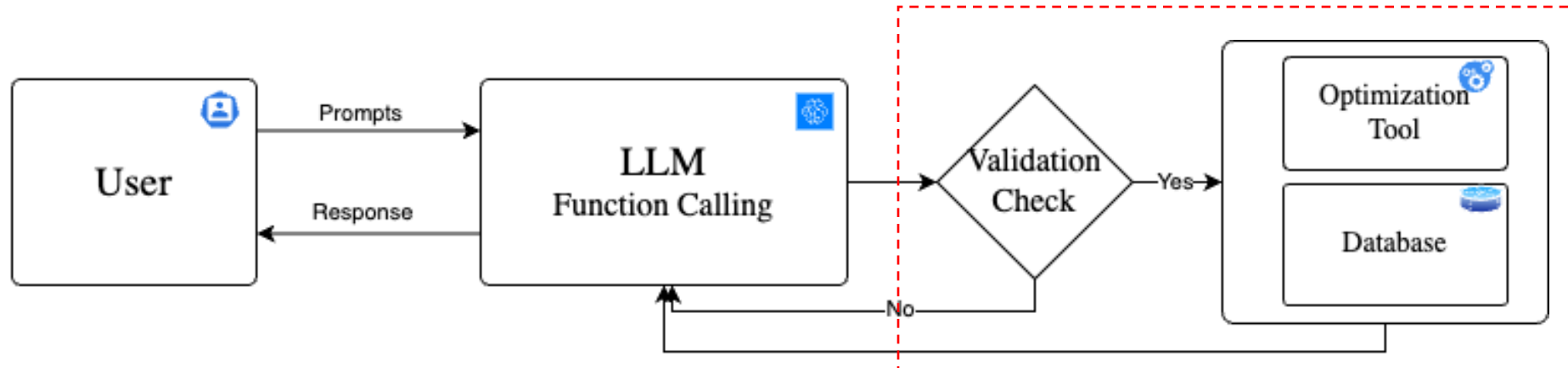
- Leverage LLM function-calling to break down the standardized prompt to actionable objects.
- Append history to follow-up calls for multi-turn conversations.



Co-Pilot Design

❑ Tool execution and response generation

1. Perform data validation checks.
2. Retrieve relevant information from databases/external sources.
3. Execute tool (e.g., optimization scripts with solvers).
4. Process through LLM for user-friendly responses.

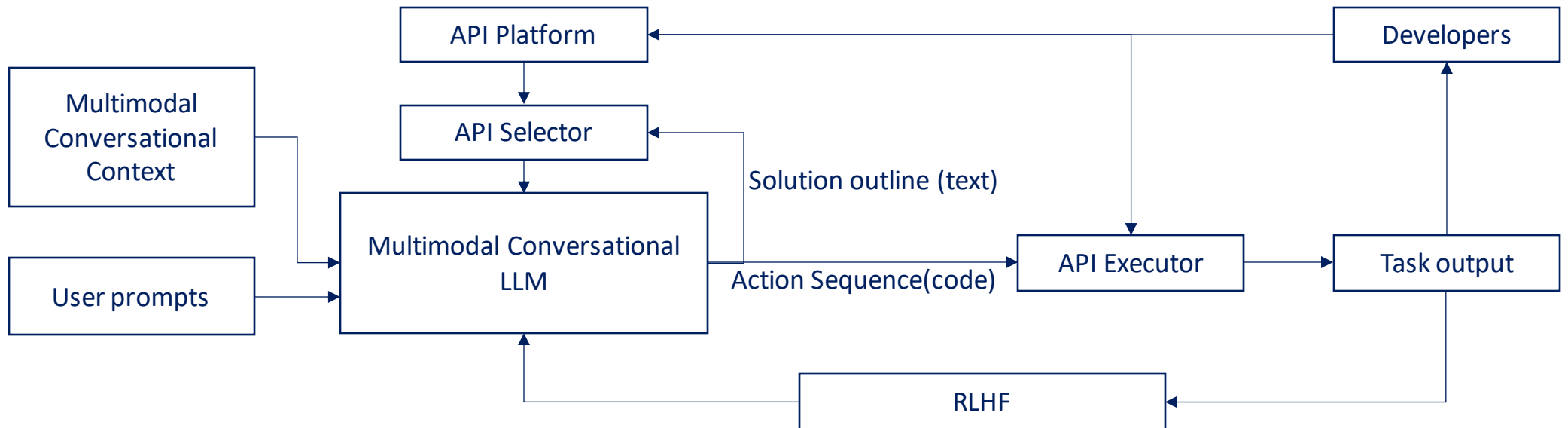


Select Prior Works

Hugginggpt*



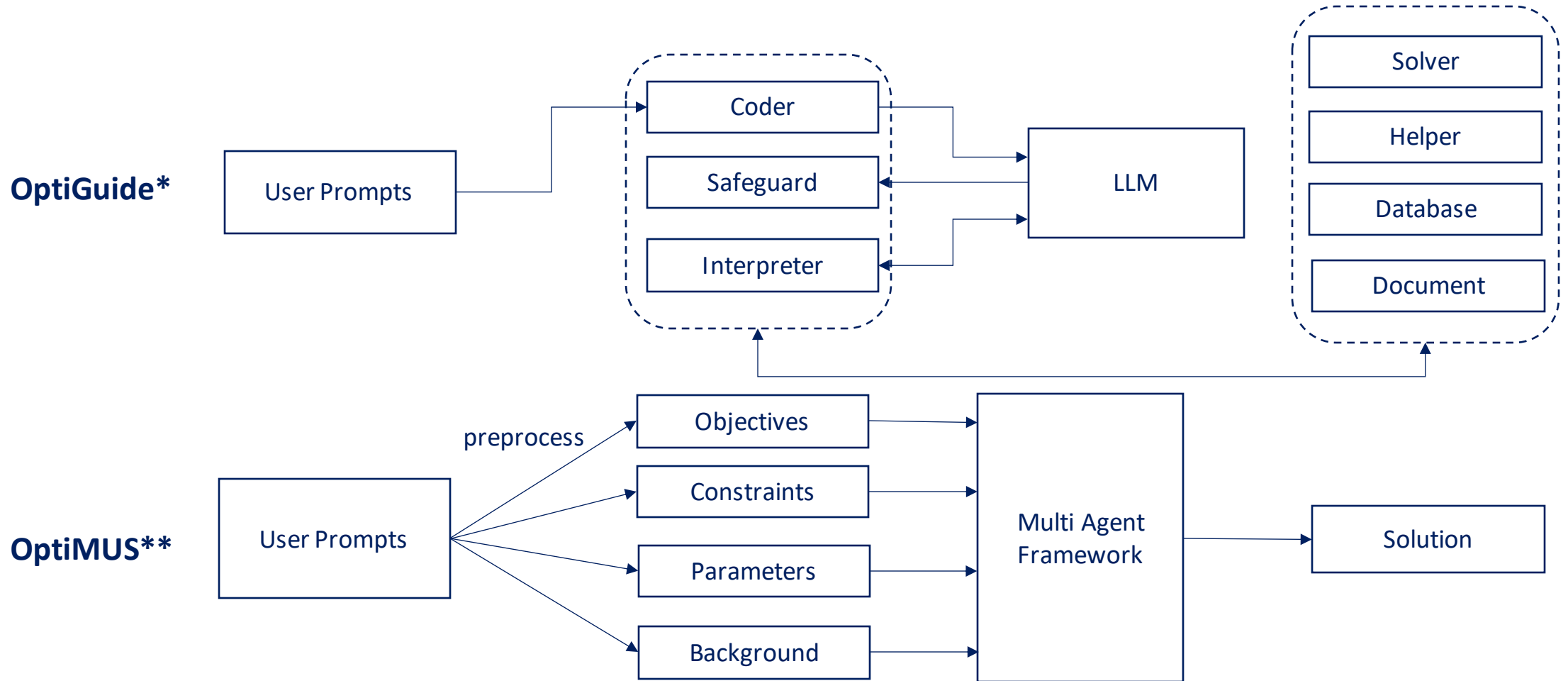
TaskMatrix.AI**



*Shen, Yongliang, et al. "Hugginggpt: Solving ai tasks with chatgpt and its friends in hugging face." *Advances in Neural Information Processing Systems* 36 (2024).

**Liang, Yaobo, et al. "Taskmatrix.ai: Completing tasks by connecting foundation models with millions of apis." *arXiv preprint arXiv:2303.16434* (2023).

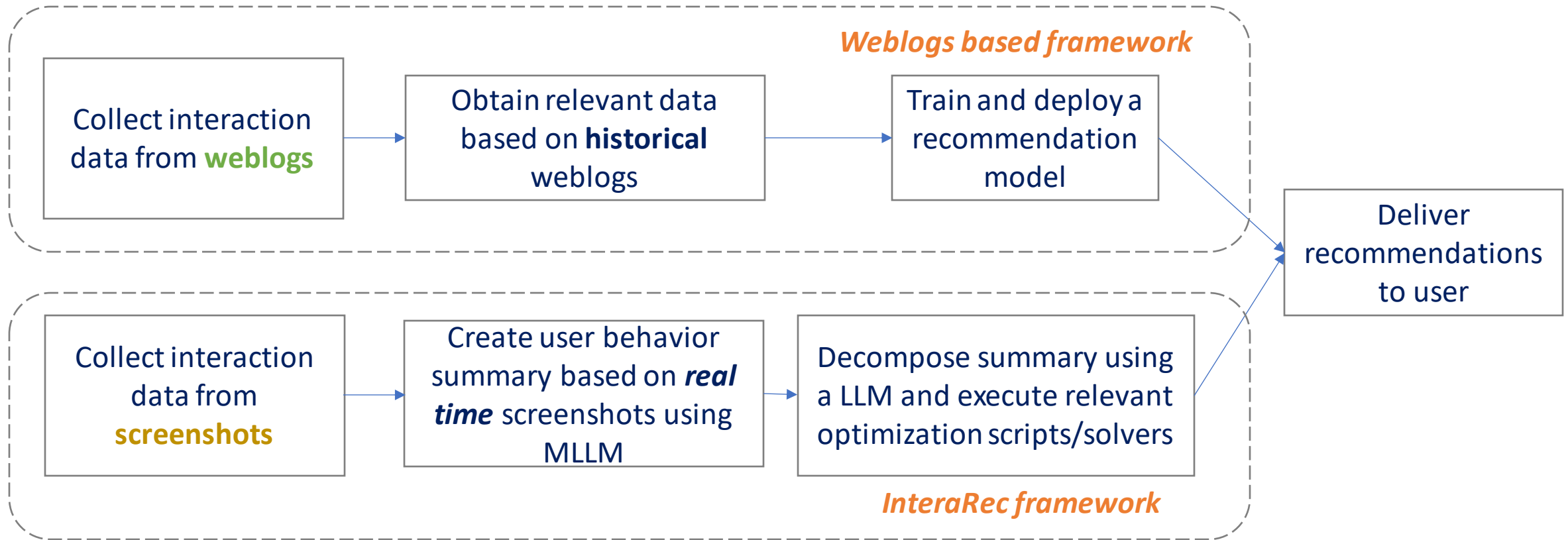
Select Prior Works



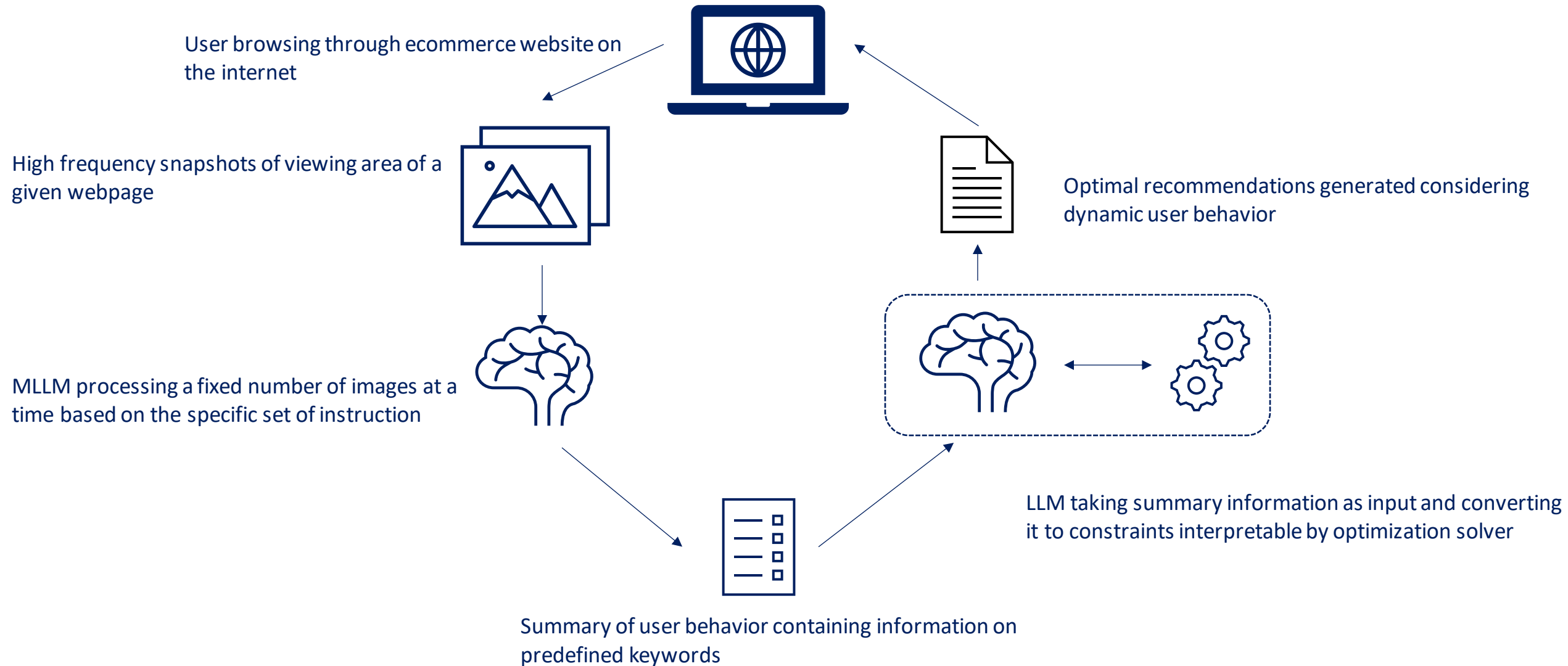
*Li, Beibin et al. "Large Language Models for Supply Chain Optimization" Arxiv: 2307.03875 (2023).

**AhmadiTeshnizi, Ali et al. "OptiMUS: Optimization Modeling Using MIP Solvers and large language models" Arxiv: 2310.06116 (2023).

Screenshot-based Recommendation Framework: *InteraRec*



Screenshot-based Recommendation Framework: *InteraRec*



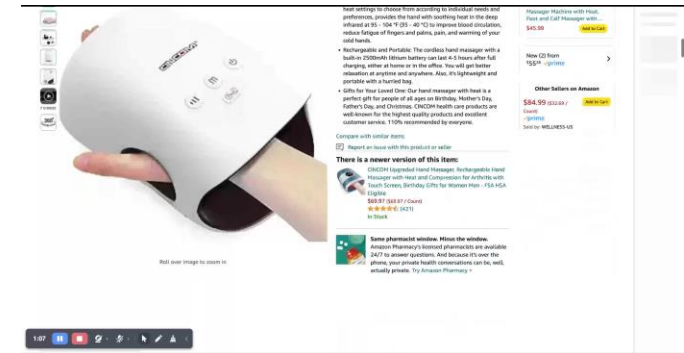
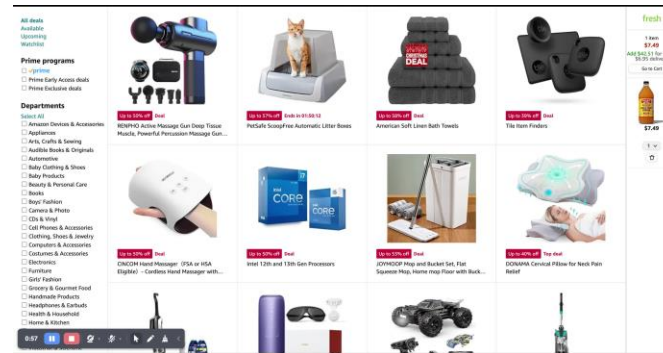
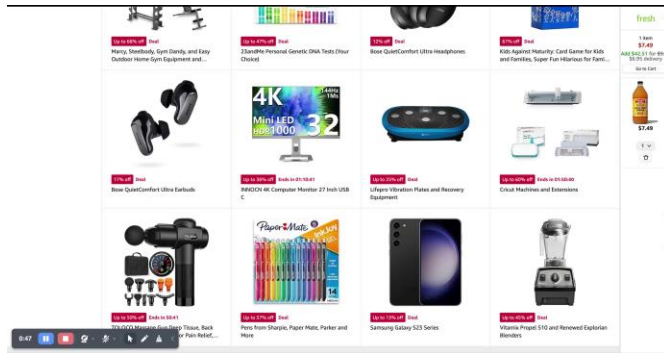
Recommendations with *InteraRec*: Three Key Steps

1. Screenshot generation
2. Behavioral summarization
3. Response generation

Recommendations with *InteraRec*

❑ Screenshot generation

- Capture high-frequency, high-quality **screenshots** of visible area of the web pages.
- Frequency can vary depending on browsing habits of users.



Three screenshots showcasing user interactions with the Amazon.com website, each captured at a 10-second interval.

Recommendations with *InteraRec*

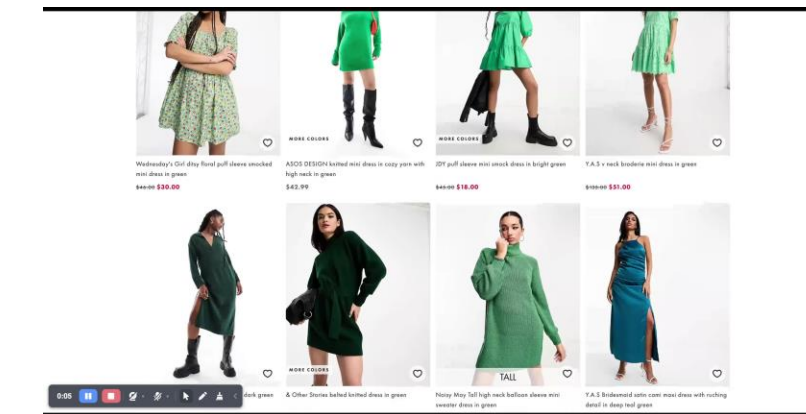
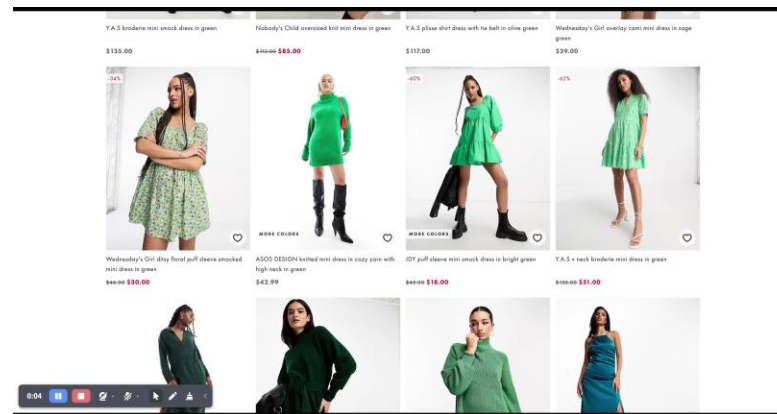
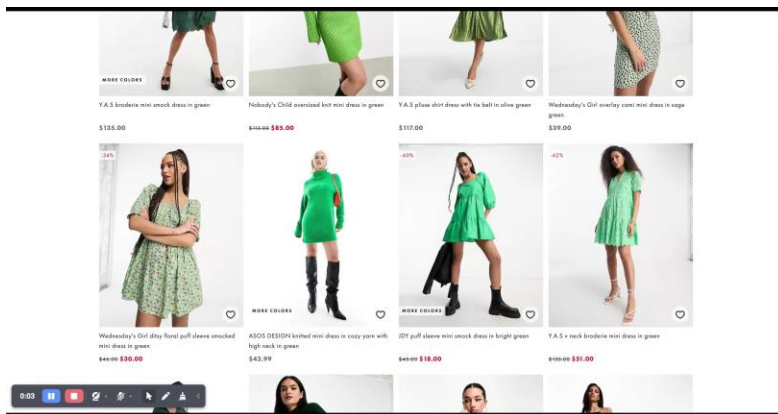
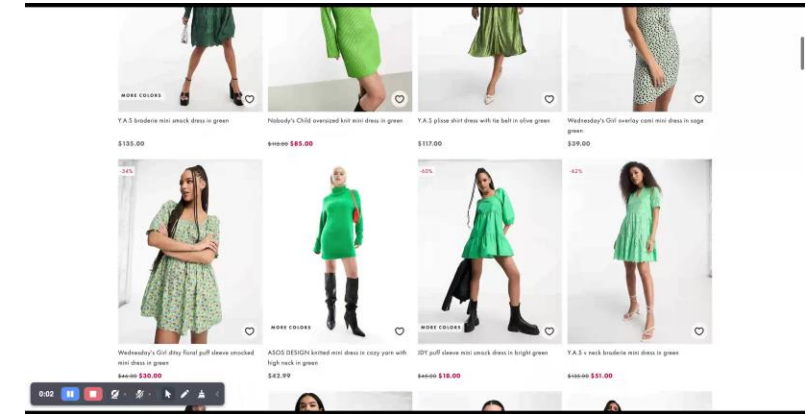
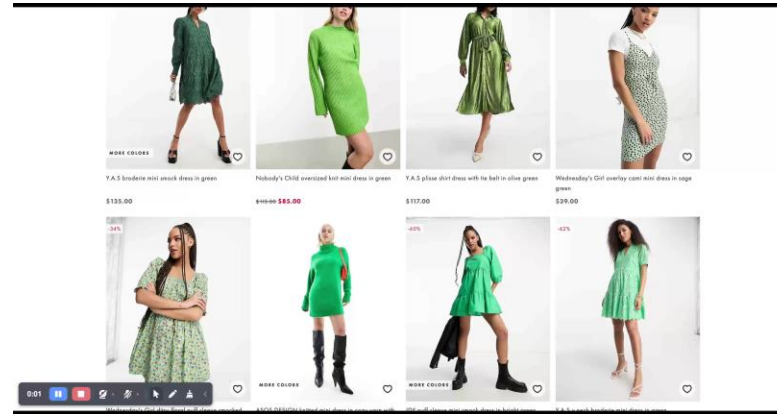
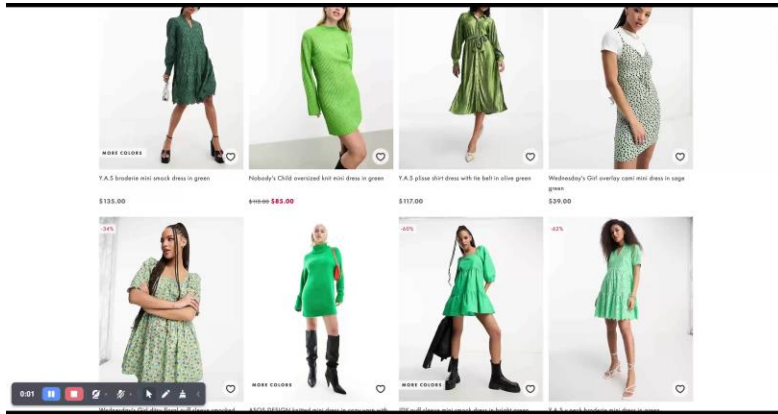
□ Behavioral summarization

- Process a finite number of **screenshots** in *near* real-time using a foundational MLLM.
- Instruct the MLLM to succinctly summarize user interaction behavior across *predefined categories*.

```
response = client.chat.completions.create(  
model="gpt-4-vision-preview"  
messages=[  
  {  
    "role": "user"  
    "content": [  
      {  
        "type": "text"  
        "text": "What can you infer from the images below \  
with regards to a user preference in the following categories? \  
Product Characteristics, \  
Lowest Price, \  
Highest Price, \  
Brand Preference, \  
Product Specifications, \  
User Reviews and Testimonials, \  
Comparisons, \  
Promotions. \  
Write a response that contains the above information. \  
if any of the categorical information is unavailable, \  
mark it as not available "}]...],...}
```

Guidelines for MLLM to generate a summary of user interactions using predefined keywords.

Screenshot Generation and Behavior Summarization: Example 1



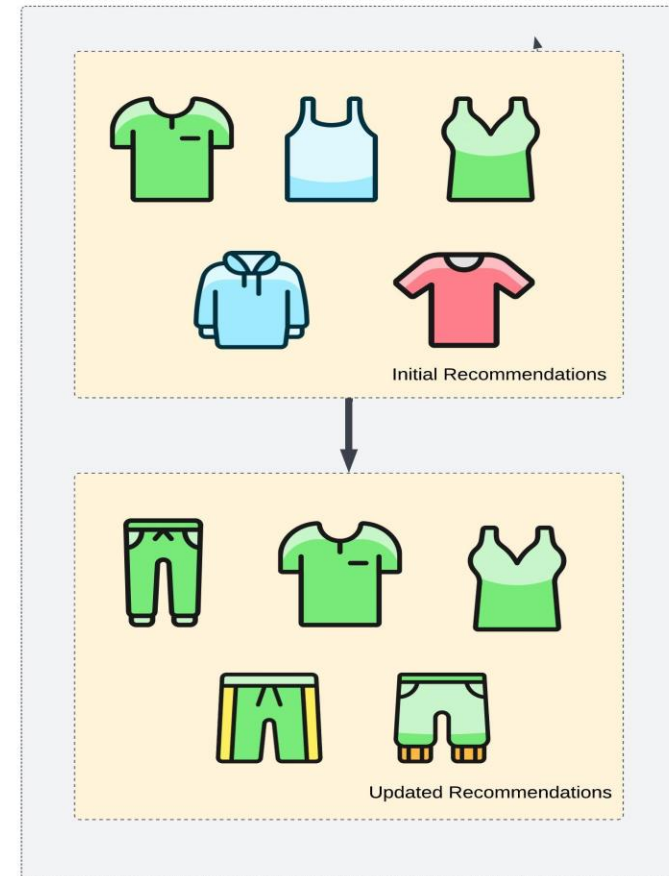
Six screenshots showcasing user interactions with the Asos.com website, each captured at a 1-second interval.

Screenshot Generation and Behavior Summarization: Example 1

Given the images shown, it's important to note that while certain inferences can be made about product preferences and price sensitivity among other things, these are assumptions based on a set of images featuring green dresses of different styles and prices. Here is a possible interpretation in JSON format: "json { "Product Characteristics": "User appears to be interested in dresses, specifically in shades of green, with various styles including mini, midi, and maxi lengths.", "Lowest Price": "18.00\$ for the JDY puff sleeve mini smock dress in bright green", "Highest Price": "144.00\$ for the Y.A.S Bridesmaid satin cami maxi dress with ruching detail in deep teal green", "Brand Preference": "Variety of brands shown, such as Y.A.S, Wednesday's Girl, Nobody's Child, suggesting no clear preference for a single brand.", "Product Specifications": "Not Available", "User Reviews and Testimonials": "Not Available", "Comparisons": "The user may be comparing dresses based on style and discounts, with several items showing reduced prices.", "Promotions": "Several dresses are listed with percentage discounts, indicating that promotions may influence user preferences." }

The response indicates an interest in green dresses, sensitivity to promotions (as indicated by discounted items), a range of price points indicating a potentially broader budget range or comparison shopping, and a variety of brands.

Extracted summary from Asos website activities revealing a user preference for green dresses, diverse brand interest, and price sensitivity.



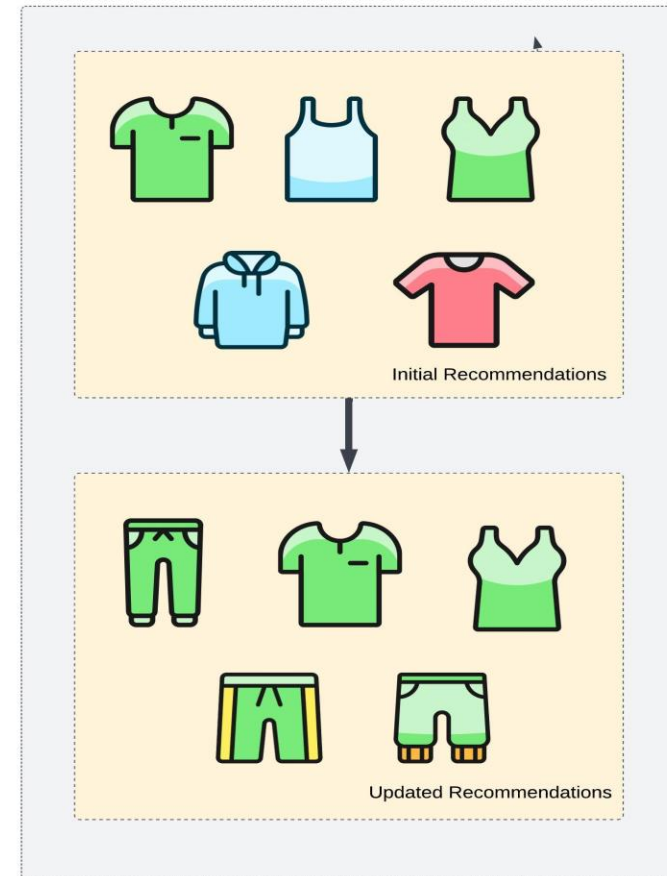
Real-time recommendations incorporating the dynamic user behavior.

Screenshot Generation and Behavior Summarization: Example 1

Given the images shown, it's important to note that while certain inferences can be made about product preferences and price sensitivity among other things, these are assumptions based on a set of images featuring green dresses of different styles and prices. Here is a possible interpretation in JSON format: "json { "Product Characteristics": "User appears to be interested in dresses, specifically in shades of **green**, with various styles including mini, midi, and maxi lengths.", "Lowest Price": "**18.00\$** for the JDY puff sleeve mini smock dress in bright green", "Highest Price": "**144.00\$** for the Y.A.S Bridesmaid satin cami maxi dress with ruching detail in deep teal green", "Brand Preference": "Variety of brands shown, such as **Y.A.S, Wednesday's Girl, Nobody's Child**, suggesting no clear preference for a single brand.", "Product Specifications": "Not Available", "User Reviews and Testimonials": "Not Available", "Comparisons": "The user may be comparing dresses based on style and discounts, with several items showing reduced prices.", "Promotions": "Several dresses are listed with percentage discounts, indicating that promotions may influence user preferences." } "

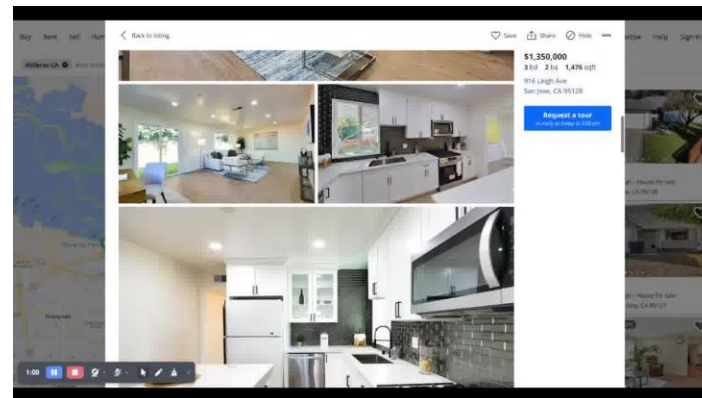
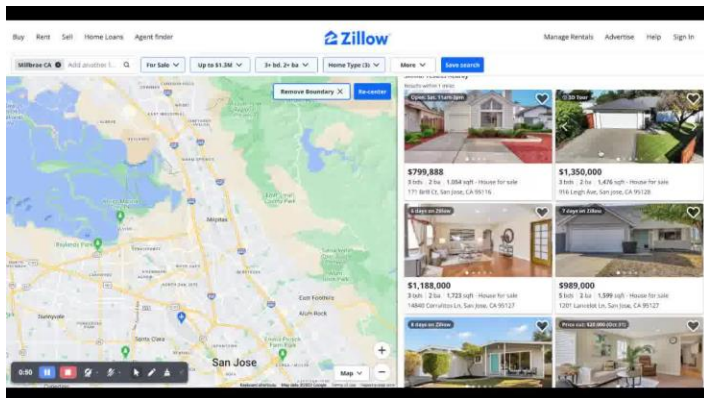
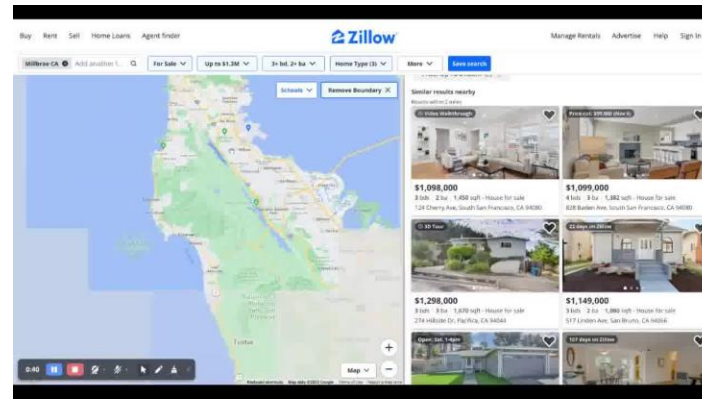
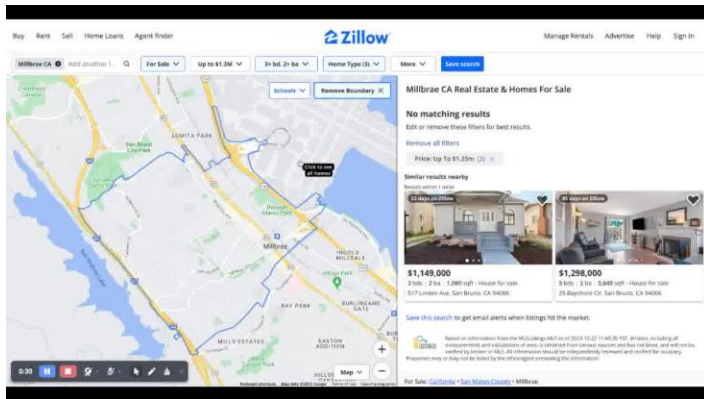
The response indicates an interest in green dresses, sensitivity to promotions (as indicated by discounted items), a range of price points indicating a potentially broader budget range or comparison shopping, and a variety of brands.

Extracted summary from Asos website activities revealing a user preference for green dresses, diverse brand interest, and price sensitivity.



Real-time recommendations incorporating the dynamic user behavior.

Screenshot Generation and Behavior Summarization: Example 2

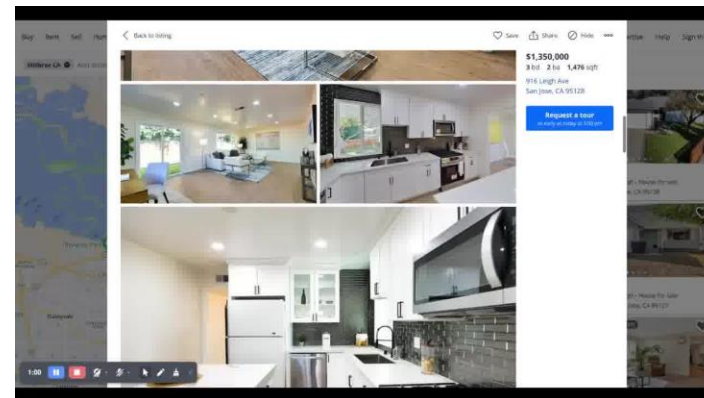
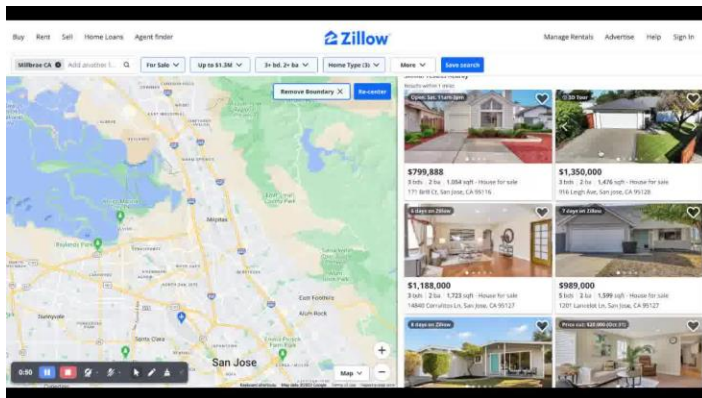
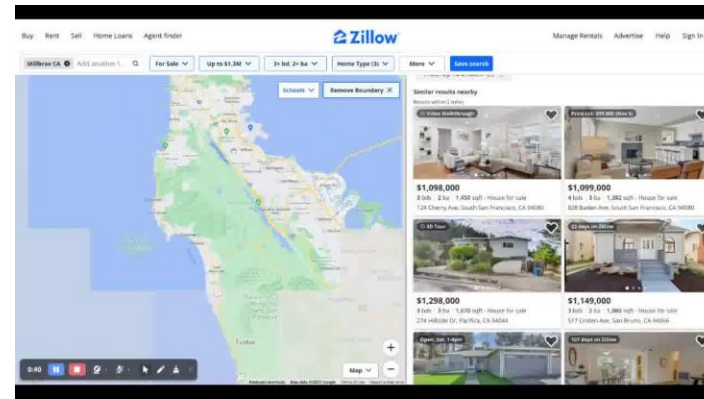
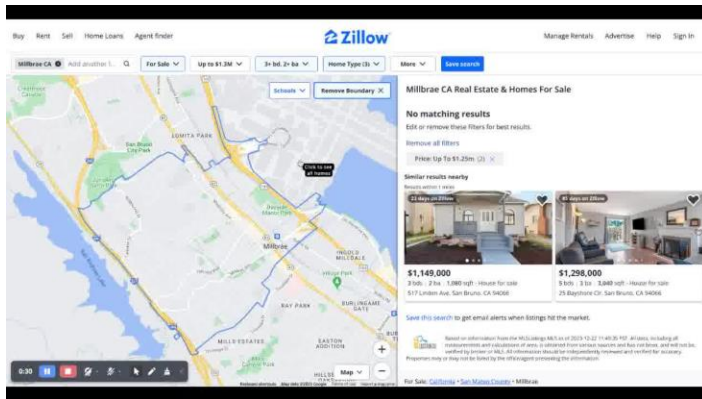


“Based on the provided images, it appears that the user is searching for real estate properties for sale on a website, likely with certain preferences. The images show a website with real estate listings that include price ranges, the number of bedrooms and bathrooms, square footage, and location. Using the data visible in the images, I will attempt to create JSON-formatted data that represents the user’s inferred preferences within the limitations of the images. “json “*Product Characteristics*”: “Location”: “Millbrae, CA; South San Francisco, CA; Pacifica, CA; San Jose, CA”, “Property Type”: “House”, “Bedrooms”: ”3+”, “Bathrooms”: ”2+”, “Lowest Price”: “\$799,888”, “Highest Price”: “\$1,350,000”, “Brand Preference”: “not available”, “Product Specifications”: “Square Footage”: “Low”: “,054 sqft”, “High”: “3,040 sqft”, “User Reviews and Testimonials”: “not available”, “Comparisons”: “Available”: true, “Criteria”: “Price, Location, Size, Bedrooms, Bathrooms”, “Promotions”: “Price Cut”: “\$20,000 (Oct 31) on a property listed for \$989,000” “Please note that the information above is constructed from the context provided by the images and might lack accuracy in details that are not visible or provide insufficient context. Some categories like Brand Preference and User Reviews and Testimonials do not apply in the context of real estate listings and therefore are marked as “not available.” Also, the promotion information is based on limited data indicating a price cut on a specific property, as seen in the images.”

Four screen captures showcasing user interactions with the Zillow.com website, each captured at a 10-second interval.

Extracted summary of user preferences shaped by activities on the Zillow.com website.

Screenshot Generation and Behavior Summarization: Example 2



“Based on the provided images, it appears that the user is searching for real estate properties for sale on a website, likely with certain preferences. The images show a website with real estate listings that include price ranges, the number of bedrooms and bathrooms, square footage, and location. Using the data visible in the images, I will attempt to create JSON-formatted data that represents the user’s inferred preferences within the limitations of the images. “json “*Product Characteristics*”: “Location”: “Millbrae, CA; South San Francisco, CA; Pacifica, CA; San Jose, CA”, “*Property Type*”: “House”, “*Bedrooms*”: “3+”, “*Bathrooms*”: “2+”, “*Lowest Price*”: “\$799,888”, “*Highest Price*”: “\$1,350,000”, “*Brand Preference*”: “not available”, “*Product Specifications*”: “Square Footage”: “Low”: “,054 sqft”, “High”: “3,040 sqft”, “*User Reviews and Testimonials*”: “not available”, “*Comparisons*”: “Available”: true, “*Criteria*”: “Price, Location, Size, Bedrooms, Bathrooms”, “*Promotions*”: “Price Cut”: “\$20,000 (Oct 31) on a property listed for \$989,000” “Please note that the information above is constructed from the context provided by the images and might lack accuracy in details that are not visible or provide insufficient context. Some categories like Brand Preference and User Reviews and Testimonials do not apply in the context of real estate listings and therefore are marked as “not available.” Also, the promotion information is based on limited data indicating a price cut on a specific property, as seen in the images.”

Four screen captures showcasing user interactions with the Zillow.com website, each captured at a 10-second interval.

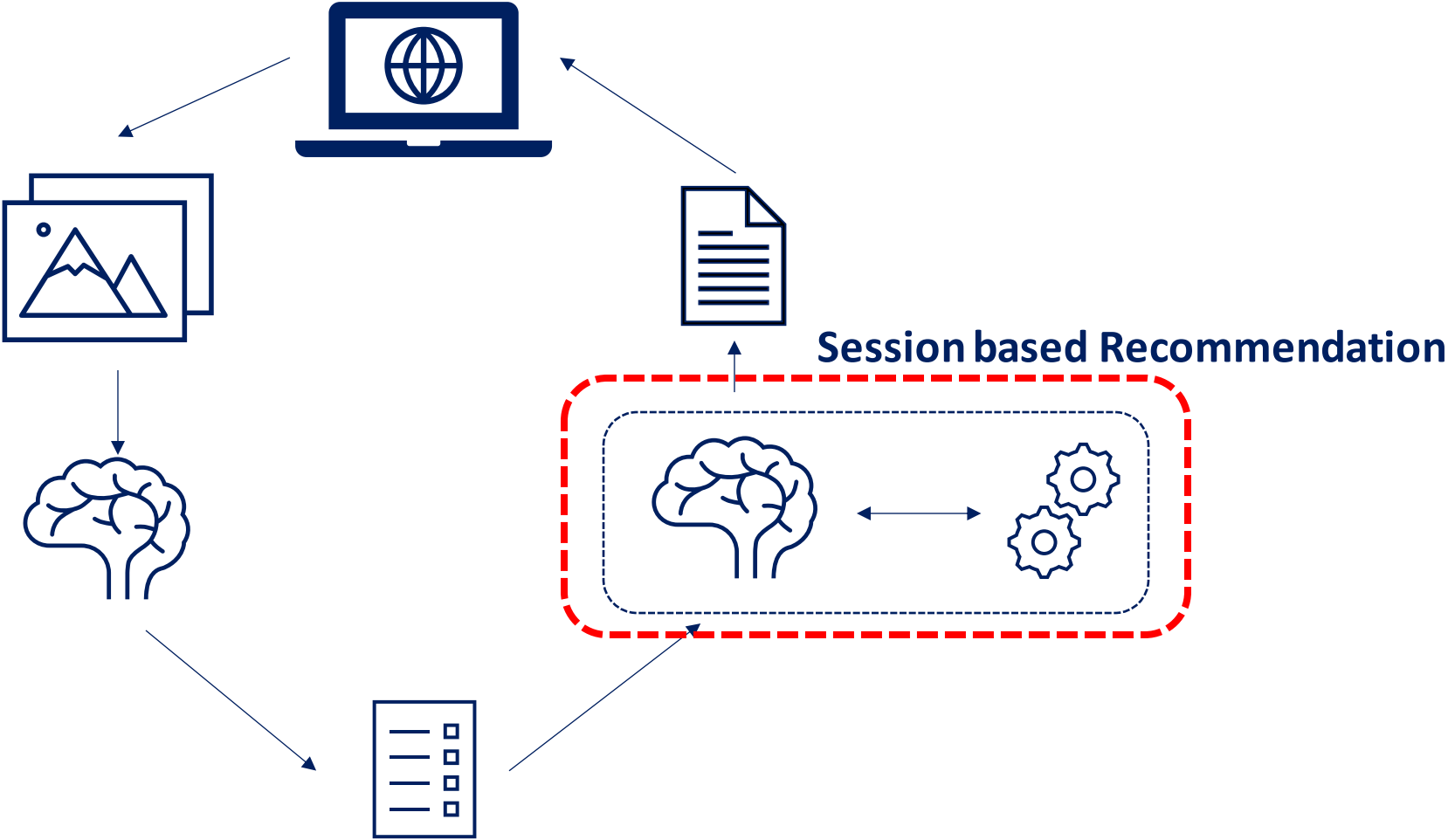
Extracted summary of user preferences shaped by activities on the Zillow.com website.

Recommendations with *InteraRec*

□ Response generation

- The information to be extracted from the behavioral summary depends on how downstream recommendation methods that can use them.
- Leverage the function-calling capabilities of the LLM to decompose the summary into relevant parameters using appropriate functions.
- Execute recommendation method to generate solutions

Using InteraRec with a Session based Recommendation Model



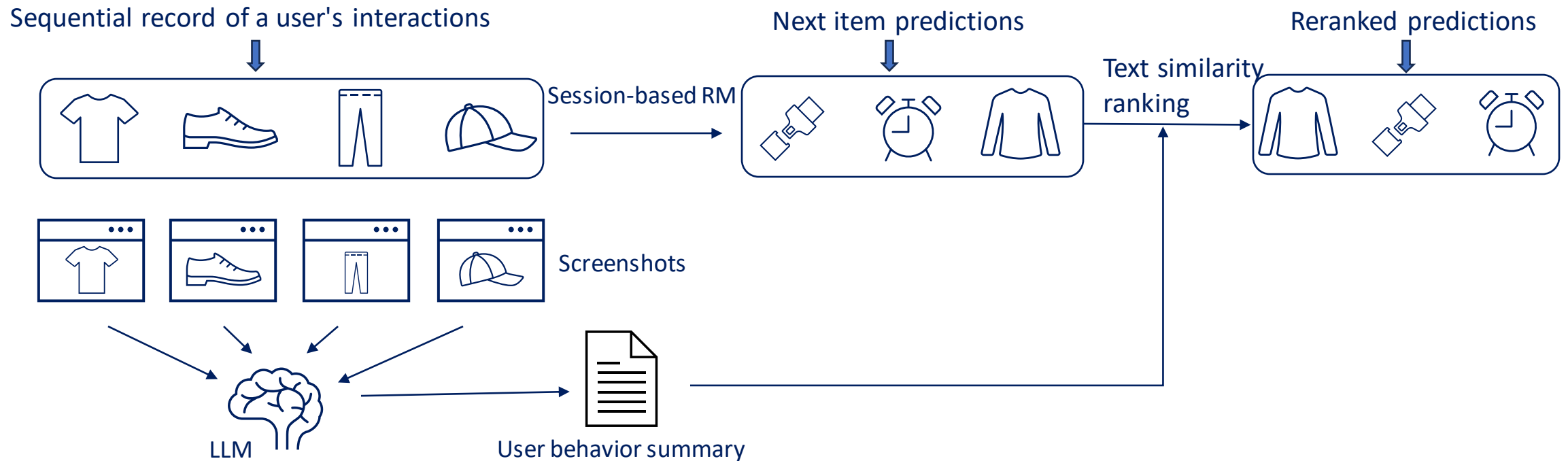
Using InteraRec with a Session based Recommendation Model

- Session based recommendation: Given a user session, recommendation on the product that the user might interact with at the next time step



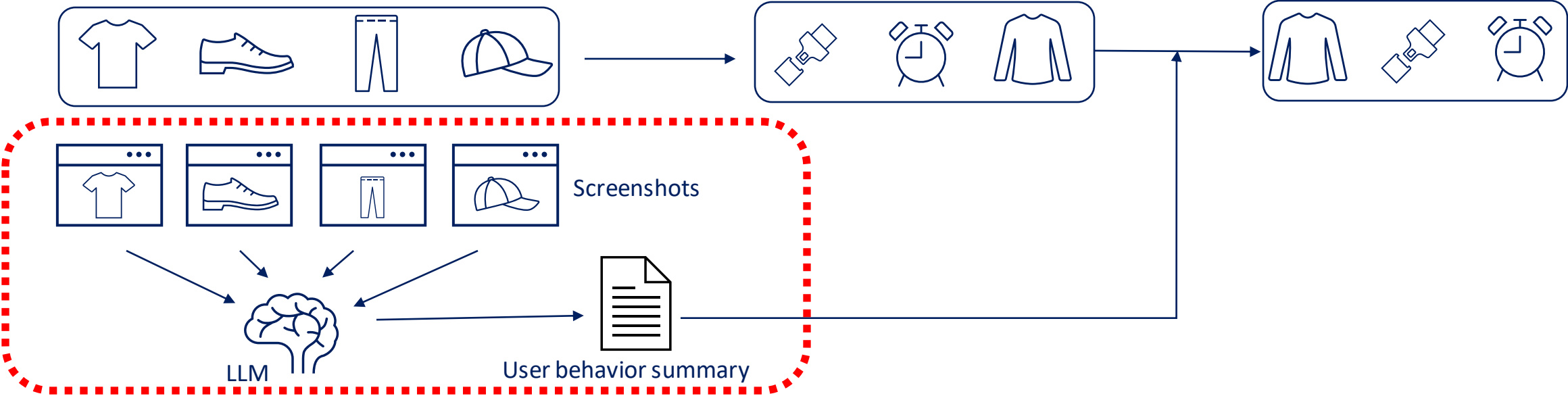
Using InteraRec with a Session based Recommendation Model

1. Screenshot generation
2. Behavioral summarization
3. Response generation (Session based recommendations + reranking)



Using InteraRec with a Session based Recommendation Model

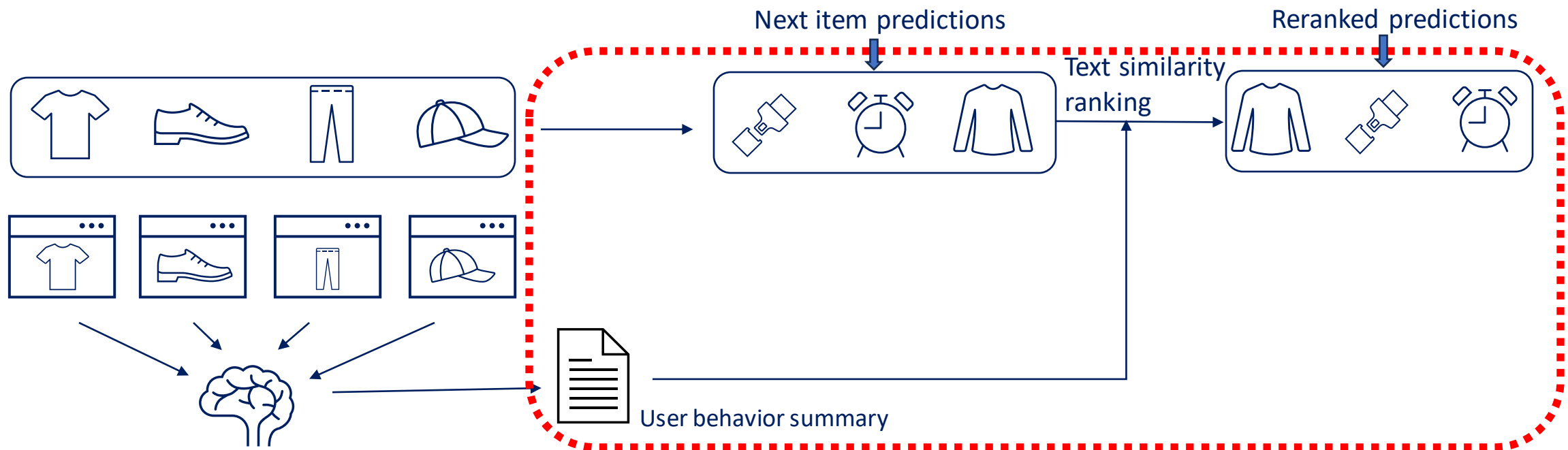
- Screenshot generation and Behavior summarization



Using InteraRec with a Session based Recommendation Model

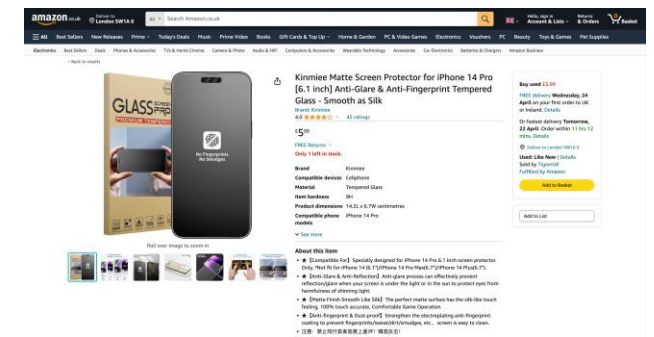
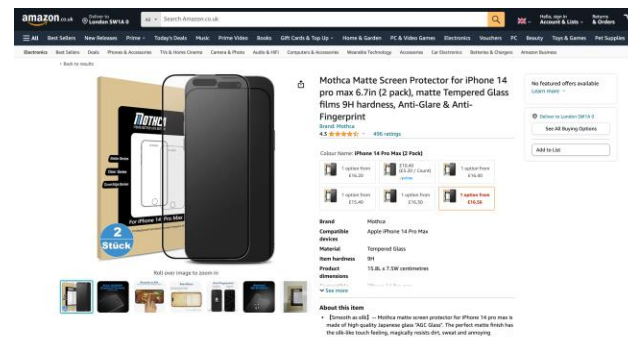
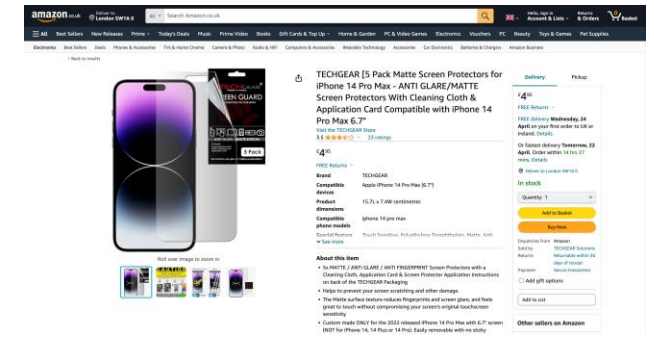
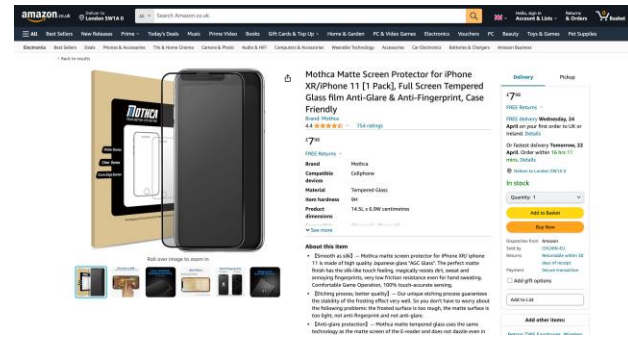
Response generation

- Re-rank the predictions based on the similarity of embeddings derived from product details (title, brand, price) and user behavior summary inferred from their interaction history.



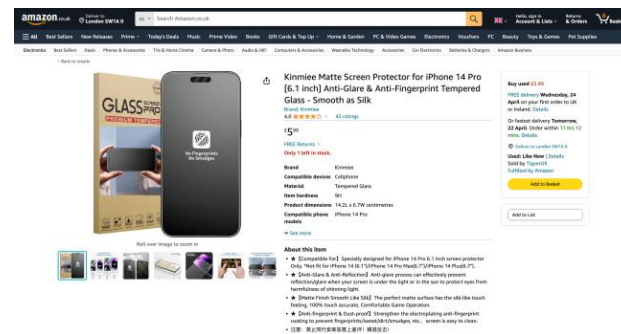
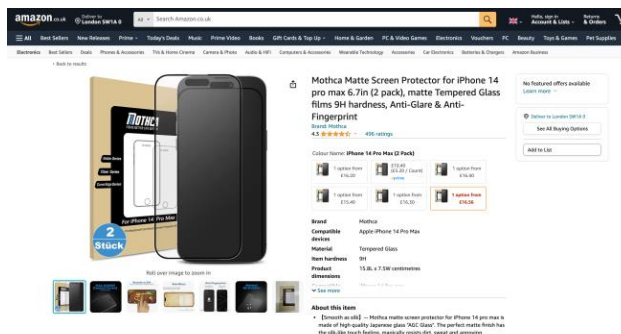
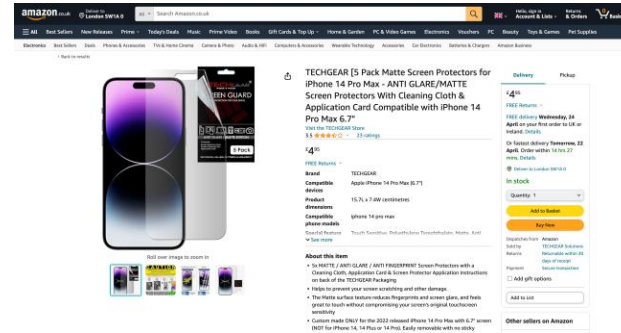
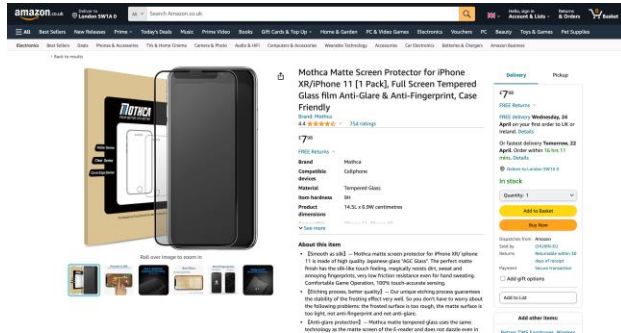
Multimodal Dataset from User Browsing Sessions on the Amazon.UK Website

▶ We construct a multimodal dataset by capturing screenshots and generating summaries of user browsing sessions from the Amazon-M2 dataset[5], leveraging the *InteraRec* framework.



Screenshots depicting a user's browsing session on Amazon.com, focused on searching for and exploring screen protector options from newly constructed multimodal dataset

Illustrative Example of Screenshot Generation and Behavior Summarization

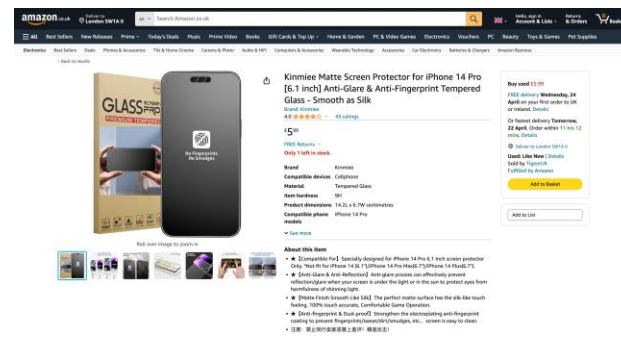
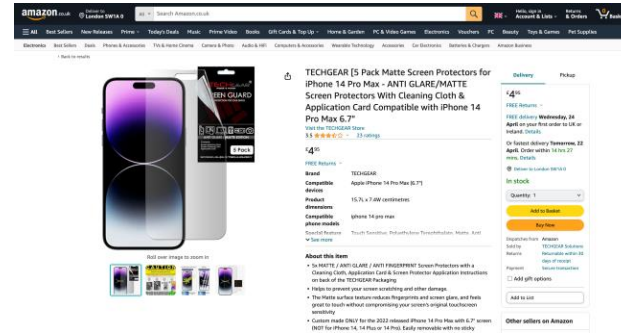
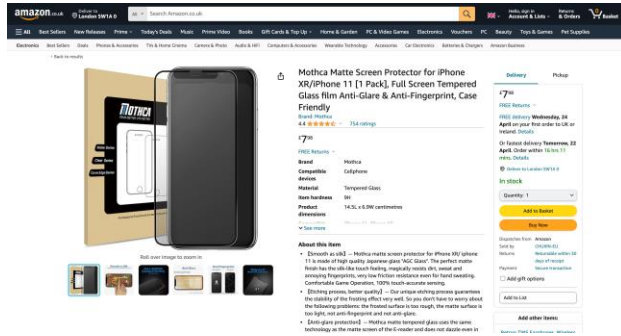


The images provided are screenshots from an online marketplace showcasing different screen protectors for smartphones. Based on what is shown, here is a summary response that addresses the categories of user preference:

- Product Characteristics:** The products are screen protectors for iPhone 14 Pro and Pro Max models. They emphasize features such as anti-glare, matte finish, and resistance to fingerprints.
- Lowest Price:** The lowest-priced item displayed is priced at £4.95, providing a budget-friendly option for users.
- Highest Price:** The highest-priced option is shown as £16.20, indicating a range of price points for the same category of product.
- Brand Preference:** The brands seen in the images include TECHGEAR, Mothca, and KINMIEE, suggesting the user is open to considering various brands.
- Product Specifications:** The product listings detail specifications such as compatibility with certain phone models, dimensions, and materials (e.g., tempered glass with a hardness rating of 9H).
- User Reviews and Testimonials:** The products have customer ratings displayed, ranging from 3.5 to 4.4 stars, indicating that user reviews are accessible and may influence the decision-making process.
- Comparisons:** Viewing multiple products allows for comparison based on features, price, brand, and reviews.
- Promotions:** There are no explicit promotions or discounts indicated in the images provided.

The InteraRec framework generated a summary of user preferences from a browsing session focused on screen protectors, indicating the user is searching for products with specific features like anti-glare and matte finishes, within a price range of \$4.95 to \$16.20, without showing a preference for any particular brand.

Illustrative Example of Screenshot Generation and Behavior Summarization



The images provided are screenshots from an online marketplace showcasing different screen protectors for smartphones. Based on what is shown, here is a summary response that addresses the categories of user preference:

- Product Characteristics:** The products are screen protectors for iPhone 14 Pro and Pro Max models. They emphasize features such as **anti-glare, matte finish, and resistance to fingerprints.**
- Lowest Price:** The lowest-priced item displayed is priced at **£4.95**, providing a budget-friendly option for users.
- Highest Price:** The highest-priced option is shown as **£16.20**, indicating a range of price points for the same category of product.
- Brand Preference:** The brands seen in the images include **TECHGEAR, Mothca, and KINMIEE**, suggesting the user is open to considering various brands.
- Product Specifications:** The product listings detail specifications such as compatibility with certain phone models, dimensions, and materials (e.g., tempered glass with a hardness rating of 9H).
- User Reviews and Testimonials:** The products have customer ratings displayed, ranging from **3.5 to 4.4 stars**, indicating that user reviews are accessible and may influence the decision-making process.
- Comparisons:** Viewing multiple products allows for comparison based on features, price, brand, and reviews.
- Promotions:** There are no explicit promotions or discounts indicated in the images provided.

The InteraRec framework generated a summary of user preferences from a browsing session focused on screen protectors, indicating the user is searching for products with specific features like anti-glare and matte finishes, within a price range of \$4.95 to \$16.20, without showing a preference for any particular brand.

(Very) Preliminary Experimental Results

	Recall@100	MRR@100	Recall@50	MRR@50
GRU4Rec	0.1667	0.007	0.0833	0.005

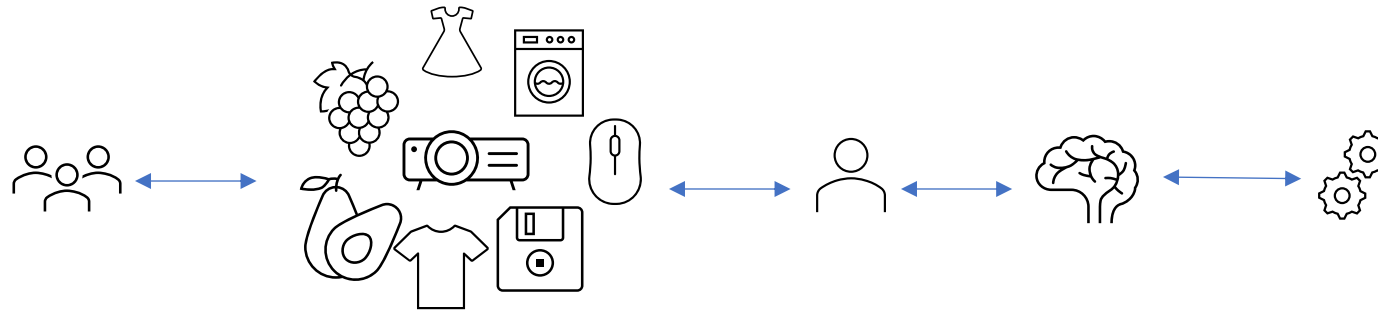


	Recall@50	MRR@50
<i>InteraRec + GRU4Rec</i>	0.1083	0.057

Performance of the recommendation model on our multimodal dataset. Higher is better.

Summary

1. A work in progress
 - Evaluating SOTA models with *InteraRec* (input modification and output re-ranking)
 - Ablation study on screenshot frequency, resolution, and MLLM token limits.
2. We discussed a new design for recommendation systems
3. Uses a non-standard sensing modality and pre-trained foundation Multimodal LLMs
4. Enables a new tech stack comparable to existing solutions, with advantages such as interpretability and use of in-built knowledge in MLLMs.



THANK YOU

