

Project Report for Cinepaw AI



Intelligent Reasoning Systems
(ISY5001)

Group Name

Pawssible

Team Members

Tan Eng Hui	(A0291201W)
Wang Tao	(A0291189R)
Lois Chee Li Ping	(A0292098R)
Dong Yuantong	(A0292041N)

CONTENTS

1. Executive Summary.....	4
2. Project Background.....	6
2.1. Business Case.....	6
2.2. Market Research	6
2.3. Value Proposition	7
2.4. Business Model.....	8
2.5. Pricing Strategy	8
3. System Design and Model	9
3.1. Recommendation Model.....	10
3.2. Stateful Interaction with Generative AI	13
3.3. Cognitive Speech.....	14
4. System Development and Implementation	16
4.1. Frontend Implementation	16
4.2. Backend Implementation	18
5. Performance and Validation	20
6. Challenges and Overcoming It	21
6.1. Time constraint	21
6.2. Bridging Experience Gaps.....	21
6.3. Integration Complication	21
7. Risks/Concerns	22
7.1. Data Source and relevancy.....	22
7.2. AI Ethnic	22
8. Future Roadmap	23
8.1. Enhance Data Sources.....	23
8.2. More Complete Ecosystem	23
8.3. Advanced Recommendation Technologies.....	23
9. Conclusion.....	24
Appendix	25
A. Installation/Deployment Guide.....	26
B. User Guide.....	28

C.	Mapped Module/System Functionalities	33
D.	Individual Project Report	34
E.	Project Proposal	39
F.	Reference.....	47

1. Executive Summary

Problem Statement and Proposed Solution

The current streaming platforms recommender systems often fail to capture individual preferences and lack personalisation elements in providing movie recommendation. This may lead to loss of users over a period of time once they feel frustrated and hop to another platform that give them better user experience to find the movies that interest them.

Thus, our proposal - CinePaw AI introduces an advanced movie recommendation system that leverages Large Language Models (LLMs) to create an intelligent, interactive chatbot to retain the users and make them royal to our platforms. This chatbot goes beyond traditional recommendation algorithms by engaging in natural language dialogues with users, enabling a personalized movie discovery experience and improve the overall user experience in sourcing for movies that they are interested and stay royal to the platform.

Differentiation from Existing Markets

CinePaw AI differentiates itself from existing market offerings such as

- its recommender model (ComiRec), which provides nuanced recommendations by categorising user interests into multiple segments.
- chatbot that remembers and reference past user interactions. This stateful interaction enhances the personalization and relevance of recommendations.
- cognitive services, allowing users to engage in human-like conversations with the system, receiving both spoken and textual responses, setting it apart from the conventional, less interactive recommendation services found in current streaming platforms.

Market Positioning

The streaming service market is highly competitive and continuously evolving with viewers seeking more personalized and engaging experiences. We position CinePaw AI to be the forefront in shifting towards the direction what we define as Interactive User-Centric Recommendation System (IUCRS). It fills a significant gap left by existing platforms, offering unique experience that emphasises on user interaction, engagement, and satisfaction.

Conclusion and Future Prospects

Moving forward, CinePaw AI plans to expand its data sources and enhance its ecosystem to maintain relevance and adapt to changing market demands. The project team is also aware of potential challenges, such as data source limitations and the need for continuous updates to the AI algorithms to handle diverse user interactions. These enhancements will be crucial for scaling operations and improving the system's accuracy and more precise user engagement.

2. Project Background

The market for streaming services is highly competitive and continuously evolving, to satisfy viewers increasingly demanding more personalised and interactive experiences. Traditional platforms typically utilise algorithms that base recommendations on users' past viewing habits, often resulting in generic suggestions that fail to meet the unique preferences of individual viewers. This gap in the market presents a significant opportunity for innovative solutions like CinePaw AI.

2.1. Business Case

Research indicates that many streaming service users prefer recommendations that are closely tailored to their individual tastes and viewing habits. As the volume of available content on streaming platforms grows, users are finding it increasingly challenging to discover content that genuinely interests them. This has led to a rising demand for more sophisticated recommendation systems that can offer personalized viewing suggestions.

The integration of Artificial Intelligence (AI) and Machine Learning (ML) in consumer applications has set new expectations for personalized technology experiences. Large Language Models (LLMs), which enable more natural and effective user interactions, are at the forefront of this technological wave. CinePaw AI leverages these advancements to provide a service that not only responds to user preferences but also engages them in meaningful conversations to continually refine those preferences.

2.2. Market Research

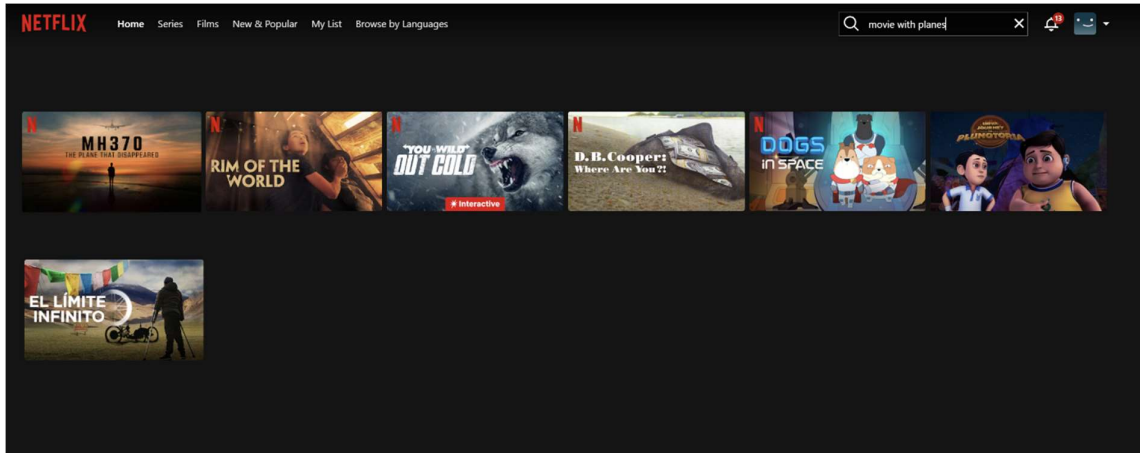
Video streaming market size was valued at USD 455.45 billion in 2022, for entertainment market it will hit 2.7 trillion. Large entertainment companies like netflix, hulu, youtube invest heavily in recommend system as it is directly related to their business operation to retain and attract their users.

While the dominant players have developed sophisticated recommendation algorithms, these often lack the interactive component that modern viewers crave. CinePaw AI's interactive chatbot differentiates itself by offering a more engaging and responsive user experience. Emerging platforms are beginning to explore similar interactive features, which suggests a shift towards more engagement-focused recommendation systems in the industry.

CinePaw AI is well-positioned to capitalise on the trend towards personalization and interaction. By offering a chatbot that dynamically adapts to user inputs and preferences, it not only meets the current market demand but also sets a new standard for user engagement

in such services. This approach not only enhances user satisfaction and retention but also positions CinePaw AI as a pioneer in the next generation of recommendation technology.

For example, Netflix search bar has limited capabilities in searching for movies. When wanting to find the movie 'Topgun' which involves fighter jets, the text input 'movie with planes' is unable to return the movie. Interpreting vague and unclear descriptions are a key limitation in the current market.



2.3. Value Proposition

We target CinePaw AI to represent a leap towards transformative engagement, turning viewers into royal users with services that are closer like human assistant through the implementation of cognitive features and Natural Language Understanding (NLU) to fulfill the requirements in order to build an Interactive User-Centric Recommendation System (IUCRS) product.

User benefits:

- Better and happier using the system as it provides personalised search experience
- Reduce frustration (no need to describe exact keyword, phrase)
- Ability to understand nuanced preferences reduce time spent on getting good recommendations
- Save time and effort
- Convenience to users when performing search (using natural language)

Business/Commercial Benefits:

- Better understanding and more engaged users increase conversion rate
- Enhanced users' satisfaction and loyalty through personalised services
- System can smartly suggest content other than movie leading to potential purchases

2.4. Business Model

We would primarily operate under a business-to-business (B2B) model, targeting existing streaming platforms, media companies, and other entities in the entertainment industry.

We aim to provide these businesses with an advanced solution to upgrade their existing systems or augment their current offerings.

By integrating our AI-driven recommendation engine, these companies can enhance user engagement, personalize user experiences, and ultimately, increase viewer satisfaction and retention.

Our service can be offered as a standalone API or integrated as part of a larger suite of services, depending on the client's needs, providing flexibility and scalability to accommodate different business sizes and requirements.

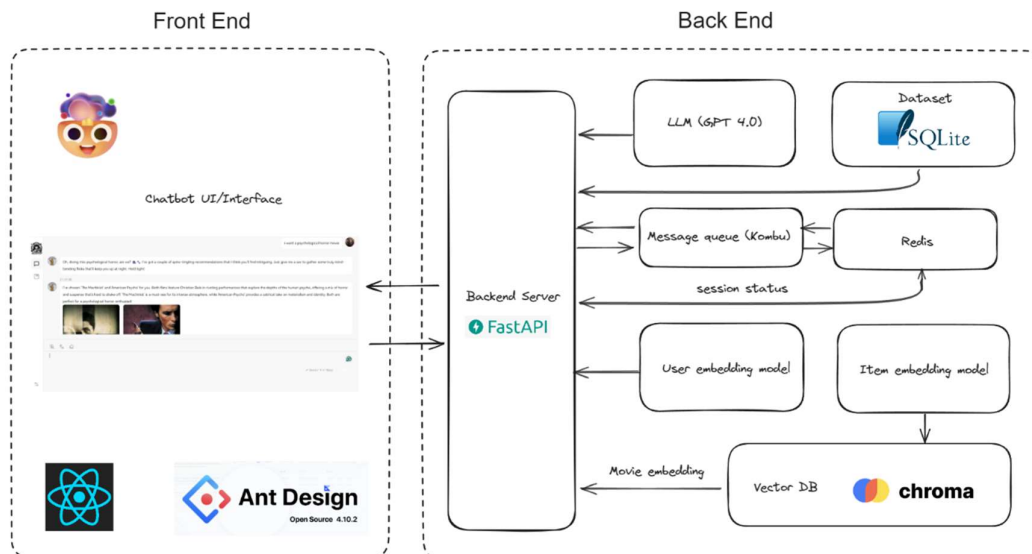
2.5. Pricing Strategy

Considering the initial high setup costs of such a system, we plan to offer a collaborative Proof of Concept (POC) at minimal cost. This approach aims to attract potential clients by demonstrating the tangible benefits and potential ROI of integrating our solution into their existing offering.

Then we would implement a tiered pricing strategy based on usage volume and the level of customisation required. This could include monthly or annual subscription models, which would provide ongoing support and updates, ensuring clients continually benefit from the latest improvements in AI technology and recommendation algorithms.

This pricing model not only facilitates scalability and adaptability to different client needs but also establishes a long-term partnership framework conducive to mutual growth and success.

3. System Design and Model



CinePaw AI is a chatbot based on web application technologies that consuming services both from local (e.g. recommender systems) and remote services provided by various service providers (e.g. Azure speech, TMDB API).

React and Next.js have been selected as the foundational framework for our interactive web application, providing a robust and scalable architecture. The chatbot interface has been meticulously crafted utilizing Ant Design and Lobe-ui, leveraging their customizable features to facilitate dynamic interactions with users.

To ensure seamless communication with the backend, we have integrated SWR and Axios for efficient data fetching and state management, enhancing the overall responsiveness of our application.

In the backend, FastAPI serves as the framework of choice, distinguished for its rapid performance and comprehensive API capabilities. Its ability to facilitate real-time interactions and ensure responsiveness aligns seamlessly with the demands of our project.

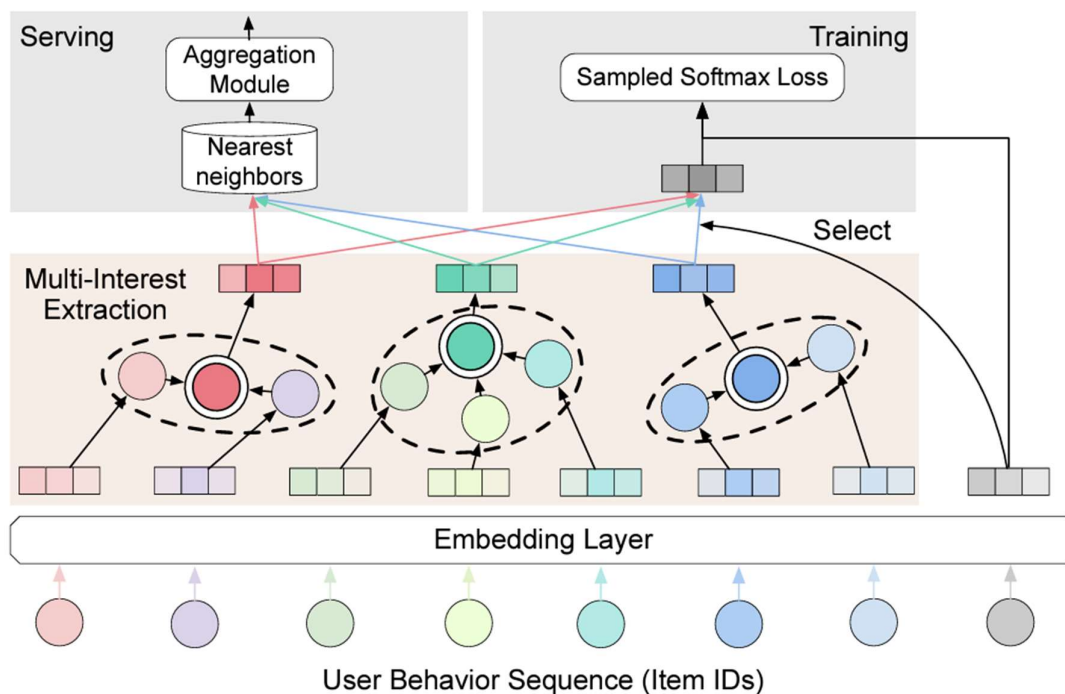
The subsections will highlight and explain the core technologies deployed.

3.1. Recommendation Model

ComiRec is an abbreviation derived from **CO**ntrollable **MU**lti-Interest model for sequential **RE**commendation. This model is adept at capturing users' diverse interests from their behaviour sequences and aim to predict what might be the next items the user would be interested in. Since this multi-interest module can capture the multiple interests of users, which can be exploited for retrieving candidate items from a large dataset. Upon receiving enough data on the user's behaviour, the model is able to categorize the data into different sections for training and serving. Afterwards, the results undergo further refinement in our aggregation module, where a unique controllable factor strikes the perfect balance between recommendation accuracy and diversity. This model is special as it is able to combine items from different interests and outputs the overall recommendation.

From the architecture perspective, the two reasons why this model was chosen is due to the ability to perform sequentially recommendation and real time user embedding.

Here is an illustration of ComiRec's model architecture:



Reference: [\[2005.09347\] Controllable Multi-Interest Framework for Recommendation \(arxiv.org\)](#)

Sequential Recommend

ComiRec is a sequential recommendation model, which means it takes user's interactive history as a sequence and tries to predict the next item user will interact with. In conversation, it is more natural to treat the history as a sequence than treating all history movies equally. Keeping the sequential information will be important to predict.

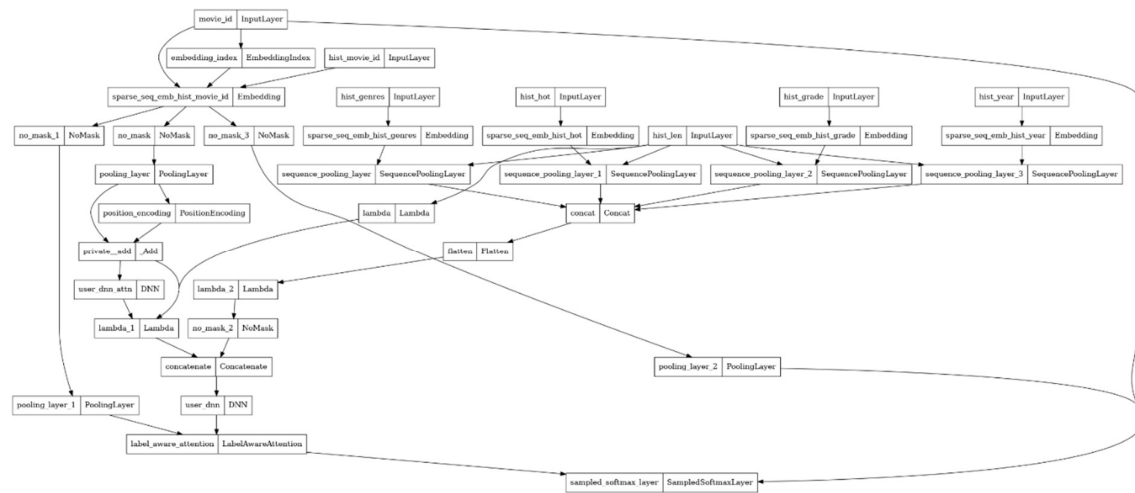


Realtime Recommendation

During conversational interaction, it is important for the system to react to user's input and change the recommendation result based on what user just said and their preference.

Our approach is to deliver the dynamic ranking of the search results by calculating the item embedding in offline, and calculating user embedding online in real-time, then match the user with items to get the latest recommendation. We identified ComiRec model fit well for this task, and it is easy to be implemented into our process flow.

Model Implement Detail



Performance

There is the performance test result from Controllable Multi-Interest Framework for Recommendation as reference, from this test we could also see that the ComiRec have advantages on comparing with popular algorithms like YouTubeDNN.

This is the result from tests we did on movielen-1m dataset:

	Hit Rate@50	Hit Rate@20	NDCG@50	NDCG@20
ComiRec-SA-K_2	25.21%	12.86%	15.52%	14.29%
MIND	24.55%	11.60%	14.80%	13.88%
YoutubeDNN	22.35%	11.47%	13.30%	12.29%

This is the result from ComiRec paper:

	Amazon Books						Taobao					
	Metrics@20			Metrics@50			Metrics@20			Metrics@50		
	Recall	NDCG	Hit Rate	Recall	NDCG	Hit Rate	Recall	NDCG	Hit Rate	Recall	NDCG	Hit Rate
MostPopular	1.368	2.259	3.020	2.400	3.936	5.226	0.395	2.065	5.424	0.735	3.603	9.309
YouTube DNN	4.567	7.670	10.285	7.312	12.075	15.894	4.205	14.511	28.785	6.172	20.248	39.108
GRU4Rec	4.057	6.803	8.945	6.501	10.369	13.666	5.884	22.095	35.745	8.494	29.396	46.068
MIND	4.862	7.933	10.618	7.638	12.230	16.145	6.281	20.394	38.119	8.155	25.069	45.846
ComiRec-SA	5.489	8.991	11.402	8.467	13.563	17.202	6.900	24.682	41.549	9.462	31.278	51.064
ComiRec-DR	5.311	9.185	12.005	8.106	13.520	17.583	6.890	24.007	41.746	9.818	31.365	52.418

Reference: [\[2005.09347\] Controllable Multi-Interest Framework for Recommendation \(arxiv.org\)](#)

3.2. Stateful Interaction with Generative AI

We attempt to enhance user experience by making interactions with Large Language Model (LLM) Generative AI model like ChatGPT more coherent and contextually aware over time, rather than treating each interaction session in isolation. To achieve the above, the following techniques were used to create longer persistent memory feature:

- **Vector embeddings:** When storing chat logs, each input or conversation can be transformed into a high-dimensional vector using techniques from natural language processing (NLP). These vectors represent the semantic content of the texts.
- **Vector database:** These vectors are stored in a specialised type of database known as a vector database, which is optimised for handling high-dimensional data (e.g. ChromaDB). This database allows for efficient querying and retrieval of vectors that are similar in context or content for subsequent chat queries.
- **Retrieval for continuity:** For future chat interactions, the system would be able to query the vector database to retrieve past conversations that are contextually relevant to the new query. The past chats would be attached as context supplementing the new query enabling the LLM (e.g. ChatGPT 4) to remembering past interactions and adjusting its responses accordingly.

This approach significantly enhances the user experience by creating a sense of continuity in interactions, making the users feeling that the chat is more personalised and relevant. It also improves the overall efficiency and accuracy of the AI's responses, as it can draw upon past interactions to better understand user preferences and context.

3.3. Cognitive Speech

We have also integrated cognitive features like Speech-to-Text (STT) and Text-to-Speech (TTS) to significantly enhance user interaction and overall experience. These features streamline how users engage with the CinePaw AI chatbot, making the service more accessible and interactive.

- **Speech-to-Text (STT):** Allows interaction with the chatbot through voice commands, which is particularly beneficial for users who may find typing cumbersome or are visually impaired. This feature ensures that the chatbot is more inclusive and can serve a wider audience. It also enhances convenience, as users can easily ask for movie recommendations or discuss their preferences without the need to type their queries manually.
- **Text-to-Speech (TTS):** Conversely, TTS technology converts text data into spoken audio. In CinePaw AI, this feature reads out the chatbot's responses aloud, which not only aids users who are visually impaired but also enriches the interactive experience for all users. Hearing the chatbot's recommendations spoken aloud can make the interaction feel more natural and engaging, similar to speaking with a human assistant.

To further enrich the user experience and cater to diverse preferences, we incorporated multiple voice options. Users can choose from a variety of voices, each with distinct characteristics and accents. We have used 'Polar Prince' as the default voice option setting. Other voices included are 'Sun Bear Soprano', 'Grizzly Groove' and 'Panda Pitch'.

- **Polar Prince:**

Inspired by the icy landscapes of the Arctic, the Polar Prince voice setting exudes an air of regal elegance. Its tone is crisp and clear, with a hint of frosty refinement.

- **Sun Bear Soprano:**

Imagine the warm rays of the sun shining down on a lush forest, and you will capture the essence of the Sun Bear Soprano setting. It carries a bright and cheerful tone.

- **Grizzly Groove:**

With the soulful vibe of a rhythm-filled forest, the Grizzly Groove voice setting brings depth and richness to every word spoken. Its tone is warm and resonant.

- **Panda Pitch:**

The Panda Pitch voice setting is characterized by its gentle and soft touch, much like the soothing presence of a panda lounging in bamboo forests. This setting lends a cozy charm to spoken words, like a comforting hug from nature itself.

Bear Config ×

Voice

☒ Polar Prince ☐ Sun Bear Soprano ☐ Grizzly Groove ☐ Panda Pitch

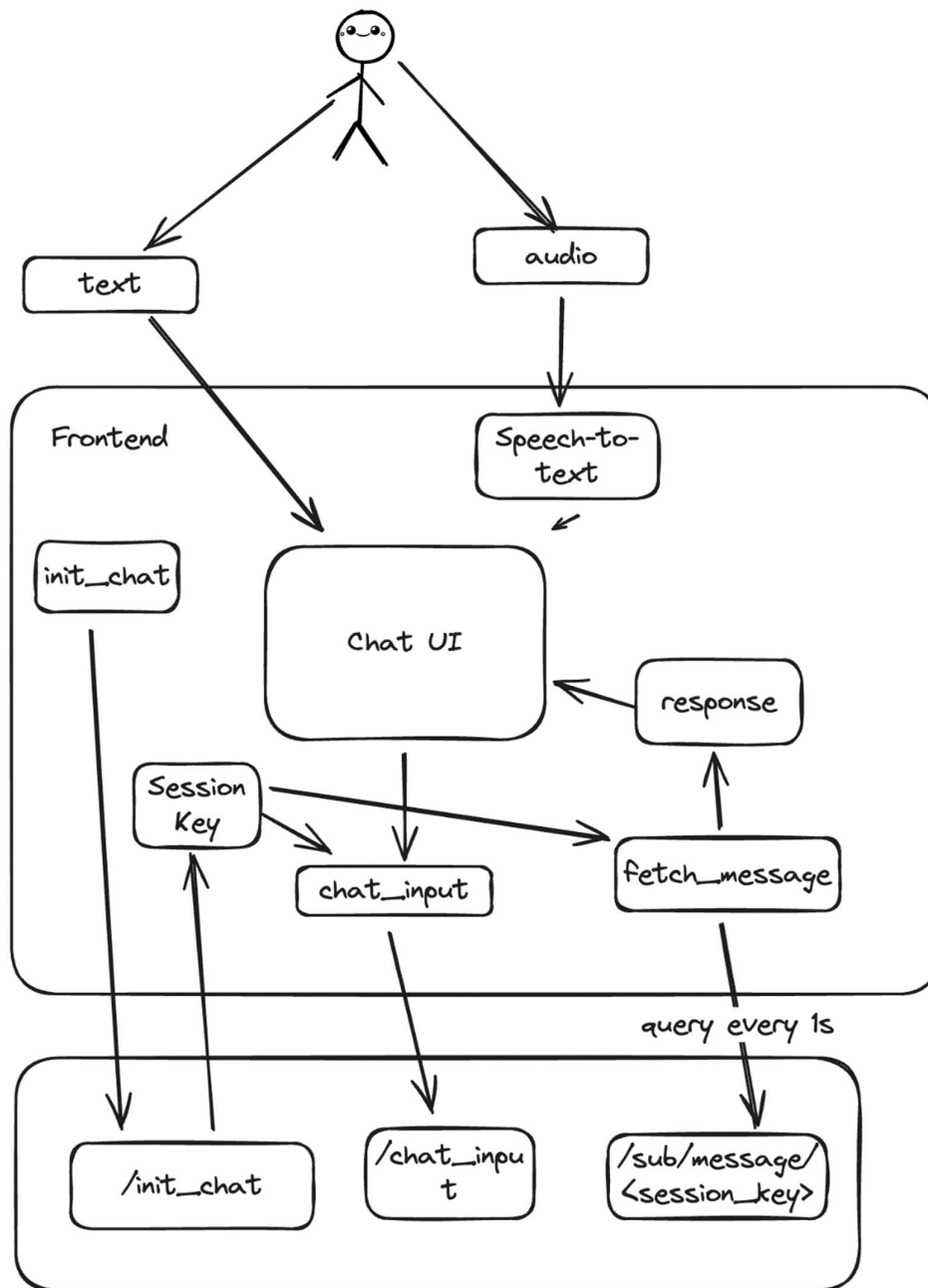
Together, these cognitive features make the CinePaw AI more user-friendly and accessible, mimicking a natural conversation flow and catering to personal user preferences. This not only improves user satisfaction but also increases the usability of the system across different user demographics, ultimately broadening the scope and appeal of the project.

4. System Development and Implementation

4.1. Frontend Implementation

This section shows the workflow implemented for the frontend, on how the text and audio are processed and communicated to backend to generate necessary response back to the user.

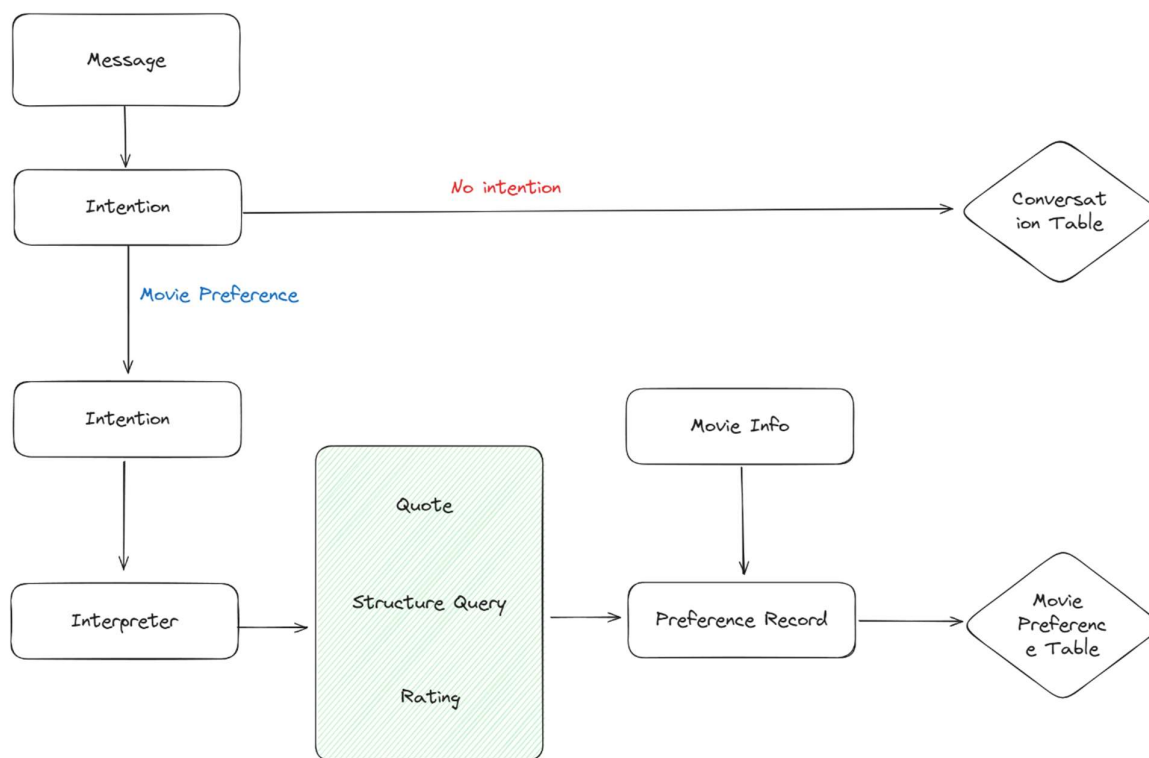
Here is the user flow in frontend:



When user star a new conversation, the frontend will first start a request to “init_chat”, to get a session_key and store it, the session_key will be the unique key for this conversation.

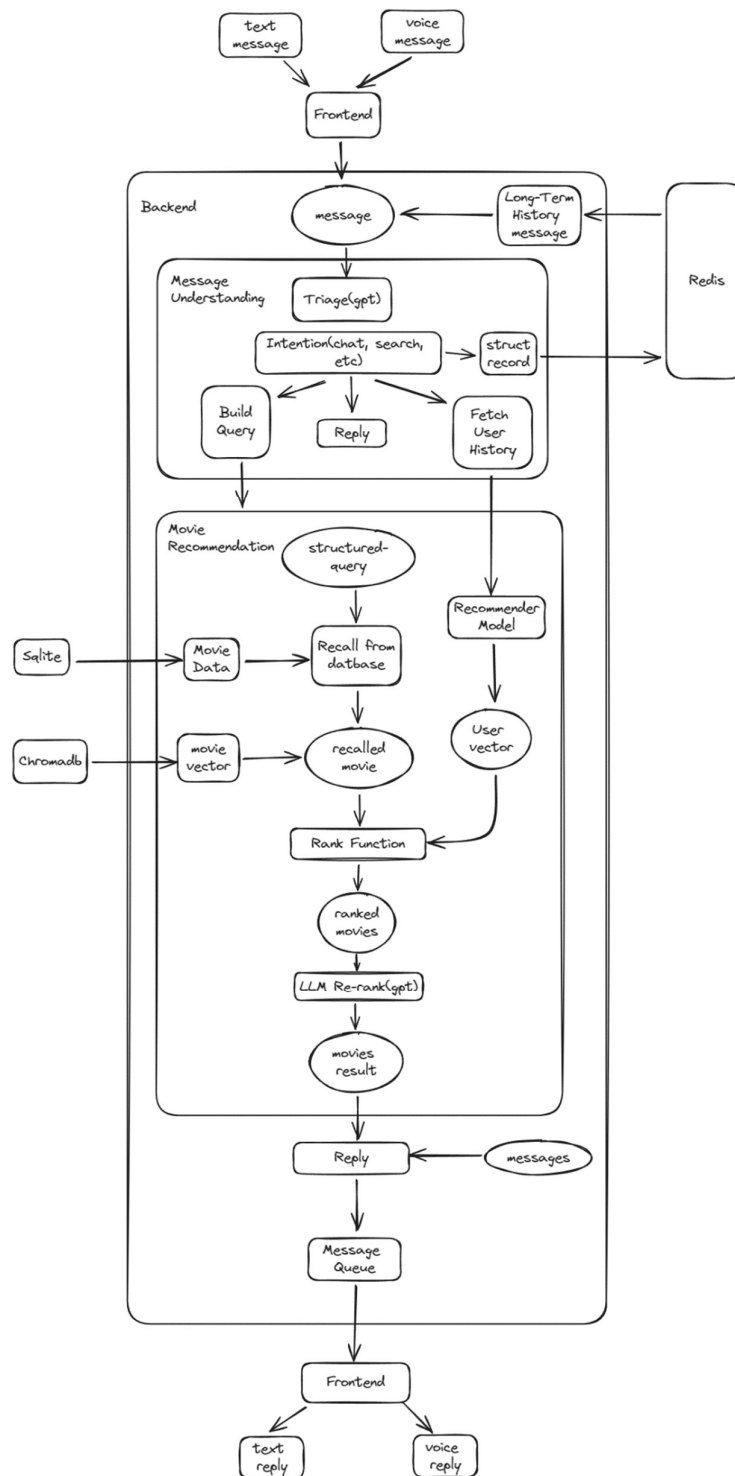
When user input text or use audio input, they will be transformed into text and merge with history and display on Chat UI. Then frontend will send the new input and history to backend “chat_input” with session_key, the frontend will only have a submit success result from chat_input instead of waiting for the response.

For getting bot’s response, the frontend will keep querying /sub/message/<session_key>, all backend response will be sent to the message query and can be consumed by this api, once frontend get non-empty result from this api, it will be appended to chat history and display on Chat UI.



4.2. Backend Implementation

Here is a breakdown for message processing in backend:



Backend consists of two main parts: interpreting the user's message and the recommendation process. More details are elaborated below:

Understanding the Message

When backend service receives user's input, it will fetch user's history conversation and data, they will be all composed into a prompt and send to LLM (gpt-4-turbo) to triage to get a struct result for user's input.

There will be intentions key that represent user's intent and decide what next the system will do.

Based on the message intention, the data structuring will be different. If the intention is detected as conversational, it will be considered as 'normal_chat' and the result will reply directly to the user. If intention is detected to consist of a search or recommend task, the structured result will be queried with "title", "genres," "tags" for searching.

If intention trigger a long-time task, it will send a reply to message queue to tell user to hold on and the system is working on this. Then it will add task to run background like parse user's preference from text, save data into database, etc.

Subsequently, the task will be run in background to parse user's preference from text.

Recommendation process

With the struct-query and user's movie history from message understanding model, the system can start the recommend process.

First the system will send struct query to recall function, recall function will recall all movies fit the query from database, it is kind like use sql(struct query) to fetch data from data table.

With all recalled movies, we will fetch pre-calculated movie_vectors for them from the chromadb, and we will calculate user_vector by using recommend model(user_embed part) in real-time. with user_vector and movie_vectors, we could calculate every movies' distance to user, and rank all movie by this.

With all ranked movies, we get top-20 movies as candidates, put these movies and user's information (messages and movie preference) into one prompt, then the system sends the prompt to LLM, to re-rank and generate the final top 5 movies recommendation for the user and the reasons why we recommend these.

5. Performance and Validation

To conclude that objectives are met, this section will detail the definition of the success criteria and how we measure in various aspects.

Technical Features/Systems	Observation/Results
Requests are queued and are processed	Yes
Speech is captured and converted accurately	Yes
Text to speech work accordingly	Yes
Persistent Memory is working	Yes (system is able to recall recent history)
Chatbot understand the intent of users when they utter movie and non-movie	Yes (system is able to differentiate the intent and trigger respective workflow)
System is able to do NLU with LLM	Yes (chatgpt 3.5 does not perform well, but GPT 4.0 is able to NLU with context provided)

We also assign internal users to do ad hoc testing on the system ensure that the chatbot and recommendations response and give relevant results. Following are some examples:

Chats [Context]	Observation/Results
[Favourite singer is Taylor Swift] Has my idol acted in any movie?	System recalled Taylor Swift and responded on the movies recommendation
By the way, do you know the guy in conjuring has any DC superhero movies? I think there is but can't remember, can you find some for me?	System responded with " Patrick Wilson " and the movie " Aquaman " he acted in.
[speech input] Give me some horror movie that kind of like footage	System captured the speech to text by the user and responded with " Grave Encounters " and " The sacrament " which are both horror movies in found footage styles
[speech input, found footage style] Have you heard of "Blair the witch"?	System responded with the movie description and associate it to found footage style.
[horror movie profile] I like the main character in the Witch, can you find me some movies played her?	System responded with the main character description, actress name and some of the horror movies she acted in.

As of any artificial intelligence (AI) project, we could only conclude that the initial solution could fulfil our basic criteria, it would still take iterations and expanding the user base to actual clients to get the better sense on how good the product is.

6. Challenges and Overcoming It

6.1. Time constraint

Time constraint is the main key issue as all team members are working full-time and the project require integration of multiple technologies and services.

Thus, establishing clear communication and regular meeting to keep every team member informed and engaged are utmost important. It is required for us to spot potential issues early, allowing for swift resolutions and to maintain momentum toward our project deadlines. Through strong teamwork and a united effort, we have successfully navigated these challenges and achieve our project objectives.

6.2. Bridging Experience Gaps

As not all team members initially possess knowledge and/or hands-on experience with AI solutioning, particularly in software development. In response, we have cultivated a collaborative environment where knowledge sharing/transfer and mentorship are prioritised. More experienced members guide those with less familiarity with AI or application development, helping to fill gaps and ensure no aspect of the project is overlooked.

For the members who lack technical experience, they heavily contribute with their insights and perspectives from their past work experience in their specific domain.

This has allowed us to utilise our unique strength to improve the collective problem-solving capabilities to resolve issues and meet the project goals.

6.3. Integration Complication

We employ diverse technologies such as Large Language Models (LLMs), ComiRec recommendation model, and list of comprehensive full-stack web application architecture (using React and FastAPI). The primary challenge for this part was ensuring seamless communication and functionality between the stateful interactions of the LLMs, the dynamic recommendation logic of ComiRec, and the real-time user interface updates on the frontend.

To address these integration complexities, the team have to implemented robust APIs to facilitate efficient data exchange and used containerisation technologies like Docker for consistent deployment environments. Additionally, extensive testing needs to be conducted to ensure that all components worked harmoniously, providing a smooth and responsive user experience.

These strategies helped overcome the integration hurdles, allowing us to deliver the final application without issue.

7. Risks/Concerns

This section highlights the risks and concerns we ought to consider should this MVP need to advance to the next phase or release.

7.1. Data Source and relevancy

MovieLens dataset might not be entirely representative of the general population's movie preferences. Users who contribute to MovieLens might have different tastes or demographics compared to the broader population. Be cautious when drawing conclusions or making predictions based solely on this dataset.

7.2. AI Ethics

Addressing AI ethics in the CinePaw AI project involves critical considerations around user data privacy and algorithmic transparency. Risks such as data misuse and biased recommendations require stringent measures including robust data encryption, explicit user consent protocols, and regular audits to ensure compliance with privacy laws like Personal Data Protection Act (PDPA).

To address this concern, we should implement features that clearly explain to users how their data is utilized and the ways in which it impacts the AI's recommendations. This approach not only ensures compliance with PDPA regulations but also enhances user trust and satisfaction by promoting transparency and understanding.

8. Future Roadmap

8.1. Enhance Data Sources

To continue improving the accuracy and relevance of our recommendations, CinePaw AI plans to diversify and update its data sources regularly. By integrating a broader range of databases, including newer and more varied content from global cinema and user-generated content platforms, the system will be better equipped to understand and predict emerging viewer preferences. We will also implement real-time data streaming, allowing the AI to incorporate current trends and feedback into the recommendation process. This will ensure that the suggestions remain fresh and in sync with current viewing trends.

8.2. More Complete Ecosystem

CinePaw AI aims to create a more interconnected and comprehensive ecosystem by integrating with additional platforms and services that extend beyond movie recommendations. This includes partnerships with streaming platforms, social media networks for sharing recommendations, and possibly e-commerce links for movie merchandise. Enhancing connectivity with these platforms will not only broaden the user base but also enrich the data pool with more diverse user interactions, further refining our AI's learning and prediction capabilities.

8.3. Advanced Recommendation Technologies.

CinePaw AI has implemented relatively new experimental methods such as combining LLM and re-ranker in recommender systems. we have also adopted the industry's more mature online-offline model separation pattern in recommender system implementation. However, in the selection of models, although the selected ComiRec architecture is also a relatively new model that has achieved practical results, there may be other model such as graph recommendation models, Bert4Rec, etc offering certain niche. In sequence recommendation, there are also newer results such as CL4SRec etc. that we have yet to test out due to time and effort required which may yield better results.

9. Conclusion

CinePaw AI has effectively demonstrated its capability to enhance user engagement by making each interaction more coherent and contextually aware. This system leverages vector embeddings and a specialized vector database, enabling it to recall and utilise past interactions, thus delivering a highly personalized and efficient user experience.

As we look to the future, the expansion of our data sources will ensure that the system's responses remain relevant and up-to-date, directly enhancing business operations by maintaining high engagement levels. Further, by building a more comprehensive ecosystem, we aim to integrate our AI more seamlessly with existing business processes and systems, increasing the overall utility and effectiveness of our technology.

These enhancements are crucial for scaling our operations and improving our service offerings. They will allow us to maintain a competitive edge in the market by continually adapting to the evolving needs of our users.

Overall, this MVP has not only met our initial goals but also laid a solid foundation for significant growth and continued improvement in delivering personalised user experiences.

Appendix

A. Installation/Deployment Guide

Pre-requisite:

- Git clone from <https://github.com/eng-hui/IRS-PM-2024-01-13-IS06PT-GRP-Cinepaw-AI> . **Please take note that you will need to obtain the follow API keys from the listed resources :**
 - <https://platform.openai.com/>
 - <https://www.themoviedb.org/>
 - <https://azure.microsoft.com/en-us/products/ai-services/ai-speech>
- Docker
- Internet Access
- Chrome Browser

Instruction for Environmental variables and API access tokens

1. Modify the **.env** file in the **backend** folder and update the following keys with your own keys (refer to pre-requisite)

Key	Default Value/Source
DEFAULT_LLM_ENDPOINT	openai or azure_openai
DEFAULT_MODEL	gpt-4-turbo-preview or other Model Name accordingly
OPENAI_API_ENDPOINT	# can be empty if not azure openai
OPENAI_API_KEY	[According to the OpenAI service]
AZURESPEECH_API_KEY	[According to the Azure AI speech service]
TMDB_API	[API Key,According to the TMDB API service]

Format reference: <https://www.dotenv.org/docs/security/env.html>

[illegible]

Instruction for Docker deployment

2. Launch terminal from the **SystemCode** folder
3. Run **docker build . -f docker/dockerfile -t=cinepaw** to build a cinepaw docker image

```
PS D:\azure\IRS-PM-2024-01-13-IS06PT-GRP-Cinepaw-AI\SystemCode> docker build . -f docker/dockerfile -t=cinepaw
[+] Building 6.4s (34/34) FINISHED
=> [internal] load build definition from dockerfile
=> => transferring dockerfile: 3.00kB
=> [internal] load metadata for docker.io/pytorch/pytorch:2.1.0-cuda11.8-cudnn8-devel
=> [internal] load .dockerignore
=> => transferring context: 2B
=> [ 1/29] FROM docker.io/pytorch/pytorch:2.1.0-cuda11.8-cudnn8-devel@sha256:558b78b9a624969d54af2f13bf03fbad279
=> [internal] load build context
=> => transferring context: 135.57MB
```

4. Run cinepaw image by **docker run -p 8111:9880 --name cinepaw cinepaw**

```
PS D:\azure\IRS-PM-2024-01-13-IS06PT-GRP-Cinepaw-AI\SystemCode> docker run -p 8111:9880 --name cinepaw cinepaw

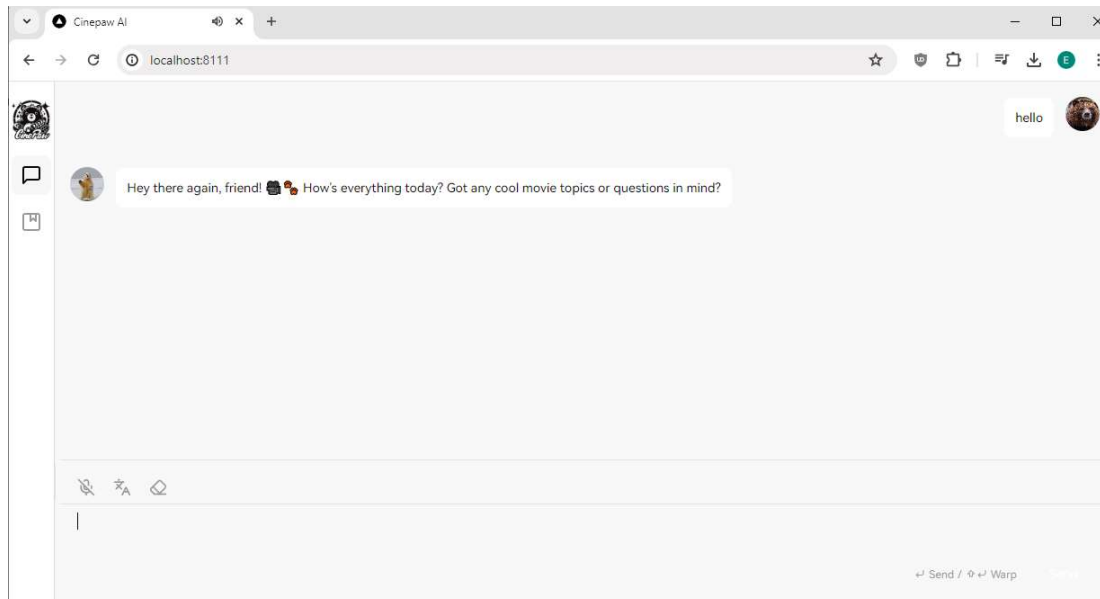
=====
==  CUDA  ==
=====

CUDA Version 11.8.0

Container image Copyright (c) 2016-2023, NVIDIA CORPORATION & AFFILIATES. All rights reserved.
```

5. Launch browser and navigate to **http://localhost:8111** to use cinepaw*

*Do take note that it will take a few seconds to warm up the system, if no response from the chatbot, please refresh after a few seconds.



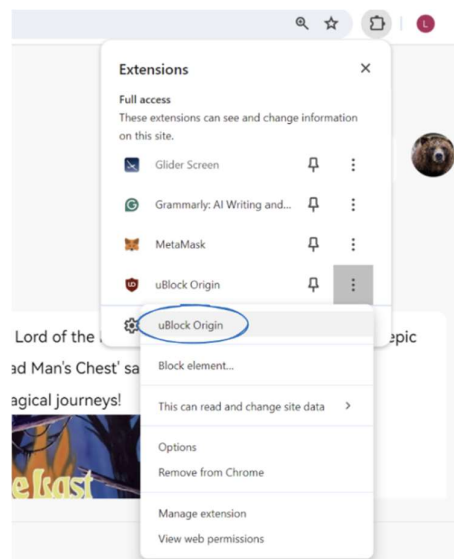
B. User Guide

Recommended browser : Chrome (requires internet access)

1. Navigate to <http://pawsible.fun/> (for ready demonstration)
 - For docker/self-deployment, go to <http://localhost:8111> , please refer to installation guide
2. Arrival at the landing screen as shown below:



3. Disable ad-blocker (if any) installed in your browser:
 - Chrome browser
 - Under plugins in the browser
 - Navigate to ad blocker installed eg. uBlock origin
 - Click onto the three dots to open up options for the ad blocker
 - Select 'unblock origin'



4. Treat insecure link as secure (for chrome browser)

- Paste this string into the search bar in chrome: `chrome://flags/#unsafely-treat-insecure-origin-as-secure`
- Under experiments, paste in the URL of cinepaw:

Experiments

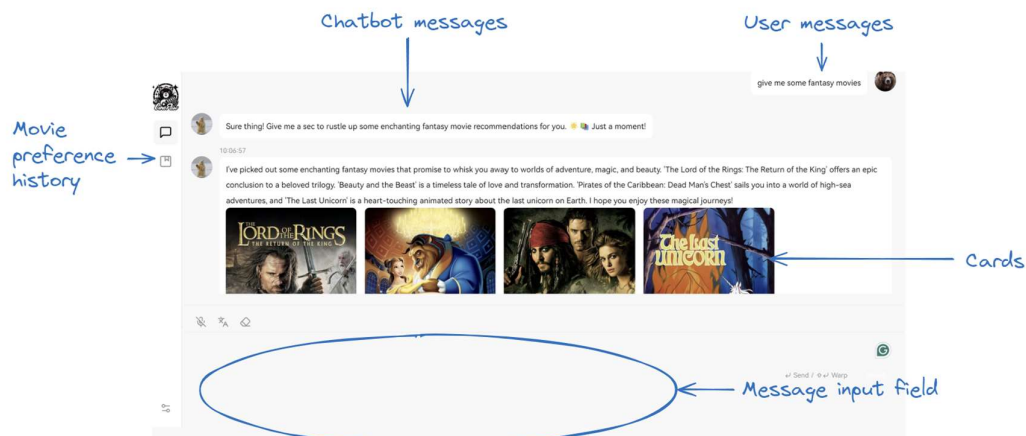
124.0.6367.63

WARNING: EXPERIMENTAL FEATURES AHEAD! By enabling these features, you could lose browser data or compromise your security or privacy. Enabled features apply to all users of this browser. If you are an enterprise admin you should not be using these flags in production.

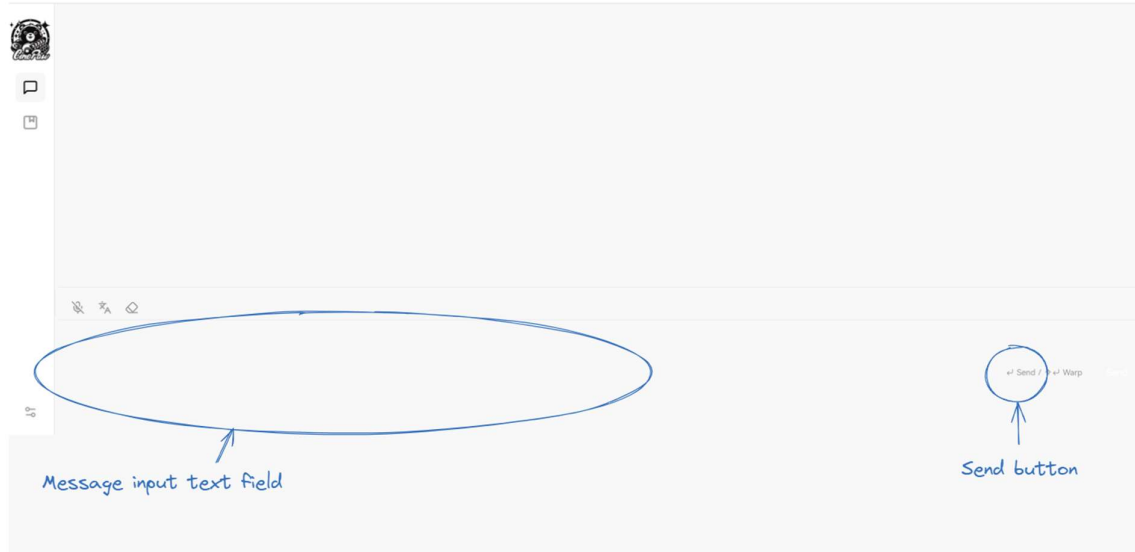
Interested in cool new Chrome features? Try our [beta channel](#).

Available	Unavailable
<p>Insecure origins treated as secure</p> <p>Treat given (insecure) origins as secure origins. Multiple origins can be supplied as a comma-separated list. Origins must have their protocol specified e.g. "http://example.com". For the definition of secure contexts, see https://w3c.github.io/webappsec-secure-contexts/ – Mac, Windows, Linux, ChromeOS, Android, Fuchsia, Lacros</p> <div> <input type="text" value="http://hikaru:9880,http://149.28.143.192:8111"/> </div> <p>#unsafely-treat-insecure-origin-as-secure</p>	<p>Enabled</p>

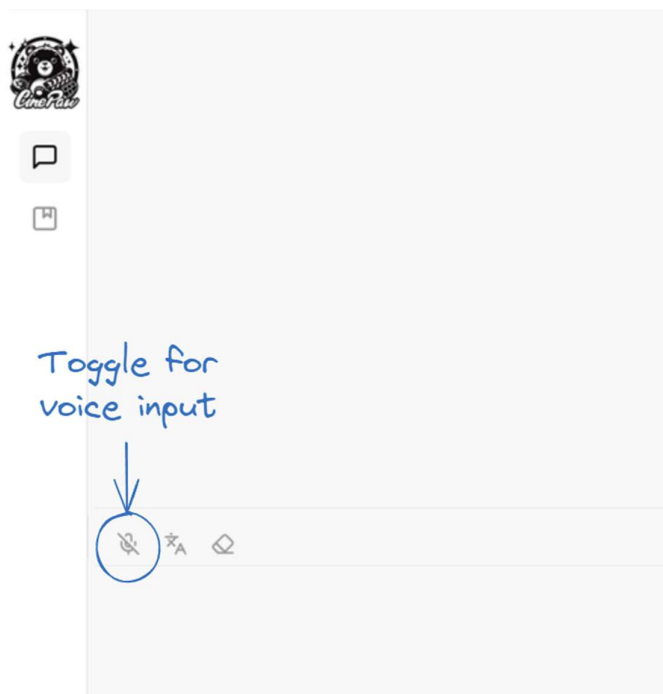
5. Description of the interface as shown below:



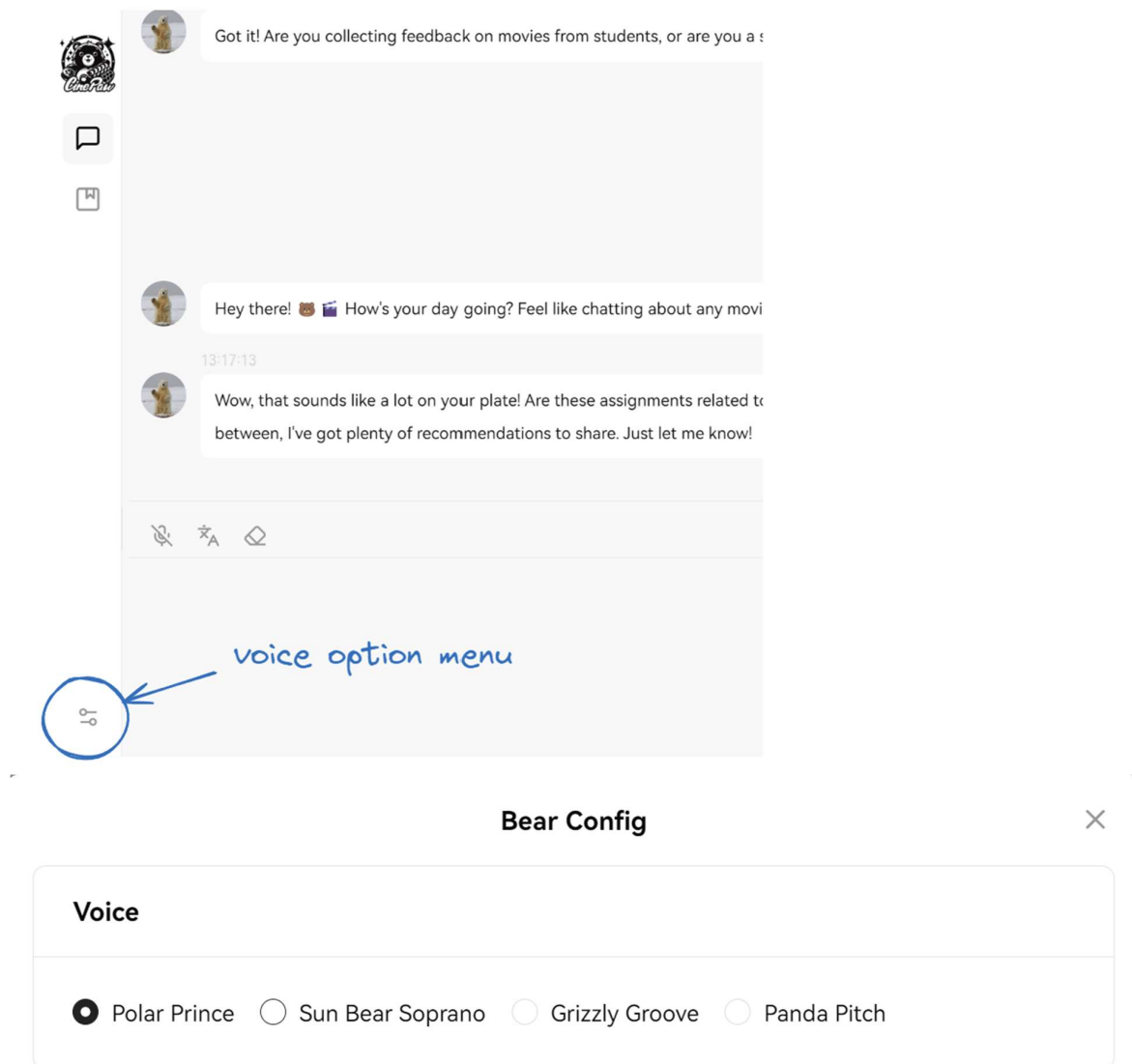
- For text inputs, type in the message input text field and press on 'Send' in the user interface or the 'Enter' key on your keyboard.



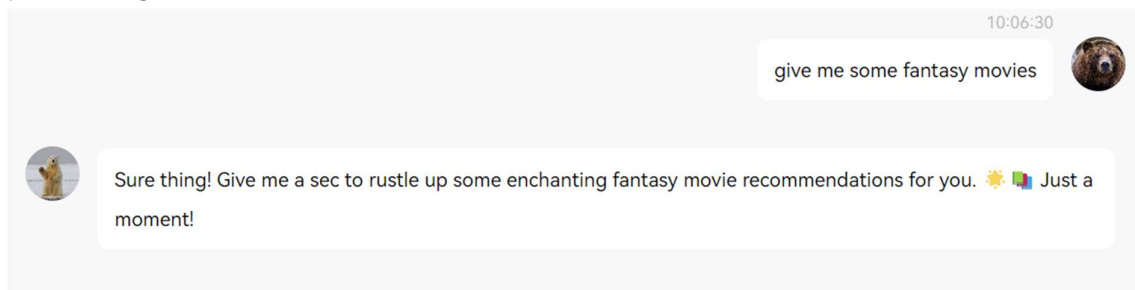
- For voice inputs, toggle the voice input until the microphone icon is enabled. You may start talking for voice inputs to be processed. Do note that this may take some time. In the event that voice input capability does not work, it may be due to ad-blocker enabled. Refer to step 3 and 4 for instructions on how to disable ad-blocker or treat webpage as secure.



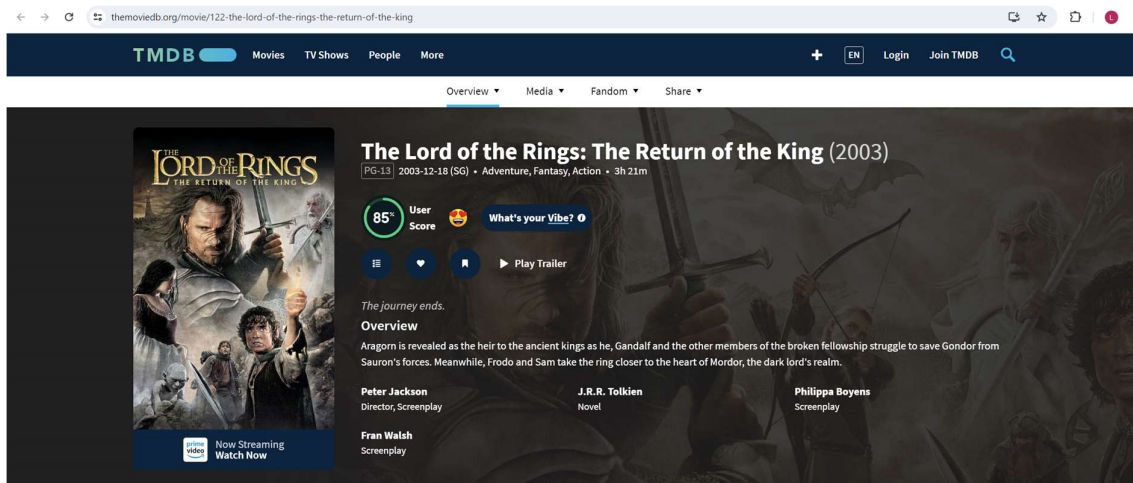
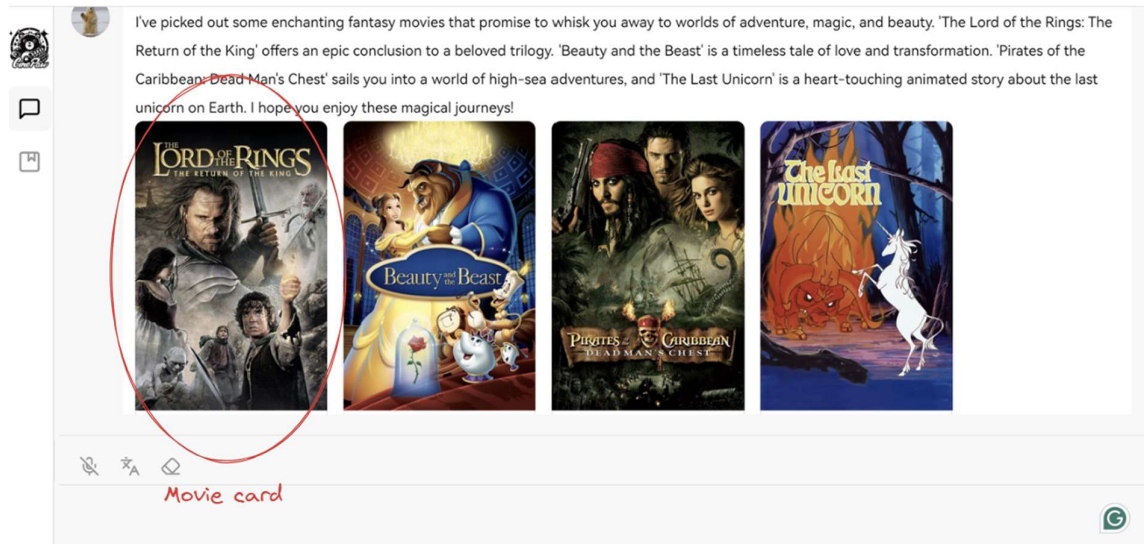
- To change voice options, click onto the voice options menu



9. Repeat steps 6 and 7 to receive replies from CinePaw. Replies from any inputs that involve movie recommendations may take slightly longer, Cinepaw will ensure that your requests are processing.



10. For movies recommended by Cinepaw, a card will be displayed on the screen. Click into the cards to access the TMDB site for the movie.



C. Mapped Module/System Functionalities

Module	System/Feature Mapping
Machine Reasoning	Persistent Memory for LLM Generate new knowledge from Chat Context Model Training (Machine Learning) Chatbot (Proactive Reasoning Systems)
Reasoning Systems	Recommendation System (Reasoning using Predictive Models) Re-ranking System
Cognitive Systems	Speech to Text Text to Speech LLM with GPT 4 (Logical Reasoning/NLU) Prompt Engineering (Chain of Thought)

D. Individual Project Report

Student Name:	Tan Eng Hui
Personal Contributions:	<ul style="list-style-type: none"> • Manage a new team with members of diverse background and pacing the team toward completion of the project goals/objectives • Development and integration of the chatbot and backend • Implementation of the speech-to-text (STT) and text-to-speech (TTS) module • Exploration of recommendation model and integration • Common Tasks: Ideation, Architecture/Design, Testing, Front/Backend Development, Landscape Exploration, Knowledge Sharing, Project Documentation
Learning Points:	<ul style="list-style-type: none"> • Gained hands-on experience implementing the STT and TTS and understand more about the function and limitation • Using chain of thought concept to query more complex to improve response from LLM
Skills Applicable in Workplace:	<ul style="list-style-type: none"> • Retrieval Augment Generation (RAG) – The technique grounding the generative content with information retrieve would greatly enhance accuracy and relevancy while reducing hallucination when using LLM for generative content. • The technology stack (e.g. react, FastAPI, ChromaDB) used to build the chatbot and enabling LLM stateful interaction would be helpful for use case of similar nature.

Student Name:	Wang Tao
Personal Contributions:	<ul style="list-style-type: none">• Led the design, development and integration of the application including both front and backend• Led the recommendation model training and integration• Implementation of the docker build and related deployment strategy• Knowledge sharing and coaching of team members on technical implementation• Common Tasks: Ideation, Architecture/Design, Testing, Front/Backend Development, Landscape Exploration, Knowledge Sharing, Project Documentation
Learning Points:	<ul style="list-style-type: none">• Learned the new progress in recsys field, tested new sequential recommendation model and learned the performance. Learned some new open-source tools for recommending like recommenders and deepmatch.• Applied the PubSub architecture in chatbot architecture, earned the implication experience.• Learned more about how to arrange tasks and cooperate with teammates.
Skills Applicable in Workplace:	<ul style="list-style-type: none">• The chatbot MQ architecture directly applies to my work chatbot project, saving a lot of time for early validation and development.• The prompt engineering skill used in this project also applied to my work project, Llm's excellent results in this project have broadened my horizons and imagination about the scope of llm's applications.

Student Name:	Lois Chee Li Ping
Personal Contributions:	<ul style="list-style-type: none"> • Co-led the market research and value proposition to ensure that the project aligns to market demand • Backend Development: Focusing on creating robust and scalable server-side logic. Key responsibilities included database setup in SQLite and recommender model exploration, training and implementation • Video Results Presentation: Developed and managed the presentation of system design video, showcasing the project's architecture design in a simple but informative manner. • Common Tasks: Ideation, Architecture/Design, Testing, Front/Backend Development, Landscape Exploration, Knowledge Sharing, Project Documentation
Learning Points:	<ul style="list-style-type: none"> • System design: Gained deeper insights into advanced backend development practices, including asynchronous programming and advanced data structures, crucial for handling high-load, real-time data processing. • Understanding how recommender models are designed and implemented, as well as choosing the most suitable model based on the situation and demands. • Understanding how various APIs are used across different applications • Understanding how queries are built for various NLP techniques. This also includes the use of vector embeddings and usage of vector databases in storing text data.
Skills Applicable in Workplace:	<ul style="list-style-type: none"> • Implementation and research into recommender models • Setup of SQLite via python • Usage of LLM for natural language processing • System architecture design for backend and frontend

Student Name:	Dong Yuantong
Personal Contributions:	<ul style="list-style-type: none"> • Backend Development: Spearheaded the backend development, focusing on creating robust and scalable server-side logic. Key responsibilities included optimizing database interactions and ensuring secure API endpoints. • Problem Solving: Actively involved in diagnosing and resolving technical issues, which included debugging complex software bugs and optimizing system performance under various load conditions. • Video Results Presentation: Developed and managed the presentation of video results, showcasing the project's capabilities and success stories in a visually compelling format. • API Design and Development: Led the design and implementation of crucial APIs that facilitated seamless integration between different software components and external services.
Learning Points:	<ul style="list-style-type: none"> • Advanced Backend Techniques: Gained deeper insights into advanced backend development practices, including asynchronous programming and advanced data structures, crucial for handling high-load, real-time data processing. • Problem-Solving Skills: Enhanced problem-solving skills by tackling real-world software development challenges, learning to implement more efficient and effective solutions. • Integration of Multimedia Content: Learned the intricacies of integrating and managing multimedia content within web applications, improving user engagement and satisfaction. • Natural Language Processing (NLP): Acquired substantial knowledge in NLP techniques, particularly in how they apply to building intelligent chat interfaces and processing user inputs for personalized responses. This included understanding the principles of vector embeddings and the operation of vector databases for maintaining conversation context and history.
Skills Applicable in Workplace:	<ul style="list-style-type: none"> • System Architecture Design: Ability to design complex system architectures, considering both scalability and security, which can be directly applied to developing enterprise-level solutions. • API Development: Developed proficiency in creating and managing APIs, which is a critical skill in software

	<p>development, especially for systems that rely heavily on external integrations.</p> <ul style="list-style-type: none">• Troubleshooting and Debugging: Improved troubleshooting and debugging skills, essential for maintaining high-quality software products in any software development or IT role.• Presentation Skills (Video): Enhanced ability to present technical data and project results effectively, which is invaluable for roles that require communicating complex information to stakeholders or clients.
--	---

E. Project Proposal






CINEPAW AI




TAN ENG HUI (A0291201W) | WANG TAO (A0291189R) | LOIS CHEELI PING (A0292098R) | DONG YUANTONG (A0292041N)

© National University of Singapore

1

Introduction to Cinepaw AI



A conversation-based movie recommendation system that understands and remembers user's need to deliver the best movie suggestion

Interactive Chat:

Users can engage in meaningful conversations to identify their unique movie preferences, moving beyond traditional, impersonal algorithms.

Customized Recommendations:

Deliver movie suggestions that resonate with individual user tastes.

Adaptive Technology:

Evolves alongside with our users, dynamically refining movie selections based on continuous feedback and interactions, ensuring a personalized experience for each user.

© National University of Singapore

2

Cinepaw AI Goal



Creating a **unique** journey across movie discovery through **simple conversations**

Keeping things personal and relatable, just like a friend whom knows you well

© National University of Singapore

3

Project Background / Market Context



Current market dominated by giants – Netflix, IMDB, Douban

The need to **stand out** against competitors is crucial in attracting viewers to their platform as well as retain viewership.

Website	http://www.imdb.com
Revenue	\$170 million
Employees	1,027 (907 on RocketReach)
Founded	1991
Phone	

Netflix spends \$150 million on content recommendations every year

Website	http://www.douban.com
Revenue	\$70 million
Employees	260 (190 on RocketReach)
Founded	2005
Address	no.53 Wudaoying Hutong, Beijing, Beijing 54 100007, CN
Phone	+86 10 6402 8881
Technologies	JavaScript, HTML, Twitter +34 more (view full list)
Category	Technology, Information and Internet, Broadcasting, Social Media, Social/Platform Software, Media & Internet, Internet Services, Social C& Entertainment, Social Network

Netflix Information	
Website	http://www.netflix.com
Ticker	NFLX
Revenue	\$30.4 billion
Funding	\$16.7 billion
Employees	22,580 (22,580 on RocketReach)
Founded	1997
Address	100 Winchester Circle, Los Gatos, California, US
Phone	(408) 540-3700
Fax	(408) 540-3737

Why Netflix thinks its personalized recommendation engine is worth \$1 billion per year

Nathan McAfee Jun 15, 2016, 5:36 AM GMT-8

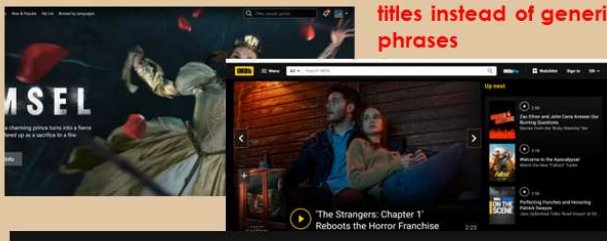
© National University of Singapore

Project Background / Market Context



Recommender system plays a crucial role in the entertainment industry – more specifically screening services

Search bar capability limited to finding movie titles instead of generic phrases



Your search for "the movie starred by the main character of Titanic" did not have any matches.

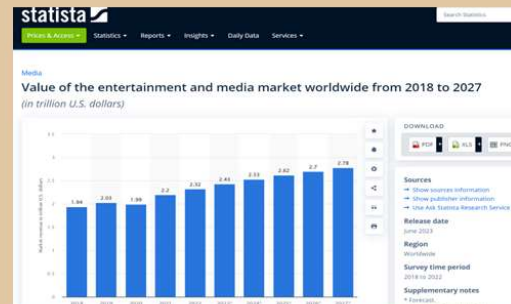
Suggestions:

- Try different keywords
- Looking for a film or TV programme?
- Try using a film, TV programme title, an actor or director.
- Try a genre, such as comedy, romance, sports or drama.

Market Landscape Overview:

All players like Netflix, Hulu, YouTube invest heavily in recommend system, it's directly related to their business performance.

video streaming market size was valued at USD 455.45 billion in 2022, and for Entertainment market it will hit 2.7 trillion.



© National University of Singapore

5

Market Research – Lack of personalization



From Spotify to Netflix and Amazon, we're surrounded by extreme personalization every day. Consumers have come to expect that same level of personalization from companies of all sizes. Investing in personalization efforts to build relationships and create better experiences can pay off with serious rewards for brands. And in a world where the vast majority of companies are focused on improving personalization, companies that don't prioritize creating a tailored experience run the risk of being overlooked.

"74% of customers feel frustrated when website content is not personalized"

Current State Of Personalization

71% of consumers feel frustrated when a shopping experience is impersonal.

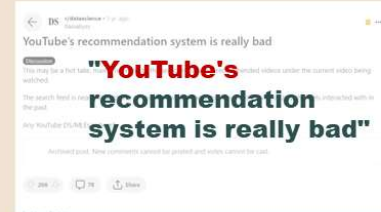
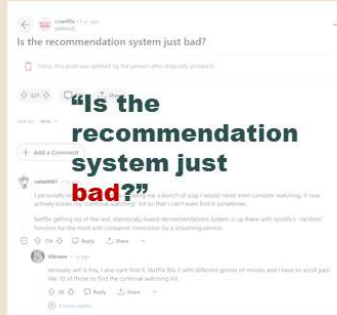
Segment: Millennials

70% of millennials are frustrated with brands sending irrelevant emails.

Source: HubSpot

74% of customers feel frustrated when website content is not personalized.

Source: HubSpot



Machine Review: new | post | comments | edit | share | delete | archive

Ask HN: Why do recommender systems not seem good?

Hi everyone, I've been thinking about this for a while.

I've noticed that many of the recommendations I get from YouTube, Netflix, and Amazon seem to be quite generic and not very personalized.

I've also noticed that many of the recommendations I get from YouTube, Netflix, and Amazon seem to be quite generic and not very personalized.

I've also noticed that many of the recommendations I get from YouTube, Netflix, and Amazon seem to be quite generic and not very personalized.

I've also noticed that many of the recommendations I get from YouTube, Netflix, and Amazon seem to be quite generic and not very personalized.

I've also noticed that many of the recommendations I get from YouTube, Netflix, and Amazon seem to be quite generic and not very personalized.

I've also noticed that many of the recommendations I get from YouTube, Netflix, and Amazon seem to be quite generic and not very personalized.

I've also noticed that many of the recommendations I get from YouTube, Netflix, and Amazon seem to be quite generic and not very personalized.

I've also noticed that many of the recommendations I get from YouTube, Netflix, and Amazon seem to be quite generic and not very personalized.

I've also noticed that many of the recommendations I get from YouTube, Netflix, and Amazon seem to be quite generic and not very personalized.

I've also noticed that many of the recommendations I get from YouTube, Netflix, and Amazon seem to be quite generic and not very personalized.

I've also noticed that many of the recommendations I get from YouTube, Netflix, and Amazon seem to be quite generic and not very personalized.

I've also noticed that many of the recommendations I get from YouTube, Netflix, and Amazon seem to be quite generic and not very personalized.

I've also noticed that many of the recommendations I get from YouTube, Netflix, and Amazon seem to be quite generic and not very personalized.

I've also noticed that many of the recommendations I get from YouTube, Netflix, and Amazon seem to be quite generic and not very personalized.

I've also noticed that many of the recommendations I get from YouTube, Netflix, and Amazon seem to be quite generic and not very personalized.

I've also noticed that many of the recommendations I get from YouTube, Netflix, and Amazon seem to be quite generic and not very personalized.

I've also noticed that many of the recommendations I get from YouTube, Netflix, and Amazon seem to be quite generic and not very personalized.

I've also noticed that many of the recommendations I get from YouTube, Netflix, and Amazon seem to be quite generic and not very personalized.

I've also noticed that many of the recommendations I get from YouTube, Netflix, and Amazon seem to be quite generic and not very personalized.

I've also noticed that many of the recommendations I get from YouTube, Netflix, and Amazon seem to be quite generic and not very personalized.

I've also noticed that many of the recommendations I get from YouTube, Netflix, and Amazon seem to be quite generic and not very personalized.

I've also noticed that many of the recommendations I get from YouTube, Netflix, and Amazon seem to be quite generic and not very personalized.

I've also noticed that many of the recommendations I get from YouTube, Netflix, and Amazon seem to be quite generic and not very personalized.

I've also noticed that many of the recommendations I get from YouTube, Netflix, and Amazon seem to be quite generic and not very personalized.

I've also noticed that many of the recommendations I get from YouTube, Netflix, and Amazon seem to be quite generic and not very personalized.

I've also noticed that many of the recommendations I get from YouTube, Netflix, and Amazon seem to be quite generic and not very personalized.

I've also noticed that many of the recommendations I get from YouTube, Netflix, and Amazon seem to be quite generic and not very personalized.

I've also noticed that many of the recommendations I get from YouTube, Netflix, and Amazon seem to be quite generic and not very personalized.

I've also noticed that many of the recommendations I get from YouTube, Netflix, and Amazon seem to be quite generic and not very personalized.

I've also noticed that many of the recommendations I get from YouTube, Netflix, and Amazon seem to be quite generic and not very personalized.

I've also noticed that many of the recommendations I get from YouTube, Netflix, and Amazon seem to be quite generic and not very personalized.

I've also noticed that many of the recommendations I get from YouTube, Netflix, and Amazon seem to be quite generic and not very personalized.

Source: <https://www.forbes.com/sites/blakemorgan/2020/02/18/50-stats-showing-the-power-of-personalization/?sh=7b8a437d2a24>

© National University of Singapore

6

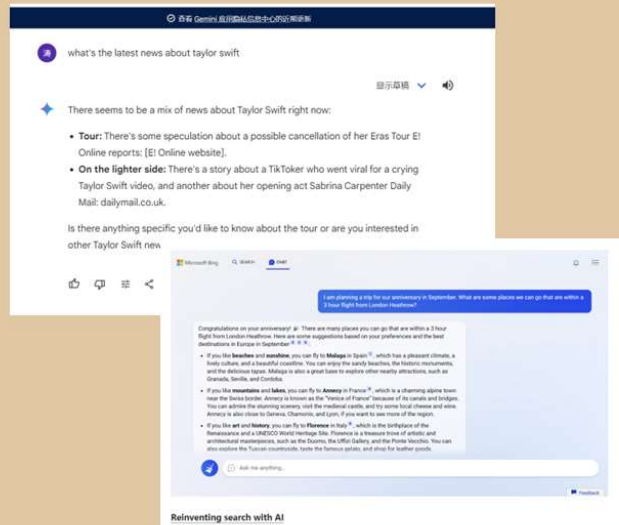
Emerging trend of incorporating AI into search bar function



- Google and Microsoft have incorporated AI into their search chat interface

- Robust system but results may not be industry specific

- Traditional players (Netflix, HBO, IMDB) may not be as fast to catch the trend – showcasing opportunities for our product as an integration to their services



<https://blogs.microsoft.com/blog/2023/02/07/reinventing-search-with-a-new-ai-powered-microsoft-bing-and-edge-your-copilot-for-the-web/>

© National University of Singapore

7

Market Values



User Benefits

- Better and happier using the system as it provide personalised search experience
- Reduce frustration (no need to describe exact keyword, phrase)
- Ability to understand nuanced preferences reduce time spent on getting good recommendations
- Save time and effort
- Convenience to users when performing search (using natural language)

Company Benefits

- Better understanding and more engaged users increase conversion rate
- Enhanced users' satisfaction and loyalty through personalised services
- System can smartly suggest content other than movie leading to potential purchases

© National University of Singapore

8

Data Collection and Preparation

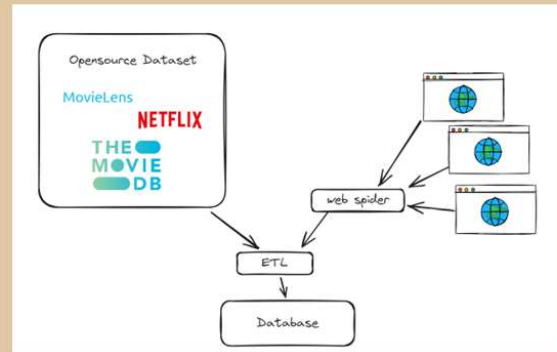


Sources :

- Publicly available movie databases like IMDb, TMDb, MovieLens, as well as local sources like Singapore Film Database.
- Datasets from social media and forums for user reviews and sentiments.

Acquisition and Processing:

- Readily available public data sources.
- Potential use of web crawlers to fetch necessary data from internet.
- Processing and management of data via ETL methods to format data used for modelling.



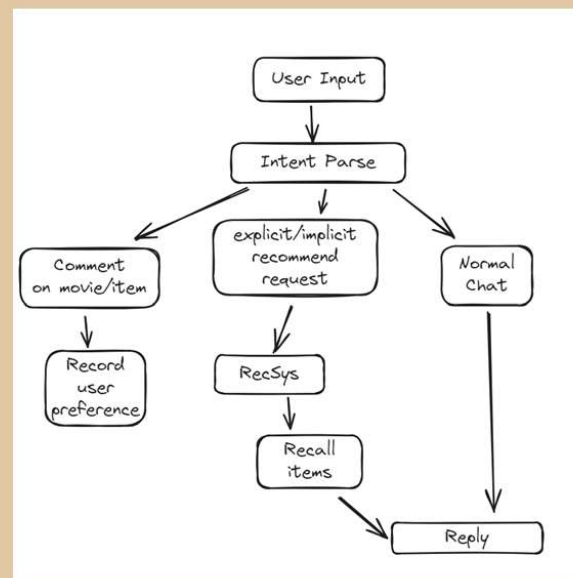
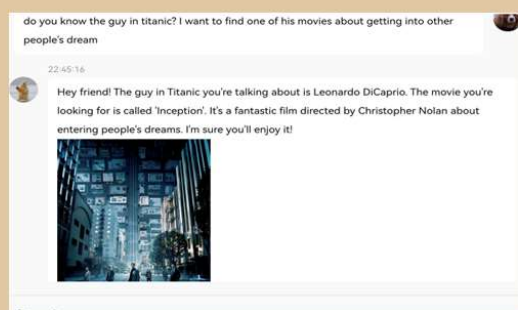
© National University of Singapore

10

System Workflow/Process



With a conversational interact interface, the search and recommending process more **natural and efficient**



© National University of Singapore

11

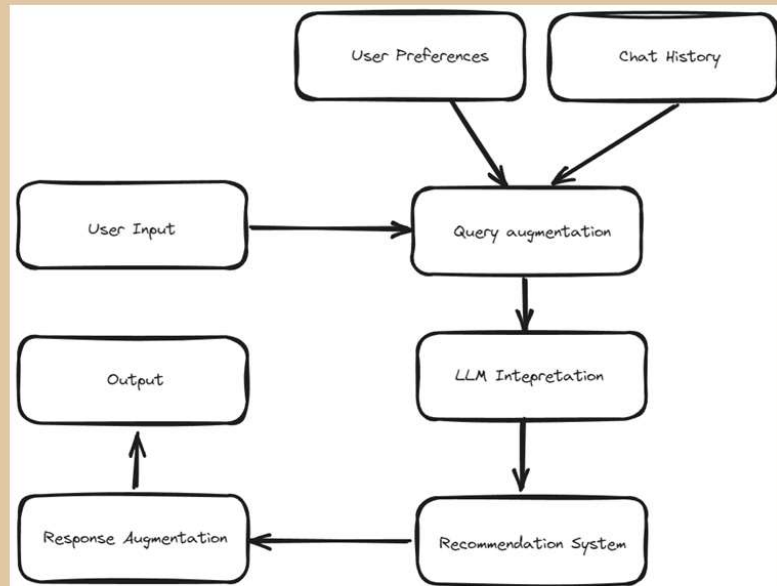
System Workflow/Process



User input fed into system & enhanced by historical data with stated preferences

LLM does processing of information from user input and stored information is then channelled to the recommendation system to give output.

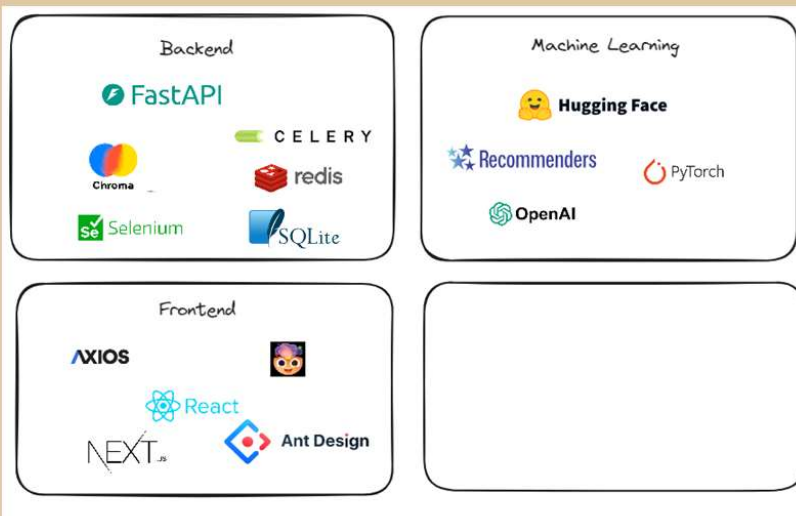
Final output is also embedded and stored in database for future use.



© National University of Singapore

12

Technology Selection



Frontend

React, Axios, Next.js, Antd, LubeUI

Backend

Flask, FastAPI

Vector Database

Chromadb/Redis

Relational Database

SQLite

Web Crawl

Selenium, BeautifulSoup

For models, we will refer to sources from huggingface or other open source to use either pre-train, fine tune, train new model or mixed of the approaches as deemed necessary during development. The list is not exhaustive and may change overtime base on requirements and performance.

© National University of Singapore

13

ANNEX

© National University of Singapore

14

Recommendation Systems Exploration

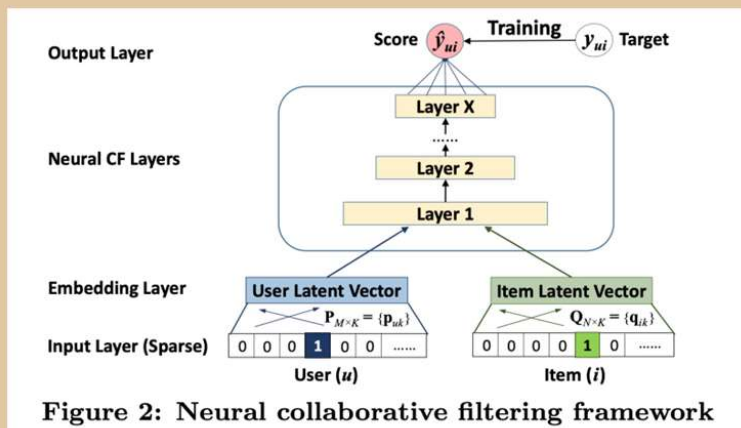


Figure 2: Neural collaborative filtering framework

Neural Collaborative Filtering

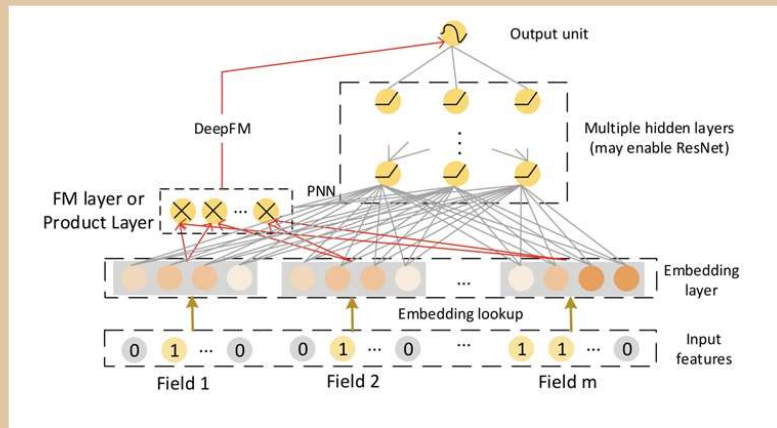
This research introduces the Neural network-based Collaborative Filtering (NCF) framework, a novel approach in recommender systems emphasizing implicit feedback. Traditional methods often rely on matrix factorization for user-item interactions, focusing mainly on auxiliary data like item descriptions. NCF diverges by using deep learning to directly model these interactions, employing a multi-layer perceptron for enhanced learning capabilities. This method has shown superior performance in experiments compared to existing techniques, particularly benefiting from deeper neural network structures.

Reference : He, X., Liao, L., Zhang, H., Nie, L., Hu, X., & Chua, T.-S. (2017). Neural Collaborative Filtering. Proceedings of the 26th International Conference on World Wide Web (WWW '17), April 3–7, Perth, Australia. <https://arxiv.org/abs/1708.05031>

© National University of Singapore

15

Recommendation Systems Exploration



xDeepFM: Combining Explicit and Implicit Feature Interactions for Recommender Systems

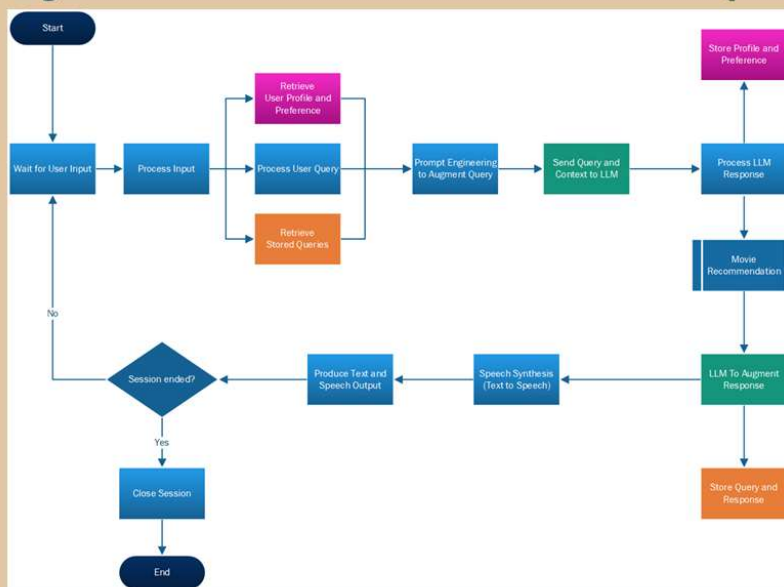
The paper introduces xDeepFM, a model that merges Deep Neural Networks (DNNs) with a Compressed Interaction Network (CIN) to capture both explicit and implicit feature interactions for recommender systems. Unlike traditional factorization models, xDeepFM can explicitly model feature interactions at a vector-wise level, enhancing its ability to understand complex data patterns. This approach leads to superior performance over existing models on various datasets. The source code is available for further exploration.

Reference : Lian, J., Zhou, X., Zhang, F., Chen, Z., Xie, X., & Sun, G. (2018). xDeepFM: Combining Explicit and Implicit Feature Interactions for Recommender Systems. arXiv preprint arXiv:1803.05170. Retrieved from <https://arxiv.org/abs/1803.05170>

© National University of Singapore

16

System Interaction Workflow (Draft)



The drafted workflow depicts the overview of the system processes. It will adapt and evolve based on further requirement analysis and system objectives.

The draft serves as a reference to the components we need to take into consideration for implementation and integration.

Subsequently, technical specification and implementation detail would derive from the high-level design to fulfill.

This is also used to determine the area of works and focus for each member.

© National University of Singapore

17

F. Reference

Cen, Y., Zhang, J., Zou, X., Zhou, C., Yang, H., & Tang, J. (2020). Controllable Multi-Interest Framework for Recommendation. Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining.

<https://arxiv.org/abs/2005.09347>

Li, C., Liu, Z., Wu, M., Xu, Y., Huang, P., Zhao, H., Kang, G., Chen, Q., Li, W., & Lee (2019). Multi-Interest Network with Dynamic Routing for Recommendation at Tmall. Proceedings of the 28th ACM International Conference on Information and Knowledge Management.

<https://arxiv.org/abs/1904.08030>

Covington, P., Adams, J.K., & Sargin, E. (2016). Deep Neural Networks for YouTube Recommendations. Proceedings of the 10th ACM Conference on Recommender Systems.

<https://dl.acm.org/doi/10.1145/2959100.2959190>

He, X., Liao, L., Zhang, H., Nie, L., Hu, X., & Chua, T.-S. (2017). Neural Collaborative Filtering. Proceedings of the 26th International Conference on World Wide Web (WWW '17), April 3–7, Perth, Australia.

<https://arxiv.org/abs/1708.05031>

Lian, J., Zhou, X., Zhang, F., Chen, Z., Xie, X., & Sun, G. (2018). xDeepFM: Combining Explicit and Implicit Feature Interactions for Recommender Systems. arXiv preprint arXiv:1803.05170.

<https://arxiv.org/abs/1803.05170>