

# BOU THE BOT

## Multilingual AI Chatbot

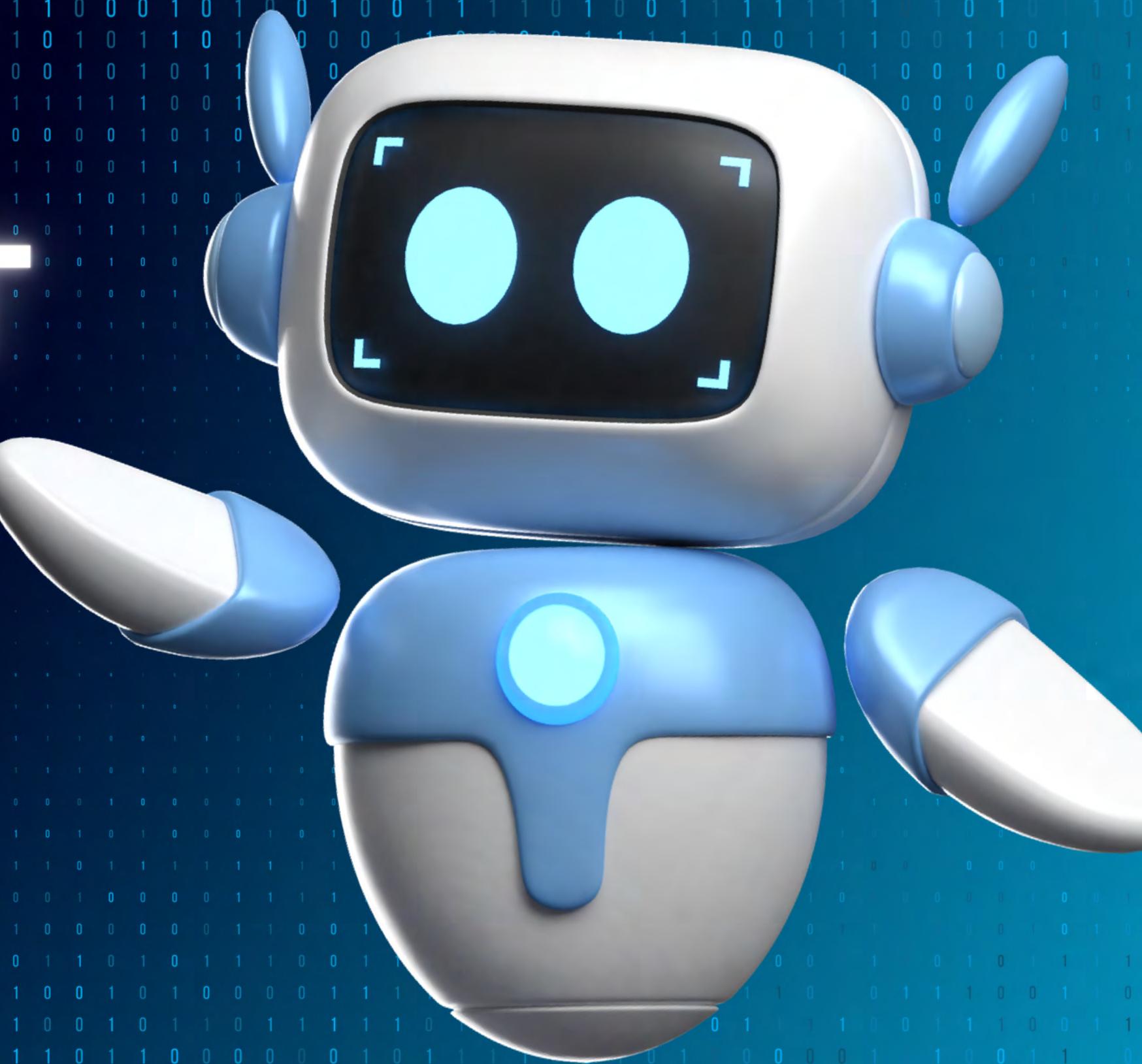
*Germaine Luah*

*Gracier Chua*

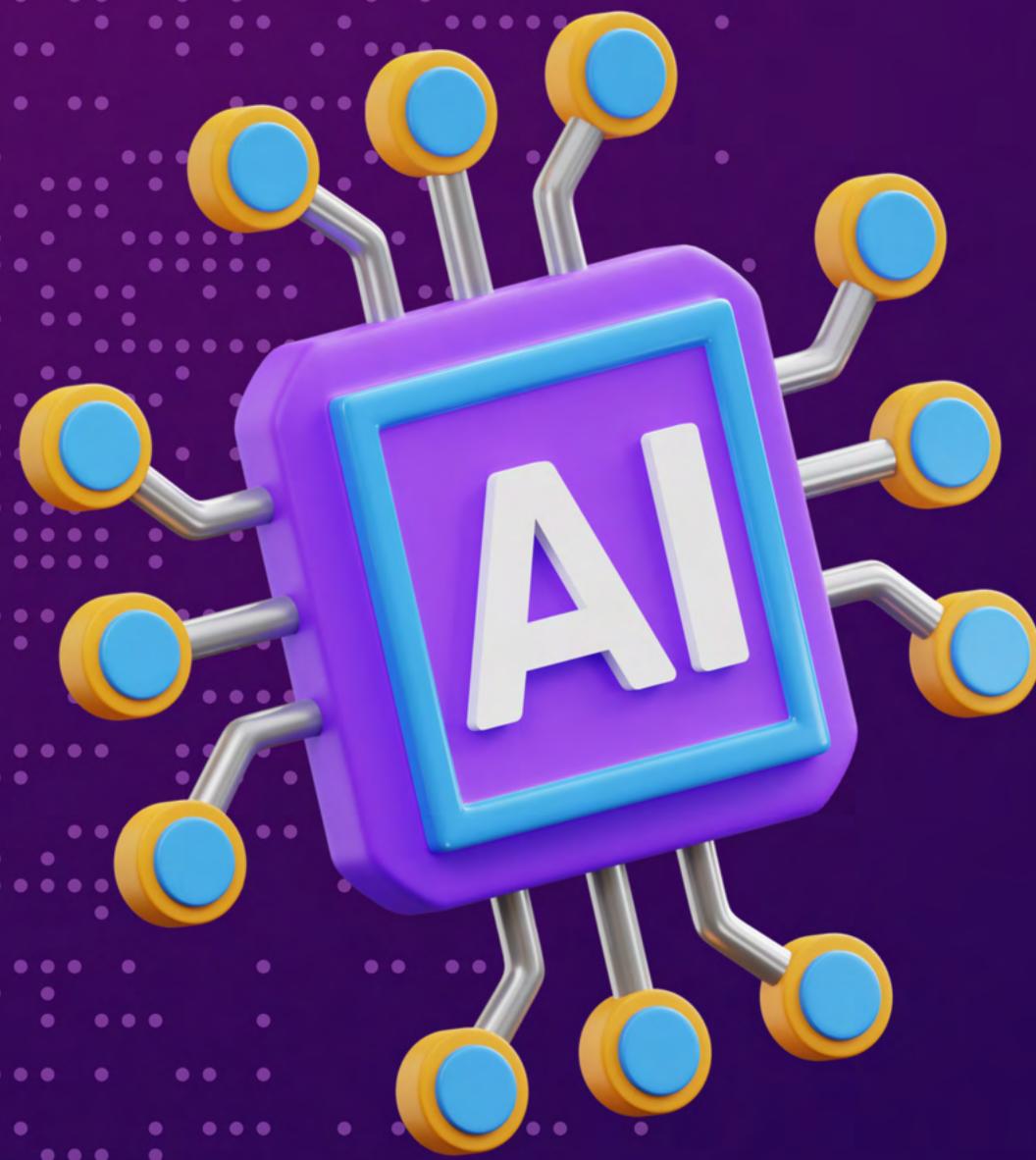
*Kevin Tan YueJun*

*Parv*

*Taufik*



# CONTENTS



- ① Problem Statement
  - ② Multilingual Customer Service Chatbot
  - ③ Technology Stack and Workflow
  - ④ Potential Challenges
  - ⑤ Current Impact & Future Iterations
- + Live Demo



# PROBLEM STATEMENT

UOB's existing chatbot struggle to understand and respond effectively to complex customer inquiries, leading to frustration.

This is due to its limitations in recognizing specific prompts, understanding context, and providing tailored responses.



1

# BIG 3 COMPARISON

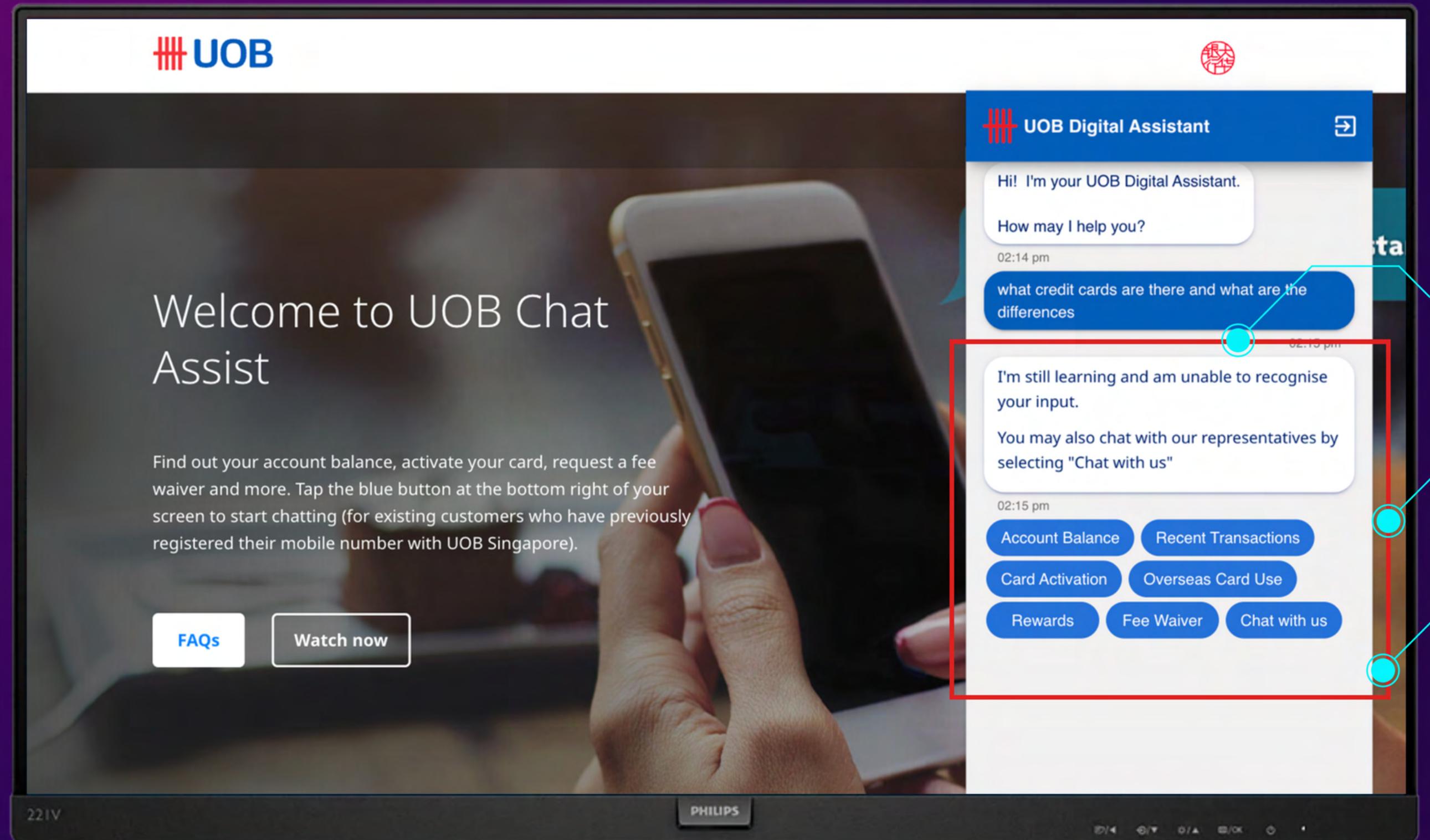
The image displays three side-by-side screenshots of mobile banking chatbots for DBS, UOB, and OCBC, illustrating their functionality and user interface.

**DBS Chatbot:** The interface is white with red accents. It features a red header bar with the text "digibot" and a close button "X". Below the header, there's a "Report Scam" button. A message bubble contains the text: "To see all the topics I can help you with, simply tap the menu." A red button labeled "qu'est-ce qu'un compte d'épargne" is shown. Another message bubble states: "Unfortunately we are unable to process your request now. Please try again later, or you may visit our Customer Service Help & Support for further assistance." At the bottom, there are two buttons: "Help and Support" and "Leave-Us-A-Message". A text input field at the very bottom says "message...".

**UOB Digital Assistant:** The interface has a blue header bar with the text "UOB Digital Assistant" and a share icon. The main message bubble says: "Hi! I'm your UOB Digital Assistant." followed by "How may I help you?". A timestamp "03:00 pm" is shown. A blue button labeled "Qu'est-ce qu'un compte épargne ?" is displayed. A timestamp "03:01 pm" is shown. The response message: "I'm still learning and am unable to recognise your input. You may also chat with our representatives by selecting "Chat with us"" is timestamped "03:01 pm". Below the message are several blue buttons: "Account Balance", "Recent Transactions", "Card Activation", "Overseas Card Use", "Rewards", "Fee Waiver", and "Chat with us".

**OCBC Chatbot:** The interface has a pink header bar with the text "Need help?" and a close button "-". It says: "Our chatbot is here to assist with your Personal Banking enquiries." A timestamp "22 Oct 24, 11:40 am" is shown. A blue button labeled "qu'est-ce qu'un compte d'épargne" is shown. A timestamp "22 Oct 24, 11:40 am" is shown. The response message: "Sorry, I do not understand. Could you please rephrase your question? If you need information about a product or service, please tell me its name." is timestamped "22 Oct 24, 11:40 am". At the bottom, there is a text input field with the placeholder "Type your question (e.g Credit Card)" and a send arrow icon.

# RULE-BASED CHATBOT



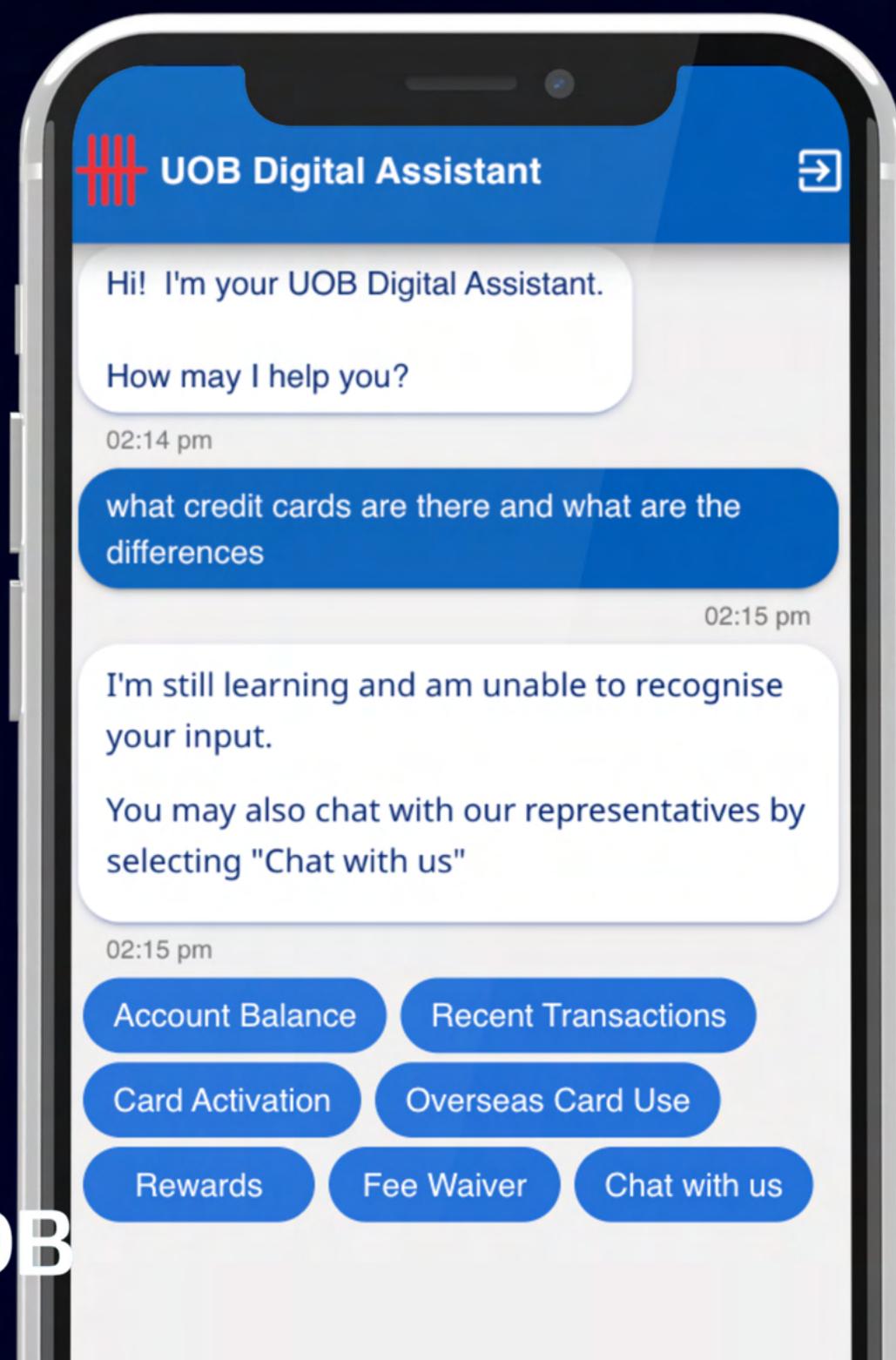
Limited flexibility

Limited knowledge

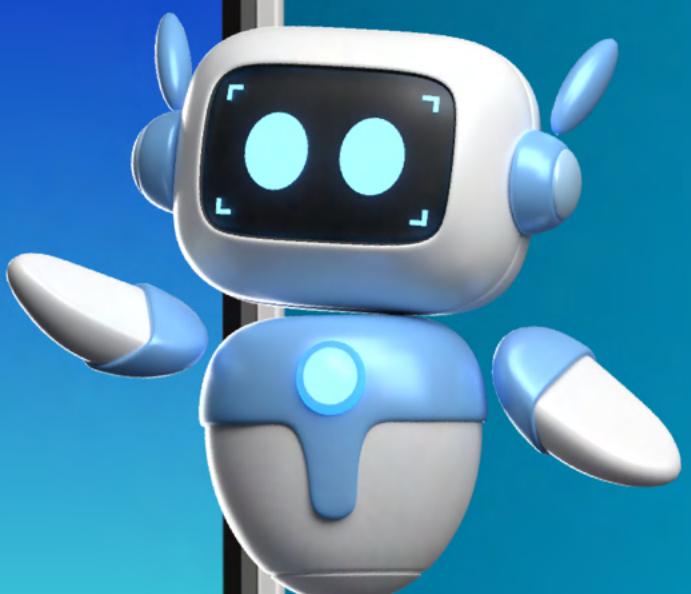
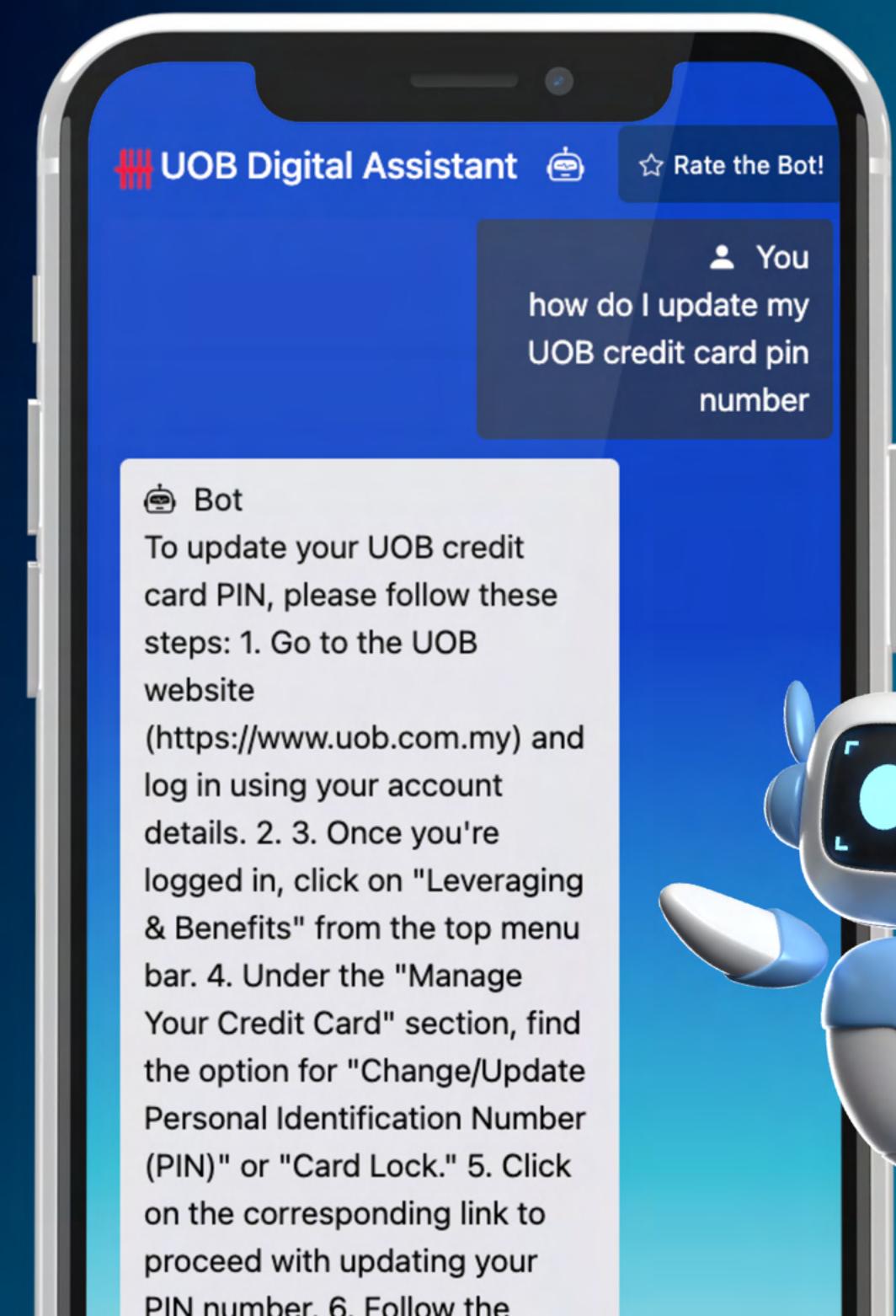
Inability to learn

+ multilingual

## RULE-BASED CHATBOT



## CONVERSATIONAL CHATBOT



2

## CHATBOT POTENTIAL AIMS

PERSONALISED USER EXPERIENCES

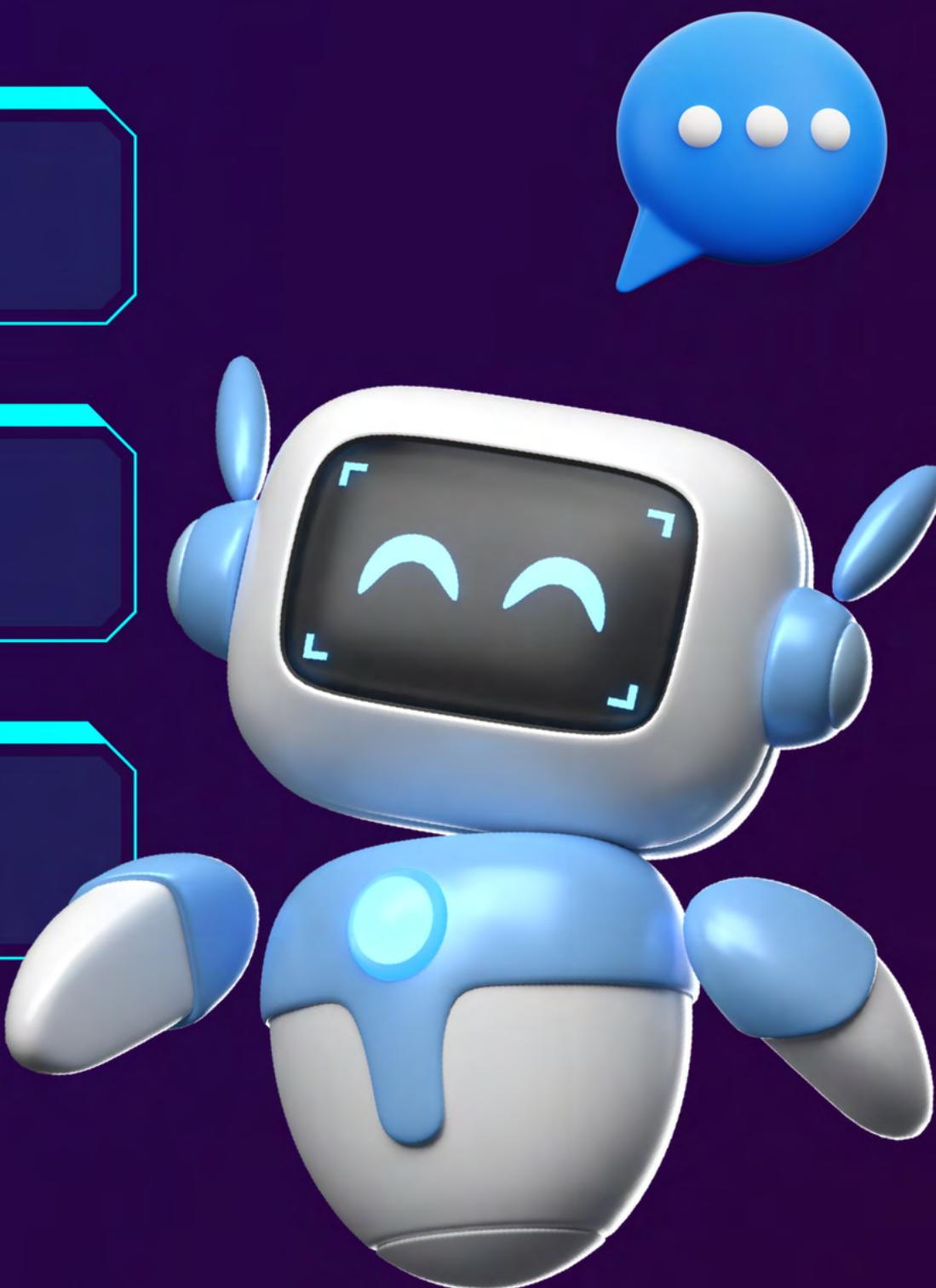
EASILY SCALABLE

MULTILINGUAL SUPPORT

24/7 AVAILABILITY

CONTINUOUSLY IMPROVING CUSTOMER EXPERIENCES

OPTIMIZING MANPOWER

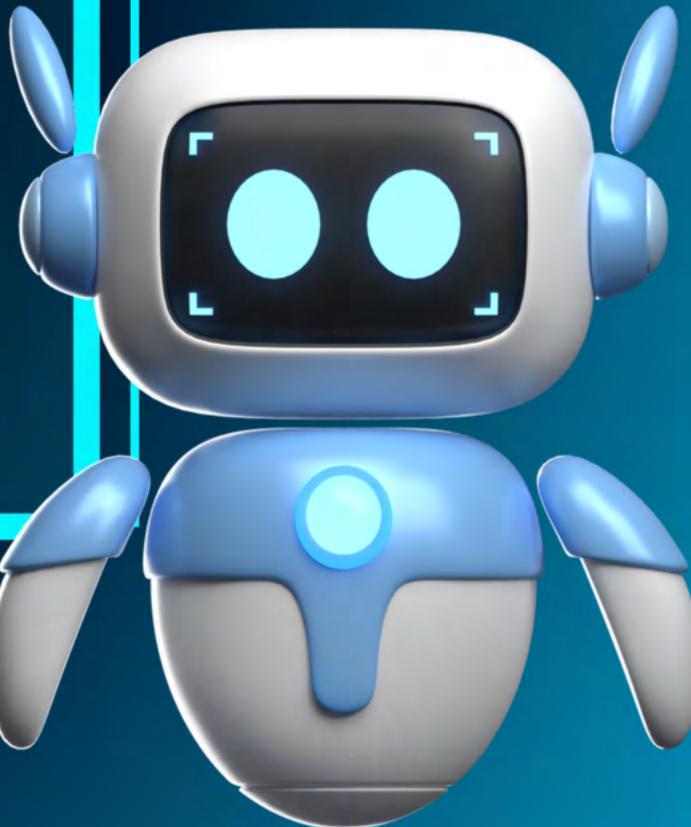


# CONVENTIONAL CHATBOTS

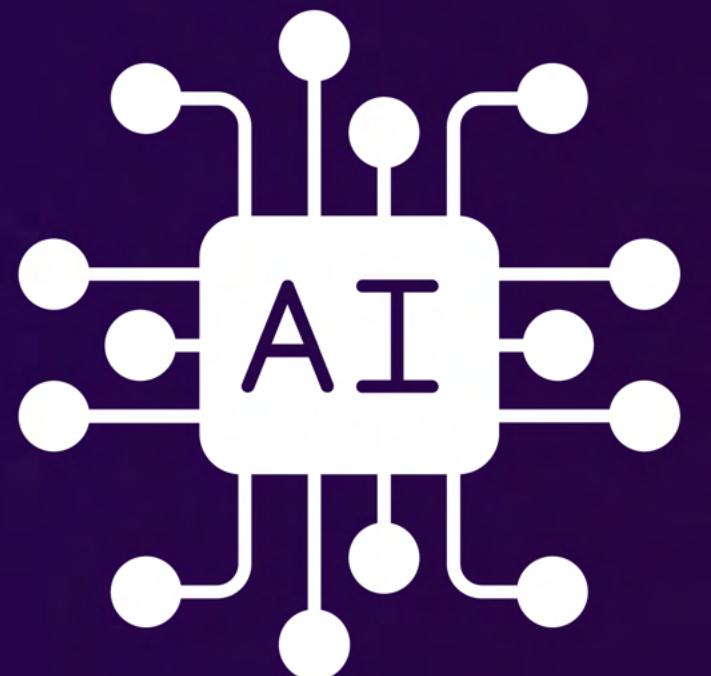
Existing chatbots usually rely on pre-defined scripts and rule-based systems.

Their performance is often poor in understanding context and responding to nuances in multiple languages.

There is no aspect of continuous improvement with greater interaction of the user inputs.



# HOW LLM CHANGES THE SCOPE



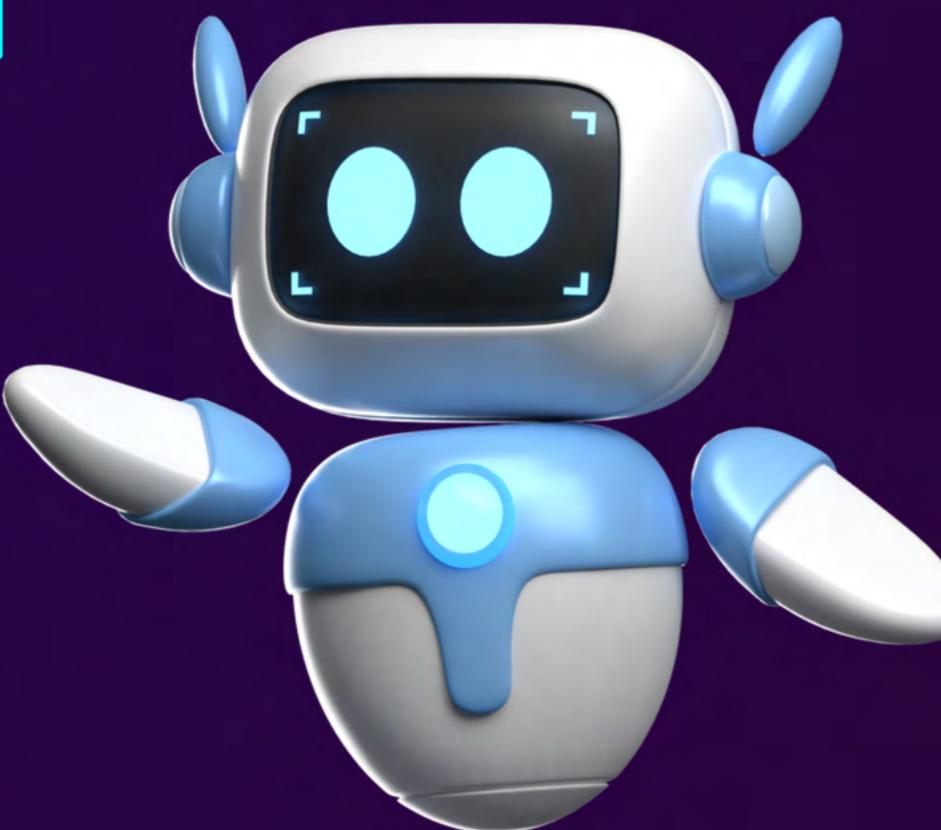
TRAINED ON VAST AMOUNTS OF DATA

SELF-LEARNING CAPABILITY

BETTER MULTILINGUAL PERFORMANCE

NO NEED FOR MANUAL CREATION OF RULES

FASTER RESPONSE TIME



### 3

# TECHNICAL OVERVIEW & TECH STACK



## python™

Powerful libraries and easy syntax



## Flask

Lightweight & flexible for building APIs



## Weights & Biases

ML platform for model management



## GitHub

Version control and collaboration



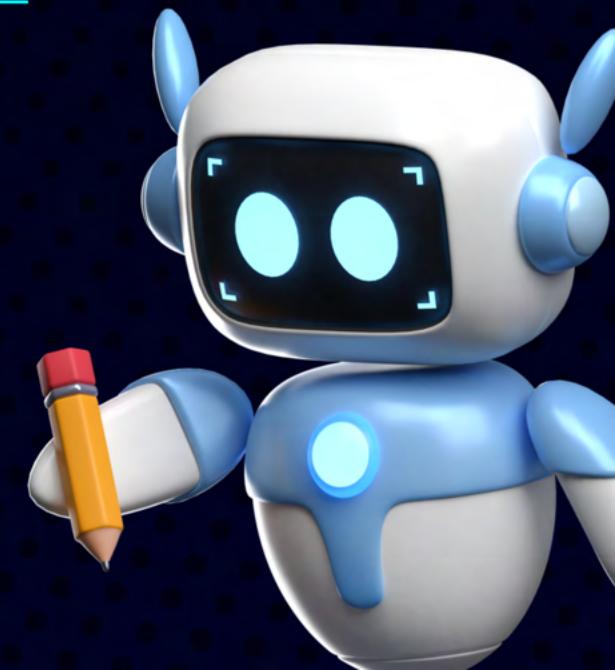
## React JS

Efficient development of dynamic user interfaces



## mongoDB®

Database used to store and retrieve data.



# LARGE LANGUAGE MODEL SELECTION

## Ollama Stable LM 2

### MULTILINGUAL SUPPORT

The model is pre-trained to be fluent in seven languages. It can also seamlessly engage in conversations across many other languages.

### DOMAIN SPECIFIC OPTIMISATION

Ease of fine-tuning to tailor responses to specific use cases (e.g. banking-related scenarios like transaction inquiries, account assistance etc)

### SCALABILITY AND 24/7 AVAILABILITY

Able to handle large-scale customer interactions in a responsive and consistent manner. Delivery of high-quality and continual customer support in a time-sensitive environment.



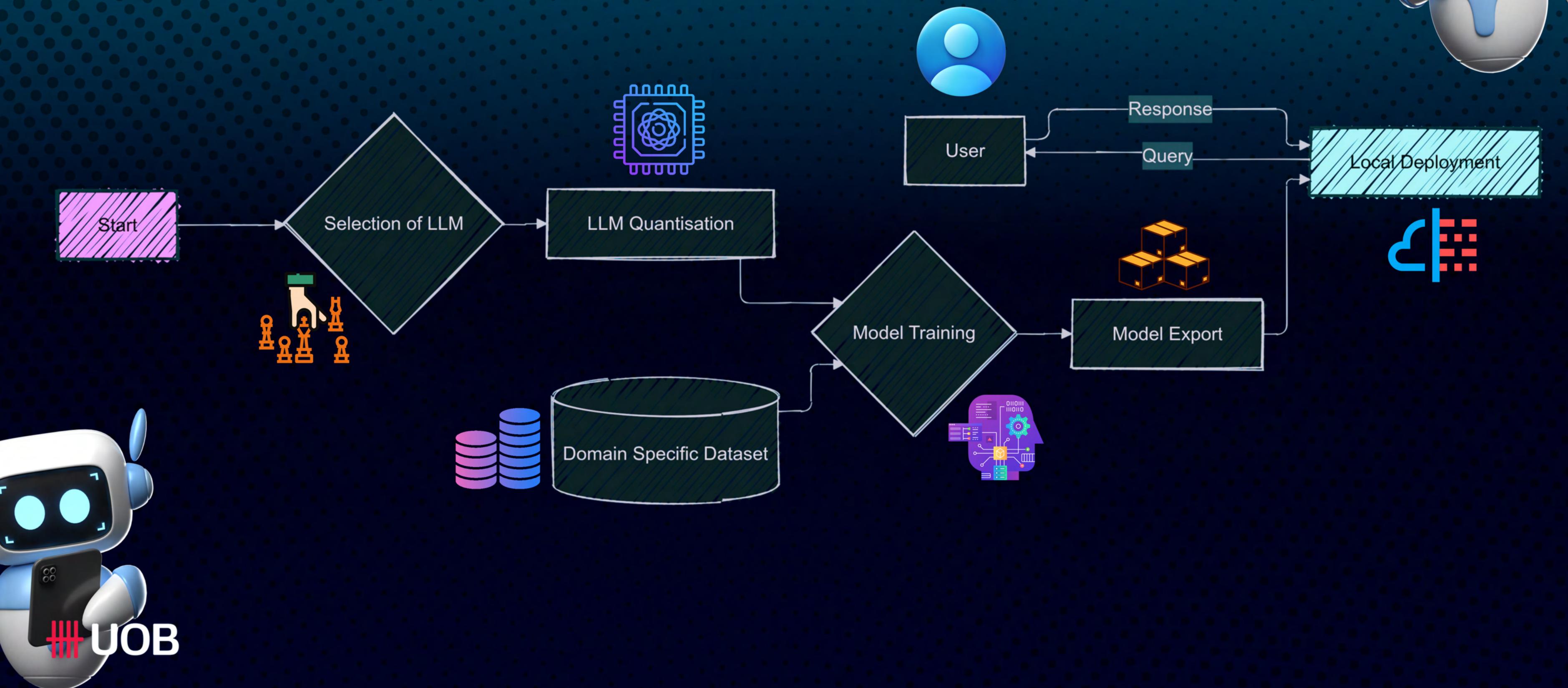
**Ollama by Meta**



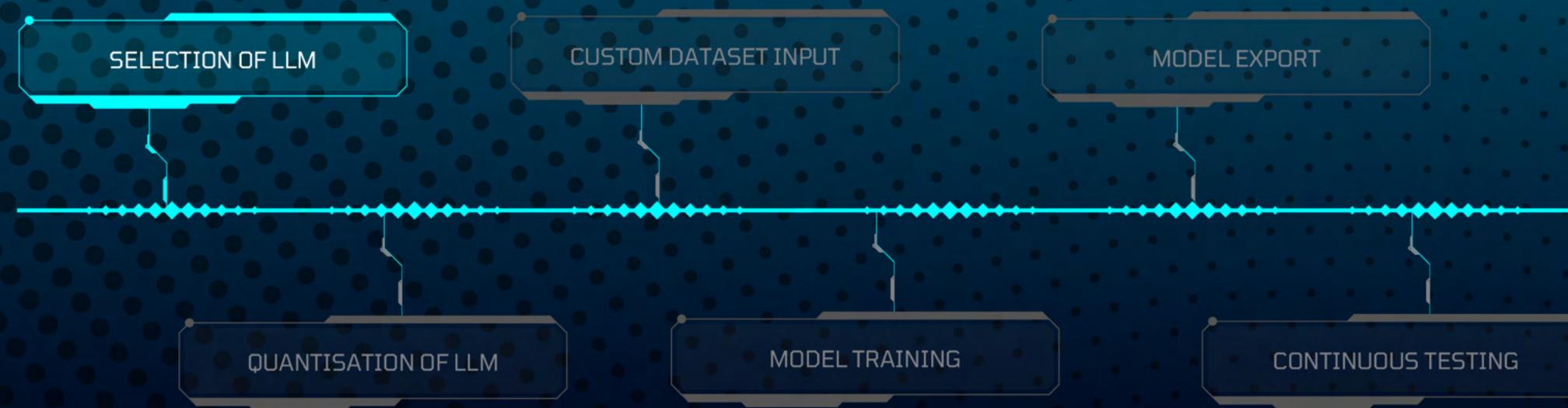
# COMPARISON TO OTHER OPEN SOURCE LLM

	<b>StableLm2</b>	<b>Other LLMs</b>
<b>COST-EFFECTIVENESS</b>	Being a 1.6 billion parameter model, this makes computational training less resource intensive.	LLMs like Llama3 and Mistral are larger in nature (with over 60 billion parameters), making finetuning computationally intensive
<b>PRIVACY &amp; SECURITY</b>	Being a localised LLM, model fine-tuning can be done in a secured environment, where that proprietary information will not be openly shared or available to the public.	Decentralised hosting of open-source models (i.e across multiple environments) without strict controls, which increases the risk for potential breaches or mishandling of sensitive information.

# WORKFLOW FOR FINE-TUNING LLM



# WORKFLOW FOR FINE-TUNING LLM



- Vast selection of open-source models to choose from.
- Pre-trained by the community of users, which enables easy off-the-shelf usage if one's computational resource for training is limited.

The screenshot shows the Hugging Face Model Hub interface. At the top, there is a navigation bar with links for Models, Datasets, Spaces, Posts, Docs, Solutions, and Pricing. Below the navigation bar, a search bar contains the text "Search models, datasets, users...". The main content area displays a model card for "stablelm-2-1\_6b". The card includes the following information:

- Model Name:** stablelm-2-1\_6b
- Downloads last month:** 10,683
- Model Description:** Stable LM 2 1.6B is a 1.6 billion parameter decoder-only language model pre-trained on 2 trillion tokens of diverse multilingual and code datasets for two epochs.
- Usage:** Get started generating text with Stable LM 2 1.6B by using the following code snippet:
- Model Details:** Text Generation, Transformers, Safetensors, 6 datasets, 7 languages, stablelm, causal-lm, Inference Endpoints, arxiv:12 papers, License: other
- Community:** 12
- Actions:** Train, Deploy, Use this

# WORKFLOW FOR FINE-TUNING LLM

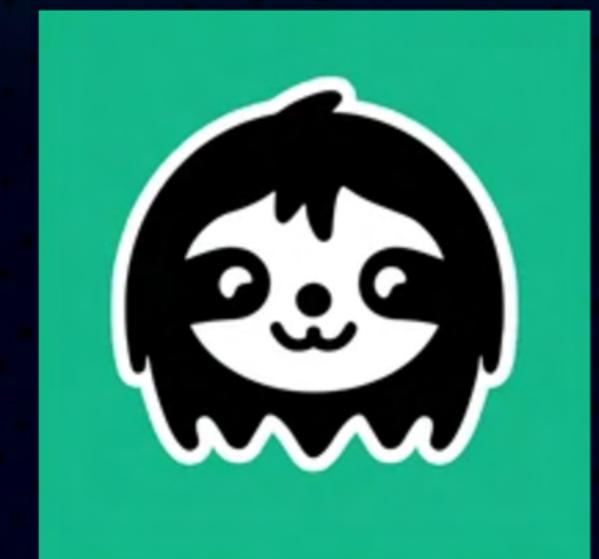
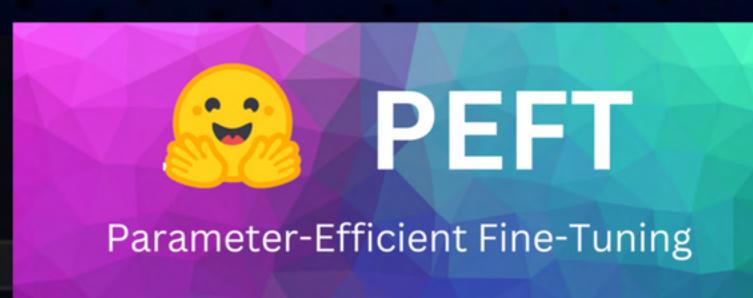


```
# LoRA config
config = LoraConfig(
    r=8,
    lora_alpha=32,
    target_modules=["q_proj"],
    lora_dropout=0.05,
    bias="none",
    task_type="CAUSAL_LM"
)

# LoRA trainable version of model
model = get_peft_model(model, config)

# Trainable parameter count
model.print_trainable_parameters()

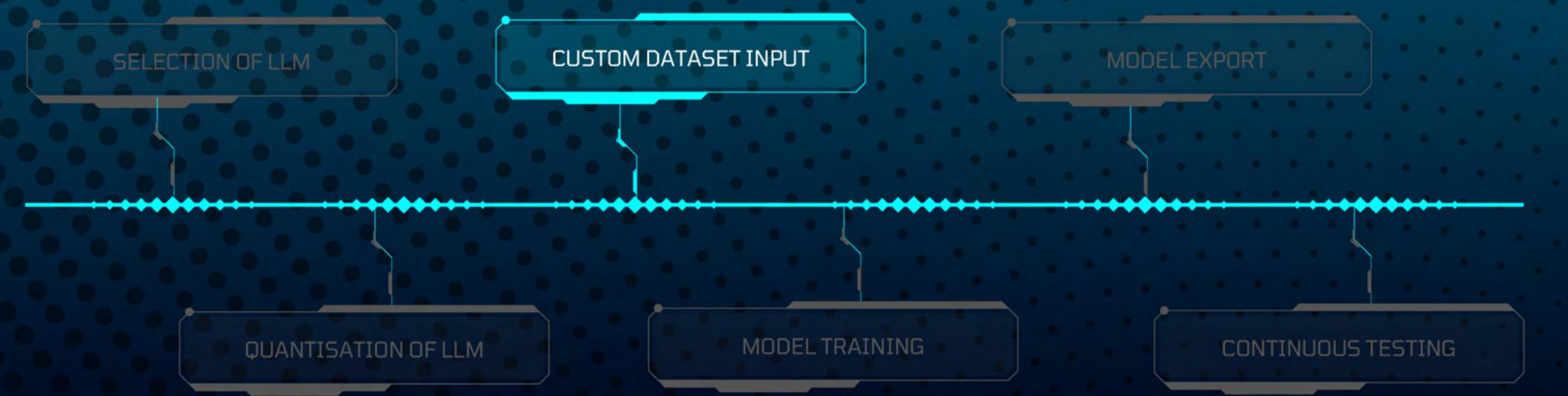
trainable params: 786,432 || all params: 1,645,301,760 || trainable%: 0.0478
```



- **LoRA**

- LoRA (Low-Rank Adaptation) reduces the memory and computational cost of fine-tuning large models by introducing small, trainable low-rank matrices into specific layers, while keeping most pre-trained parameters frozen.

# WORKFLOW FOR FINE-TUNING LLM



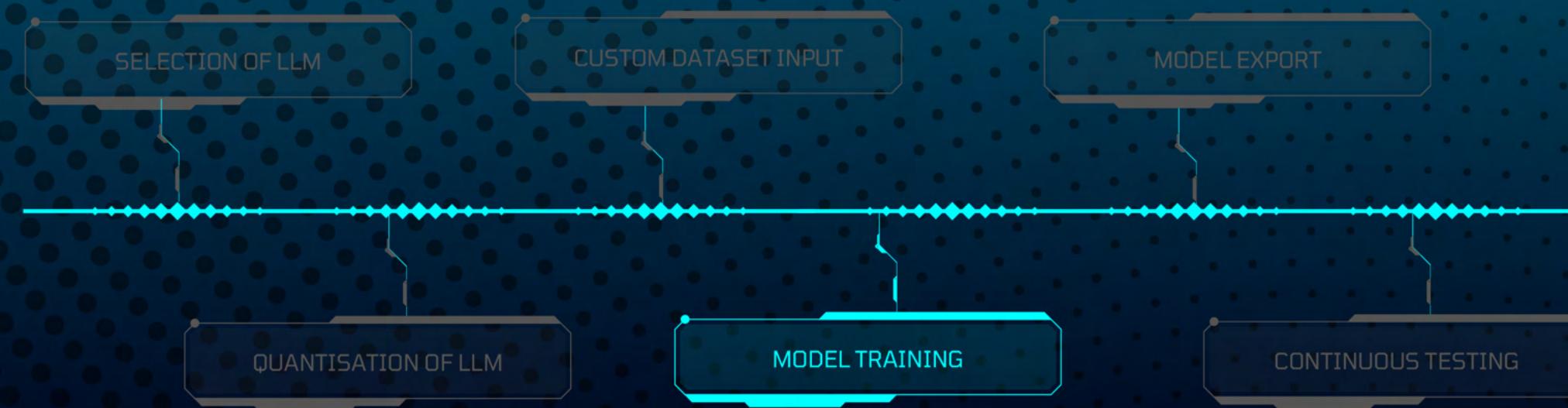
“role-content” format

```
{
  "content": "What documents do I need to apply for a credit card?",  
  "role": "user"  
},  
  
{
  "content": "You'll need your NRIC, latest income documents, and proof of residence to apply for a credit card.",  
  "role": "assistant"  
},  
  
{
  "content": "How can I apply for a personal loan?",  
  "role": "user"  
},  
  
{
  "content": "You can apply for a personal loan through our online banking portal or by visiting a UOB branch.",  
  "role": "assistant"
}
```

- **Custom Dataset**
  - The choice of formatting depends on the framework (Hugging Face, PyTorch, TensorFlow) and the type of model training (e.g., dialogue systems, translation, QA).
  - JSONL and CSV are popular for instruction-based fine-tuning tasks due to their simplicity and structure.

```
DatasetDict({  
    train: Dataset({  
        features: ['conversations'],  
        num_rows: 425  
    })  
    test: Dataset({  
        features: ['conversations'],  
        num_rows: 48  
    })  
})
```

# WORKFLOW FOR FINE-TUNING LLM

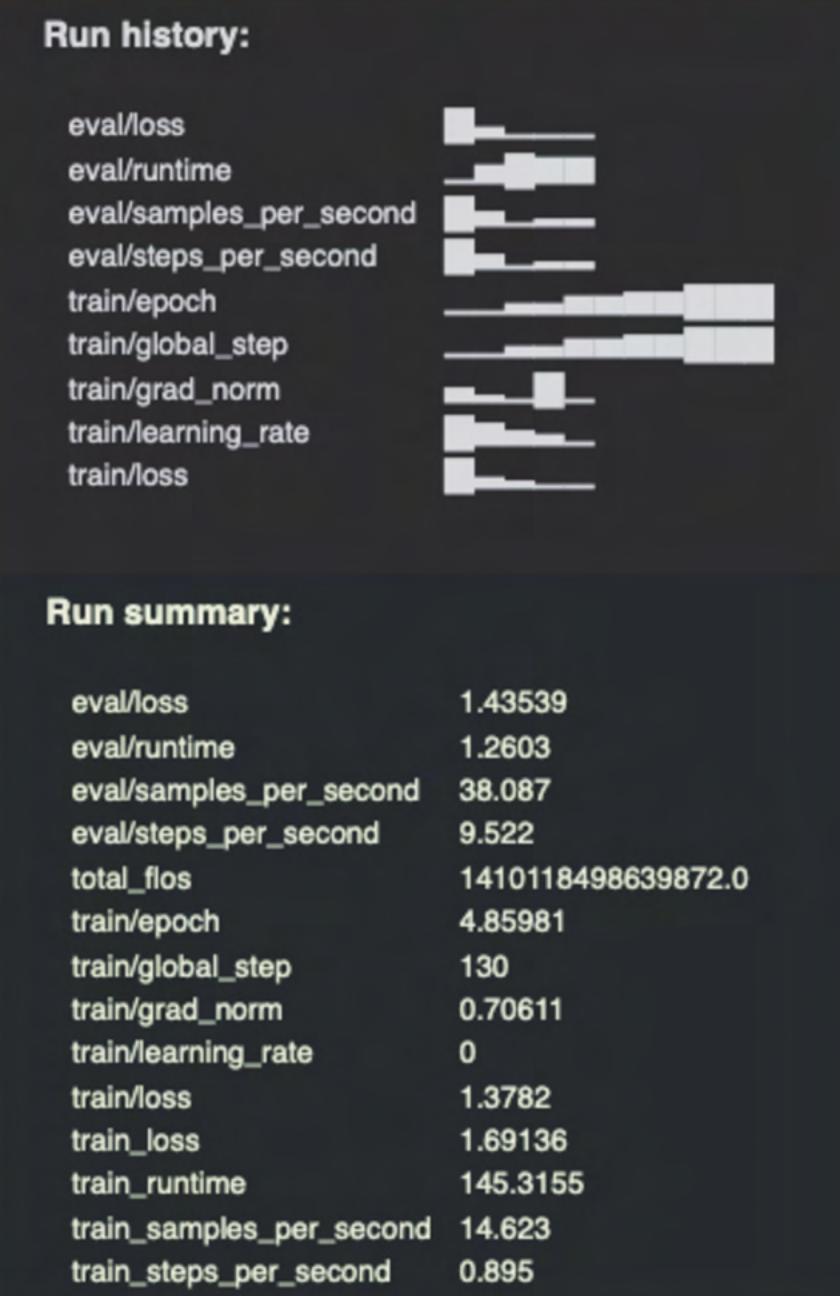


## Model Fine-tuning

- Specification of training arguments in the `TrainingArguments` function.
- Due to limited computational resources, a smaller epoch and learning rate were chosen as runtime parameters.
- All fine-tuning iterations produces training logs, which are uploaded to weights-and-biases (wandb) via an API call.

```
# Define training arguments
training_args = transformers.TrainingArguments(
    output_dir= output_dir,
    learning_rate=lr,
    per_device_train_batch_size=batch_size,
    per_device_eval_batch_size=batch_size,
    num_train_epochs=num_epochs,
    weight_decay=0.01,
    logging_strategy="epoch",
    eval_strategy="epoch",
    save_strategy="epoch",
    load_best_model_at_end=True,
    gradient_accumulation_steps=grad_acc_steps,
    warmup_steps=2,
    fp16=True,
    optim="paged_adamw_8bit",
)

trainer = transformers.Trainer(
    model=model,
    args=training_args,
    data_collator=collate,
    train_dataset=dataset_tokenized["train"],
    eval_dataset=dataset_tokenized["test"],
)
```



# WORKFLOW FOR FINE-TUNING LLM



## **Conversion of trained model to gguf format**

```
!python llama.cpp/convert_hf_to_gguf.py "../LLM Model Training/qLoRA/tdp_stablelm2_ft_merged"
INFO:hf-to-gguf:Loading model: tdp_stablelm2_ft_merged
INFO:gguf.gguf_writer:gguf: This GGUF file is for Little Endian only
INFO:hf-to-gguf:Exporting model...
INFO:hf-to-gguf:gguf: loading model part 'model.safetensors'
INFO:hf-to-gguf:output.weight, torch.float16 --> F16, shape = {2048, 100352}
INFO:hf-to-gguf:token_embd.weight, torch.float16 --> F16, shape = {2048, 100352}
INFO:hf-to-gguf:blk.0.attn_norm.bias, torch.float16 --> F32, shape = {2048}
INFO:hf-to-gguf:blk.0.attn_norm.weight, torch.float16 --> F32, shape = {2048}
INFO:hf-to-gguf:blk.0.ffn_down.weight, torch.float16 --> F16, shape = {5632, 2048}
INFO:hf-to-gguf:blk.0.ffn_gate.weight, torch.float16 --> F16, shape = {2048, 5632}
INFO:hf-to-gguf:blk.0.ffn_up.weight, torch.float16 --> F16, shape = {2048, 5632}
INFO:hf-to-gguf:blk.0.ffn_norm.bias, torch.float16 --> F32, shape = {2048}
INFO:hf-to-gguf:blk.0.ffn_norm.weight, torch.float16 --> F32, shape = {2048}
INFO:hf-to-gguf:blk.0.attn_k.bias, torch.float16 --> F32, shape = {2048}
INFO:hf-to-gguf:blk.0.attn_k.weight, torch.float16 --> F16, shape = {2048, 2048}
INFO:hf-to-gguf:blk.0.attn_output.weight, torch.float16 --> F16, shape = {2048, 2048}
INFO:hf-to-gguf:blk.0.attn_q.bias, torch.float16 --> F32, shape = {2048}
INFO:hf-to-gguf:blk.0.attn_q.weight, torch.float16 --> F16, shape = {2048, 2048}
INFO:hf-to-gguf:blk.0.attn_v.bias, torch.float16 --> F32, shape = {2048}
INFO:hf-to-gguf:blk.0.attn_v.weight, torch.float16 --> F16, shape = {2048, 2048}
INFO:hf-to-gguf:blk.1.attn_norm.bias, torch.float16 --> F32, shape = {2048}
INFO:hf-to-gguf:blk.1.attn_norm.weight, torch.float16 --> F32, shape = {2048}
INFO:hf-to-gguf:blk.1.ffn_down.weight, torch.float16 --> F16, shape = {5632, 2048}
INFO:hf-to-gguf:blk.1.ffn_gate.weight, torch.float16 --> F16, shape = {2048, 5632}
INFO:hf-to-gguf:blk.1.ffn_up.weight, torch.float16 --> F16, shape = {2048, 5632}
...
INFO:gguf.gguf_writer:Writing the following files:
INFO:gguf.gguf_writer:/content/drive/My Drive/TDP Capstone Grp 7/LLM Model Training/qLoRA/tdp_stablelm2_ft_merged.gguf
INFO:gguf.gguf_writer:Writing 100% 3.29G/3.29G [00:43<00:00, 76.4Mbyte/s]
INFO:hf-to-gguf:Model successfully exported to /content/drive/My Drive/TDP Capstone Grp 7/LLM Model Training/qLoRA/tdp_stablelm2_ft_merged
```

**gguf quantisation with Llama.cpp's Q4\_K\_M**

# WORKFLOW FOR FINE-TUNING LLM



# Modelfile to specify pretrained model and general instructions for the LLM

```
# Modelfile generated by "ollama show"

FROM tdp_stablelm2_ft_merged/ggml-model-Q4_K_M.gguf
TEMPLATE "{{ if .System }}<|im_start|>system
{{ .System }}<|im_end|>
{{ end }}{{ if .Prompt }}<|im_start|>user
{{ .Prompt }}<|im_end|>
{{ end }}<|im_start|>assistant
{{ .Response }}<|im_end|>
"
SYSTEM "You are a helpful assistant that is conversant in responding to user queries.
You are trained to only answer UOB related queries, and after each response, gather feedback on the
usefulness of the content provided.
"
PARAMETER stop <|im_start|>
PARAMETER stop <|im_end|>

UOB
SOURCE PATH TO MODEL
INPUT-RESPONSE STRUCTURE OF STABLELM2
SYSTEM COMMAND
```

## Exporting Ollama model locally

```
!llama create betaStableLm2Tdp --file stablelm2.modelfile
```

```
ferring model data :: ferring model data :: ferring model data :: ferring model c  
using existing layer sha256:ba27dcf43c2c2173dff0ba9f16a046e04d2bee1ed6cb0a98f5  
using existing layer sha256:62fbfd9ed093d6e5ac83190c86eec5369317919f4b149598d2  
creating new layer sha256:120976c3fe61e2690c3f3d98567aebb8d98ffcd011403b08a75e  
using existing layer sha256:f02dd72bb2423204352eabc5637b44d79d17f109fdb510a7c5  
creating new layer sha256:bcf18541c31d9e84a767fe9723cf6c1e331920af3d7651fa4bba  
writing manifest  
success
```

NAME	ID	SIZE	MODIFIED
betaStableLm2Tdp:latest	f478475d58b8	1.0 GB	10 seconds ago
mistrail:latest	f974a74358d6	4.1 GB	7 days ago
myOwnStablelm2:latest	1df5ba03896a	982 MB	9 days ago
stablelm2:latest	714a6116cffa	982 MB	4 weeks ago
llama3:latest	365c0bd3c000	4.7 GB	5 weeks ago
llama3.1:latest	91ab477bec9d	4.7 GB	5 weeks ago

4

# POTENTIAL CHALLENGES OF MODEL



## Multilingual Precision & Content Retention

- StableLM2 may not perform equally well across all languages
- Regional slang, idioms or specific dialects may be mis-interpreted
- Limitation in memory retention to track long user conversations



## Bias and Ethical Concerns

- StableLM2 may carry inherent biases from its training data
- Affects fairness in handling customer queries from diverse demographic groups

4

# POTENTIAL CHALLENGES OF MODEL



## Latency in Real-Time Responses

- Due to size and complexity of StableLM2, response times could be slower
- Consider implementing caching mechanisms for frequently asked questions to reduce model's processing speed



## Emotion Detection Intelligence

- Most LLMs lack emotional intelligence
- Results in responses that feel impersonal, especially in high-stake scenarios like fraud detection or financial distress
- Diminishes user satisfaction

4

# POTENTIAL CHALLENGES OF MODEL



## Limited Domain-Specific Knowledge (Fine-Tuning)



- Requires relevant training datasets
- High computational resources (e.g. GPUs/TPUs) and time
- Regular re-training needed to stay-up-to-date with evolving regulations, products or services



## Regulatory, Security and Privacy Concerns

- Data privacy risks such as data breach, or regulatory non-compliance
- Varying regulations across regions adds complexity to the chatbot's deployment
- Implement end-to-end encryption or tokenization to protect sensitive information

# 5

# CURRENT IMPACT & FEASIBILITY



## Automatic Language Detection & Speech Recognition

Highly fluent in English, Spanish, German, Italian, French, Portuguese, and Dutch



**UOB Digital Assistant** ☆ Rate the Bot!

PIN number.

5. Click on "Change PIN" or find it under the "Account Information" section of your credit card details page.

6. Enter a new 4-digit PIN and confirm its validity by entering another code displayed in the next field.

7. Once you have entered the correct information, click on "Save".

If you need further assistance with changing your credit card PIN number or any other queries related to your online banking account, please don't hesitate to let me know. I'm here to help you at all times.

Please make sure that you follow these steps carefully and accurately for the best possible outcome.

**You** 如何在大华银行网站上更改我的信用卡密码?

**Bot** 要更改你的信用卡密码, 在大华银行网站上, 请按这些简单步骤:

1. 登录到您的大华银行 online帐户。
2. 一旦您已经成功进入了您的帐户, 点击"Credit Cards"或"Manage Card"。
3. 在选择你要更改的信用卡之前, 点击"Change PIN"或在您的信用卡详细信息页面中查找"Change PIN"按钮。
4. 尝试输入一个 4 寸字的密码, 并确认其可用性, 通过输入另一次示范的代码。然后, 只需点击"Save"。

如果你在进行更改过程中遇到任何问题, 请随时告诉我, 我将竭力帮助你。

Account Balance Recent Transactions Card Activation Overseas Card Use  
Rewards Fee Waiver Chat with us

Type your message... Speak Send

**UOB Digital Assistant** ☆ Rate the Bot!

**You** Comment puis-je changer le code PIN de ma carte de crédit?

**Bot** Pour changer votre code PIN de la carte de crédit, suivez ces étapes simples : 1. Connectez-vous à votre compte bancaire en ligne. 2. Dans le panier ou la section "sécurité", recherchez un bouton ou une option correspondant au "lancement de PIN". 3. Cliquez sur ce bouton pour déclencher l'opération de changer le code PIN. 4. Entrez votre numéro de carte de crédit actuel lorsque l'appareil vous demande d'indiquer le code PIN. 5. Créez un nouveau code PIN unique et personnalisé. Vous pouvez le trouver en explorant

Type your message... Send

# ⑤ CURRENT IMPACT & FEASIBILITY



# User Feedback for Continuous Improvement



The screenshot shows the UOB Digital Assistant interface. At the top, there's a blue header bar with the UOB logo and the text "UOB Digital Assistant". On the right side of the header, there's a button labeled "Rate the Bot!" with a cursor pointing at it. The main content area has a light gray background. On the left, there's a sidebar with a "Bot" section containing Chinese text and three blue buttons labeled "Account Balance", "Recent T", and "Rewards". The main area contains a large white pop-up window titled "Chatbot Experience Ratings". Inside the pop-up, there's a 5-star rating system with four yellow stars and one gray star. Below the rating is a text input field with the placeholder "What's your experience?". At the bottom of the pop-up are two buttons: a green "Submit" button and a red "Close" button. In the bottom right corner of the main screen, there's a text input field with a microphone icon and a "Send" button.

# 5

# CURRENT IMPACT & FEASIBILITY



**Report Incorrect  
Prompts or Escalate**

The screenshot shows a multi-step interaction with the UOB Digital Assistant. Step 1: A user asks about changing their PIN. Step 2: The bot provides instructions on how to change the PIN. Step 3: An AI response window appears, detailing the steps to check rewards. Step 4: The user reports an incorrect prompt by clicking the 'Report' button, which is highlighted with a red circle and a cursor arrow. The interface includes standard chat controls like 'Type your message...', 'Speak', and 'Send'.

UOB Digital Assistant

PIN number.  
5. Click on "Change PIN" or find it under the "Account Information" section of your credit card details page.  
6. Enter a new 4-digit PIN and confirm its validity by entering another code displayed in the next field.  
7. Once you have entered the correct information, click on "Save".  
If you need further assistance with changing your credit card PIN number or any other query related to your banking account, please don't hesitate to let me know. I'm here to help you at all times.  
Please make sure that you follow these steps accurately for the best possible outcome.

Bot  
要更改你的信用卡密码，在大华银行 online 账户。  
1. 登录到您的大华银行 online 账户。  
2. 一旦您已经成功进入了您的账户，导航到“Credit Cards”或“Manage Card”。  
3. 在选择你要更改的信用卡之前，点击“Rewards & Benefits”或在“Account Information”部分找到它。  
4. 尝试输入一个 4 寸字的密码，并确认。然后，只需点击“Save”。  
如果你在进行更改过程中遇到任何问题，请随时向我寻求帮助。

AI Response  
To check the rewards that UOB credit card offers, follow these simple steps: 1. Log in to your UOB online banking account. 2. Navigate to the "Credit Cards" or "Manage Card" section of your account. 3. Select the credit card for which you want to learn about the rewards. 4. Click on "Rewards & Benefits" or find it under the "Account Information" section of your credit card details page. 5. Browse through the list of rewards and benefits offered by the credit card, including cashback, discounts at retail stores, and travel perks. 6. If you have any questions about a specific reward, click on the corresponding link to learn more. If you need further assistance with checking the rewards or any other queries related to your UOB credit card, please don't hesitate to let me know. I'm here to help you at all times.

Escalate (Speak to Human Agent)

Type your message...

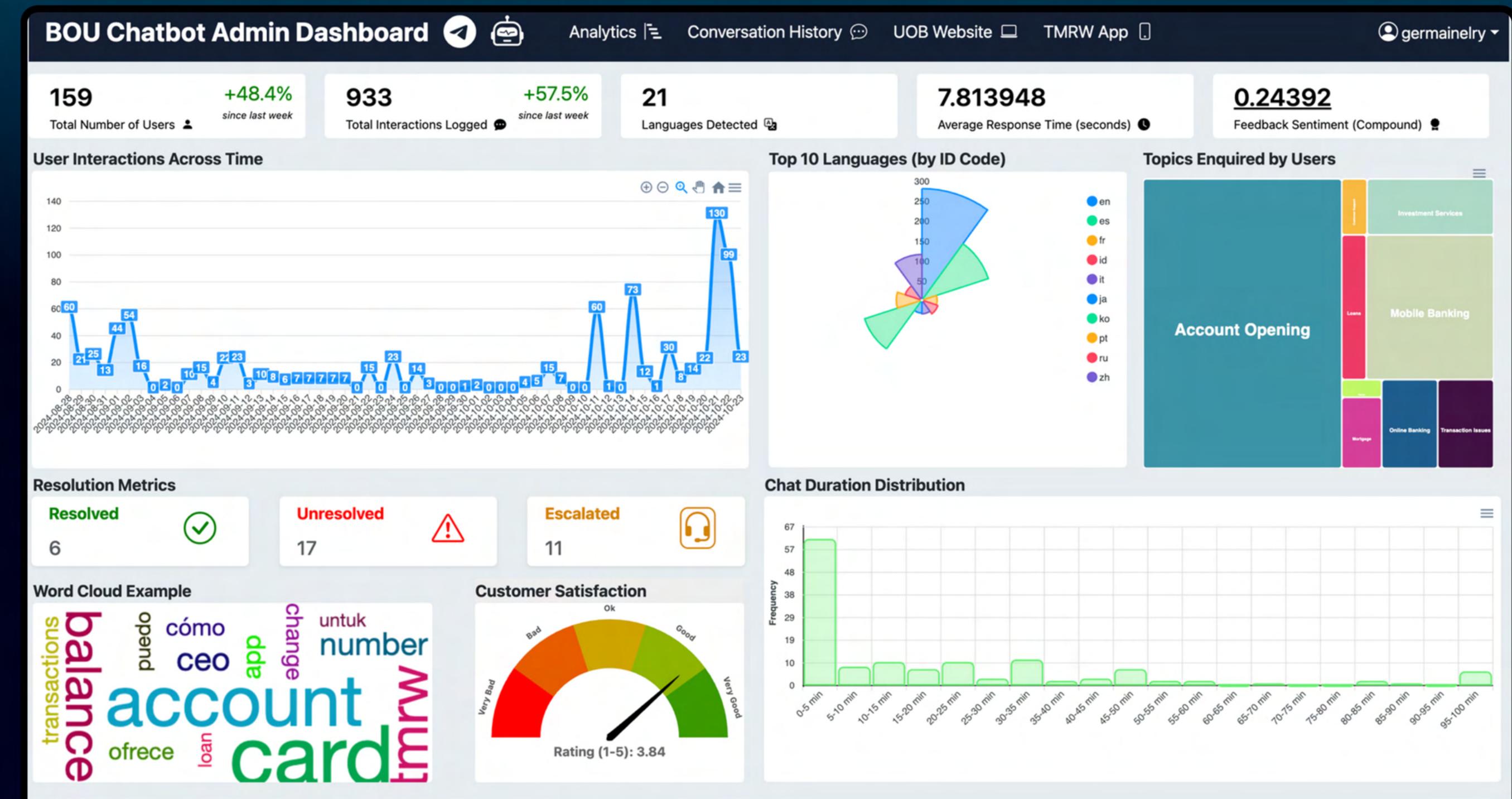
5

# CURRENT IMPACT & FEASIBILITY



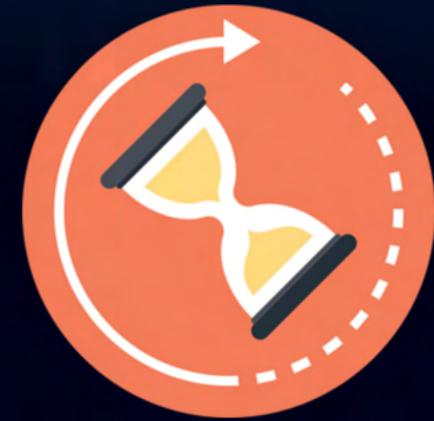
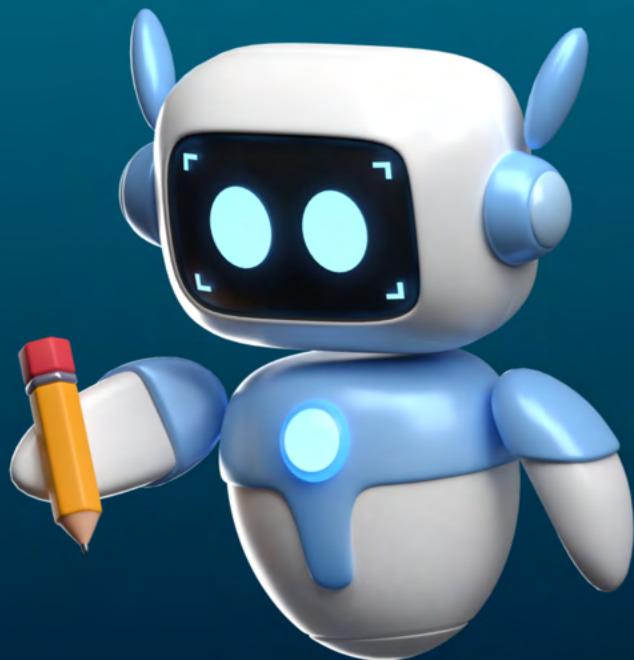
## Real-time Monitoring & Analytics

Quick detection and resolution of issues



# 5

# CURRENT IMPACT & FEASIBILITY



## Fall-back Mechanism

### Implicit



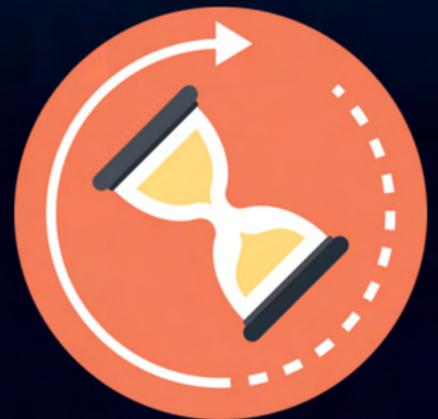
- Happens quietly
- Reformat user input to match known patterns
- If its confidence score drops below a threshold, it will give safer, generic responses

### Explicit



- Predefined error messages to inform user that the chatbot cannot fulfil a request
- Prompts user to escalate issue to live agents via "Speak to Human Agent".

# 5 CURRENT IMPACT & FEASIBILITY



## Fall-back Mechanism

### Explicit fall-back mechanism via LLM's modelfile

- Specification of instructions under the 'SYSTEM' command
- Level of detail and depth of prompt can impact how the LLM will respond to user queries

In the modelfile, we can specify the context:

```
"  
SYSTEM """ You are a helpful United Overseas Bank (Singapore) AI chatbot that  
is capable of handling customer queries. Every response must be detailed and informative.  
In addition, you should avoid answering questions that are not related to banking with UOB.  
"""  
PARAMETER after_error_attempts
```

# 5 FUTURE ITERATIONS



**ML Algorithms for Training**

- Automate training using new data from user inputs and feedback
- To meet evolving customer and compliance needs

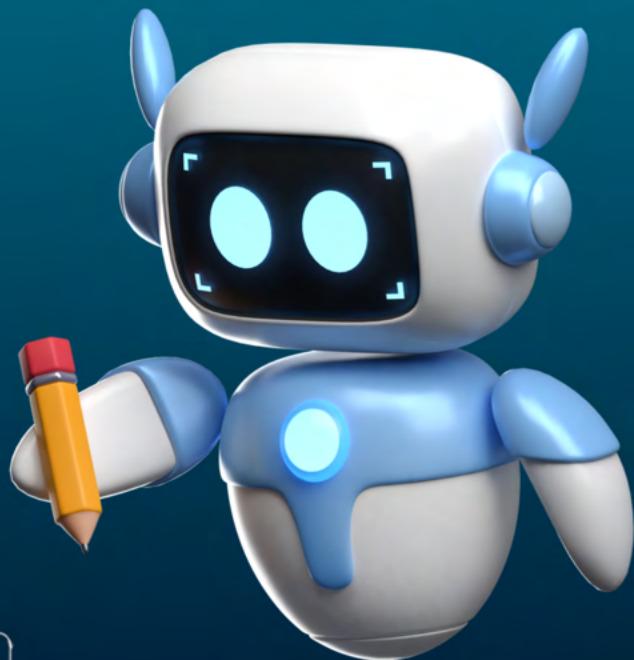
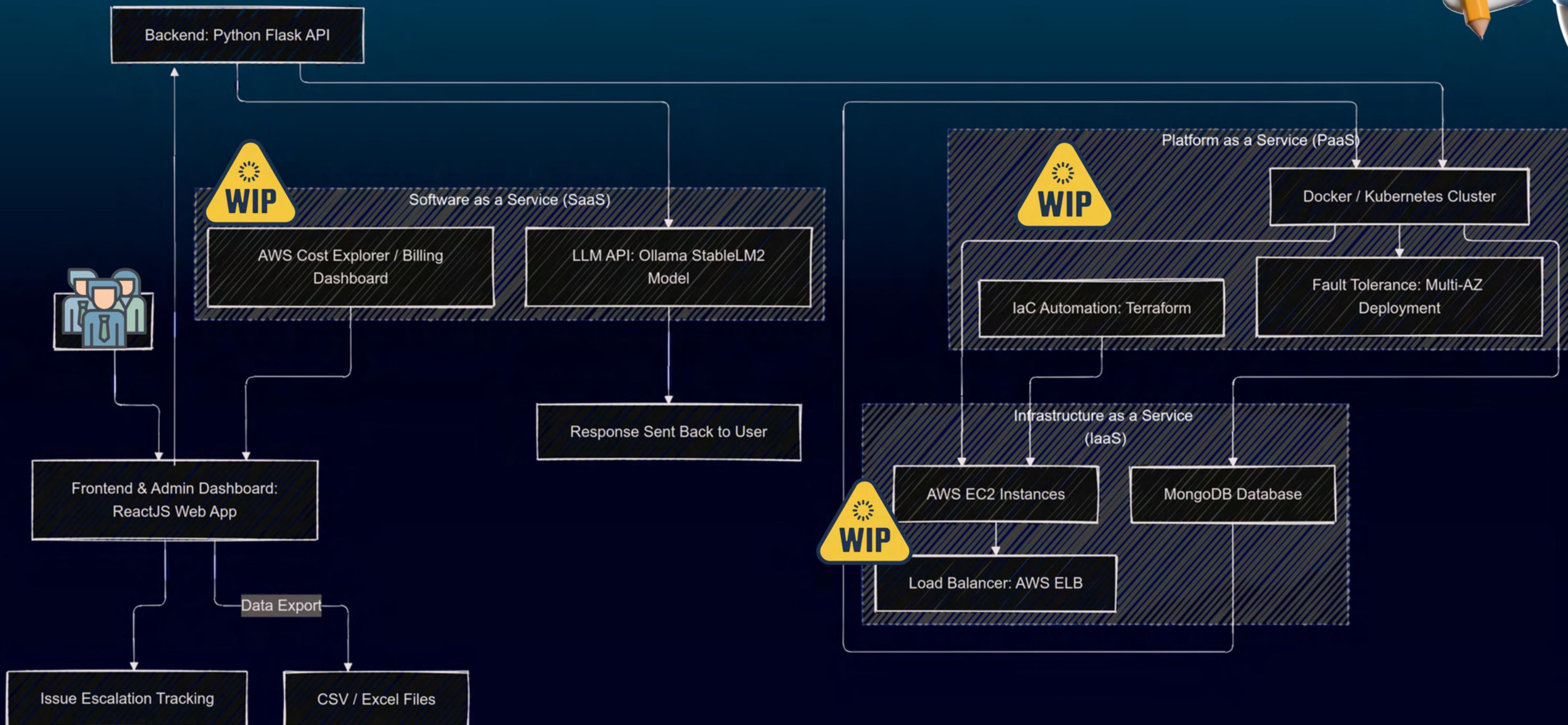


**Enhancing Dashboard Insights with new KPIs**

- Inclusion of more Key Performance Indicators (KPIs) suiting business needs
  - *Uptime and Availability (% of time the chatbot is operational and available)*
  - *Success Rate (% of interactions resolved without human agent escalation)*
  - *Abandonment Rate (% of conversations dropped or exited before completion)*

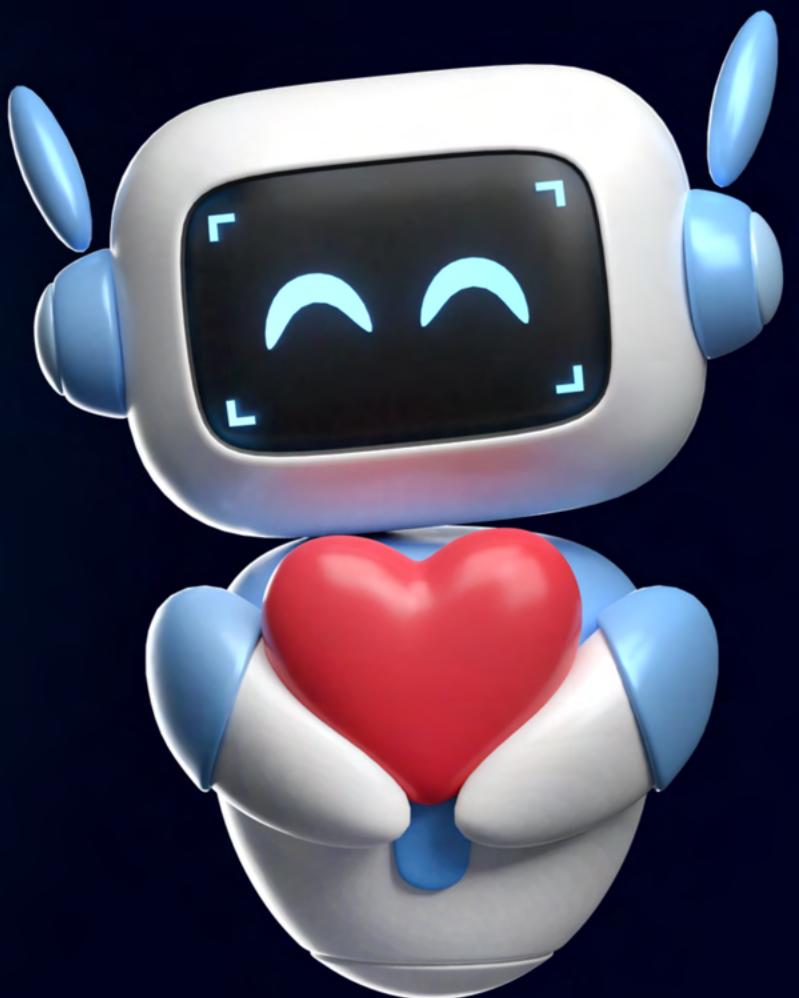
## 5

## CHATBOT ARCHITECTURE SCHEMA



LIVE DEMO

# THANK YOU!



# QUESTIONS?

