

Advanced-level Workshop of LTTC LLM Studio

Developing GenAl-empowered Customised Workflows on Model-as-a-Service (MaaS) Platform

Date: 3 Dec 2024

Time: 2:30 pm – 4:30 pm

Venue: E-P-12 (Metaverse Education Lab)

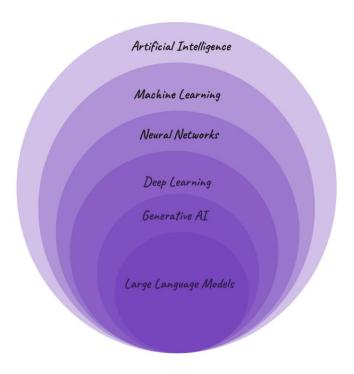


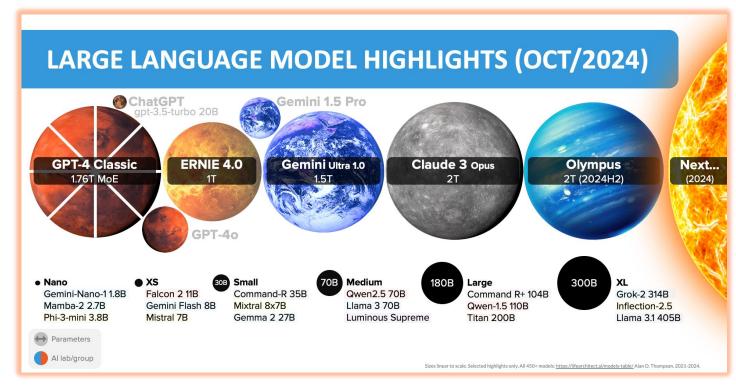


Background

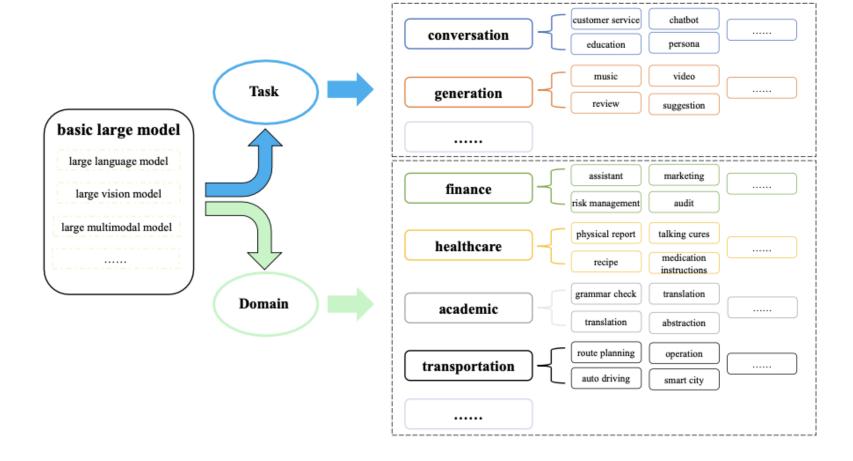


Large Language Models(LLMs)





Service



Beyond chatting with LLMs - MaaS platform





Domain-Specific Al Assistant

Create Al-powered chatbots for specific domains by training models with imported documents or Q&A pairs.



Automated Data Preprocessing

Save time and improve efficiency with automated text preprocessing, vectorization, and QA segmentation.



Workflow Orchestration

Support Al Workflow orchestration,
Design complex workflow using a visual
drag-and-drop interface, integrating tasks
like database queries and inventory
checks.



Seamless API Integration

Seamlessly connect with existing GPT applications and platforms like Discord, Slack, and Telegram using OpenAlaligned APIs.

Common MaaS Platforms











ollama

Coze

vellum













Agents; Model-as-a-Service; LLMOps; Workflow;

Google Vertex AI

Lark

Azure

Fastgpt













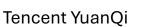
Xunfei Spark

Dify

Beam



腾讯元器





Zhi Pu Qing Yan



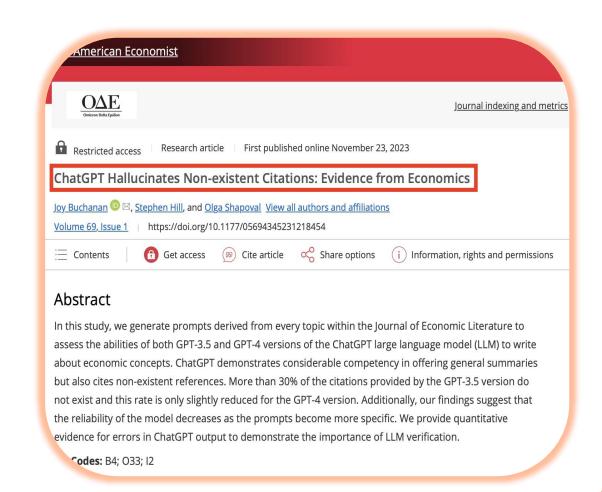
Langflow

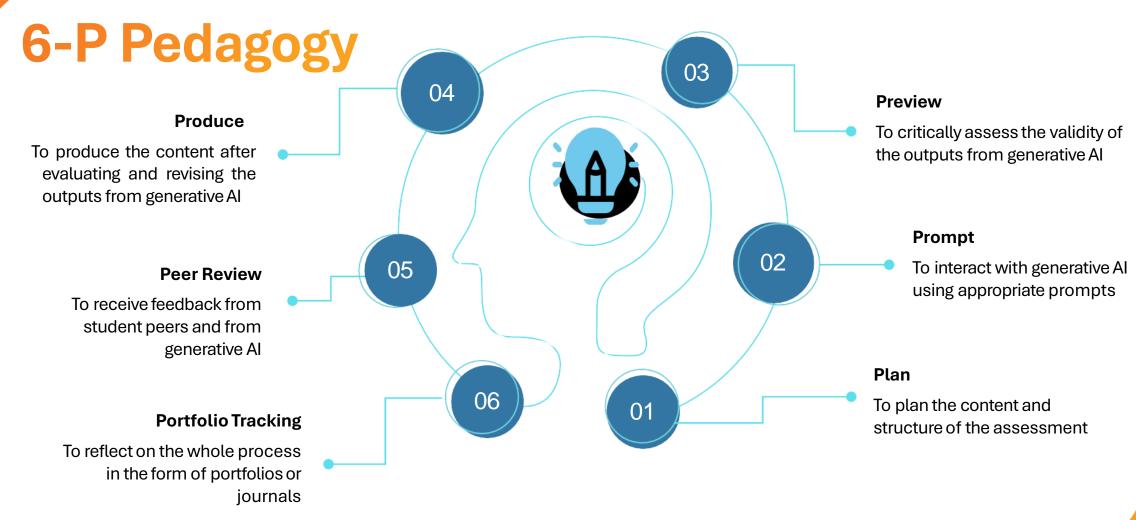


WenXin

Hallucination

- LLMs may produce inaccurate responses, fabricating their own reality.
- All outputs generated by LLMs require additional verification and validation.





Kong, S.-C., Lee, J. C.-K., & Tsang, O. (2024). A pedagogical design for self-regulated learning in academic writing using text-based generative artificial intelligence tools: 6-P pedagogy of plan, prompt, preview, produce, peer-review, portfolio-tracking. Research and Practice in Technology Enhanced Learning, 19(030). https://doi.org/10.58459/rptel.2024.19030

General Guidelines of Using AI for Academic at EdUHK

- Students can appropriately use AI tools to support their personal study and research, but they should be mindful of the possible issues associated with AI tools.
- Students should note the class policies provided by their course instructors and discuss any concerns or questions with them as necessary.
- Students should clearly declare the use of AI Tools in their assessments.
- There is no one-size-fit-all assessment strategy for all courses. It is suggested that the Departmental Learning and Teaching Committee (DLTC) guides course instructors in discussing and designing discipline-specific assessments

Sample Declaration Content

- Al Tools: Specify which Al tools were used in the assessment.
- **Purpose**: Describe the intended used of these AI tools in the assessment.
- Prompts Used: Detail the prompts entered into the AI tools.
- Integration of Outputs: Explain how the AI-generated outputs were integrated into the submitted work

Example 1: Enhancing Academic Language with AI

I acknowledge using EdUHK's ChatGPT (https://chatgpt.eduhk.hk) to enhance the academic language of my own work. I submitted my entire report to the AI tool with the following prompt: 'Check the accuracy of language use in the report'. The output generated by the AI tool was then utilized to correct my grammatical mistakes and improve my style of writing in the report.

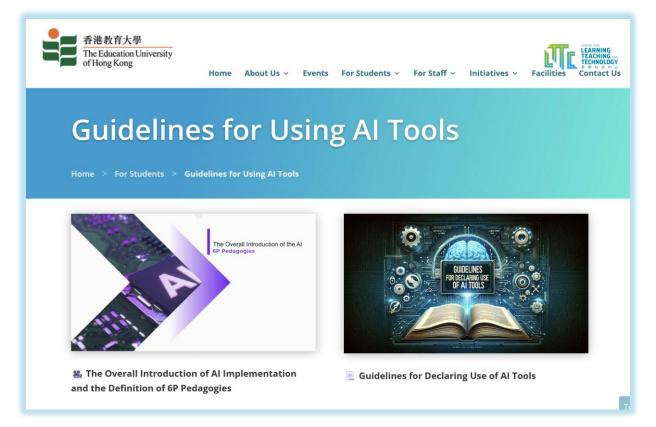
Example 2: Brainstorming Project Ideas with AI

I acknowledge using Poe (https://poe.com/) to serve as a brainstorming aid for generating initial project ideas. I sent the following prompt to the AI tool: 'Generate 5 good project ideas on using technology to support SEN students'. The ideas generated by the AI tool was then critically evaluated for their feasibility, potential for enhancing learning outcomes, and alignment with the current educational needs of SEN students. The best idea, after refinement, was further developed into an actionable project proposal.

CHENG, G. (2024). Guidelines for Using Al Tools. Retrieved from https://www.lttc.eduhk.hk/for-students/guidelines-for-using-ai-tools-students/

Find more information via LTTC website

https://www.lttc.eduhk.hk/for-students/guidelines-for-using-ai-tools-students/



Way to interact with LLM - Prompt Engineering

- Develop and optimise prompts to efficiently use LLMs.
- Apply for a wide variety of applications and research topics.
- Improve the **capacity** of LLMs on a wide range of common and complex tasks(e.g., question answering and arithmetic reasoning).
- Design robust and effective techniques that interface with LLMs and other tools.

Generated Knowledge (Liu et al., 2022) ReAct (Yao et al., 2022) Zero-Shot **Prompting (Wei et Chain-of-Thought** (Wei et al., 2022) **Meta Prompting** (Zhang et al., 2024) **Few-Shot Prompting Graph prompting Tree of Thoughts** (Brown et al., 2020) (Liu et al., 2023) (Yao et al., 2023)

Prompting







Step 2: Double-check of correctness/relevance/reliabili ty of generated contents and their sources



Step 3: Optimise your prompt and ask for more specificity and clarity several times



Step1: Construct TOSOR Structure

- Topics/contents e.g., subject or field
- Operations
 e.g., compose, rewrite, correct, polish,
 summarise
- Sources

e.g., direct input, uploaded image file, open pdf file in the Edge browser, URL opened as a tab in the Edge browser, asking for searching on the Internet, or combined information sources

- Output format/templates
- e.g., directly provide a template/indicate some common layouts, such as "a 200-word paragraph in the introduction section of a journal paper"/indicate formats, such as paragraphs, bullet points, and tables."
- Requirement of the accuracy of information with sources provided
- e.g., must include sources from sources (webpage, pdf), "list all sources in the APA in-text citation format with hyperlink."



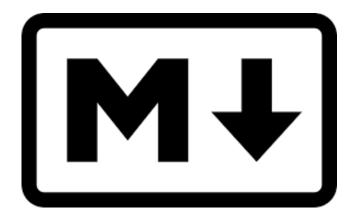
Prompt Example of TOSOR

[Draft the abstract] (Operations) based on [the opened page/the uploaded file] (Source) in [language assessment] (Contents/Topics). The 400-word abstract should include a brief background, research questions, methodology, key results, limitations and conclusion. Do not give me bullet points; instead, several structured paragraphs(Output format/Template). [List all the referred contents of the article for creating the abstract sentence by sentence] (Requirement of the accuracy of information with sources provided).



Prompt formatting

- # Topic: language assessment]
- # Operations:
- Draft the abstract
- If any, find the opened page/uploaded file
- # Output format
- ## 400-word abstract should include
- brief background,
- research questions,
- methodology,
- key results,
- limitations and conclusion.
- ## Do NOT give me bullet points;
- ## Instead, several structured paragraphs
- # Requirement
- List all the referred contents of the article for creating the abstract sentence-by-sentence





Double-check of

- Correctness
- Relevance
- Reliability

of generated contents and their sources (Important!)

Step 3: Optimise the prompt

Review and optimise the input

Specificity and Clarity

What areas are not covered and do they need to be followed up?

Structured Inputs and Outputs

Revise TOSOR, or other prompt structure?

Delimiters for Enhanced Structure

Adopt Markdown,
JSON, or XML?

Advanced Prompting Strategies

Few-Shot Prompting: Providing the LLM with a **few examples** of desired input-output pairs guides it towards generating higher-quality responses by demonstrating the expected pattern.

Chain-of-Thought: Encouraging the model to "think step-by-step" by explicitly prompting it to break down complex tasks into **intermediate reasoning steps** enhances its ability to solve problems that require logical deduction.

ReAct (Reason + Act): This method activates advanced reasoning, planning, and even tool use from the LLM.

Step 3: Optimise the input

Review and optimise the input

Specificity and Clarity

What areas are not covered and do they need to be followed up?

Structured Inputs and Outputs

Revise TOSOR, or other prompt structure?

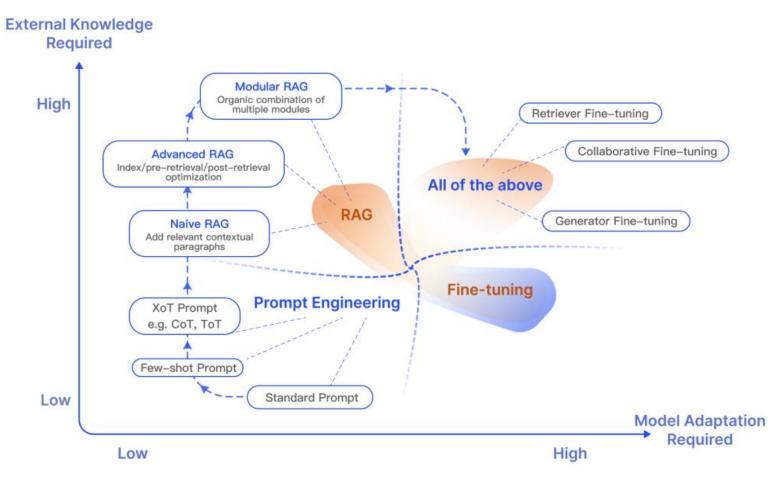
Delimiters for Enhanced Structure

Adopt Markdown, JSON, or XML?

Task Decomposition for Complex Operations

Use advanced techniques?

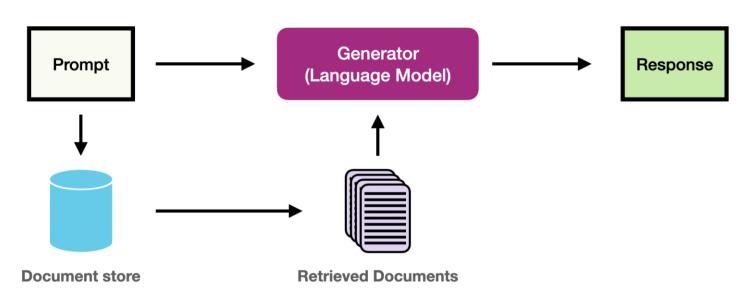
Way to Reduce Hallucinations - RAG



Way to Reduce Hallucinations - RAG

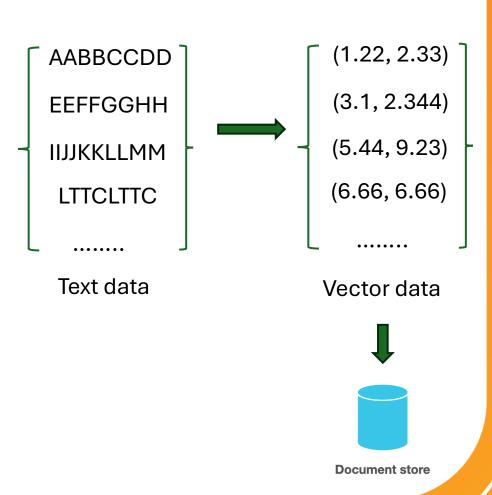
- Provide more contextual and related information for generation
- Retrieval Augmented Generation (RAG)

Why RAG?
No source
Out of date



Knowledge Base for RAG

- Knowledge base is a centralized collection of information that an AI system can access to retrieve accurate and relevant data.
- RAG combines this knowledge base with generative AI models.
- When a user asks a question
- (1) retrieves relevant information from the knowledge base
- (2) uses that information with LLM to generate a more precise and contextually appropriate response.

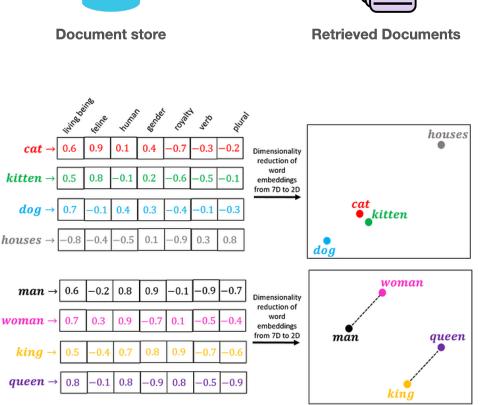


Embedding

Word2Vec is a technique for converting words into numeric vectors that capture the **semantic relationships** between words.

It "understands" the **similarities** and associations between words.

It helps the system more accurately find information that is relevant to the user's query, thus **generating more targeted responses**.



reduction

Word embedding

Visualization of word

embeddings in 2D

Limitations

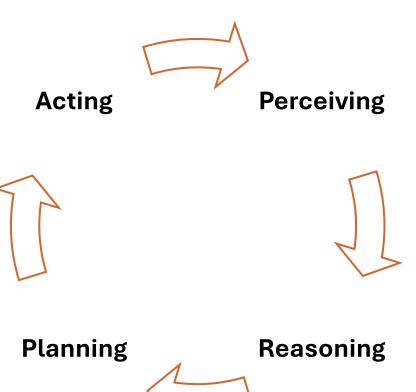


- **1.Bias and Fairness:** Al models can carry forward biases inherent in their training data, leading to unequal outcomes and raising concerns about fairness and equity.
- **2. Complexity and Transparency:** Many AI systems operate as 'black boxes', making it challenging to understand how certain conclusions are derived.
- **3. Dependency:** There is a risk of over-relying on AI solutions, which might overshadow the human aspects of data analysis. Important qualitative factors may be overlooked in favour of algorithmic outcomes.
- **4. Hallucination:** Al models may produce inaccurate responses, fabricating their reality. All outputs generated by Al require additional verification and validation.
- **5. Data Quality:** Ensuring the quality of data used for AI is crucial. Poor data quality can lead to inaccurate results.

How to resolve more specific tasks

- Autonomous intelligent systems to perform many tasks
- Generate diagrams and pictures
- Assessment with designated criteria
- Interactive with real-time news and social media
- Provide guidance and instruction for students using AI step-by-step.
- Apply AI ethically and maintain integrity

•



Features of LTTC LLM Studio MaaS Platform







Large Language Model Studio 大語言模型工作室



MaaS Platform

Create your own smart workflows on the Model-as-a-Service (MaaS) platform by using LLMs



Consultation

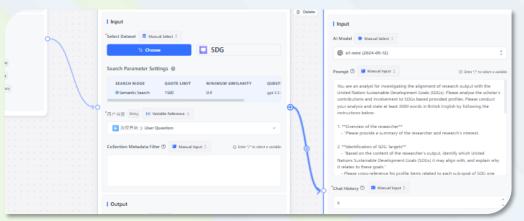
Find solutions and strategies to optimise LLM applications for specific needs



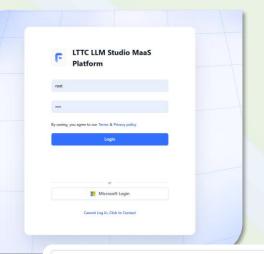
Workshop

Equip practical skills for using LLM-powered tools to meet diverse educational needs

LTTC LLM Studio Model-as-a-Service (MaaS) Platform







Microsoft Login

- Features: Open API to enable utilising LLM, workflow design, data embedding and vectorisation
- access via https://maas.eduhk.hk/
- for non-code customising LLM-included workflows to establish web services for learning, teaching and research purposes
- Enable knowledge-based datasets and embedded URL access to Internet resources

An Open-Source Platform

Develops as an open-source platform with a large user community, complete features and stable performance

Unique Design of Q&A Hierarchy

Splits datasets to create a knowledge base into a paired Q&A format to increase RAG quality

Easy-to-Test

Contains multiple debugging methods such as search testing, reference modification, and full dialogue preview



Multiple LLMs Supported

• //////

Supports one-stop access to various LLMs, including deployment and utilisation of major, custom and localised LLMs



Multiple File Types Supported

Supports mainstream file formats such as Word, Excel, PDF, and Markdown and can quickly build a knowledge base via vectorisation using existing files.







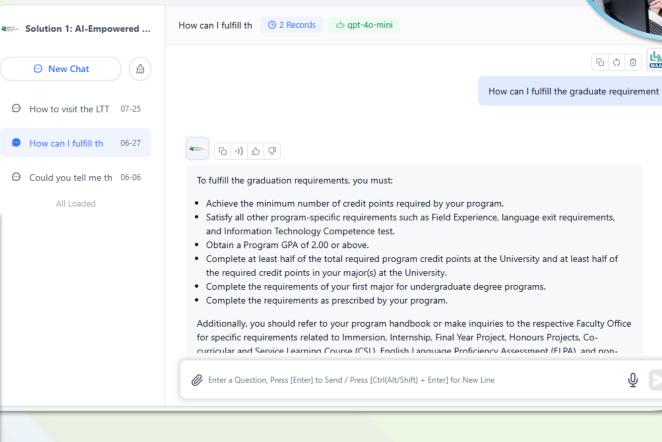
Visualised Workflow

Makes it easy to implement complex workflows, making your own comprehensive AI applications

LTTC LLM Studio Solutions and

Consultations





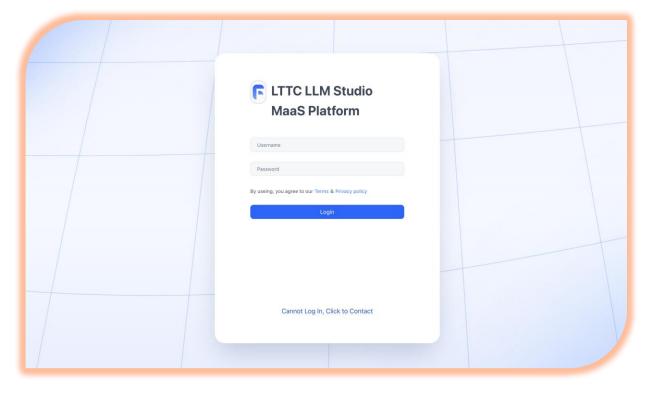


0 0 0

• /////

LTTC LLM Studio MaaS Platform

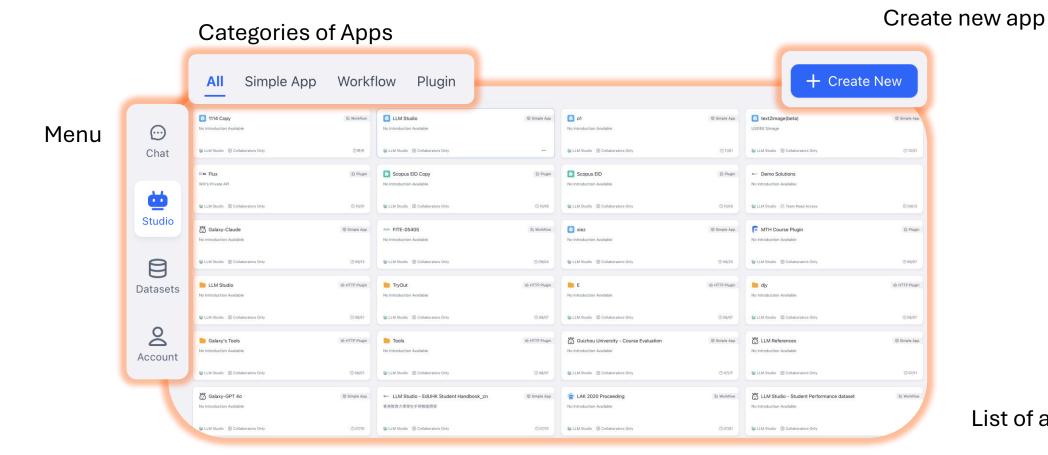
- Private chatbots and Al assistant
- Seamless integration of External Knowledge
- Orchestration of tools and workflow



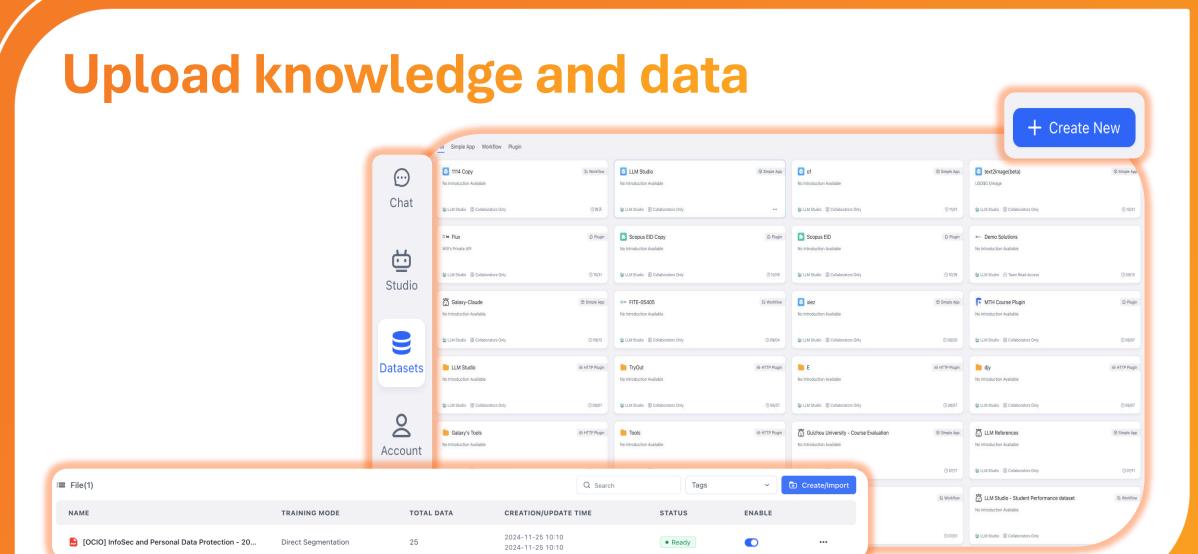
https://maas.eduhk.hk

Account: workshop Password: 123456

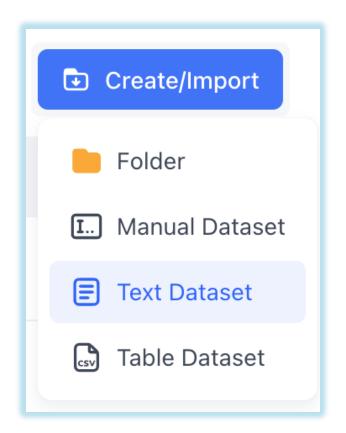
Basic Features of MaaS Platform



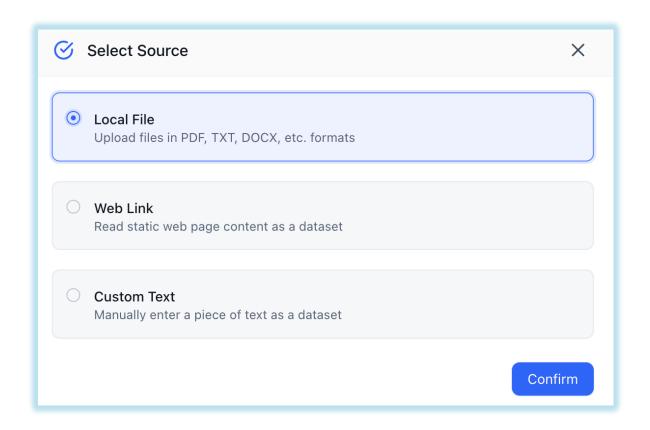
List of apps



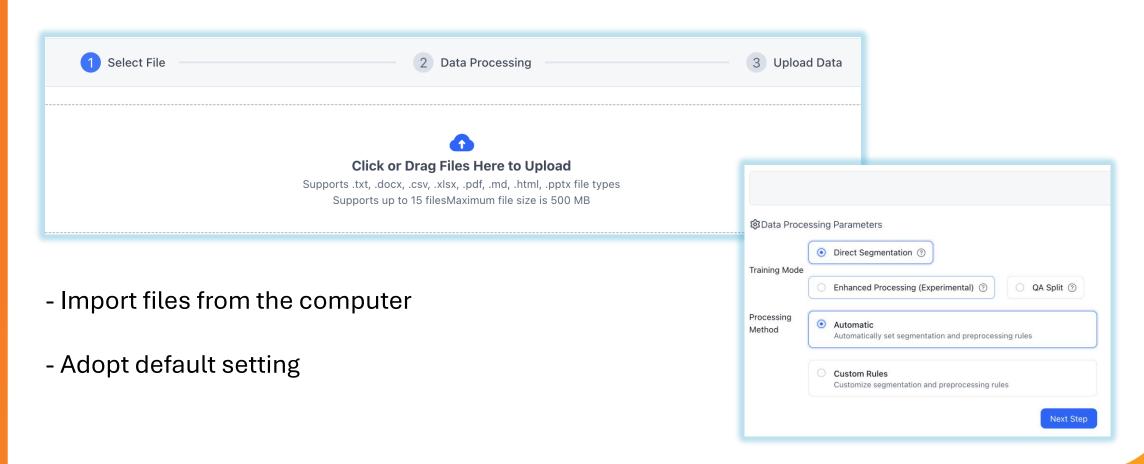
Create a knowledge base for LLM



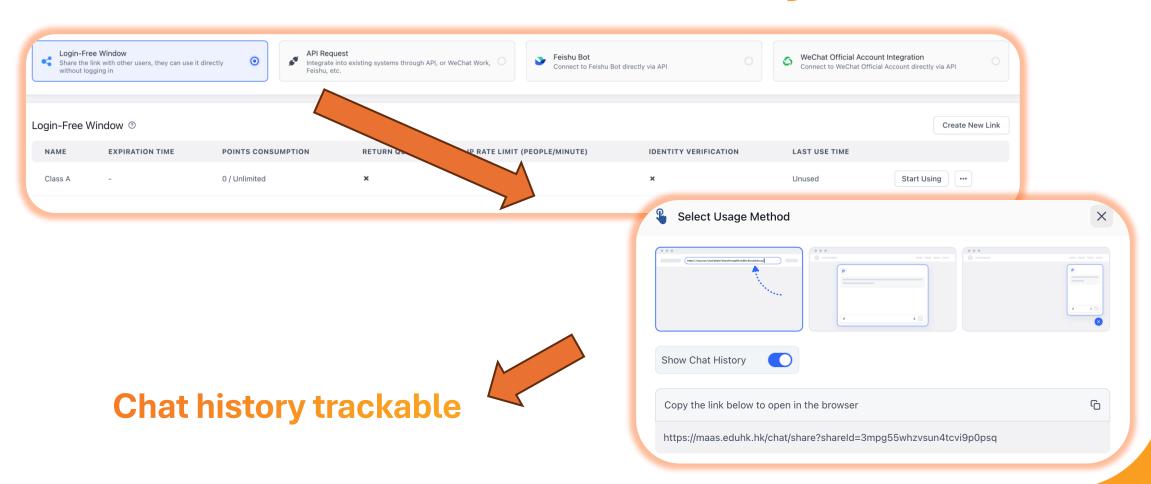




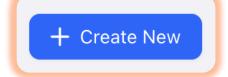
Import the data in the spreadsheet format



Distribute customised Al to anyone

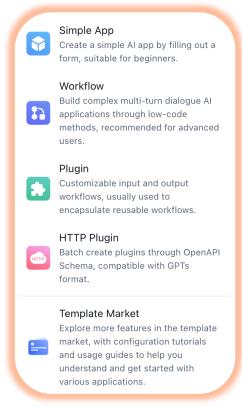


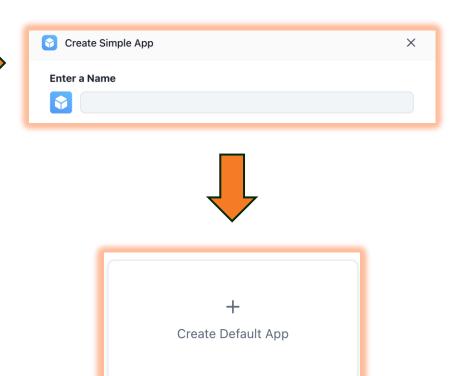
Simple application





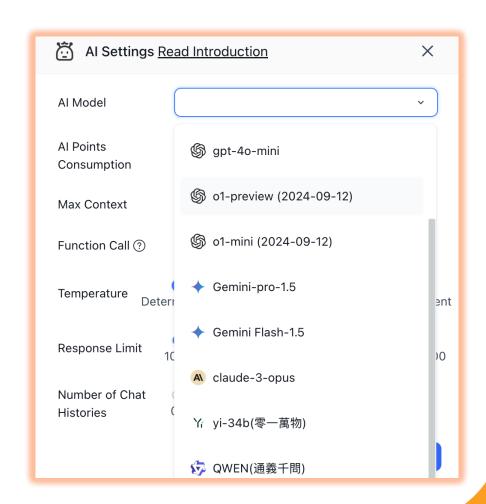
Create your app with name



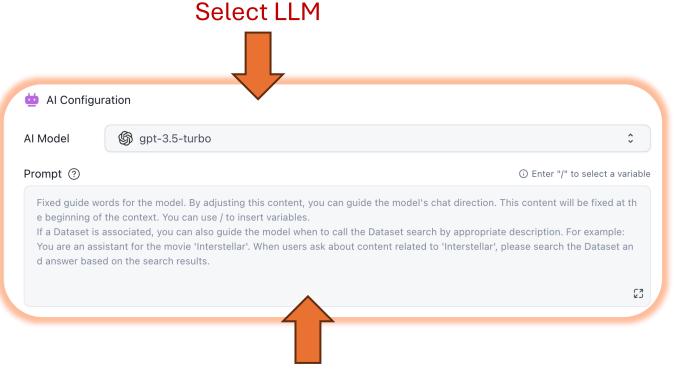


LLM Configuration

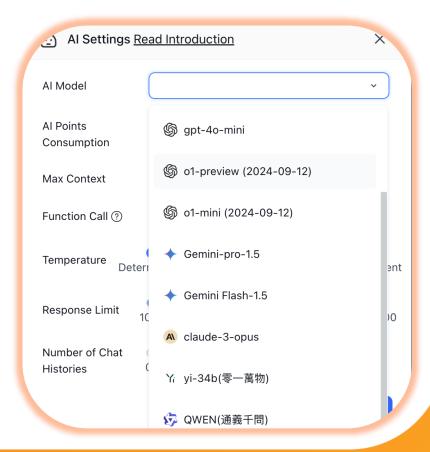
- Temperature: In short, the lower the temperature, the more deterministic the results in the sense that the highest probable next token is always picked.
- Max context: Maximum text to be contained in the LLM (including input and output)
- Response limit: Specifying a max length helps you prevent long or irrelevant responses and control costs.



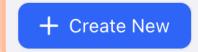
Config LLM by system prompt



System prompt for guiding the LLM

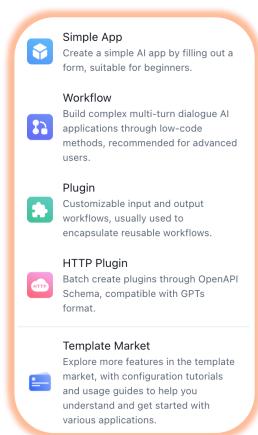


Workflow

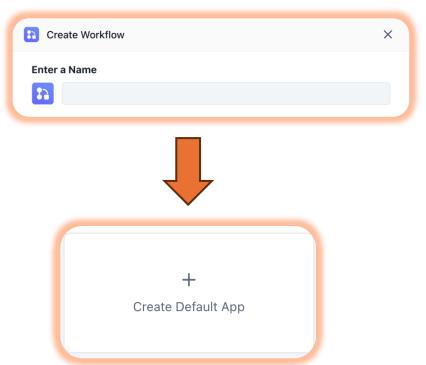




Create your app with name





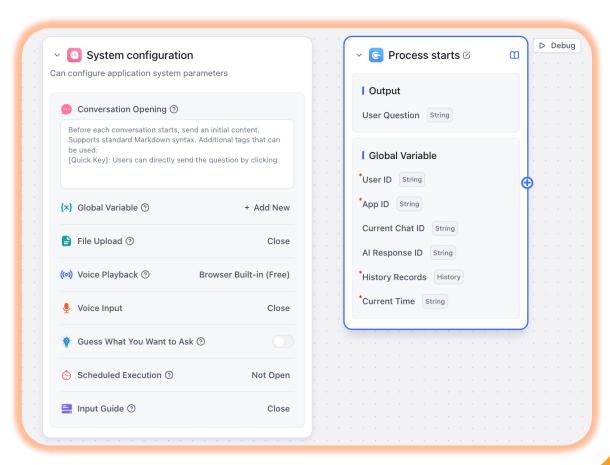


→ ////////

Workflow configuration

Configuration

Start

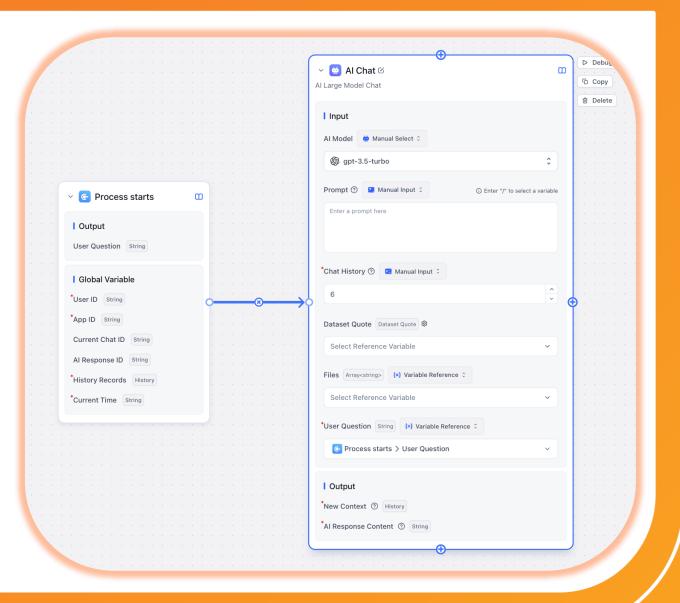


LLMs

Configuration

Start

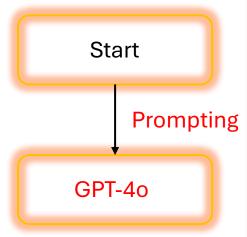
GPT-3.5

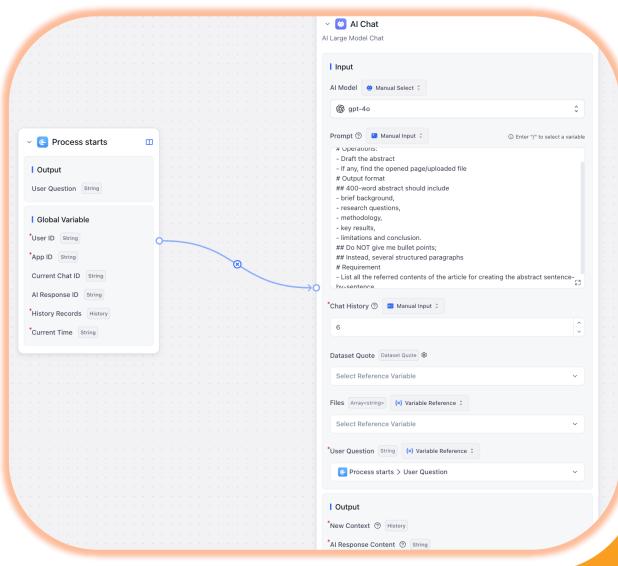


→

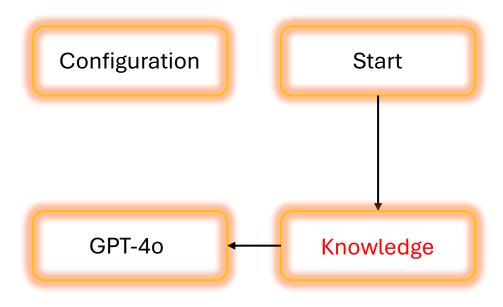
LLMs

Configuration

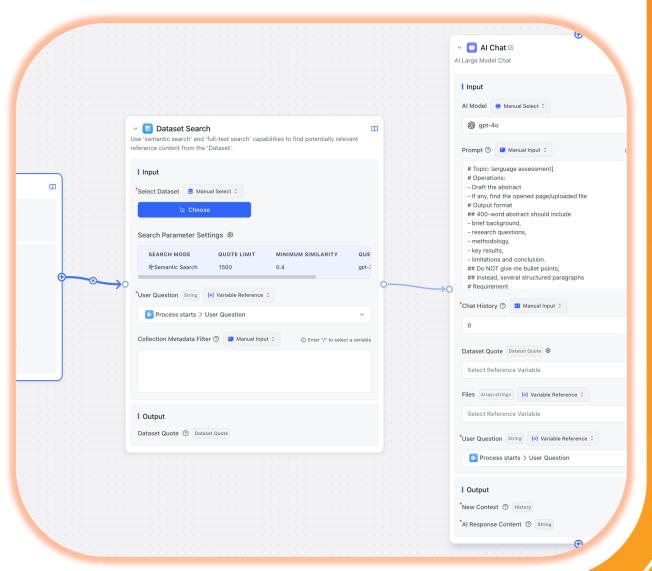




LLMs with RAG



Prompting+Retrieving



Hands-on Practices



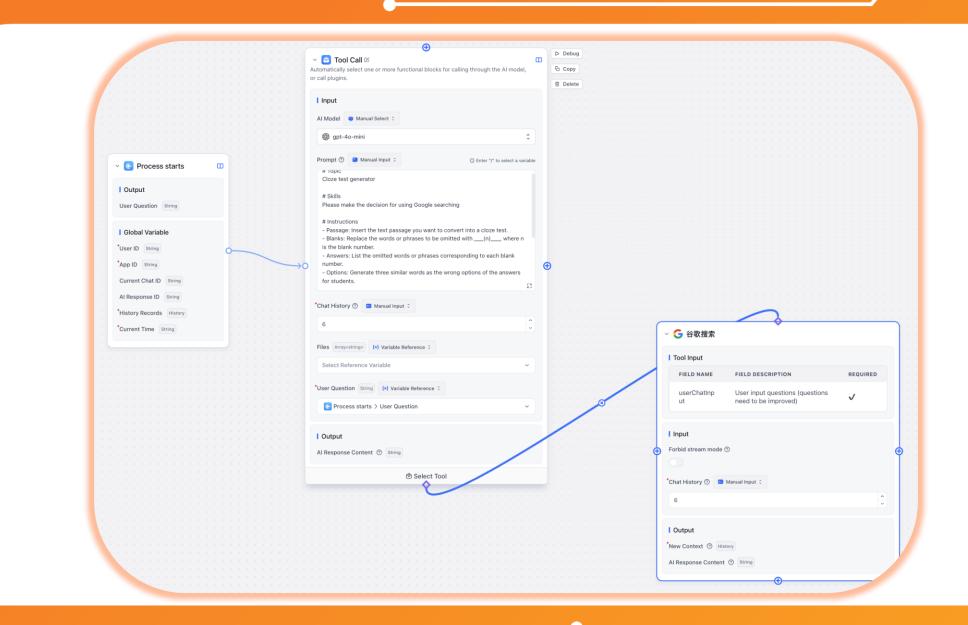
Hands-on Practices

Hands-on Practice 1 (20 mins)

Automated Assessment Generation: cloze tests

Hands-on Practice 2 (10 mins)

Build a knowledge base to enable RAG for your data analysis need



System prompt

Topic
Cloze test generator
Skills

Please make the decision for using Google searching

- # Instructions
- Passage: Insert the text passage you want to convert into a cloze test.
- Blanks: Replace the words or phrases to be omitted with ___(n)___, where n is the blank number.
- Answers: List the omitted words or phrases corresponding to each blank number.
- Options: Generate three similar words as the wrong options of the answers for students.
- # Requirements for Creating Effective Cloze Tests
- Choose Key Vocabulary: Select words that are essential for understanding the passage.
- Maintain Context: Ensure that the remaining text still makes sense and provides enough context clues.
- Vary Difficulty: Adjust the complexity of the omitted words based on your target audience.
- Limit the Number of Blanks: Avoid overloading the test with too many blanks, which can be overwhelming.
- Provide multiple options for each answer and list all the answers, then, bold the right answer (3 wrong and 1 right)

User prompt

Use Google search to search the news on December 2024 in US, and give me a cloze test Here are some of the news highlights from December 2024 in the US:

- Biden Pardons Hunter Biden President Biden pardoned his son, Hunter Biden, who faced sentencing in two criminal cases. <u>Read more</u>.
- Marquette University Ranking As of December 2, 2024, Marquette University was ranked 86th nationally in the latest U.S. News Best Colleges rankings. Read more.
- Best Car Deals Various car manufacturers are offering deals for December 2024, including purchase deals for brands like Acura, Audi, Ford, and more. Read more.
- 4. **DUI Incident** A woman pleaded guilty in a DUI crash that resulted in the death of a bride, marking a tragic event covered in the news. Read more.
- Financial News Various financial institutions are promoting their best private student loans and balance transfer credit cards for December 2024. Read more about student loans and balance transfer cards.
- Visa Bulletin The Visa Bulletin for December 2024 has been released, providing updates for U.S. permanent residents and immigrant visas. Read more.

In December 2024, President (1) pardoned his son, Hunter Biden, who was facing sentencing in two criminal cases. Additionally, Marquette University was ranked (2) nationally in the latest U.S. News Best Colleges rankings. Many car manufacturers are providing attractive (3) for purchase deals, with brands like Acura and Audi featured prominently. A tragic incident involving a DUI crash led to a woman pleading guilty for the death of a (4). On a financial note, institutions are promoting their best private (5) options and balance transfer credit cards, while the Visa Bulletin for December was also released, updating regarding U.S. permanent residents and immigrant (6).

Answers:

- 1. Biden
- 2. 86th
- 3. deals
- 4. bride
- 5. student loans
- 6. visas

Options for Each Blank:

- 1. A. Obama B. Biden C. Trump D. Clinton (Correct: B)
- 2. A. 56th B. 76th C. 86th D. 96th (Correct: C)
- A. sales B. promotions C. deals D. offers (Correct: C)
- A. child B. bride C. friend D. stranger (Correct:
 B)
- 5. A. loans B. grants C. student loans D. scholarships (Correct: C)
- 6. A. travel B. tickets C. visas D. permits (Correct: C)

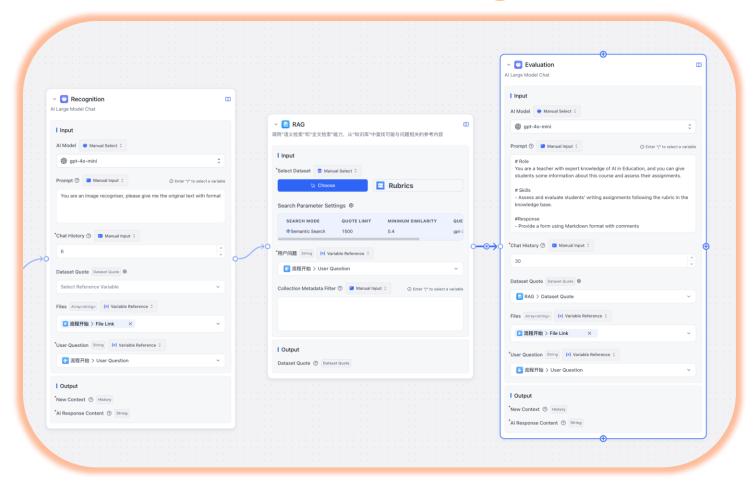
Hands-on Practices

Hands-on Practice 1 (20 mins)
Automated Assessment Generation: cloze tests

Hands-on Practice 2 (20 mins)

Handwriting Recognition and Evaluation

Workflow on Recognition and Evaluation



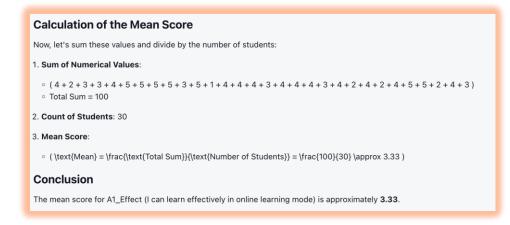
Double-check and review

Hallucination

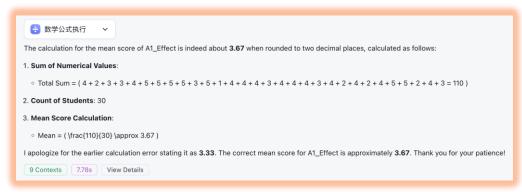
Double-check of

- Correctness
- Relevance
- Reliability

of generated contents and their sources (Important!)



Revised





Evaluation & Post-survey



(1) Control State (1) (The Control State (1) Control State (1)

Thank You

https://eduhk.au1.qualtrics.com/jfe/form/SV_8lbPtoUfvfc0VPo