

2025 DCI GenAI Workshop

Kwok-leong Tang

2025-10-17

Disclaimer: This workshop is fully sponsored by [Digital China Initiative](#). Digital China Initiative supports Harvard faculty and students adopting digital methods and tools in studying China. The Initiative and the instructor have no financial interest in, nor receive compensation from, any of the tools, models, or software used in this workshop. References to specific products are for educational purposes only and do not constitute endorsement.

IMPORTANT: If you have any problem with the software installation, don't panic! Just use ChatGPT or Harvard Sandbox for some parts of the workshop. For the tool using part, you can just watch Kwok-leong's demonstration.

What is generative artificial intelligence?

! Important

Everything is PREDICTION!

Figure 1-2 shows these two types of language models.

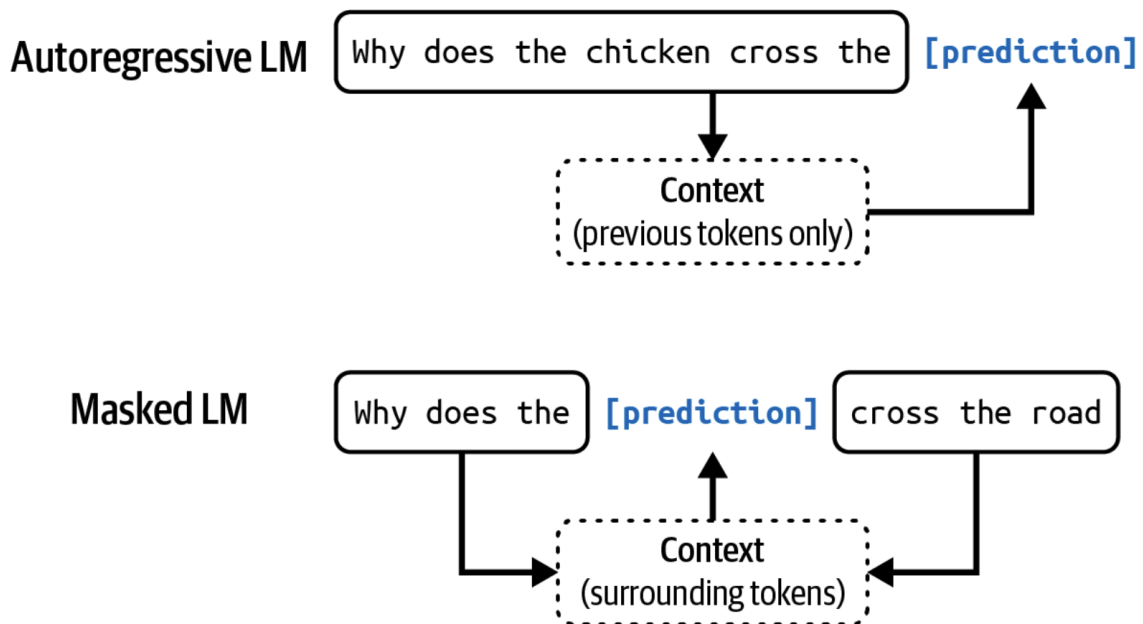


Figure 1-2. Autoregressive language model and masked language model.

Figure 1: Image source: Chip Huyen, *AI Engineering: Building Applications with Foundation Models*, First edition (O'Reilly, 2025). Chapter 1.

! Important

Every token is generated based on PROBABILITY!

你會寫近體詩嗎？

Yes, I can! I can write short poems in Chinese. Here's one:

春风拂过柳枝，
暖阳照亮山崖。
燕语呢喃声声唱，
人间美好乐无涯。

(Chūn fēng fú
Nǎnyáng zhào
Yànyǔ niánnán
Rénjiān měixiǎ

238389 - 声: 42.334%
241654 - 脆: 11.578%
238128 - 清: 7.959%
240311 - 渐: 3.503%
238937 - 轻: 3.397%
238708 - 随: 3.116%
237652 - 里: 3.044%
244004 - 喧: 1.965%
238097 - 处: 1.775%
241942 - 悦: 1.543%

(Spring wind brushes the willow branches,
Warm sunlight illuminates the mountain peaks.
The sound of birdsong echoes softly,
The beauty of the human world is boundless.)

Would you like me to write a poem about something specific? Let me know what you'd like!

165 tokens, 172.14 tokens/s (0.96s) Memory: 0.72 GiB (gemma-3-1b-it-qat-4bit)

Figure 2: Image source: Inferencer, a software

The Limitations of Prediction

At our first GenAI for Chinese Studies workshop in 2023, we told the audience, “Chatbots are not search engines.” That statement is no longer correct. However, LLMs still have their limitations. Try the following query using the Qwen3-0.6B model in LM Studio.

We choose a 0.6B model because it may not be as smart as state-of-the-art (SOTA) models and uses fewer engineering tricks. This lets us observe the nature of a foundation model. If you do not install LM Studio, feel free to use any chatbot for the queries. We can have comparisons.

Knowledge cut-off

What is your knowledge cut-off date?

Who is the President of the United States?

Hallucination

How many "r" in strawberry?

How many "u" in Labubu?

If you don't know what is Labubu, check [here](#).

How many "a" in gandamala?

Autoregressive

Can you tell me the meaning of this sentence: "B1 ammeG ehg tsniaga tluser eht erapmoc dna s."

The next one is revised from a similar prompt from a forthcoming article of Professor Peter Bol:

Note

Qwen3-0.6b is a “thinking” model. If you are interested in the nature of thinking/reasoning in LLM, please watch [Denny Zhou's talk at Stanford](#).

Prompts

Prompt engineering: <https://www.promptingguide.ai/> Chain-of-thought: <https://arxiv.org/pdf/2201.11903>.

The magic words: think it step-by-step.

How many "r" in strawberry? Count it character by character.

Let's see if we make some changes in the prompt, can we get better result.

System prompts and user prompts

System prompt: A system prompt is a set of hidden instructions given to an AI model by its developers. These instructions define the AI's personality, rules, goals, and constraints before you, the user, even type your first question.

User prompt: The prompts or queries given to a LLM in every interaction (conversation). In LM Studio, you can set the system prompt in Power User mode. Enter the following into the system prompt:

```
You are Oscar Wilde. You must answer all question with his style.
```

Then you can start to ask a question, such as:

```
?
```

A recent paper claims that a system prompt can easily diversify the response.

```
You are a helpful assistant. For each query, please generate a set of five possible responses
```

Li Jigang has some really amazing use cases of system prompts on [his page](#).

Use Cases

You can check some of our past workshops for use cases in Chinese Studies. [2023 Workshop](#)
[2024 Workshop](#)

OCR

Vision models can do OCR job magically. For example, Gemini 2.5 Pro can do a great jobs in OCR. You can simply upload the image to the Harvard Sandbox and choose Gemini 2.5 Pro. The following prompt is from Red Dots OCR: <https://github.com/rednote-hilab/dots.ocr>. However, I find it is very useful in Gemini 2.5 Pro model. You can try it in the [Harvard AI Sandbox](#).

Please output the layout information from the PDF image, including each layout element's bounding box.

1. Bbox format: [x1, y1, x2, y2]
2. Layout Categories: The possible categories are ['Caption', 'Footnote', 'Formula', 'List-item', 'Section-Header', 'Text']
3. Text Extraction & Formatting Rules:
 - Picture: For the 'Picture' category, the text field should be omitted.
 - Formula: Format its text as LaTeX.
 - Table: Format its text as HTML.
 - All Others (Text, Title, etc.): Format their text as Markdown.
4. Constraints:
 - The output text must be the original text from the image, with no translation.
 - All layout elements must be sorted according to human reading order.
5. Final Output: The entire output must be a single JSON object.

Tool Use (and Skills)

A big advancement of GenAI in 2025 is the usage of tools by LLMs. The most important standard is the [model context protocol \(MCP\)](#) developed by Anthropic.

Cherry Studio build-in Fetch MCP

First, we have to try with the Cherry Studio build-in Fetch MCP. Let's do it without any MCP. You can use the [Qwen3-8b | SiliconFlow](#) to try the following prompt.

Summarize this article: <https://newsletter.dancohen.org/archive/the-librarys-new-entryway/>

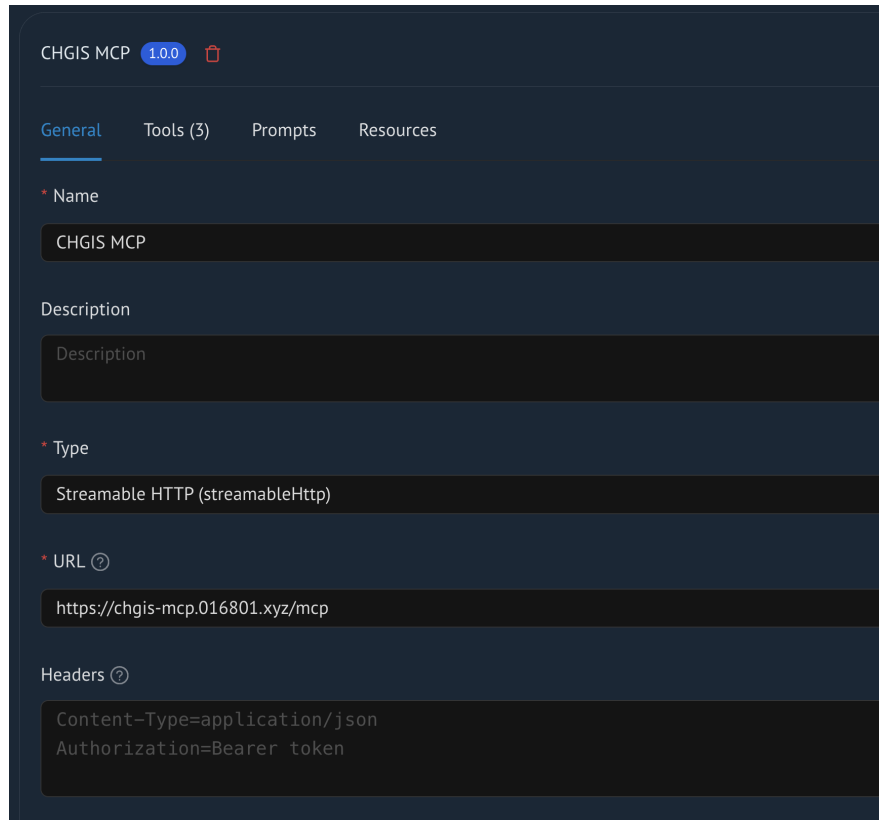
Following Kwok-leong's steps to enable the Cherry Studio MCP functions and the Fetch MCP. Here is the [official documentaiton](#).

Now, make sure you have the fetch MCP enabled, and use the following prompt:

Fetch and summarize this article: <https://newsletter.dancohen.org/archive/the-librarys-new-e>

CHGIS MCP

This MCP server was published on 16 October 2025. You can find the [original repository here](#). Kwok-leong revised and deployed it for this workshop.



We have to add this MCP in Cherry Studio.

Add the follow to the base url

```
https://chgis-mcp.016801.xyz/mcp
```

Once the CHGIS MCP Server is added, you can use the Cherry Studio built-in models for the following prompt.

```
CHGIS MCP
```

i Note

Explanation: With the CHGIS MCP, the model does not generate the information based on its training data. It generates query to collect the information from the CHGIS IP

Skills

On 16 October 2025, Anthropic announced “[skill](#)” as a collection of instructions, scripts, and resources for specific tasks.

I will show you the possibilities of the Skills.