介绍Smart Prompt框架

汇报人：庄表伟

时间：2024.11

CONTENTS

目录

如何理解这一轮AI浪潮？

01

什么是LLM?

- **定义：**
  - LLM是通过深度学习训练的大型神经网络模型，专门用于理解和生成自然语言。
- **工作原理：**
  - 通过大规模的文本数据学习语言结构、语法和语义。
  - 具备处理多种语言任务的能力，如翻译、文章生成、问题回答等。
- **应用场景：**
  - AI对话系统
  - 自动翻译
  - 内容创作
  - 自然语言处理（NLP）

什么是提示词?

- **定义：**
  - 提示词是输入给LLM的文本，用于引导模型生成相关输出或执行特定任务。
- **功能与重要性：**
  - 清晰、具体的提示词能提高模型的回答质量和准确性。
  - 模糊、不明确的提示词可能导致不准确的输出。
  - 不同的模型（开源的/闭源的，小的/大的），针对不同的任务，不同的提示词，可能有不同的效果
- **示例：**
  - 提示词1：“请解释什么是机器学习。”
  - 提示词2：“翻译以下句子为法语：‘Hello, how are you?’”
  - 提示词3：“告诉我这个出错信息是什么意思，我要怎么解决？”

什么是可编程的提示词?

- **定义：**本质上是一种传统架构与AI架构集成起来的**架构模式**
  - 并非所有的问题，都适合用LLM来解决
  - 在选择调用何种LLM时，如何才能灵活切换?

SmartPrompt的设计思路

02

Ruby DSL

- **什么是DSL?**
  - 领域特定语言（DSL）是一种专门为特定问题域设计的编程语言或语法。
  - 与通用编程语言（如Ruby、Python等）不同，DSL通常具有简化的语法，更贴近业务领域的需求。
- **Ruby中的DSL特点：**
  - 简洁性与可读性：Ruby的语法非常简洁、优雅，适合快速开发领域特定的语言。
  - 高效的抽象：通过Ruby的元编程功能，可以创建简洁的DSL，隐藏复杂性。
  - 自然语言风格：Ruby的语法使得DSL通常看起来像自然语言，易于理解和使用。
- **常见应用：**
  - 配置文件：DSL常用于编写配置文件，提升可读性和简洁性。例如，Ruby on Rails的路由配置。
  - 测试框架：RSpec是Ruby中的一个DSL，用于编写易于理解的测试用例。
  - 查询语言：ActiveRecord的查询接口提供了一个Ruby风格的DSL，用于构建数据库查询。

```ruby
# 示例：RSpec测试DSL
describe "Calculator" do
  it "adds two numbers" do
    expect(1 + 1).to eq(2)
  end
end
```

```ruby
# 使用DSL定义任务
scheduler = TaskScheduler.new

scheduler.task("task_1") do
  puts "Executing task 1..."
end

scheduler.task("task_2") do
  puts "Executing task 2..."
end

scheduler.task("task_3") do
  puts "Executing task 3..."
end

# 运行所有任务
scheduler.run
```

用ERB作为提示词模板

- **什么是ERB?**
  - ERB（Embedded Ruby）是Ruby语言中的一种模板引擎，它允许将Ruby代码嵌入到HTML或其他文本格式中。
  - 通过ERB模板，开发者可以将动态内容嵌入到静态页面中，生成动态HTML页面或其他格式的输出。
- **ERB 的工作原理：**
  - 模板语法：在ERB模板中，Ruby代码通过特定的标签（<% %>）嵌入到文本中。
  - <%= %>：执行Ruby代码并将结果插入到输出中。
  - <% %>：执行Ruby代码但不将结果插入到输出中。
  - ERB解析器会将模板中的Ruby代码解析并执行，生成最终的HTML或文本输出。
- **常见应用场景：**
  - Web开发：在Rails等Web框架中，ERB广泛用于生成动态HTML页面，特别适合与数据库或用户输入交互。
  - 邮件模板：生成动态电子邮件内容时，使用ERB模板可以嵌入动态数据（如用户名称、订单信息等）。
  - 报告生成：根据数据生成动态报告，ERB模板可以根据输入的数据结构动态生成报告内容。

模板示例：

Translate the following text from <%= source_language %> to <%= target_language %>:

<%= text %>

The translation should be accurate, maintain the original meaning, and be appropriate for the cultural context of the target language.

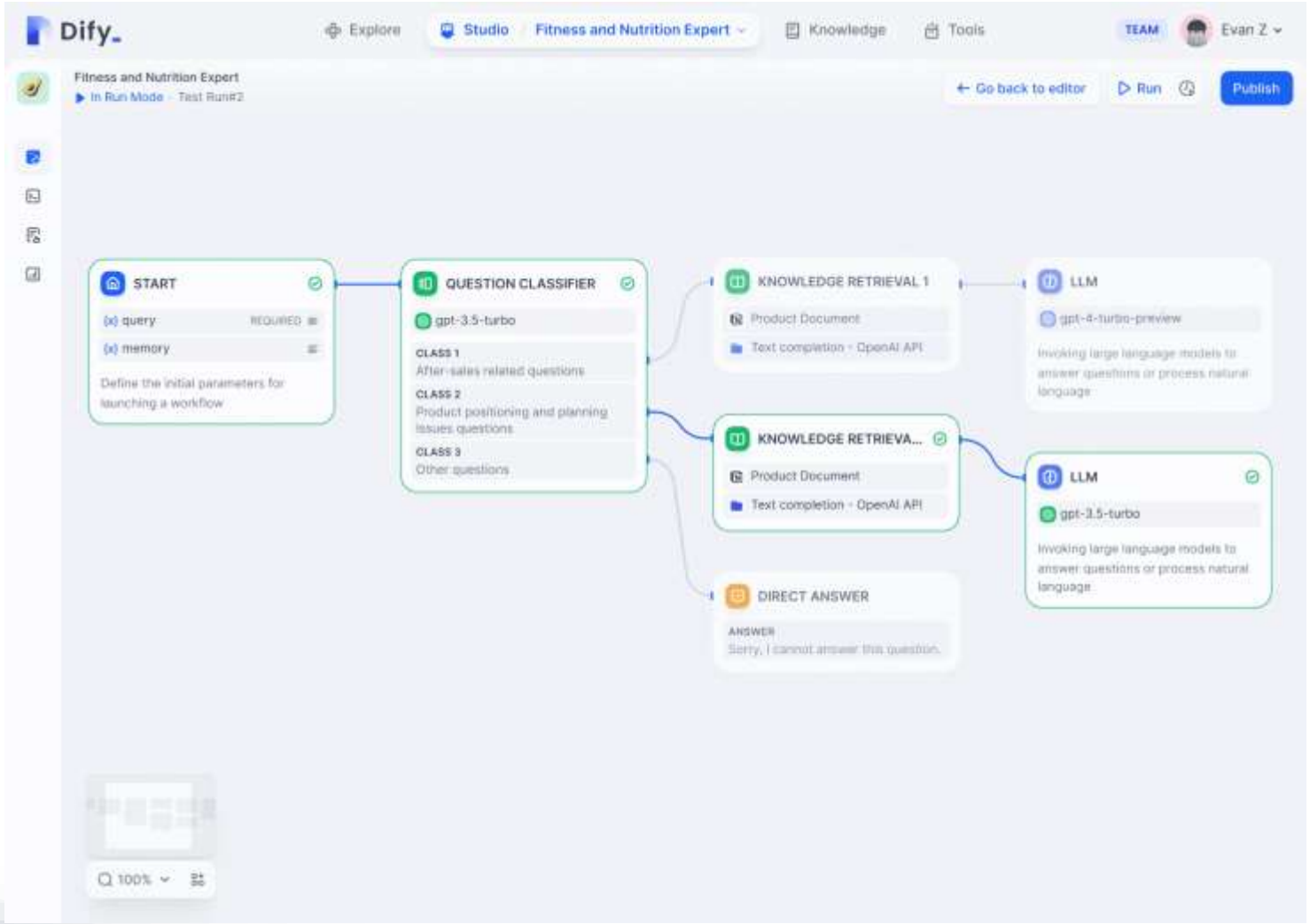
多模型切换

workers > get_code.rb

```ruby
SmartPrompt.define_worker :get_code do
  use "siliconflow"
  sys_msg "You are a helpful programmer."
  model "Qwen/Qwen2.5-Coder-7B-Instruct"
  prompt :generate_code, {
    name: params[:name],
    description: params[:description],
    input: params[:input],
    output: params[:output]
  }
  code_text = safe_send_msg
  if code_text.include?("Failed to call LLM after")
    code_text
  else
    model "meta-llama/Meta-Llama-3.1-8B-Instruct"
    prompt :get_code, {code_text: code_text}
    safe_send_msg
  end
end
```

选择不同的模型

Worker + Workflow



低代码的工作流看起来好看，其实还不如自己写代码来得快。

SmartPrompt功能示例

03

应用的目录结构

- config
  - llm_config.yml
- log
  - log.txt
- templates
  - *.erb
- workers
  - *.rb
- hello.rb
- code.rb
- translat.rb
- embedding.rb
- rag.rb

- ……

```
# gem install smart_prompt
```

config > ! llm_config.yml

```yaml
adapters:
  openai: OpenAIAdapter
  ollama: OllamaAdapter
logger_file: ./log/log.txt
llms:
  siliconflow:
    adapter: openai
    url: https://api.siliconflow.cn/v1/
    api_key: ENV["APIKey"]
    default_model: Qwen/Qwen2.5-7B-Instruct
  llamacpp:
    adapter: openai
    url: http://localhost:8080/
  ollama:
    adapter: ollama
    url: http://localhost:11434/
    default_model: qwen2.5
default_llm: siliconflow
worker_path: "./workers"
template_path: "./templates"
```

翻译

translate_txt.rb > ...

```ruby
require "../smart_prompt/lib/smart_prompt"
engine = SmartPrompt::Engine.new("./config/llm_config.yml")
Words = {}

def translate_text(engine, text)
  checked = true
  count = 0
  result = ""
  while checked == true && count < 10
    puts "try #{count}"
    result = engine.call_worker(:categorized_translation, {
        text: text,
        source_language: "English",
        target_language: "Chinese"})
    checked =
      (result.include?("{") && result.include?("}")) ||
      (result.include?("```") || result.include?("```json")) ||
      result.include?("**Output**") ||
      result.include?("**输出**") ||
      (result.split("\n").size > text.split("\n").size)
    count += 1
  end
  result
end

book = File.read("../OpenLife.txt")
new_book = File.new("../OpenLife_new.txt", "w+")

book.split("\n").each do |line|
  if line.strip.empty?
    result = line
  elsif Words.has_key?(line)
    result = Words[line]
  else
    result = translate_text(engine, line)
    Words[line] = result
  end
  new_book.puts result
end
new_book.close
```

```ruby
SmartPrompt.define_worker :categorized_translation do
  use "siliconflow"
  model "Qwen/Qwen2.5-Coder-7B-Instruct"
  prompt :cat_trans, {text: params[:text], source_language: params[:source_language]}
  json = send_msg
  model "Qwen/Qwen2.5-72B-Instruct-128K"
  if word?(params[:text])
    prompt :senior_translator, {
      json: json,
      text: params[:text],
      source_language: params[:source_language],
      target_language: params[:target_language]
    }
    result = send_msg
  else
    prompt :senior_translator2, {
      json: json,
      text: params[:text],
      source_language: params[:source_language],
      target_language: params[:target_language]
    }
    result = send_msg
  end
  result
end
```

翻译

templates > cat_trans.erb

```
# Command:
**Input**: <%= source_language %>
**Output**: Analysis
1. **Category**: Based on the content of a given text, determine to which of t
    - Abbreviations and acronyms
    - Dates and times (No translation required)
    - Numerical value (No translation required)
    - Monetary units
    - Units of measurement
    - Chapter headings
    - Proper nouns
    - Literature citations
    - Fixed expressions, phrases, idioms and expressions
    - Titles of literary and artistic works
    - Legal and regulatory provisions
    - Technical terms
    - File name (No translation required)
    - Dialogue, inflections, slang and colloquial expressions
    - Historical and cultural contexts
    - Formatted texts, formulas and symbols
    - URI/URL (No translation required)
    - Other complex texts
2. **Instances**: Under this category, enter actual examples from the text.

# Example:

**Input**: NASA's mission was launched on March 1, 2023, at 10:30 PM.
**Output**(JSON format):
{
  "category": ["Abbreviations and acronyms", "Dates and times"],
  "Instances": {
    "Abbreviations and acronyms": ["NASA (name of organization)"],
    "Date & times": ["March 1, 2023", "10:30 PM"]
  }
}

**Input**: <%= text %>
**Output**(JSON format):
```

templates > senior_translator.erb

```
# Analysis:
<%= json %>

# Command:
1. Output only translated content
2. Please follow the input file format strictly and do not add anything else.
3. Based on the results of the above analysis, translate the following <%= source_language %>, into <%= target_language %>.

# Example:
**Input**: , John Wiley & Sons.
**Output**: ，约翰·威利父子。

# Task:
**Input**: <%= text %>
**Output**(Plain text):
```

templates > senior_translator2.erb

```
# Analysis:
<%= json %>
# Command:
0. The input just one word
1. Output only translated content
2. Please follow the input file format strictly and do not add anything else.
3. Based on the results of the above analysis, translate the following <%= source_language %>, into <%= target_language %>.
# Example:
**Input**: world
**Output**: 世界
# Task:
**Input**: <%= text %>
**Output**(Plain text):
```

抓取新闻

workers > get_news.rb

```ruby
SmartPrompt.define_worker :get_news do
  url = params[:text]
  html = call_worker(:download_page, {url: url})
  text = call_worker(:html_to_text, {html: html})
  use "siliconflow"
  sys_msg "You're a journalist familiar with open source-related news coverage."
  model "Qwen/Qwen2.5-7B-Instruct"
  prompt :analyzing_news_content, {news: text}
  news_json = safe_send_msg
  f = File.open("news.json", "w+")
  f.puts news_json
  f.close
  prompt :generate_sql2, {news: news_json}
  sql = safe_send_msg
  sql
end
```

templates > analyzing_news_content.erb

```
Based on the content of the page provided below, the content of N news items
was analyzed and presented in the following format:

```json
[
    {
        original_title: "",
        chinese_title: "",
        date: "",
        summary: "",
        categories: ["",""]
    },
]
```

Input news:
<%= news %>
```

生成代码

code.rb

```ruby
require "../smart_prompt/lib/smart_prompt"
engine = SmartPrompt::Engine.new("./config/llm_config.yml")
result = engine.call_worker(:get_code, {
  name: "calculate_triangle_area",
  description: "calculates the area of a triangle",
  input: "the base and height of the triangle",
  output: "the area as a float"
})

if result.include?("Failed to call LLM after")
  puts result
else
  code_str = result + "\n" + "puts calculate_triangle_area(8, 10)"
  eval(code_str)
end
```

workers > get_code.rb

```ruby
SmartPrompt.define_worker :get_code do
  use "siliconflow"
  sys_msg "You are a helpful programmer."
  model "Qwen/Qwen2.5-Coder-7B-Instruct"
  prompt :generate_code, {
    name: params[:name],
    description: params[:description],
    input: params[:input],
    output: params[:output]
  }
  code_text = safe_send_msg
  if code_text.include?("Failed to call LLM after")
    code_text
  else
    model "meta-llama/Meta-Llama-3.1-8B-Instruct"
    prompt :get_code, {code_text: code_text}
    safe_send_msg
  end
end
```

tes > generate_code.erb

```
Write a Ruby function called "<%= name %>" that <%= description %>.
The function should accept <%= input %> as input and return <%= output %>.
Just return the source code to me, no other explanations needed.
```

tes > get_code.erb

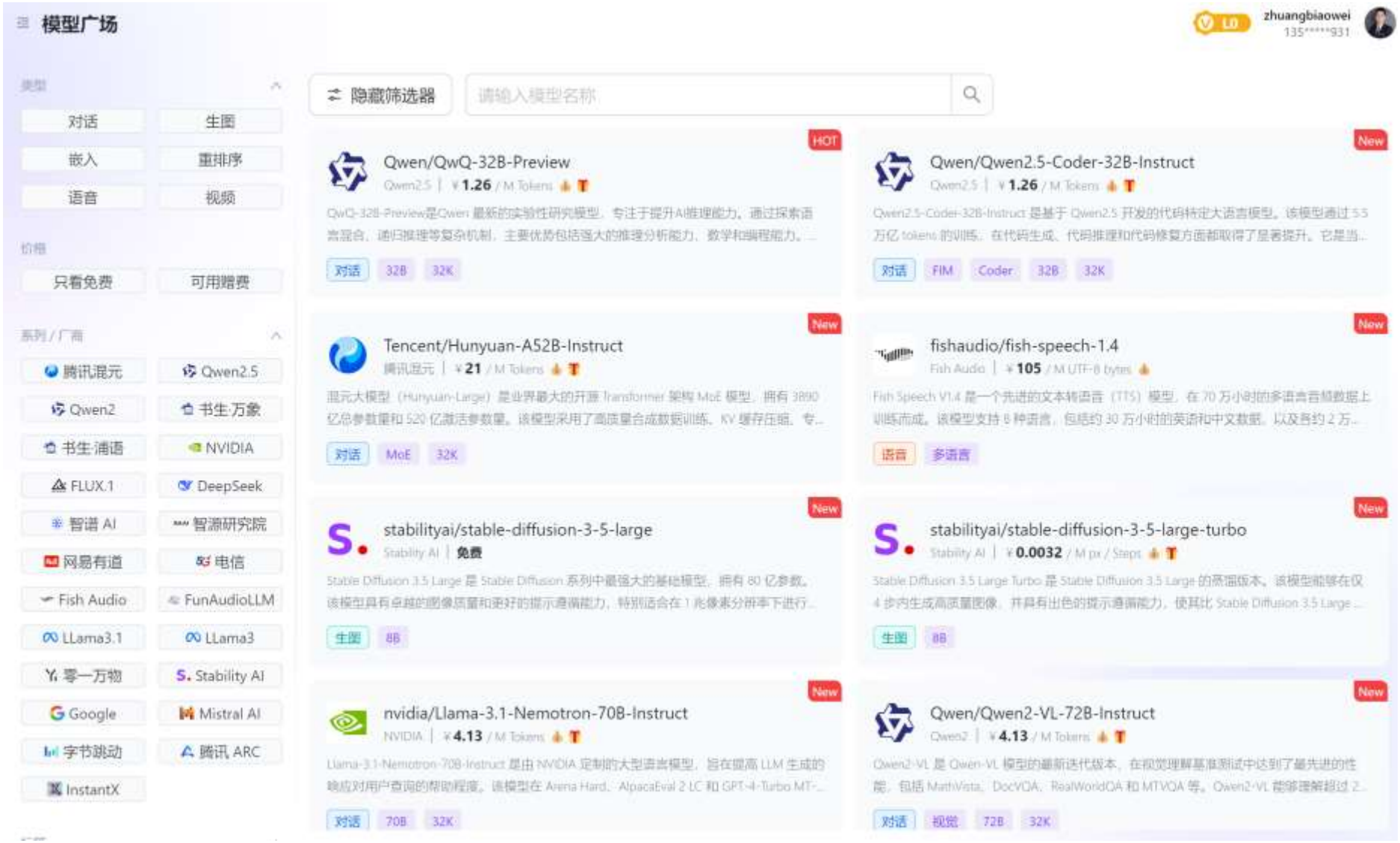
```
Remove all non-code text and return only the code. Not even characters like ```.

<%= code_text %>
```

```ruby
def calculate_triangle_area(base, height)
  (base * height) / 2.0
end
puts calculate_triangle_area(8, 10)
40.0
```

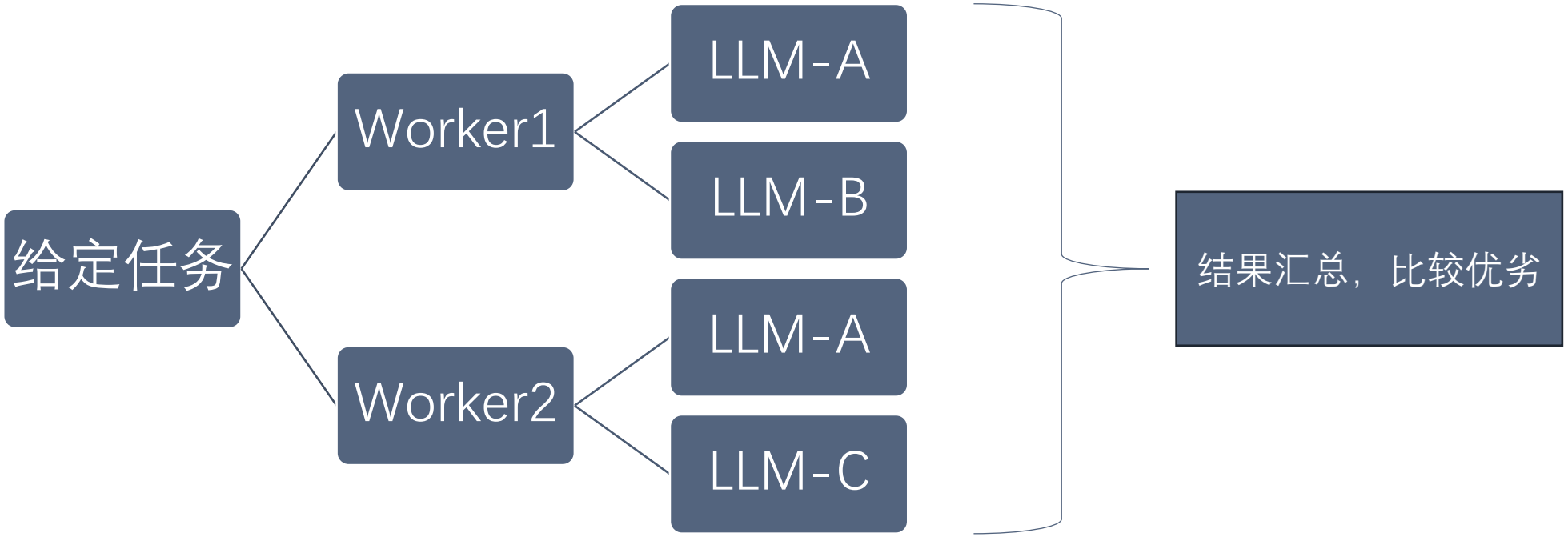
SmartPrompt的后续发展计划

04

支持多模态的大模型


模型广场
zhuangbiaowei
隐藏筛选器
请输入模型名称
对话
生图
嵌入
重排序
语音
视频
只看免费
可用赠费
腾讯混元
Qwen2.5
Qwen2
书生·万象
书生·浦语
NVIDIA
FLUX.1
DeepSeek
智谱 AI
智源研究院
网易有道
电信
Fish Audio
FunAudioLLM
LLama3.1
LLama3
零一万物
Stability AI
Google
Mistral AI
字节跳动
腾讯 ARC
InstantX
HOT
Qwen/QwQ-32B-Preview
¥1.26 / M Tokens
New
Qwen/Qwen2.5-Coder-32B-Instruct
¥1.26 / M Tokens
Tencent/Hunyuan-A52B-Instruct
¥21 / M Tokens
fishaudio/fish-speech-1.4
¥105 / M UTF-8 bytes
stabilityai/stable-diffusion-3-5-large
免费
stabilityai/stable-diffusion-3-5-large-turbo
¥0.0032 / M px / Steps
nvidia/Llama-3.1-Nemotron-70B-Instruct
¥4.13 / M Tokens
Qwen/Qwen2-VL-72B-Instruct
¥4.13 / M Tokens

多个Worker、多个模型并发



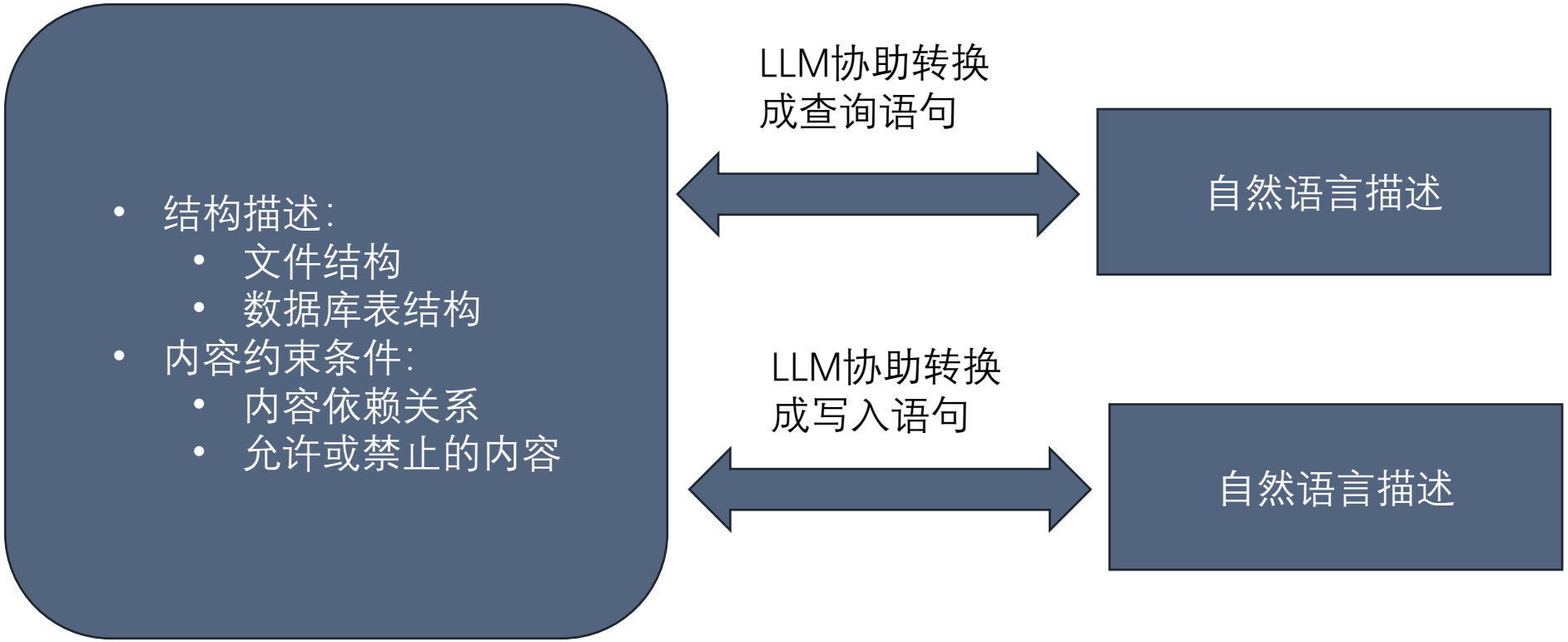
对于Worker的输出结果进行调优

现在的worker里，写死了模型名称，也有明确调用的模板。

如果能够定义：“理想的输出结果”，然后针对可选的模板，反复优化模板的提示词，直到找到最佳组合。

效果就会更加理想。

系统级Worker：文件管理、数据库管理等



结语

05

目前，这还是一个非常早期的项目，因为很多想要实现的功能，还没有找到自己觉得满意的写法。

估计在上述设想实现之后，再正式开源。

https://github.com/zhuangbiaowei/smart_prompt

https://github.com/zhuangbiaowei/SmartPromptDemo

谢谢大家

汇报人：庄表伟

时间：2024.11