

腾讯云向量数据库
RAG七天入门课 第五节

让LLM理解知识 — Prompt

腾讯云高级算法工程师

赵九州

第五节

让LLM理解知识

-Prompt



Prompt

Prompt 可以被理解为用于指导AI模型生成特定类型、主题或格式内容的文本。在NLP中，Prompt 通常由一个问题或任务描述组成，例如“给我写一篇有关RAG的文章”，这句话就是Prompt

Prompt赋予了LLM小样本甚至零样本学习的能力

- Zero-shot

```
1 Translate English to French:
2 cheese => .....
```

- One-shot

```
1 Translate English to French:
2 sea otter => loutre de mer
3 cheese => .....
```

- Few-shot

```
1 Translate English to French:
2 sea otter => loutre de mer
3 peppermint => menthe poivrée
4 plush girafe => girafe peluche
5 cheese => .....
```



Prompt

大模型的能力本质上来说是续写，通过编写更好的prompt来指导模型，并因此获得更好的结果。

提示词

The sky is

输出结果

blue
The sky is blue on a clear day. On a cloudy day, the sky may be gray or white.

无具体指令，模型只会续写。输出的结果可能是出人意料
的，或远高于我们的任务要求。

提示词

完善以下句子：
The sky is

输出结果

so beautiful today.

告知模型去完善句子，因此输出的结果和我们最初的输入是完全符合的。Prompt Engineering就是探讨如何设计出最佳Prompt，
用于指导LLM高效完成某项任务。

Prompt的进阶技巧CoT

Chain of Thought , 让模型输出更多的上下文与思考过程 , 提升模型输出下一个token的准确率

Standard Prompting

Input

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

Model Output

A: The answer is 27. ❌

Chain of Thought Prompting

Input

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls. $5 + 6 = 11$. The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

Model Output

A: The cafeteria had 23 apples originally. They used 20 to make lunch. So they had $23 - 20 = 3$. They bought 6 more apples, so they have $3 + 6 = 9$. The answer is 9. ✅

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?

A: **Let's think step by step.**

(Output) There are 16 balls in total. Half of the balls are golf balls. That means that there are 8 golf balls. Half of the golf balls are blue. That means that there are 4 blue golf balls. ✅

优化Prompt，提升模型推理能力和问答准确率

1、分布式引导提问

把解决问题的思路分成多步，引导模型分步执行

Plan Generation

Given a question, generate plan of mined actions

Question: What part of the final scene best connects to the story's opening conversation?

```
1.open_conv = FIND_ELEMENT(CTX,"opening conver..")
2.final_scene = SUMMARIZE_X(CTX, "final_scene")
3.reflection = FIND_RELATION(init_conv, final_scene)
```

Plan Execution

Execute the plan step-by-step

```
open_conv = "In the initial conversation, Phil Conover is excited about his upcoming mission to be the first man to see the other side of the moon ...."
```

	QUALITY LONG	QUALITY SHORT	ALL
PROMPTING METHODS			
GPT-4 zero-shot	64.3	79.1	68.8
GPT-3.5 zero-shot (text-davinci-003)	45.5	56.3	48.8
GPT-4 zero-shot chain-of-thought	65.9	77.2	69.3
GPT-4 PEARL	70.9	77.8	73.0

2、Prompt代码化

LLM通常都会有代码数据，prompt代码化进一步提升模型的推理能力

Input

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: Roger started with 5 tennis balls.
tennis_balls = 5
2 cans of 3 tennis balls each is
bought_balls = 2 * 3
tennis_balls. The answer is
answer = tennis_balls + bought_balls

Q: The bakers at the Beverly Hills Bakery baked 200 loaves of bread on Monday morning. They sold 93 loaves in the morning and 39 loaves in the afternoon. A grocery store returned 6 unsold loaves. How many loaves of bread did they have left?

Model Output

A: The bakers started with 200 loaves
loaves_baked = 200
They sold 93 in the morning and 39 in the afternoon
loaves_sold_morning = 93
loaves_sold_afternoon = 39
The grocery store returned 6 loaves.
loaves_returned = 6
The answer is
answer = loaves_baked - loaves_sold_morning
- loaves_sold_afternoon + loaves_returned
74

	GSM8K	GSM-HARD	SVAMP	ASDIV	SINGLEEQ	SINGLEOP
DIRECT Codex	19.7	5.0	69.9	74.0	86.8	93.1
CoT UL2-20B	4.1	-	12.6	16.9	-	-
CoT LaMDA-137B	17.1	-	39.9	49.0	-	-
CoT Codex	65.6	23.1	74.8	76.9	89.1	91.9
CoT PaLM-540B	56.9	-	79.0	73.9	92.3	94.1
CoT Minerva 540B	58.8	-	-	-	-	-
PAL	72.0	61.2	79.4	79.6	96.1	94.6

```
prompt = "Please adhere strictly to the following Q&A when answering the question. Let's think step by step, " \
"ensuring that your answers are evidence-based and not fabricated. Do not omit the " \
"! [(https://***.jpg)` style image links. Refrain from directly outputting reference Q&A. " \
"Aim to keep your response as concise as possible.\n"
```

优化后



2

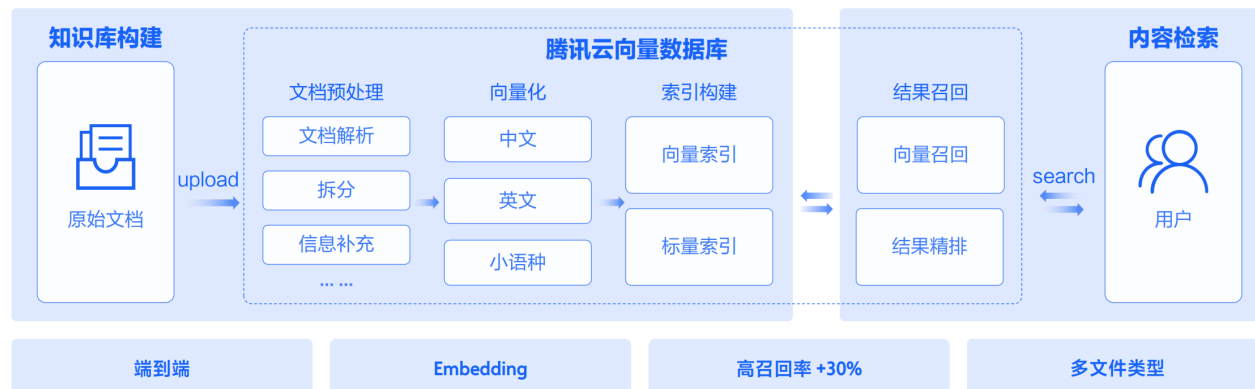
prompt = "下面提供了一个list，其中的每个元素是{Q:A}组成的dict。最后有一个用户的问题，我们需要根据list中的内容来回答用户问题。" \
"请遵守以下规则来回答问题:\n" \
"1、在列表中找到与用户问题最相似的Q，并围绕与这个Q对应的A来回答用户问题，输出答案时一定要把其中的图片、链接以markdown格式打印出来" \
"2、回答问题时不要自己编造，需要严格参照list中的内容" \
"3、回答需要清晰、简洁、明了。" \
"请注意：用户无法看到list，你需要用正常回答问题的语气来解答，而不要提到list中的内容。" \
"

1

腾讯云向量数据库：消除大模型幻觉，加速大模型在企业落地

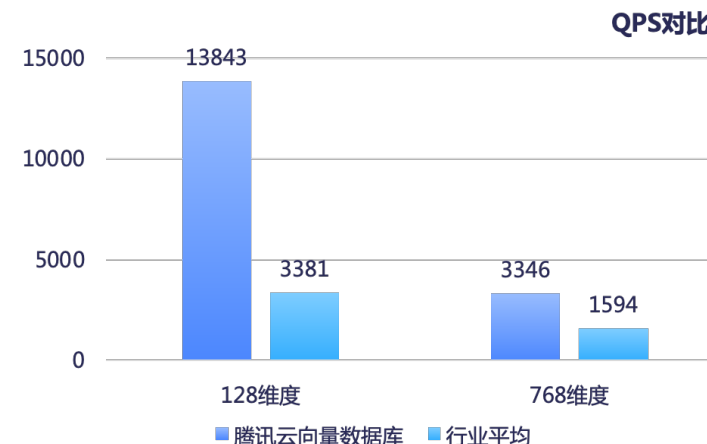
端到端AI套件，AGI时代的知识库解决方案

提供**一站式**知识检索方案，实现业界内**最高召回率**、**大幅降低开发门槛**，帮助企业快速搭建RAG应用，解决大模型幻觉问题



源自集团多年积累，产品能力行业领先

源自腾讯自研向量检索引擎OLAMA，集团内部**40+**业务线上使用，日均处理**1600亿次**检索请求



『**首家**』通过中国信通院
向量数据库标准测试



单索引支持最高**千亿级**
超大数据规模



单实例最高可达**500万 QPS**