大语言模型（LLMs）基础概念和重要技术概述

- 报告人：陈振源
- 预计时长：15mins
- 大纲：
  - 一、大语言模型基础概念
    - 什么是大语言模型（LLMs）？
    - 基础大模型 vs. 垂直领域大模型
    - 大模型的涌现能力（Emergent Abilities）
  - 二、大语言模型重要技术
    - 监督微调（SFT）
    - 基于人类反馈的强化学习(RLHF)
    - 嵌入（Embedding）
    - 幻觉（Illusion）
    - 信息检索增强(RAG)

一、大语言模型基础概念

1. 什么是大语言模型（LLMs）？

大语言模型，即Large Language Models，本质上是一个语言模型，由于其相较于过往的语言模型表现出惊人的能力，为了便于区分，我们将这类参数量极大的语言模型称为**大语言模型**。

语言模型发展历史悠久且研究积淀深厚，在自然语言处理（NLP）的各项任务上表现良好。

大语言模型一词最早出现于GPT-3技术报告中，研究人员发现当预训练语言模型（Pre-trained Language Models，PLM）参数量达到一定规模时（often to be Billions），出现了惊人的能力，即涌现能力（Emergent Ability），为了将这类新模型与普通的PLM区分，研究人员创造（coin）了一个新的词汇"Large Language Model"。

概念1：Large Language Model

- 学名：大语言模型
- 简称：大模型
- 发展：Statistical language models (SLM)→

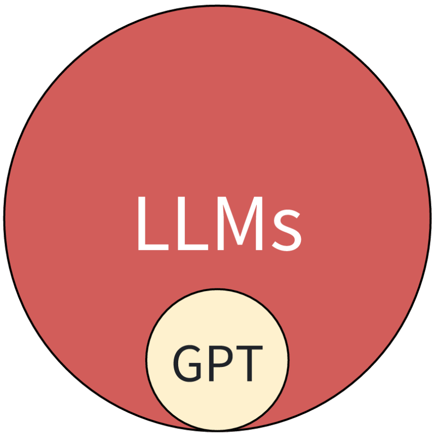
  Neural language models (NLM)→

  Pre-trained Language Models (PLM)→

  Large Language Models(LLM/PLM+)

概念2：Generative Pre-Training(GPT)

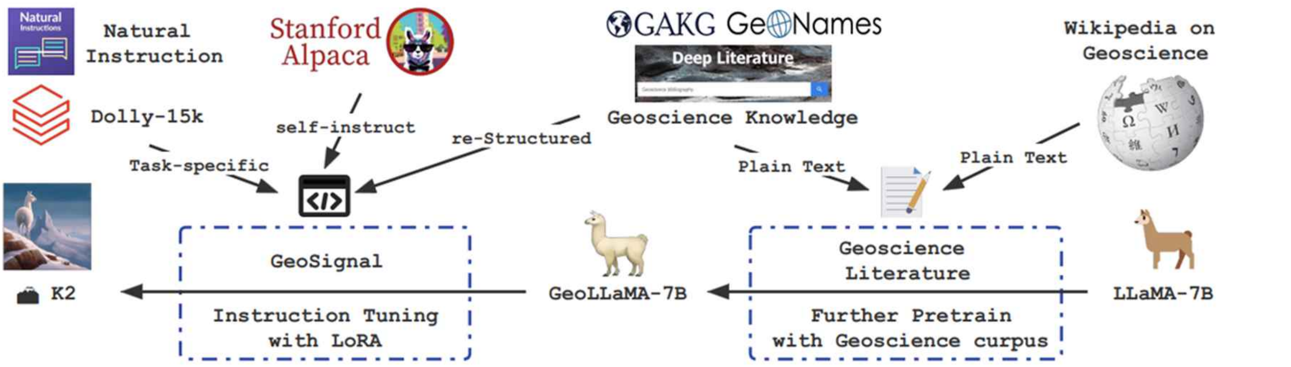
- 学名：生成性预训练（模型/技术）
- OpenAI的核心大模型产品



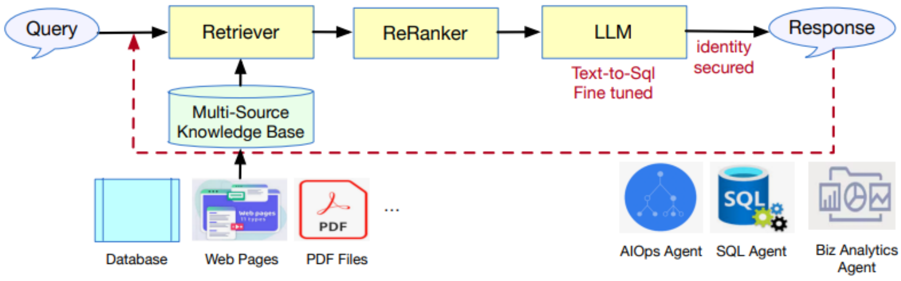
GPT series, LLaMA series, K2, GeoGalactica, **DB-GPT**

2. 基础大模型 vs. 垂直领域大模型

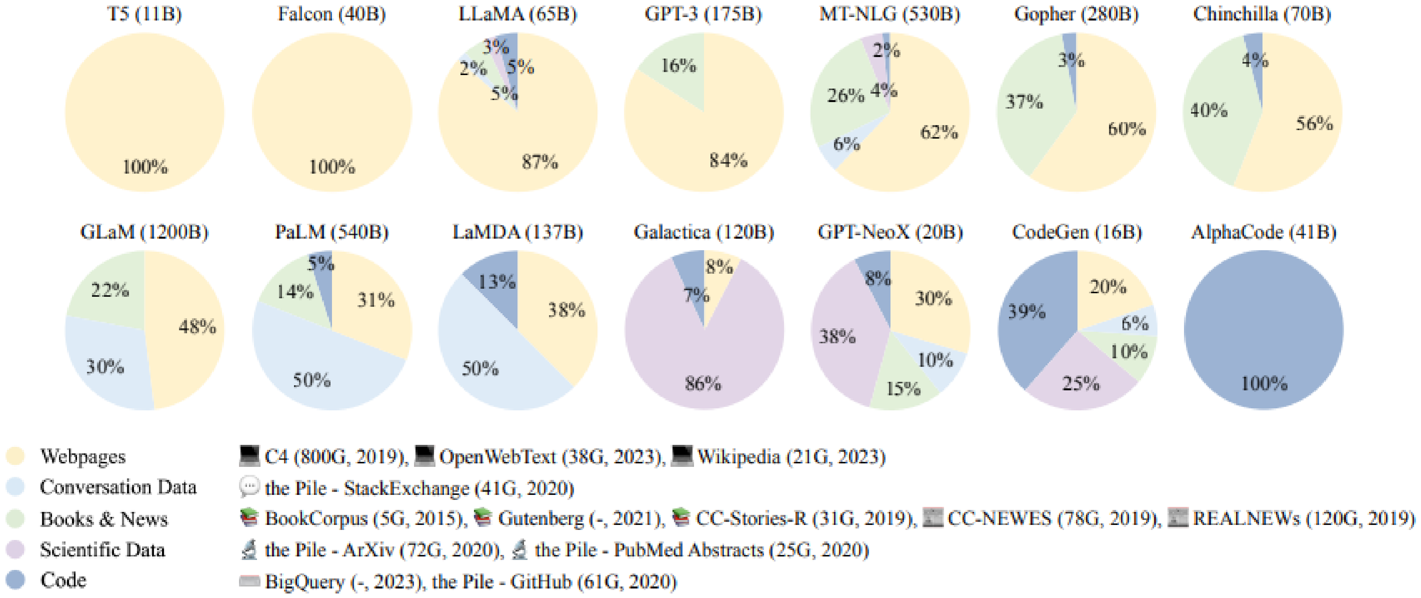
Cheng Deng, et al.



Pipeline of training K2



The architecture of DB-GPT



Ratios of various data sources in the pre-training data for existing LLMs

- 基础大模型/基模型：从零开始训练模型参数
  - 训练过程计算量大，耗时长
  - 以LLaMA-60B为例，其报告指出LLaMA-60B在训练时使用2048张80G的A100卡，训练时间为21天
  - GPT-4，LLaMA，Galactica，**Qwen，Baichuan**
- 垂直领域大模型：在基模型上调整参数
  - 训练/微调计算消耗少，耗时短
  - K2 (from LLaMa)，GeoGalactica (from Galactica), **DB-GPT**(base model could be a list of base LLMs)
  - 上述中，K2和GeoGalactica都是交大团队做的地学领域大模型，其中K2以LMaMA-**7B**为基模型使用4张40G的A100卡训练9天，而GeoGalactica和K2的训练消耗相近

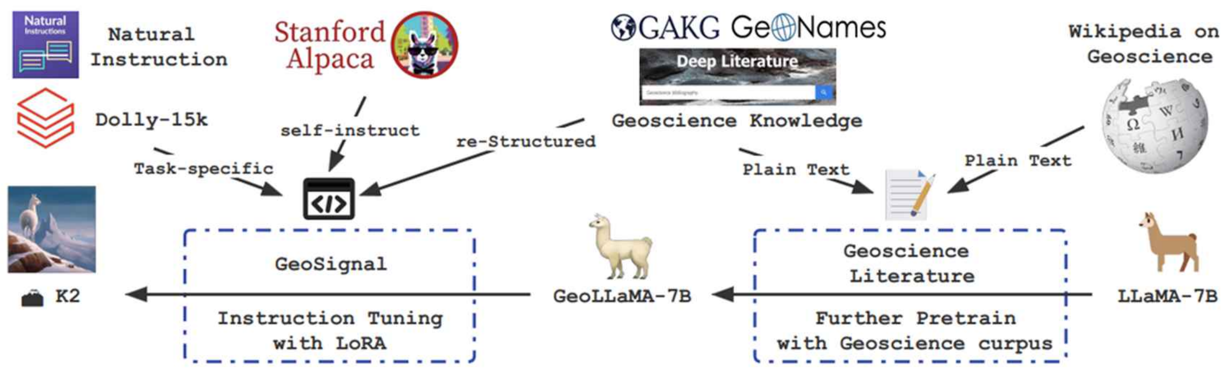
3. 大模型的涌现能力（Emergent Abilities）

https://chat.openai.com/share/89a284ac-c753-4866-b4a9-be523c5a5b30

涌现能力

- 定义：小模型中不存在但在大模型中出现的能力(Wei et al., 2022)
- 典型涌现能力
    - 上下文学习（In-context learning）
    - 指令遵循（Instruction following）
    - 逐步推理（Step-by-step reasoning）

二、大语言模型重要技术

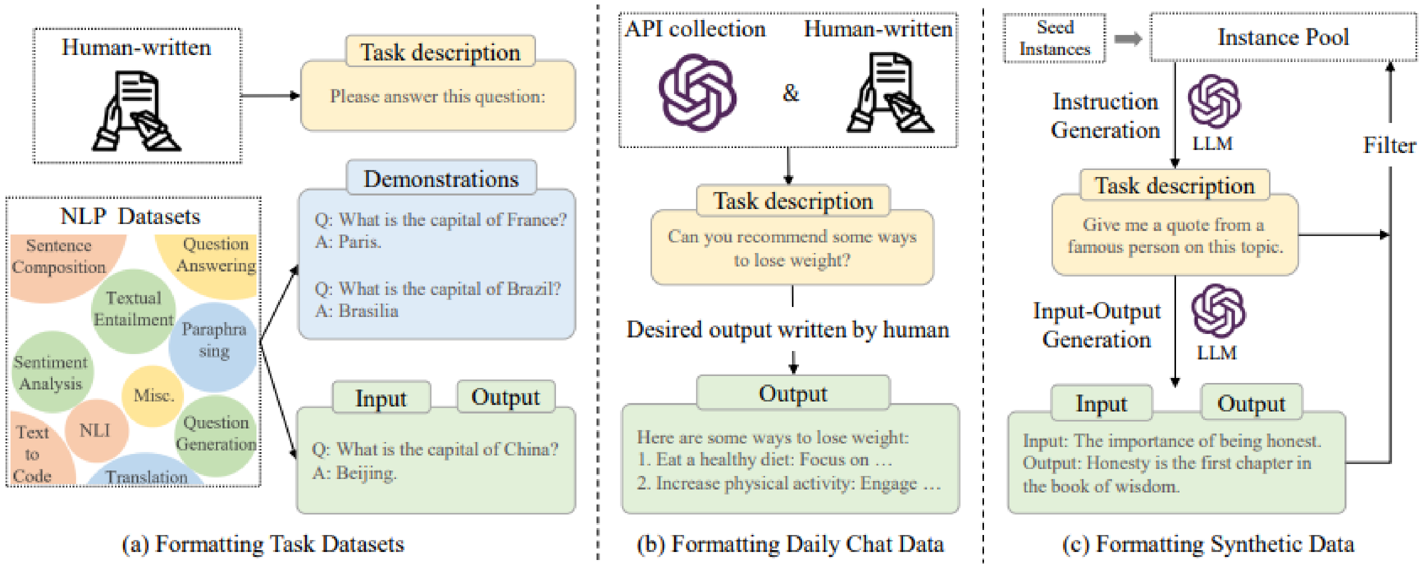


Cheng Deng, et al.

Pipeline of training K2

1. 监督微调（Supervised Fine-Tuning,SFT）

📌 The dataset for SFT is generally a Q&A Pairs（often in JSON file in project）.


Human-written
Task description
Please answer this question:
NLP Datasets
Sentence Composition
Question Answering
Textual Entailment
Paraphrasing
Sentiment Analysis
Misc.
Text to Code
NLI
Question Generation
Translation
Demonstrations
Q: What is the capital of France?
A: Paris.
Q: What is the capital of Brazil?
A: Brasilia
Input
Output
Q: What is the capital of China?
A: Beijing.
API collection
&
Human-written
Task description
Can you recommend some ways to lose weight?
Desired output written by human
Output
Here are some ways to lose weight:
1. Eat a healthy diet: Focus on ...
2. Increase physical activity: Engage ...
Seed Instances
Instance Pool
Instruction Generation
LLM
Filter
Task description
Give me a quote from a famous person on this topic.
Input-Output Generation
LLM
Input
Output
Input: The importance of being honest.
Output: Honesty is the first chapter in the book of wisdom.

(a) Formatting Task Datasets

(b) Formatting Daily Chat Data

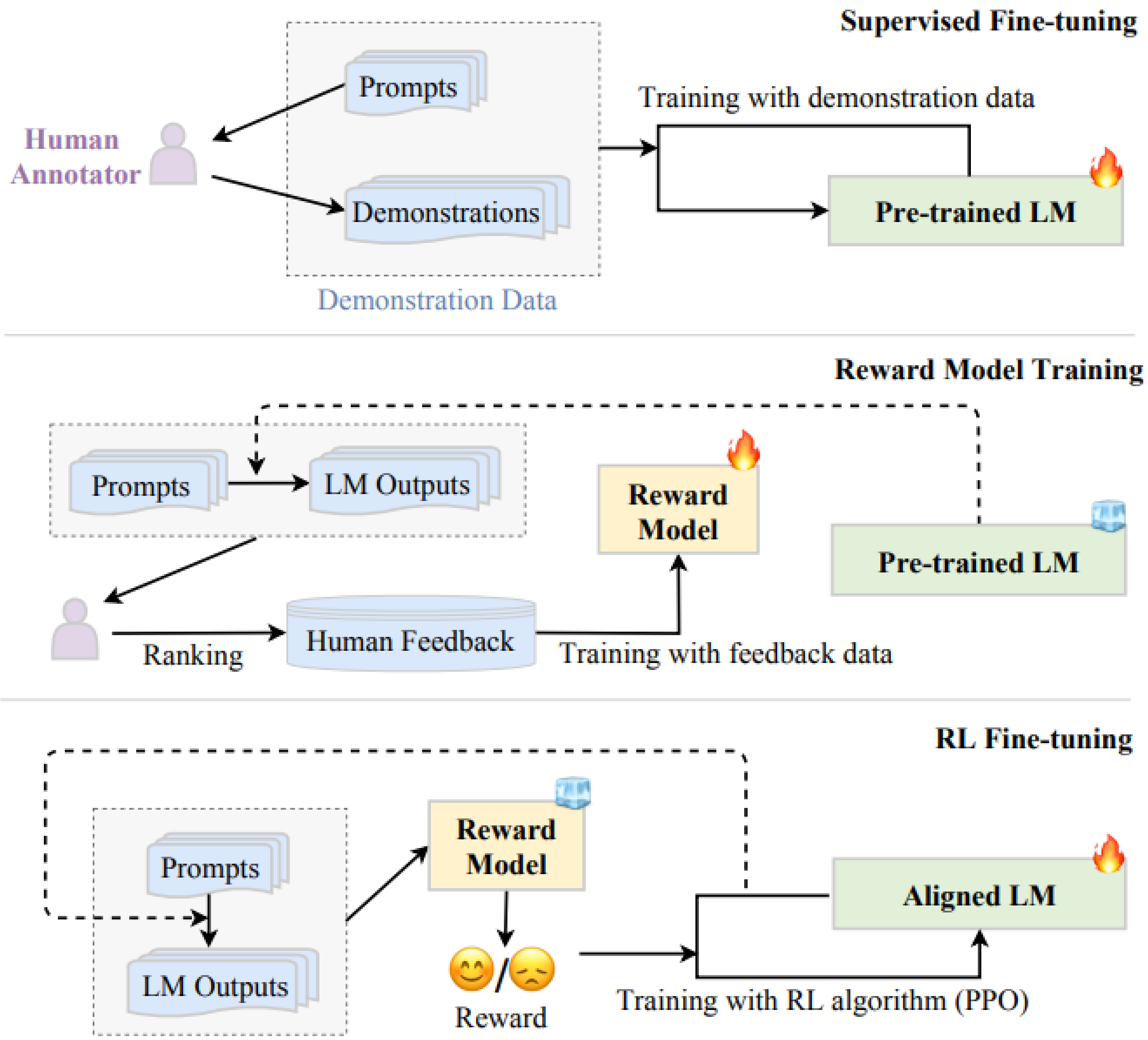
(c) Formatting Synthetic Data

An illustration of instance formatting and three different methods for constructing the instruction-formatted instances

```
// An example of SFT dataset from K2
// K2 geosinal dataset consists of 5000 Q&A pairs
{
        "instruction": "Why is California the best state?",
        "input": "",
        "output": "California has the best landscape in the country. We have some of the best mountains to ski on, most iconic beach spots, and richest soil for agriculture. What else could you ask for?",
        "type": "dolly"
},
```

2. 基于人类反馈的强化学习(Reinforcement Learning from Human Feedback,RLHF)

> 📌 The purpose of RLHF is to get a better SFT dataset.



The workflow of the RLHF algorithm

3. 嵌入（Embedding）

📌 By converting texts to vectors(embeddings), **we can predict the next word better!**

> Say you read a detective novel. It’ s like complicated plot, a storyline, different characters, lots of events, mysteries like clues, it’ s unclear. Then, let’ s say that at the last page of the book, the detective has gathered all the clues, gathered all the people and saying, "okay, I’ m going to reveal the identity of whoever committed the crime and that person’ s name is". Predict that word. ... Now, there are many different words. But predicting those words better and better, the understanding of the text keeps on increasing. GPT-4 predicts the next word better. ——Ilya Sutskever (OpenAI Chief Scientist)

```python
# An example of embedding in python codes using langchain
from langchain.embeddings import OpenAIEmbeddings
embeddings_model = OpenAIEmbeddings(
    #by default, the dimension of embedding for this model is 1536
    model="text-embedding-ada-002",

    # private info
    openai_api_base=openai_api_base_sakura,
    openai_api_key=openai_api_key_sakura,
    )
embeddings = embeddings_model.embed_documents(
    [
        "DDE",
        "Langchain",
        "LLMs",
        "GIS Lab",
        "Sakura",
    ]
)
embedding_test=embeddings[4]
print("Type of embedding_test:", type(embedding_test))
print("Length of embedding_test:", len(embedding_test))
print("First 10 elements of embedding_test:", embedding_test[:10])
```

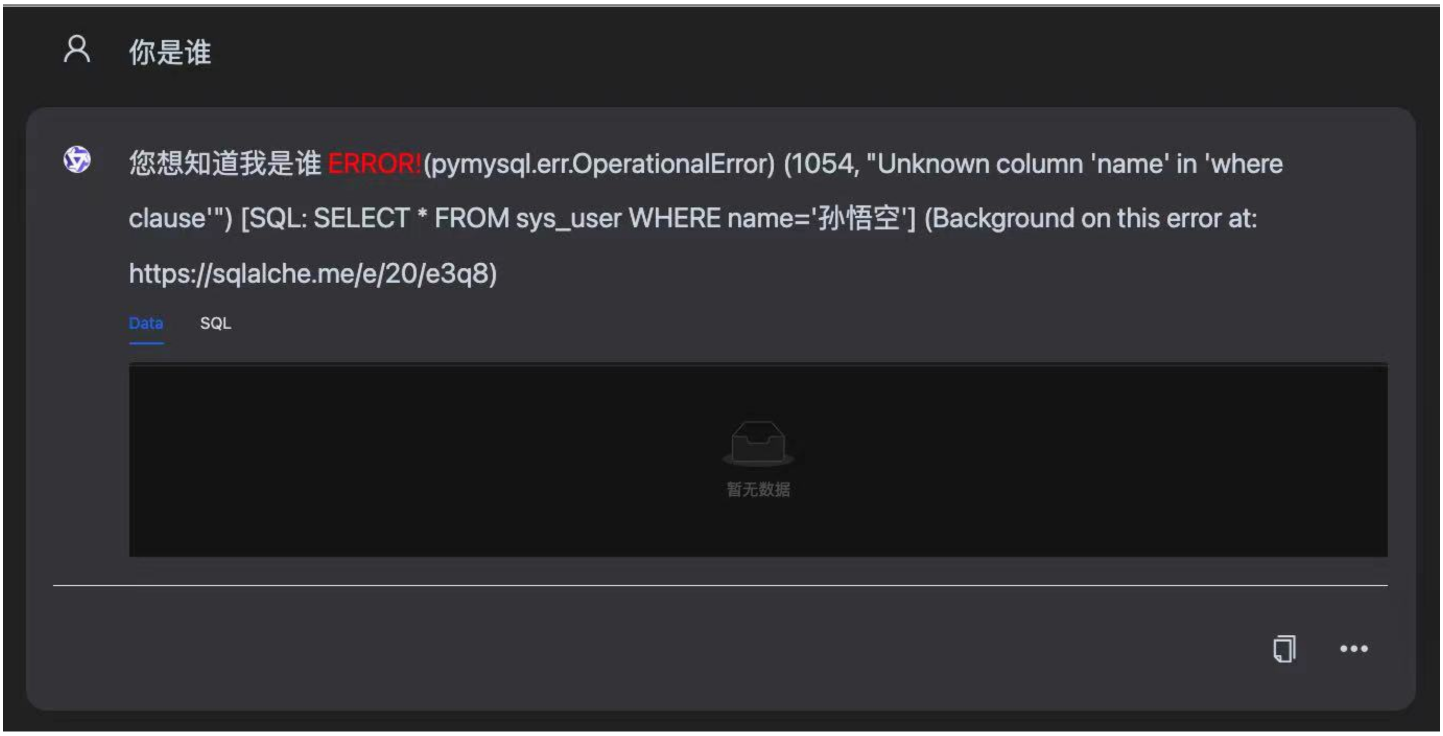
```
# Output:
Type of embedding_test: <class 'list'>

Length of embedding_test: 1536

First 10 elements of embedding_test: [0.0030393233841935568, 0.005510930198672564,
-0.015331488733466266, 0.016648842706815322, -0.00796685400601108,
```

> 0.00041990653243888107, -0.02359945237875529, -0.015494590331346678, -0.00024504349411003866, -0.03640914858176072]

4. 幻觉（Illusion）

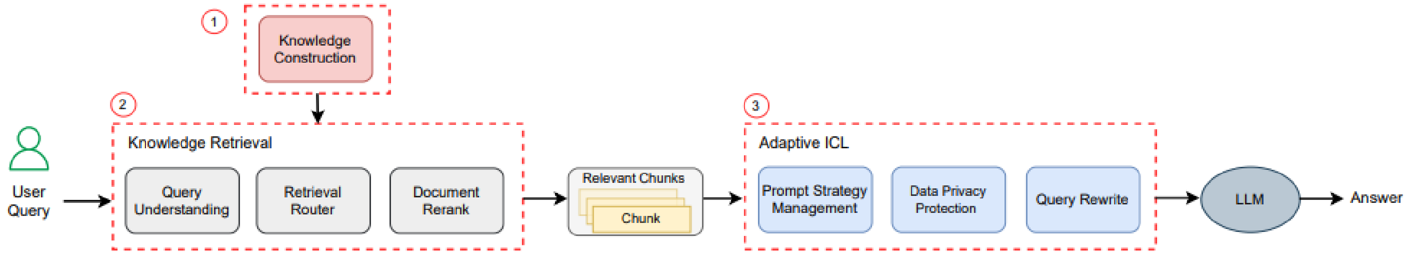


你是谁

您想知道我是谁 ERROR!(pymysql.err.OperationalError) (1054, "Unknown column 'name' in 'where clause'") [SQL: SELECT * FROM sys_user WHERE name='孙悟空'] (Background on this error at: https://sqlalche.me/e/20/e3q8)

Test on DB-GPT Chat Data Module

5. 信息检索增强(Retrieval Augmented Generation,RAG)



The detailed RAG architecture in DB-GPT

参考论文及参考材料

1. Zhao, W.X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., Du, Y., Yang, C., Chen, Y., Chen, Z., Jiang, J., Ren, R., Li, Y., Tang, X., Liu, Z., Liu, P., Nie, J., & Wen, J. (2023). A Survey of Large Language Models. *ArXiv, abs/2303.18223.*
2. Deng, C., Zhang, T., He, Z., Chen, Q., Shi, Y., Zhou, L., Fu, L., Zhang, W., Wang, X., Zhou, C., Lin, Z., & He, J. (2023). Learning A Foundation Language Model for Geoscience Knowledge Understanding and Utilization. *ArXiv, abs/2306.05064.*

3. Lin, Z., Deng, C., Zhou, L., Zhang, T., Xu, Y., Xu, Y., He, Z., Shi, Y., Dai, B., Song, Y., Zeng, B., Chen, Q., Shi, T., Huang, T., Xu, Y., Wang, S., Fu, L., Zhang, W., He, J., Ma, C., Zhu, Y., Wang, X., & Zhou, C. (2023). GeoGalactica: A Scientific Large Language Model in Geoscience.
4. Xue, S., Jiang, C., Shi, W., Cheng, F., Chen, K., Yang, H., Zhang, Z., He, J., Zhang, H., Wei, G., Zhao, W., Zhou, F., Qi, D., Yi, H., Liu, S., & Chen, F. (2023). DB-GPT: Empowering Database Interactions with Private Large Language Models.
5. Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., Chi, E.H., Hashimoto, T., Vinyals, O., Liang, P., Dean, J., & Fedus, W. (2022). Emergent Abilities of Large Language Models. *ArXiv, abs/2206.07682*.
6. https://db-gpt.readthedocs.io/en/latest/
7. https://github.com/davendw49/k2/tree/main