NLP研究杂谈

陈启源

https://qiyuan-chen.github.io/

2023/11/24

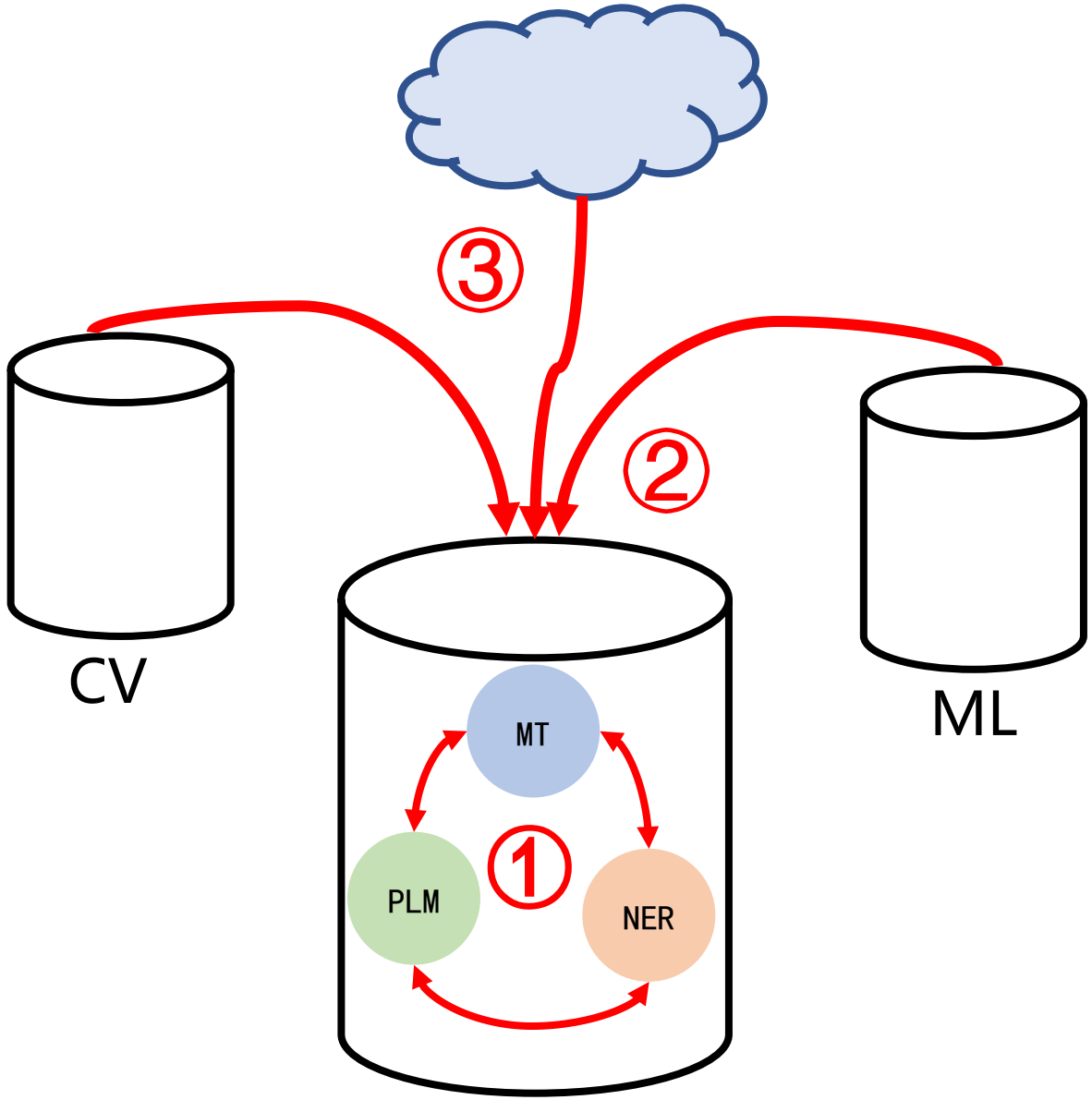
NLP研究划分

从哪来？ (别人的研究)

① 内卷形

② 外卷形

③ 天赋形



NLP研究划分

从哪来？ （别人的研究）➡ **技巧性**

① 内卷形

② 外卷形

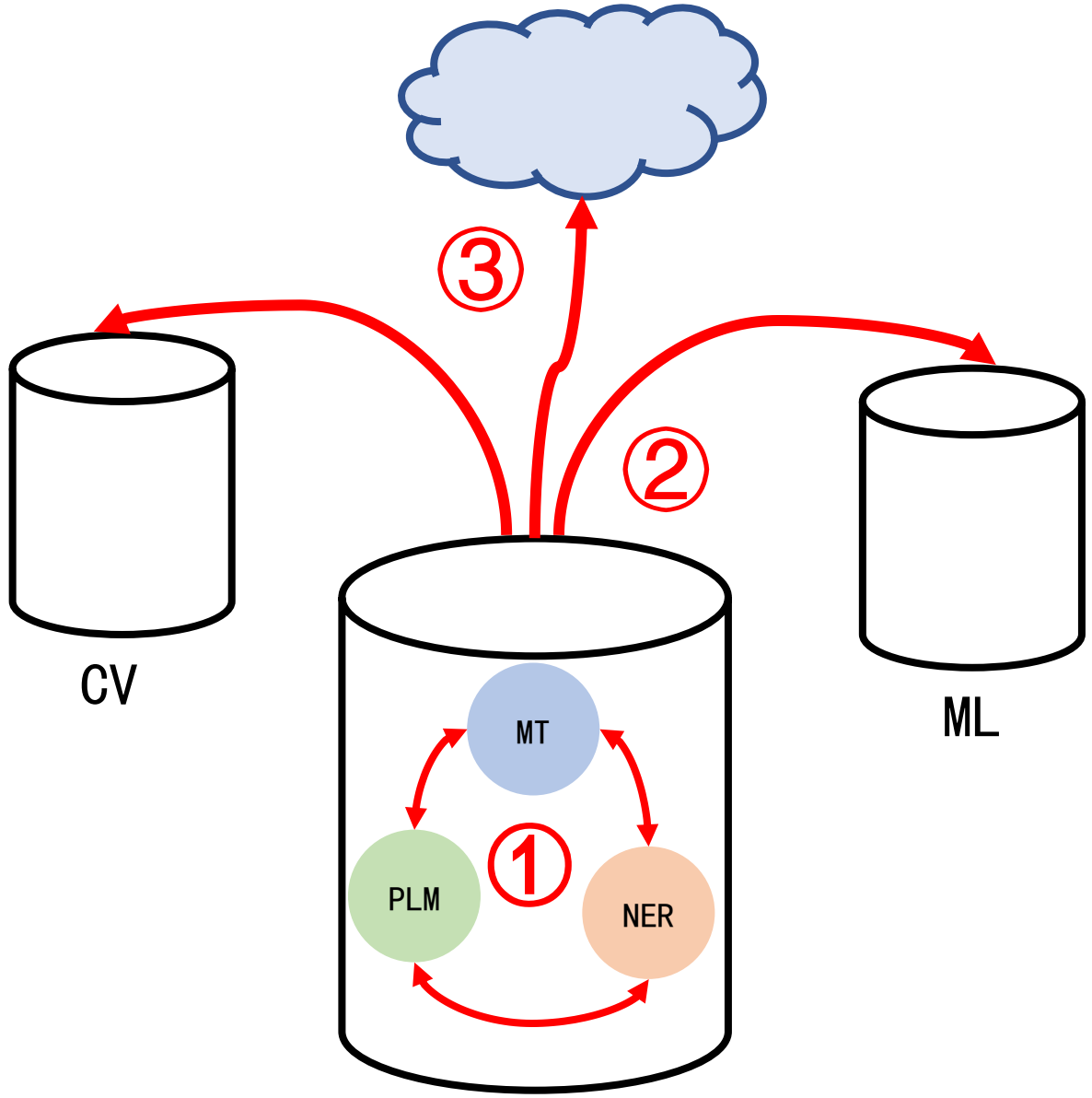
③ 天赋形

到哪去？ （别人的研究）➡ **影响力**

① NLP的其他任务

② 其他领域

③ 思维方式的启发



技巧性 V.S. 影响力

技巧性工作（巧妙引入了某种方法）

- XLNet, ELECTRA ...

影响力工作（简单实用，被广泛使用）

- BERT, ELMo ...

启发

从哪来？ （别人的研究）→ 技巧性 → **可遇不可求的**

① 内卷形

② 外卷形

③ 天赋形

到哪去？ （别人的研究）→ 影响力 → **做有意义的方向！**

① NLP的其他任务

② 其他领域

③ 思维方式的启发

什么是有意义的方向？

任务角度：能体现整个NLP领域的发展水平

- 正例：Parsing → MT/QA → PLM
- 负例：只输入不输出的领域

问题角度：包含NLP领域的本质问题

- 正例：降低语言模型的困惑度
- 负例：重复生成问题（随着困惑都下降逐渐消失的问题）

不要做只输入不输出的任务！

LLM的爆火给了NLPer巨大机会

如何训练一个LLM?

语料
(百度、知乎、
贴吧……)

Transformer

训练目标：Next Word Prediction
一个例子：
输入：华中师范在

概率分布

预训练

Base LLM

指令微调数据集

Base LLM

输入：华中师范在哪里?
希望模型得到的输出：华中师范在武汉

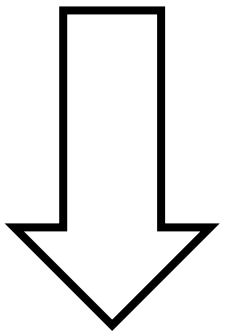
指令微调

Instruction Tuned LLM

预训练 v.s. 指令微调

数据量：1.2万亿 tokens（Baichuan 7B）

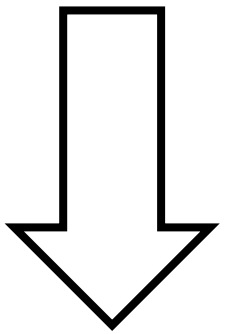
数据源：知识密集型



学知识

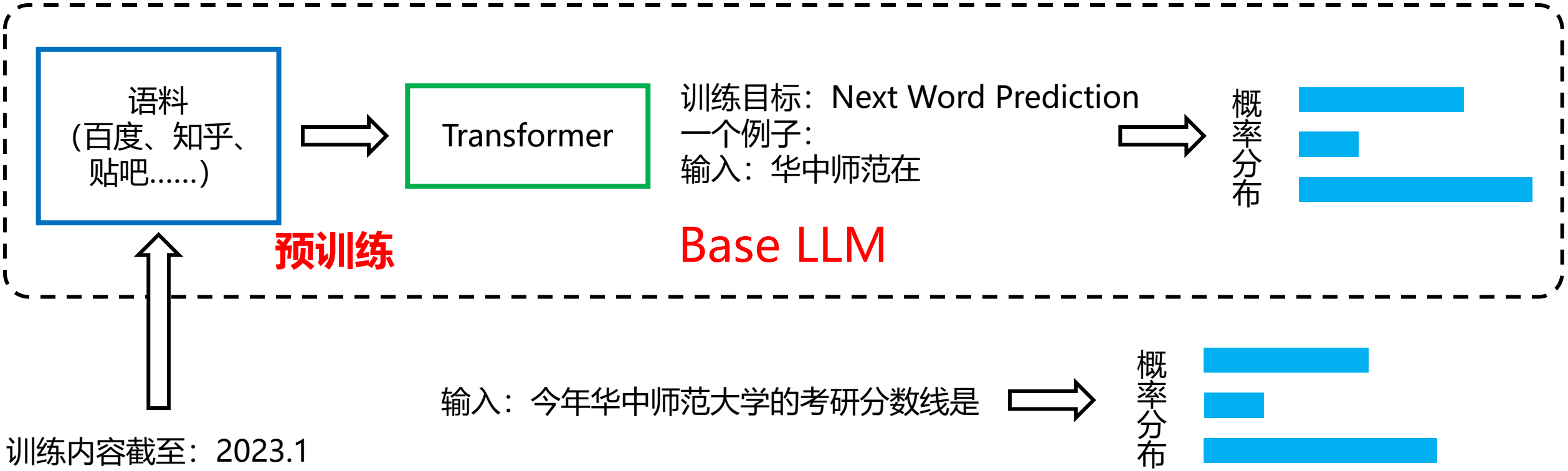
数据量：52K指令微调问答对（Alpaca）

数据源：任务密集型



学本领

预训练的天然弊端：语料不够新



理论上，模型是不知道正确答案的！

但问题不止于此……

输入：今年华中师范大学的考研分数线？

模型一定会给出一个输出！

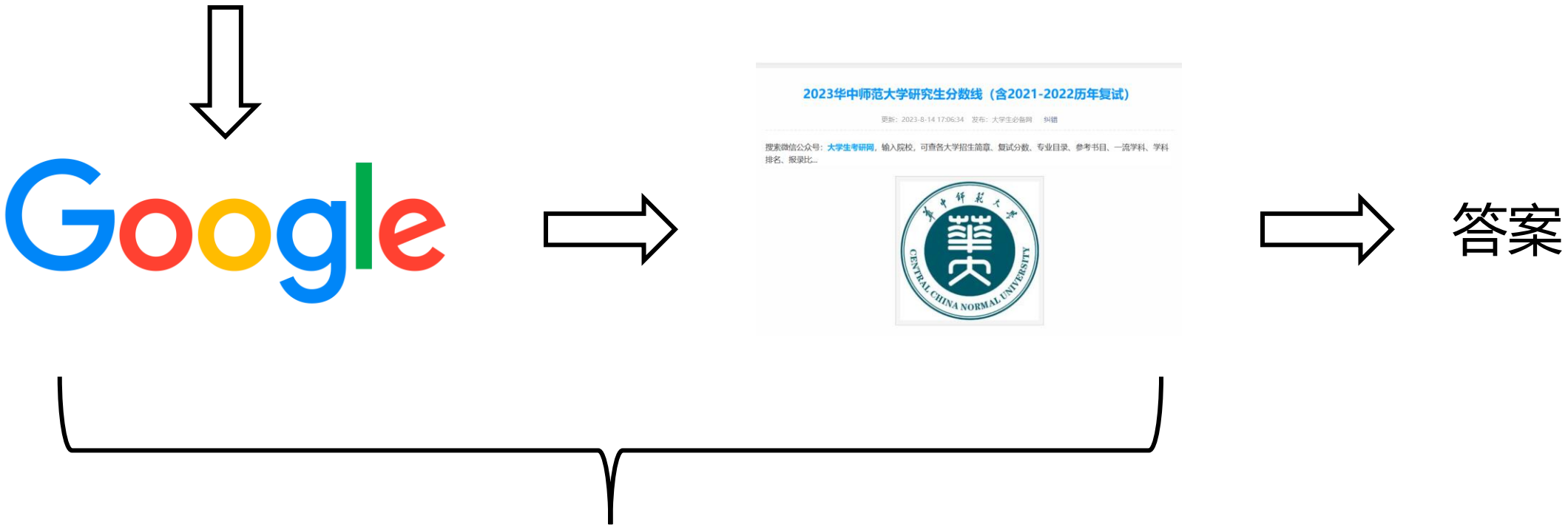
输出：今年华中师范大学的考研分数线是452

模型给出了一个看似正确，实则错误的答案

产生“幻觉”

人是如何解决这个问题的?

问题：今年华中师范大学教育学考研分数线?



检索增强的语言模型 ⟸ 内卷形（ODQA）

ACL 2023 Tutorial:
Retrieval-based Language Models and Applications

Akari Asai[1], Sewon Min[1], Zexuan Zhong[2], Danqi Chen[2]

[1]University of Washington, [2]Princeton University

Sunday July 9 14:00 - 17:30 (EDT) @ Metropolitan West

Visit this link for the Zoom recording of the tutorial

QnA: **tinyurl.com/retrieval-lm-tutorial**

ACL 2023 Tutorial: Retrieval-based LMs and Applications (acl2023-retrieval-lm.github.io)

Overview

用什么检索器?

阅读器怎么优化?

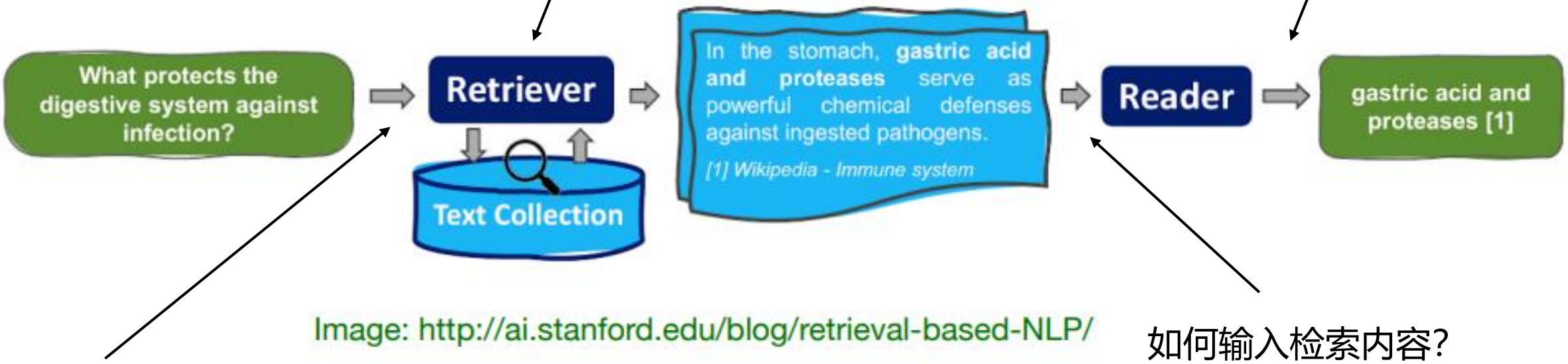


Image: http://ai.stanford.edu/blog/retrieval-based-NLP/

拿什么去检索?

如何输入检索内容?

检索部分

拿什么去检索?

研究动机：用户的问题并不一定适合检索（长度/语义）

解决方法：重写问题/多次检索

用什么检索?

常见方法：
BM25/DPR/ReRanker……

Tree of Clarifications: Answering Ambiguous Questions with R-A LLM; **EMNLP 23**

对于模糊问题，进行 query 澄清

Open-source Large Language Models are Strong Zero-shot Query Likelihood; **EMNLP 23**

用大模型做ReRanker

阅读部分

后处理是一个值得关注的方向！
我自己也在做

如何输入索内容？

研究动机：检索到的内容很长/存在错误

解决方法：精简检索证据/后处理

A Retrieval-Augmented Gaussian Mixture Variational Auto-Encoder; **EMNLP 23**

整合检索信息

阅读器如何优化？

中间层融合/动态生成-检索……

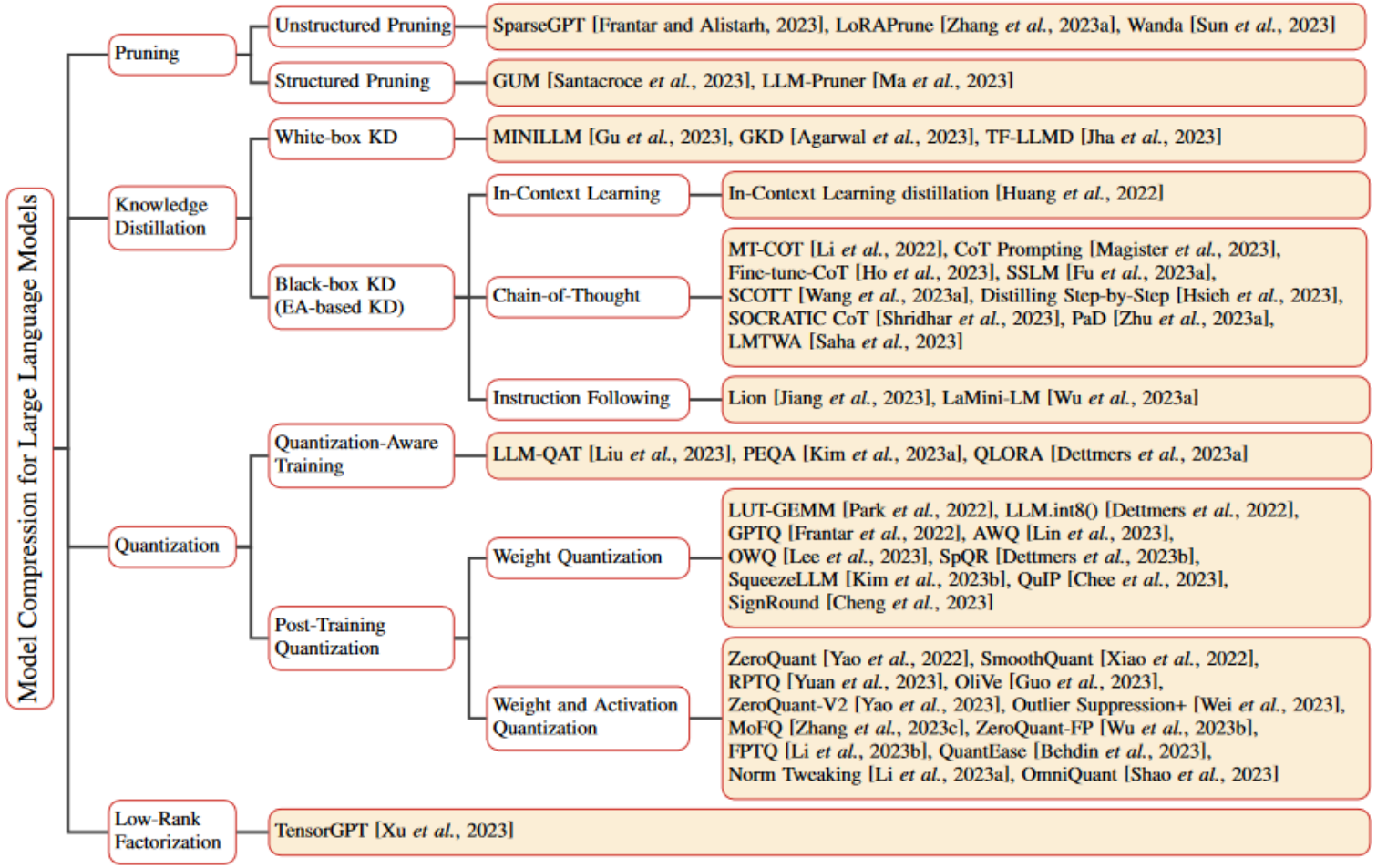
Active Retrieval Augmented Generation; **EMNLP 23**

（CMU）改进检索生成范式

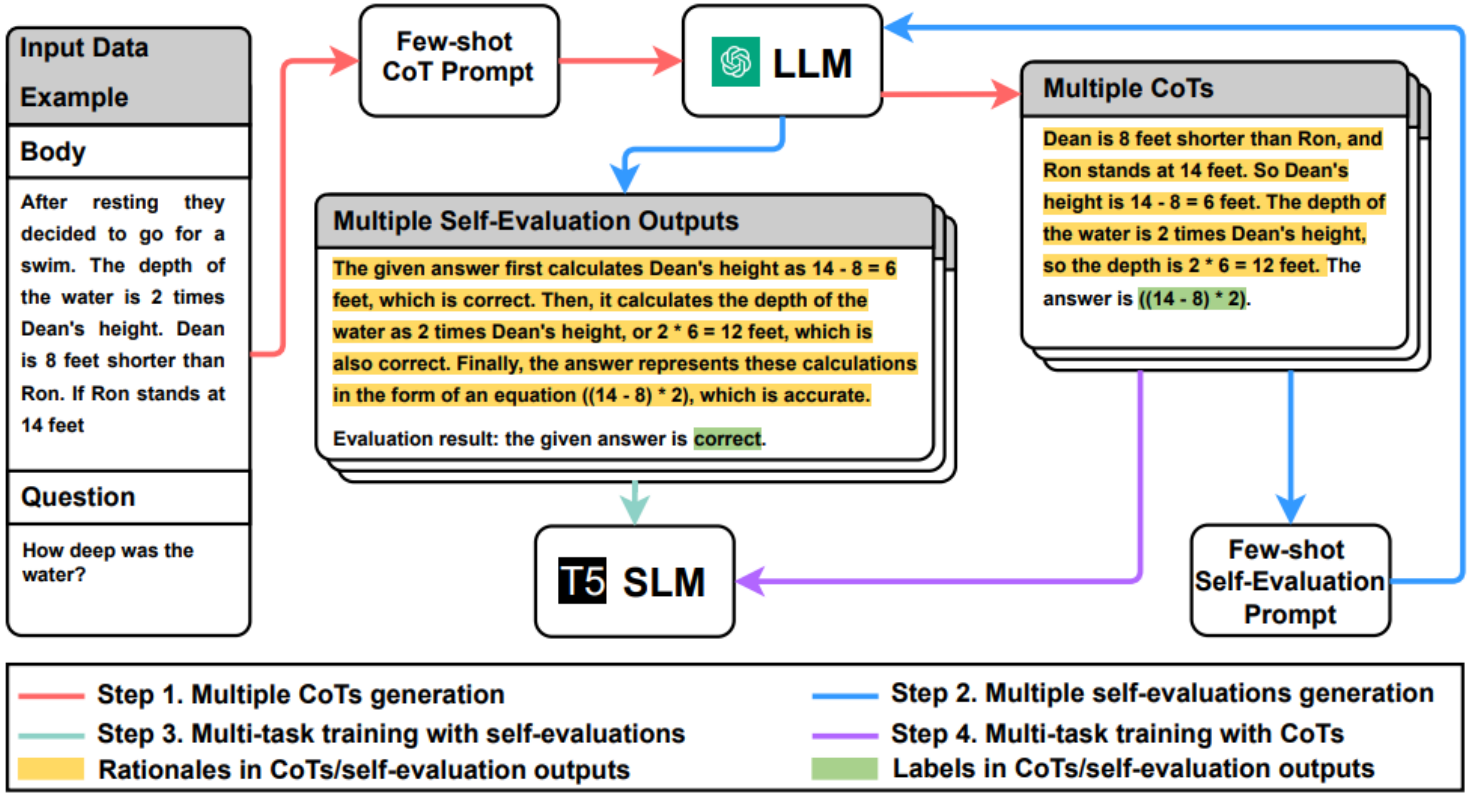
More Papers will be presented in my **SURVEY**!

欢迎关注我的个人网站

大模型的高效应用 ⟸ 外卷形（CV）



思维链蒸馏



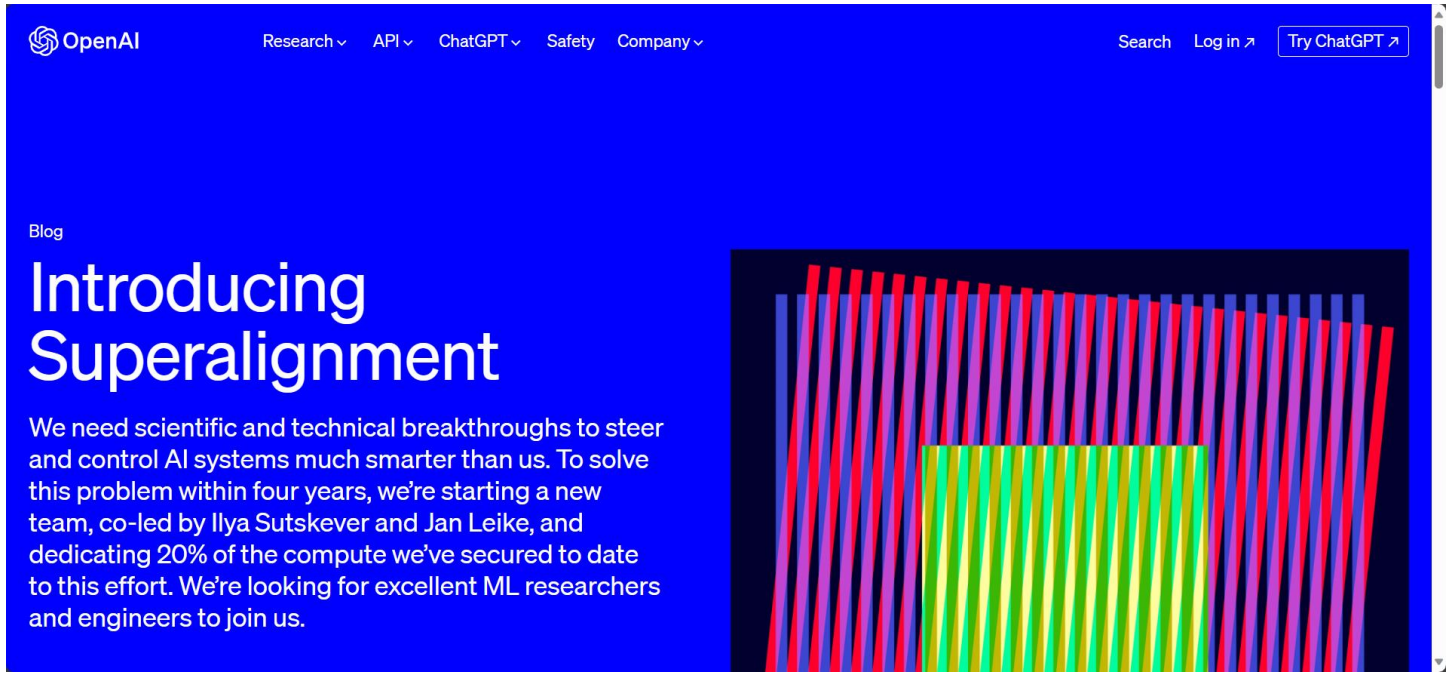
Mind's Mirror: Distilling Self-Evaluation Capability and Comprehensive Thinking from Large Language Models; Weize Liu, Guocong Li, Kai Zhang, Bang Du, **Qiyuan Chen**, Xuming Hu, Hongxia Xu, Jintai Chen, Jian Wu; **ArXiv Preprint**; 2023.

Superalignment ⟸ 天赋形

目的：和人类意图/价值观对齐（Align）

超级对齐的定义

所谓超级对齐，就是要求AI系统能够在各种复杂环境下，自发推导出符合人类价值观的行动方针。与简单的“把人类价值观硬编码进AI系统”不同，超级对齐需要AI自主推理人类的终极价值目标，在不同情形下做出判断,而不是单纯依靠设计者提供的价值观模型。



Introducing Superalignment (openai.com)

Reference

Do Models Explain Themselves? Counterfactual Simulatability of Natural Language Explanations;

Yanda Chen, ***Ruiqi Zhong***, Narutatsu Ri, Chen Zhao, He He, Jacob Steinhardt, Zhou Yu, Kathleen McKeown

RAIN: Your Language Models Can Align Themselves without Finetuning;

Yuhui Li, Fangyun Wei, Jinjing Zhao, Chao Zhang, Hongyang Zhang

A Rising Star

Ruiqi Zhong

My name is Ruiqi Zhong. I am currently a 5th year PhD student in the UC Berkeley EECS department, advised by Prof. Jacob Steinhardt and Prof. Dan Klein. I finished my undergrad at Columbia University, where I worked with Prof. Kathleen McKeown.

Email / Google Scholar / Twitter / Github

Research Overview

I work on scalable oversight -- supervising AI systems to accomplish tasks where humans alone struggle to determine the ground truth. Doing so requires human-AI collaborations, a better epistemic foundation, and new algorithmic tools. I currently work on concrete related problems in Natural Language Processing, Machine Learning, and Programming Language. See presentation slides here and my talk here to get a sense of my research interests.

谢谢大家！

陈启源

https://qiyuan-chen.github.io/