



GOTC 2023 全球开源技术峰会

THE GLOBAL OPENSOURCE TECHNOLOGY CONFERENCE

OPEN SOURCE, INTO THE FUTURE

「AI is Everywhere」专场

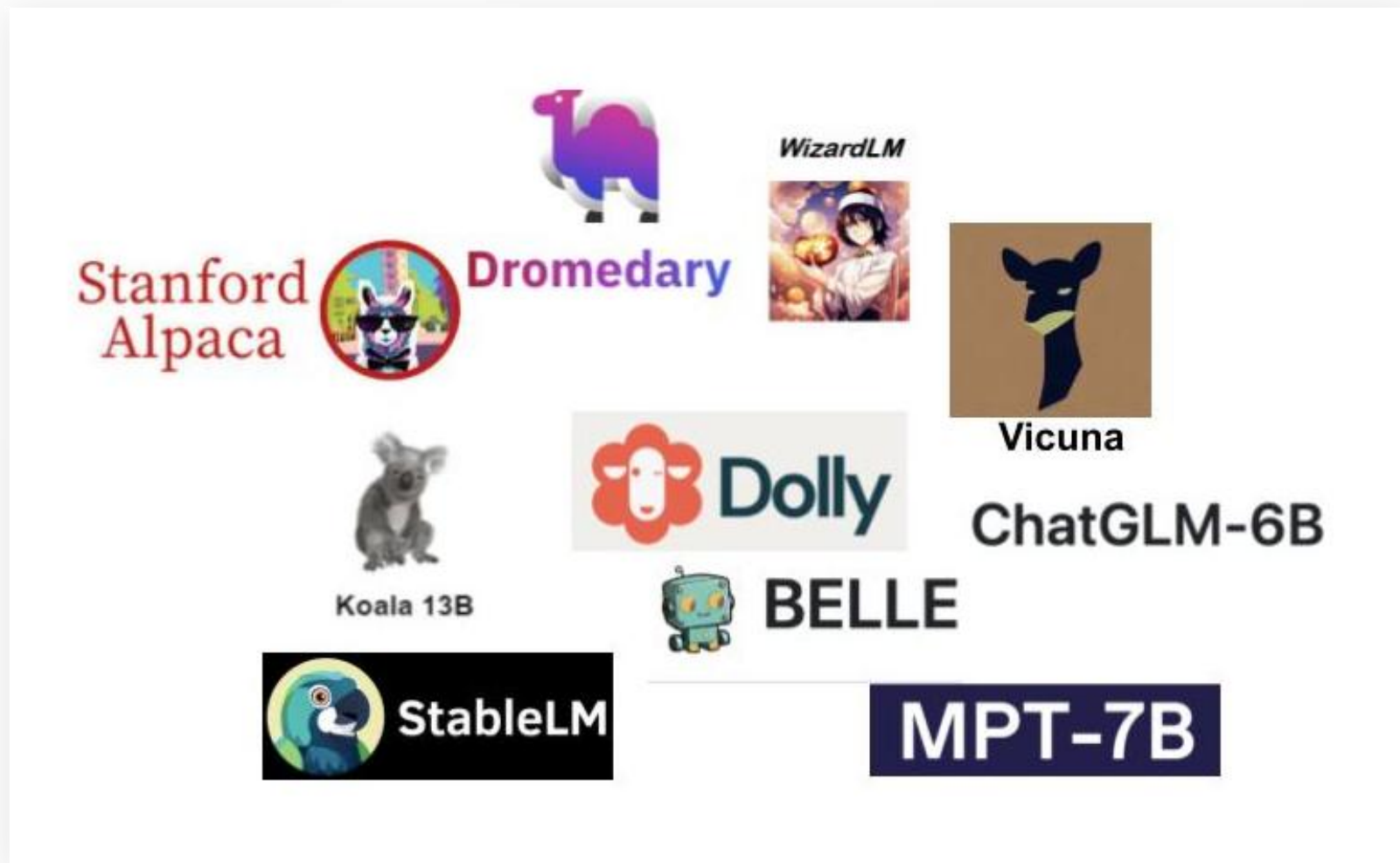
下一代知识工具：

以用户为中心的个性化语言模型和混合部署策略

常诚 2023年05月28日

- # LANGUAGE MODEL SIZES TO MAR/2023
-
- Legend:**
- Parameters (circle size)
 - AI lab/group (color)
 - Available (solid circle)
 - Closed (circle with slash)
 - Chinchilla scale (star)
- Model Data:**
- | Model | Size (Parameters) | AI Lab/Group | Status | Notes |
|-------------------|-------------------|-------------------|-----------|-------|
| BERT | 340M | Google | Available | |
| GPT-1 | 117M | OpenAI | Available | |
| GPT-2 | 155B | OpenAI | Available | |
| T5 | 11B | Google | Available | |
| Megatron-11B | 11B | Meta | Available | |
| ruGPT-3 | 11B | RU.nl | Available | |
| Plato-XL | 11B | Plato | Available | |
| Macaw | 11B | Macaw | Available | |
| Cohere | 11B | Cohere | Available | |
| GPT-NeoX-20B | 20B | OpenAI | Available | |
| CM3 | 13B | CM3 | Available | |
| VLM-4 | 10B | VLM-4 | Available | |
| mGPT | 13B | mGPT | Available | |
| BLOOM | 176B | BLOOM | Available | |
| BLOOMZ | 176B | BLOOM | Available | |
| Atlas | 11B | Atlas | Available | |
| Flan-T5 | 11B | Flan-T5 | Available | |
| Kosmos-1 | 1.8B | Kosmos-1 | Available | |
| GLM-130B | 130B | GLM | Available | |
| ChatGLM-6B | 6B | ChatGLM | Available | |
| NLLB | 54.5B | NLLB | Available | |
| OPT-175B | 175B | Meta | Available | |
| BB3 | 175B | BB3 | Available | |
| OPT-IML | 175B | OPT-IML | Available | |
| MOSS | 20B | MOSS | Available | |
| GPT-4 | Undisclosed | OpenAI | Available | |
| LLaMA | 65B | Meta | Available | |
| Alpaca | 7B | Alpaca | Available | |
| Toolformer | 10B | Toolformer | Available | |
| Galactica | 120B | Galactica | Available | |
| PaLM | 540B | Google | Available | |
| PaLM-Coder | 540B | Google | Available | |
| Minerva | 540B | Google | Available | |
| Med-PaLM | 540B | Google | Available | |
| Fian-PaLM | 540B | Google | Available | |
| U-PaLM | 540B | Google | Available | |
| Fian-U-PaLM | 540B | Google | Available | |
| Med-PaLM 2 | 540B | Google | Available | |
| MT-NLG | 530B | MT-NLG | Available | |
| XGLM | 75B | XGLM | Available | |
| Cedille | 6B | Cedille | Available | |
| Fairseq | 13B | Fairseq | Available | |
| Gopher | 280B | Gopher | Available | |
| Chinchilla | 70B | Chinchilla | Available | |
| Flamingo | 80B | Flamingo | Available | |
| 52B RL-CAI Claude | 52B | 52B RL-CAI Claude | Available | |
| Anthropic-LM | 6B | Anthropic-LM | Available | |
| GPT-J | 6B | GPT-J | Available | |
| BlenderBot2.0 | 9.4B | BlenderBot2.0 | Available | |
| LaMDA | 137B | Google | Available | |
| LaMDA 2 | 137B | Google | Available | |
| Bard | 137B | Google | Available | |
| SeeKer | 2.7B | SeeKer | Available | |
| Z-Code++ | 20B | Z-Code++ | Available | |
| Gato | 1.2B | Gato | Available | |
| FIM | 0.5B | FIM | Available | |
| AlexaTM | 200M | AlexaTM | Available | |
| VIMA | 200M | VIMA | Available | |
| WeLM | 10B | WeLM | Available | |

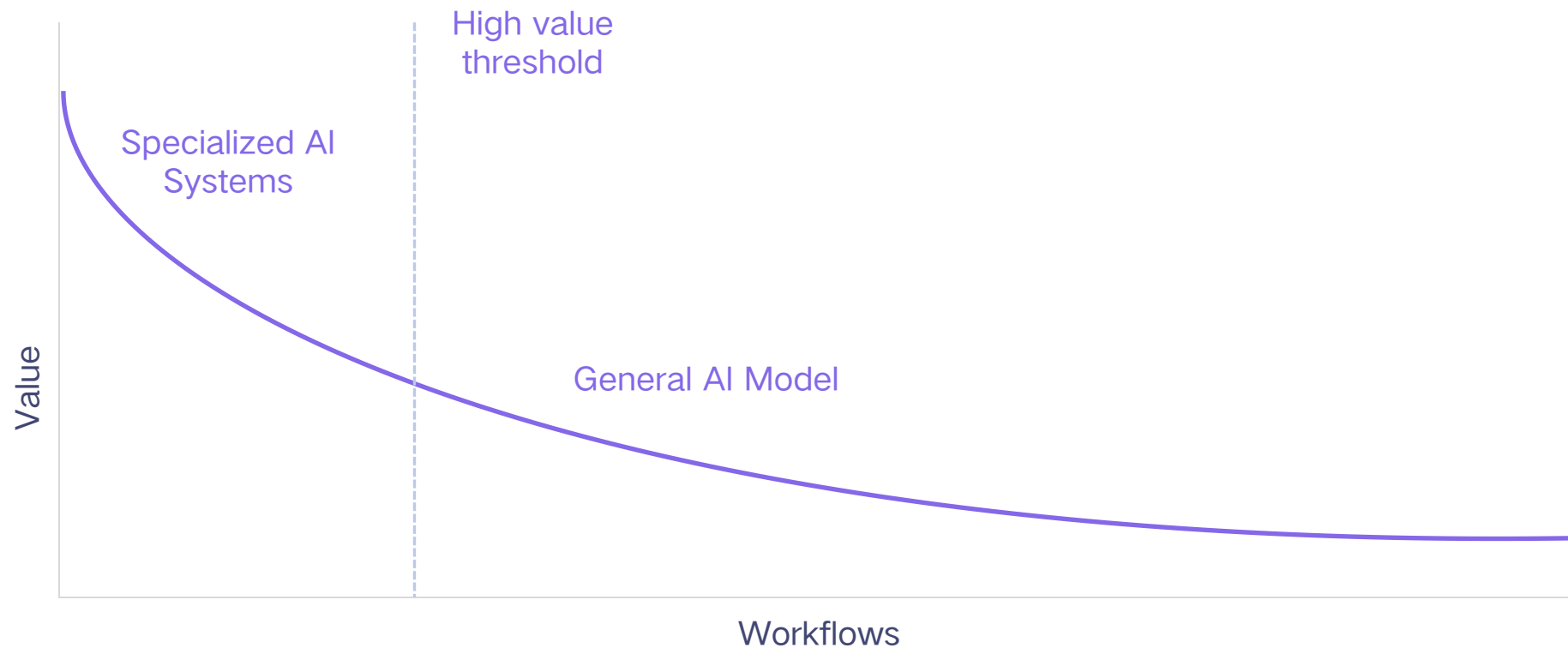
- 指令数据集微调
 - 上下文理解
 - 推理增强
- 插件增强
- 模型小型化





- 提示工程 Prompt Engineering
- 关键信息提示 Few-shot Prompting
- 检索 + 提示 Retrieval + Prompting (Embedding Indexing)
- 迭代式优化 Iterative Refinement (Prompt decomposition)
- 托管模型调优 Fine-tuning hosted model
- 开源模型调优 Fine-tuning Open Source LLM
- 开源模型零基础训练 Training Open Source LLM from scratch
- 自定义模型零基础搭建 Create custom LLM from scratch

为何要做大象GPT?



Does One Large Model Rule Them All? 2023.4

Maithra Raghu (Samaya AI) Matei Zaharia (Databricks), Eric Schmidt (Schmidt Futures)

- 专有AI系统的重要性
 - 高价值工作流程需要专业化AI的定制适应性
- 利用用户反馈
 - 专有语言模型更容易迭代，优化用户反馈利用，减少数据筛选工作量
- 专有数据和专有知识的价值
 - 专有语言模型需要独特的数据加工流程
- 专有模型与内外部服务互为补充
 - 专有模型处理特定问题更佳
 - 提供更可控的”路由器”
- AI模型的商业化
 - 开发效率和部署成本是关键竞争因素



专有大语言模型（大象 GPT）训练



基于用户行为和数据的混合部署策略

Public Data Mixture
through Self-instruct
(Code/QA/Story/Wiki等)

YX Public data
(剪藏/识堂等)

Candidate
Dataset

Instruction
Dataset

Instruction
Augmentation

Fine-tuning

大象GPT

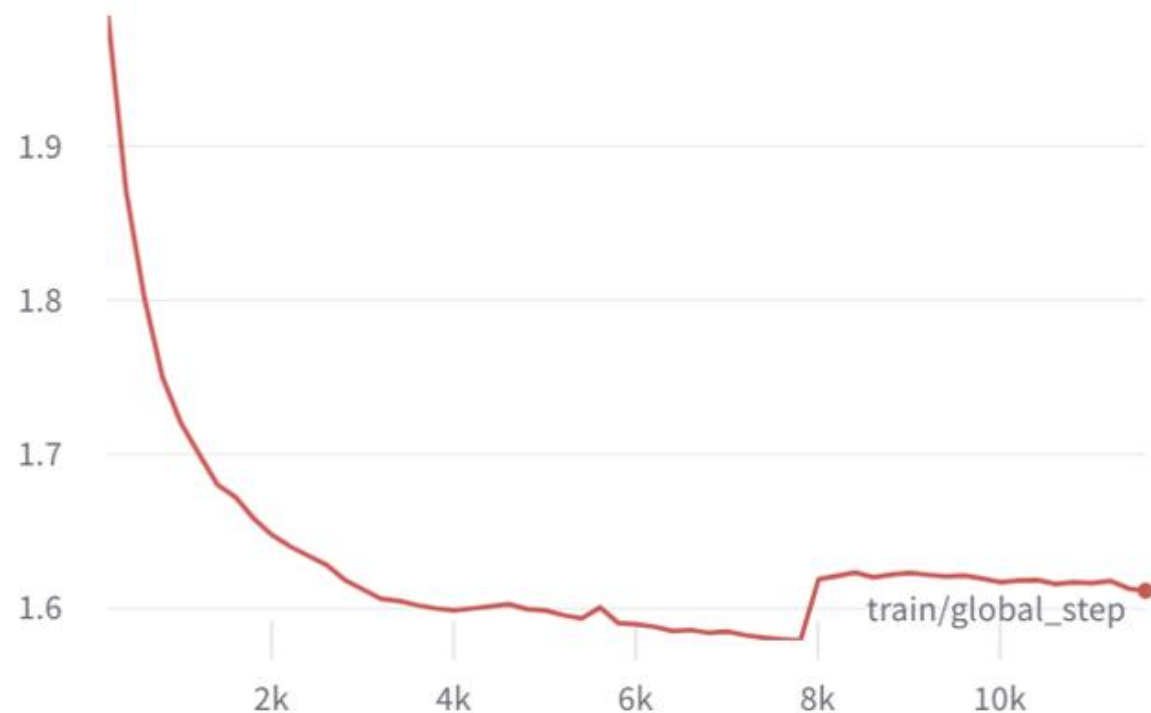
Pretrained LLMs

- 训练数据的构成及混合比例
- 指令集数据的加工

MODEL	REPRESENTED DOMAINS (%)									FILTERS		DATA	
	WIKI	WEB	BOOKS	DIALOG	CODE	ACAD	PILE	C4	M-L	TOX	QUAL	PUB	YEAR
BERT	76		24				✗	✗			H	Part	2018
GPT-2		100					✗	✗			H	Part	2019
RoBERTA	7	90	3				✗	✓			H	Part	2019
XLNet	8	89	3				✗	✓			H	Part	2019
T5	<1	99					✗	✓		H	H	✓	2019
GPT-3	3	82	16				✗	✓	7%		C	✗	2021
GPT-J/Neo	1.5	38	15	4.5	13	28	✓	Part			C	✓	2020
GLaM	6	46	20	28			✗	✓			C	✗	2021
LAMDA	13	24		50	13		✗	✓	10%	C	C	✗	2021
ALPHA CODE					100		✗	✗			H	✗	2021
CODEGEN	1	24	10	3	40	22	✓	Part			H	Part	2020
CHINCHILLA	1	65	10		4		✗	✓		H	C	✗	2021
MINERVA	<1	1.5	<1	2.5	<1	95	✗	✓	<1%		C	✗	2022
BLOOM	5	60	10	5	10	10	✓	✓	71%		C	Part	2021
PALM	4	28	13	50	5		✗	✓	22%		C	✗	2021
GALACTICA	1	7	1		7	84	✗	Part			H	Part	2022
LLAMA	4.5	82	4.5	2	4.5	2.5	Part	✓	4%		C	Part	2020

- Datasize VS Model implementation
- When to stop?

$$|D| = -\log P_f(D) + |f|$$





菜单式 “Shopping”

GUI vs LUI

HCI 人机交互界面

开放对话式 “Creation”



询问有关: Hive调优学习.pdf

这篇文档讲述了Hive调优的内容，包括建表设计和严格模式等。Hive语句执行完毕后，JVM进程会退出。

您还可以选择以下问题向我提问，或者在下方输入问题询问：

1. Hive调优包括哪些内容？

2. Hive语句执行完毕后会发什么？

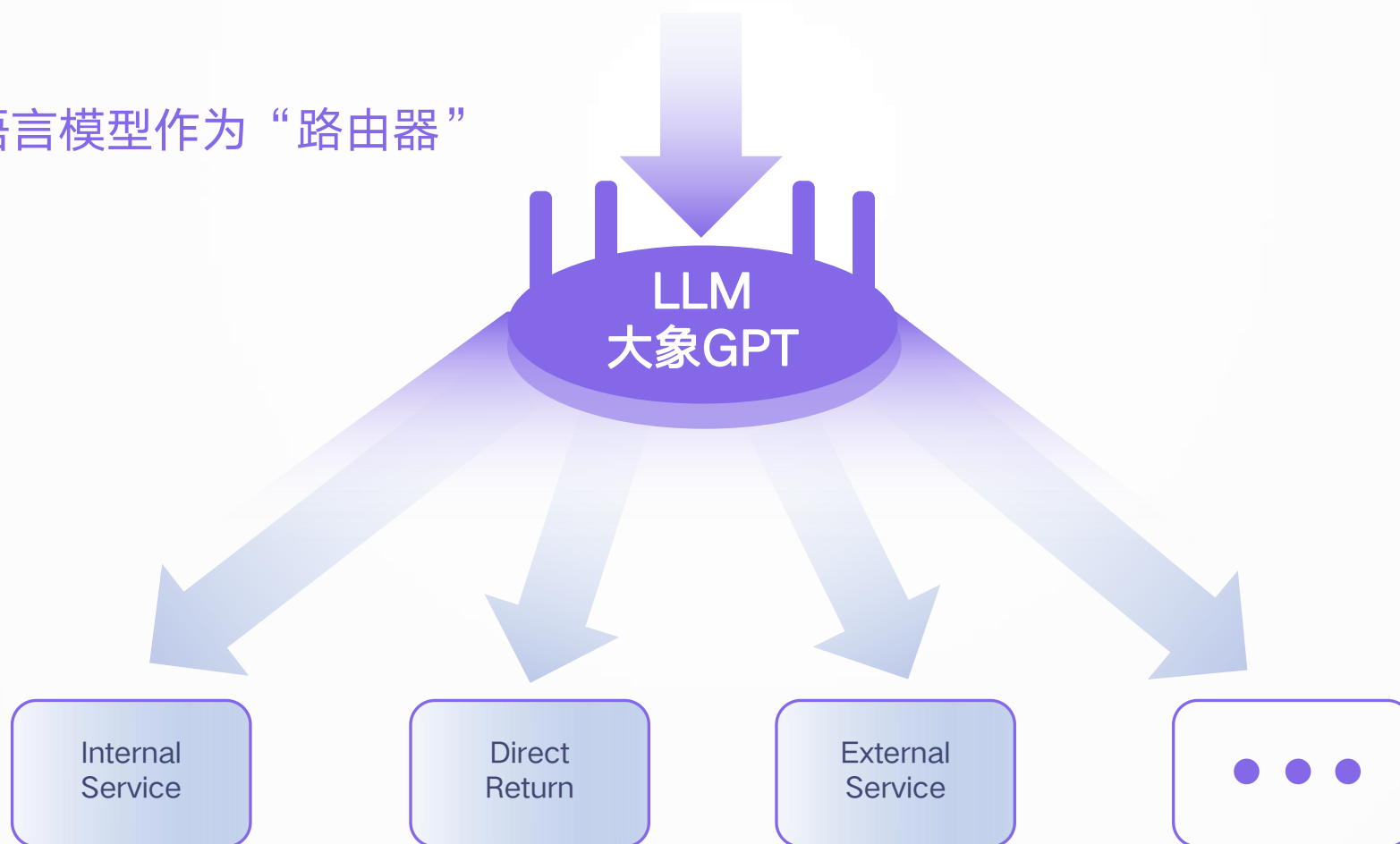
3. 用户如何使用严格模式？

1. Hive调优包括哪些内容？

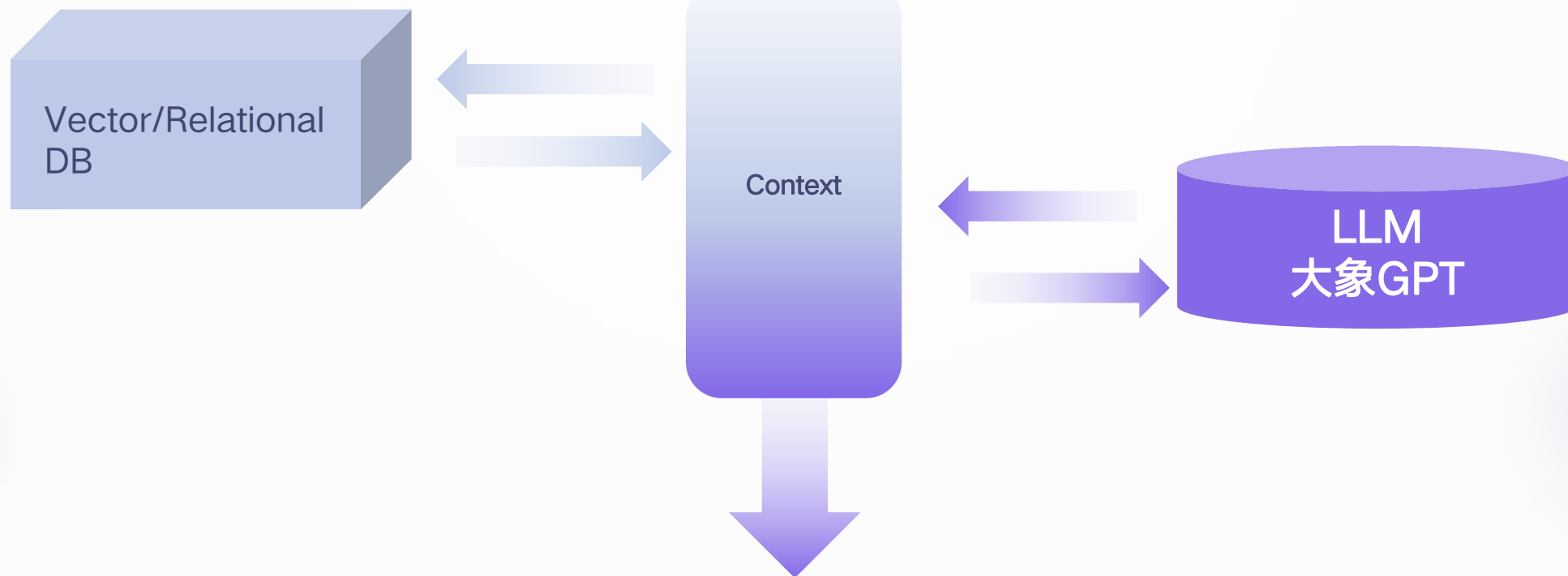
根据文章内容，Hive调优主要分为三个方面展开，包括对Hive的建表设计方面的优化、对HiveHQL语句本身的优化以及对底层引擎MapReduce方面的调整。在调优过程中，需要特别注意数据倾斜、数据冗余、Job或I/O过多、MapReduce分配不合理等问题，以降低资源使用量、减少任务执行时间，同时保证业务结果不变。值得注意的是，Hive的复杂HQL底层会转

输入更多你想询问的问题

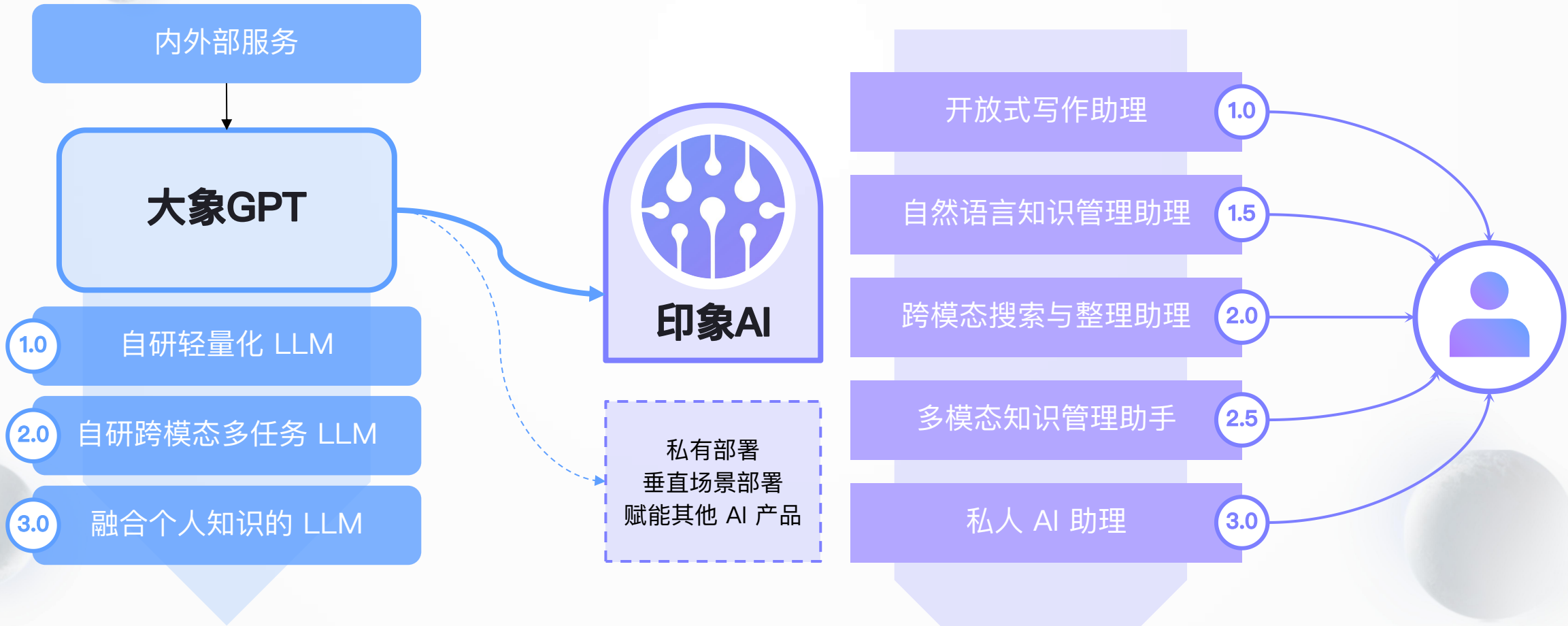
模式一：大语言模型作为“路由器”



模式二：大语言模型作为“推理器”



印象AI | 下一代知识工具的未来



THANKS