中国电信人工智能研究院
Institute of AI, China Telecom

大模型数据建设探索与实践

赵 宇

中国电信人工智能研究院
(TeleAI)

2024年 8月

TeleAI（中国电信人工智能研究院）

由中国电信集团CTO、首席科学家**李学龙**教授（Prof. Dr. Xuelong Li）发起并组建，围绕大模型、具身智能、AIGC等方向开展基础研究、技术攻关和应用落地，旨在打造人工智能研发与产业转化标杆性平台

大模型核心研发成果

- **TeleChat** TeleAI 自主研发的通用语义大模型基座，已开源1B、7B、12B、52B等多个版本
- **TeleChat-PTD** 开源 1TB 综合性中文预训练数据集，来源于网页、书籍、官方媒体
- **Tele-FLM-1T** TeleAI 与北京智源研究院共同研发，全球首个开源稠密万亿参数大模型

https://github.com/Tele-AI/Telechat
TeleChat Technical Report
52B to 1T: Lessons Learned via Tele-FLM Series

- 从工程化视角看数据建设
- 预训练数据配比
- 后训练数据筛选

01

从工程化视角看数据建设

工程化思考

- 问题：从零开始训练千亿参数大模型，任务流程是如何执行的？

设想中的训练流程

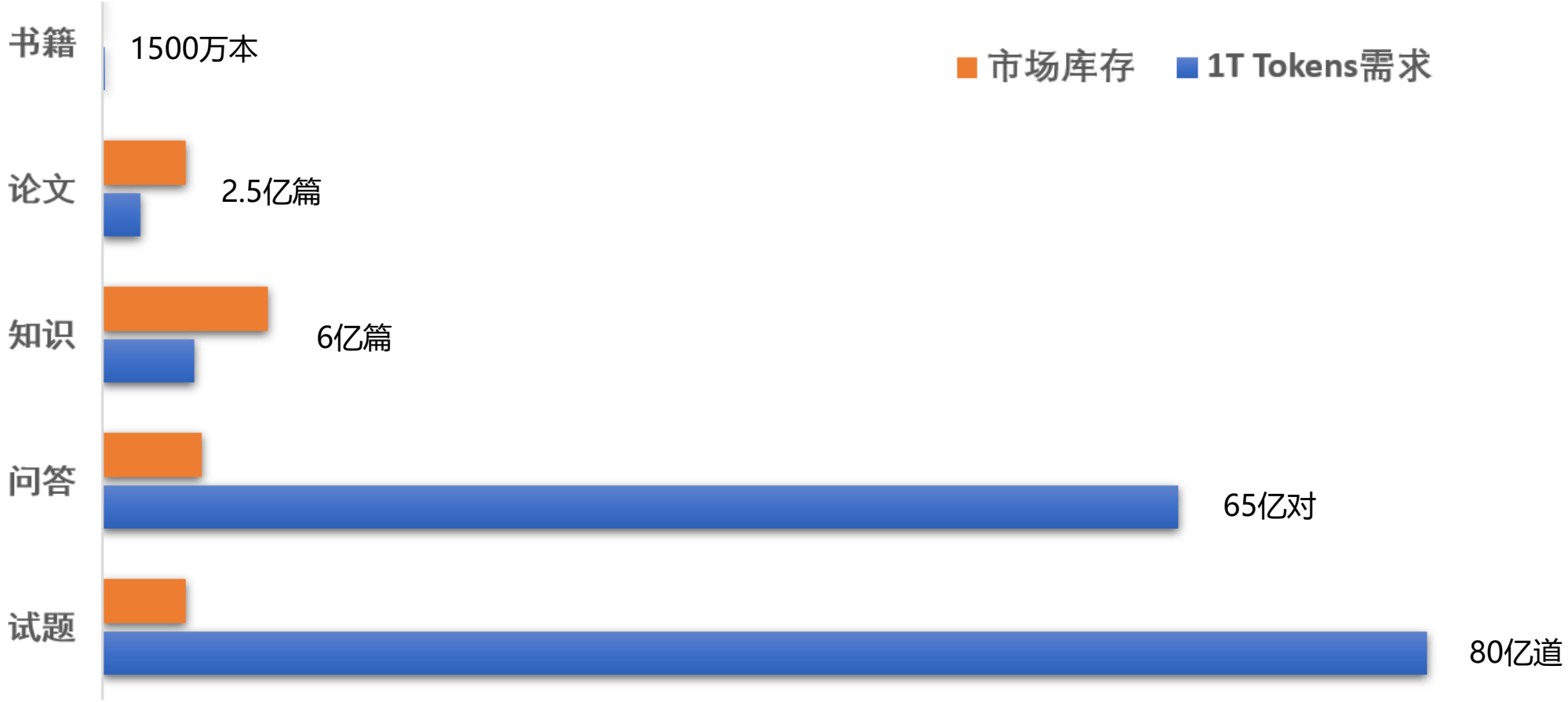
工程化思考

- 超大规模参数的模型预训练，时间跨度往往长达几个月
- 伴随模型训练进度变化，数据版本必然会持续动态调整

实际上的训练流程

影响数据版本更新的因素

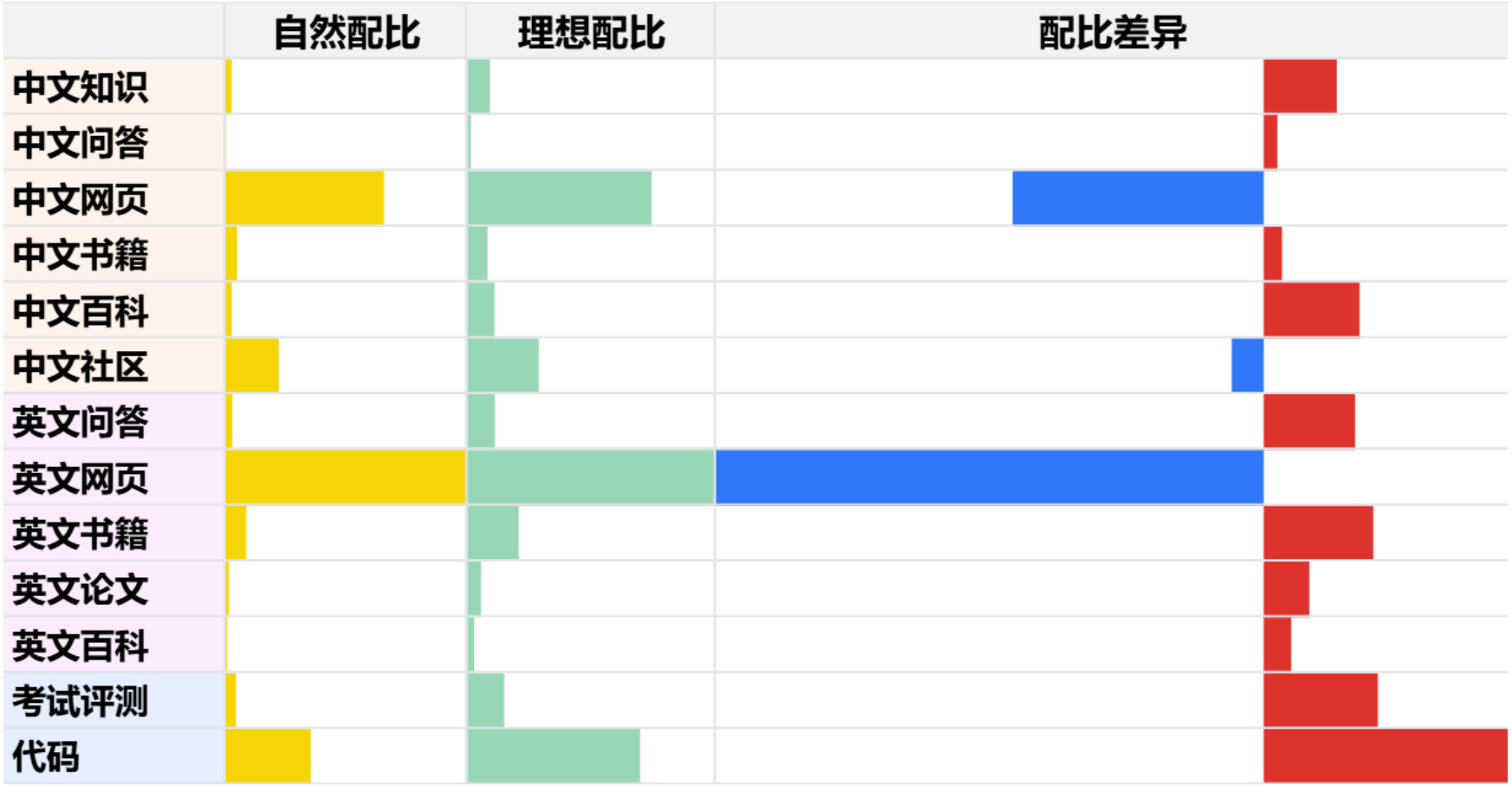
- **数据来源变更 & 新数据引入**



新增 1T Tokens 预训练数据，需要采集多少语料

影响数据版本更新的因素

- 自然配比与理想配比

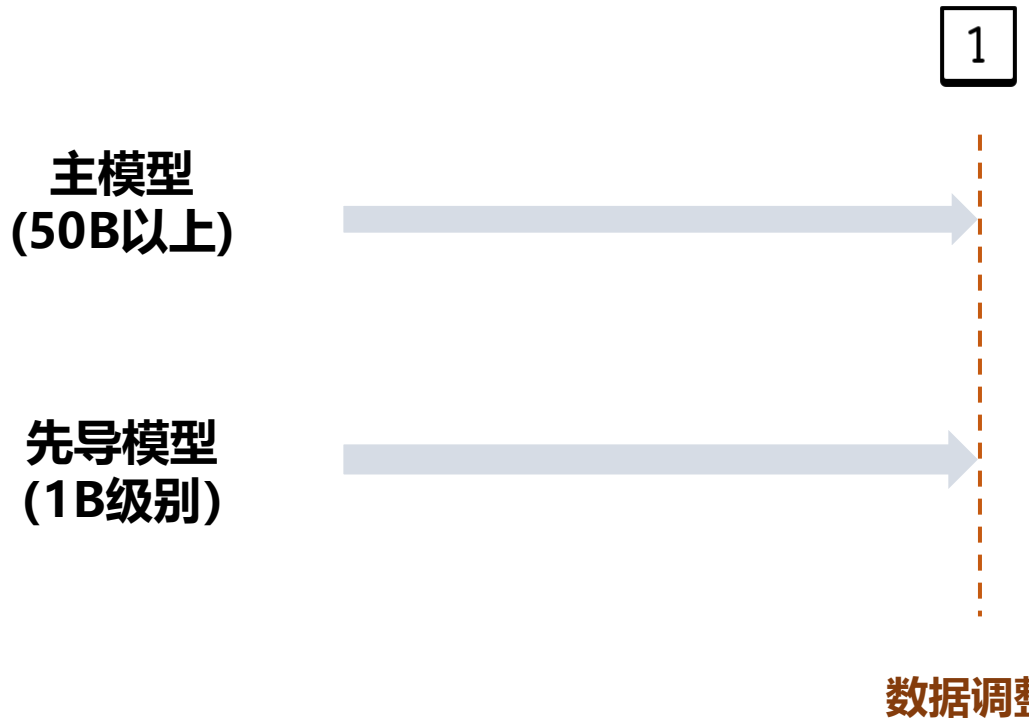


红色代表数据消耗速度过快

蓝色代表数据消耗速度偏慢

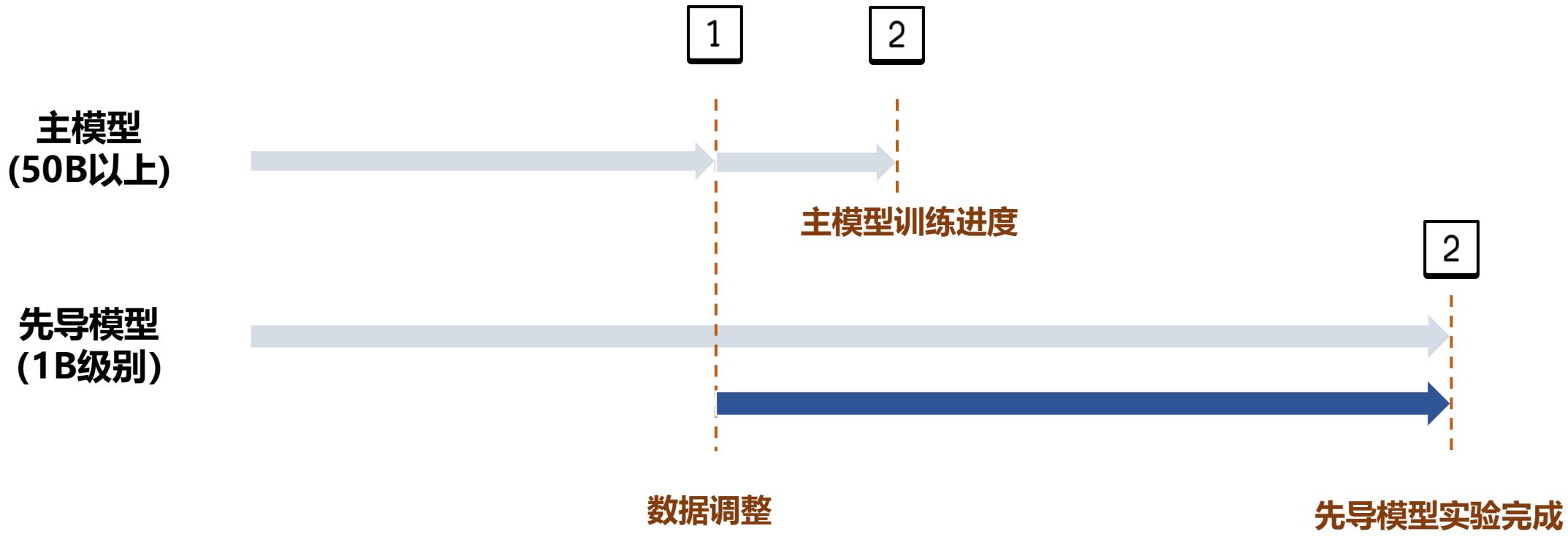
工程化实践

- 先导模型



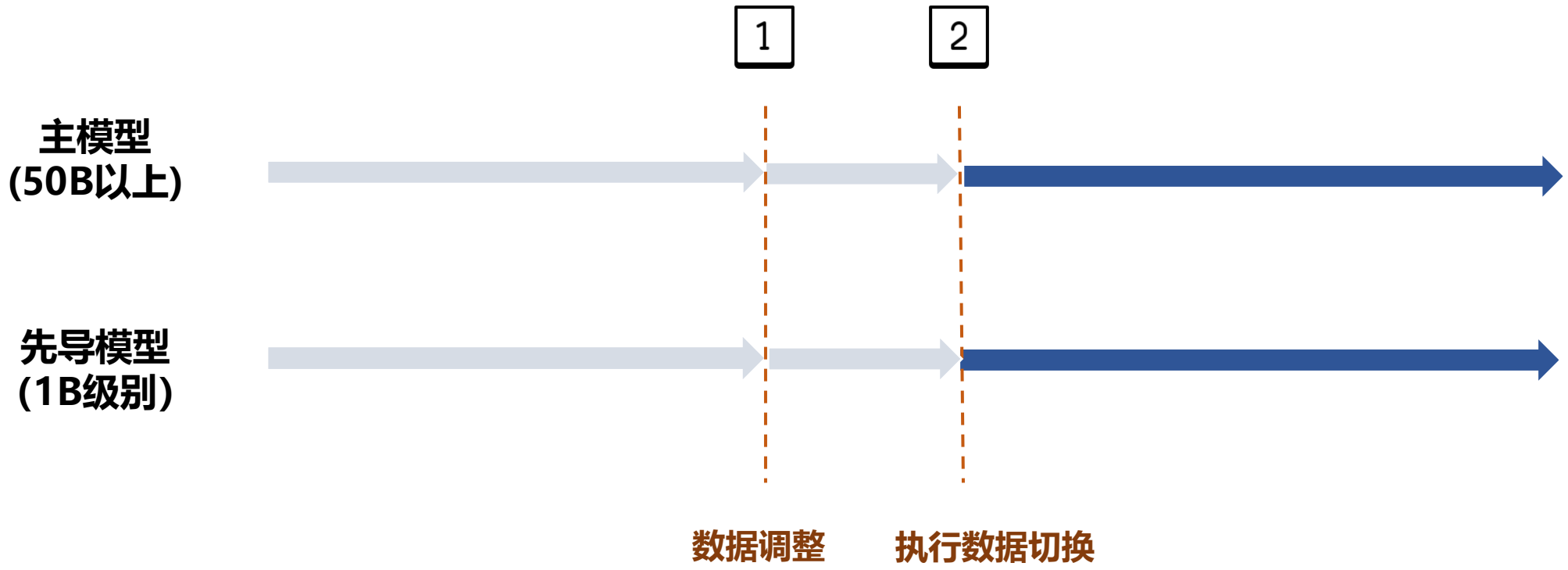
工程化实践

- 先导模型



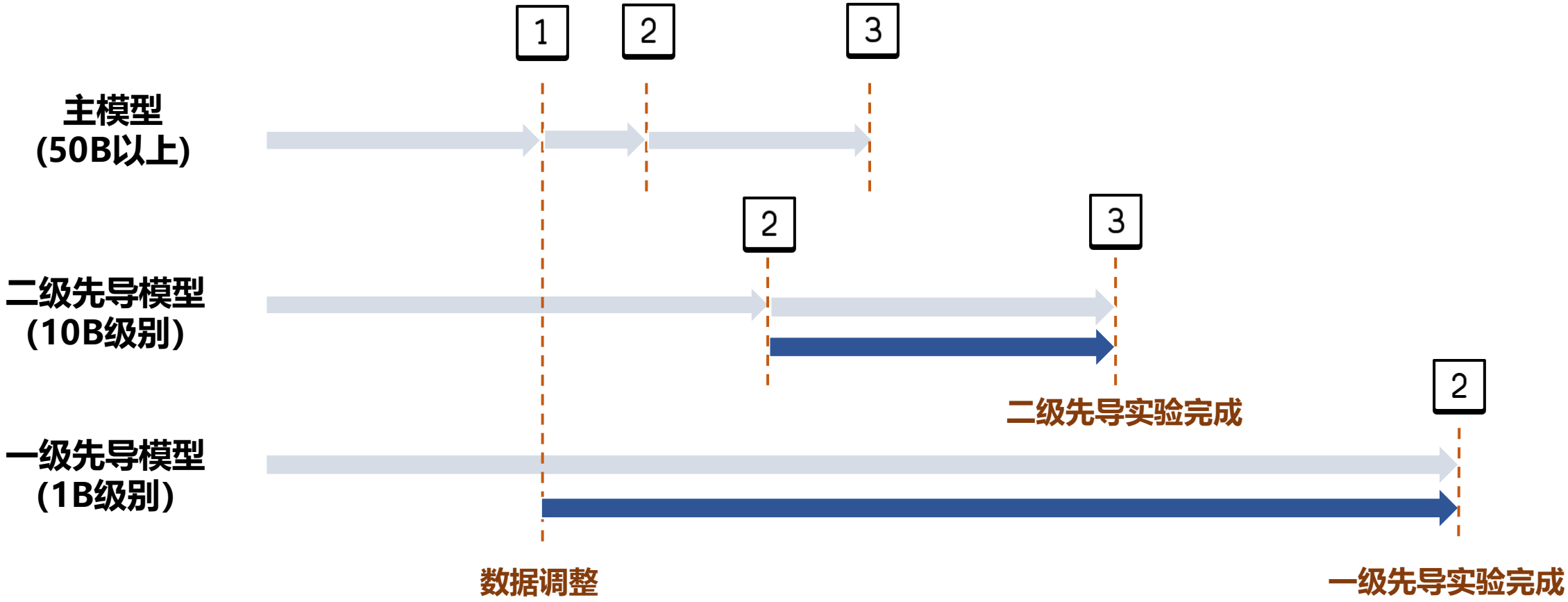
工程化实践

- 先导模型



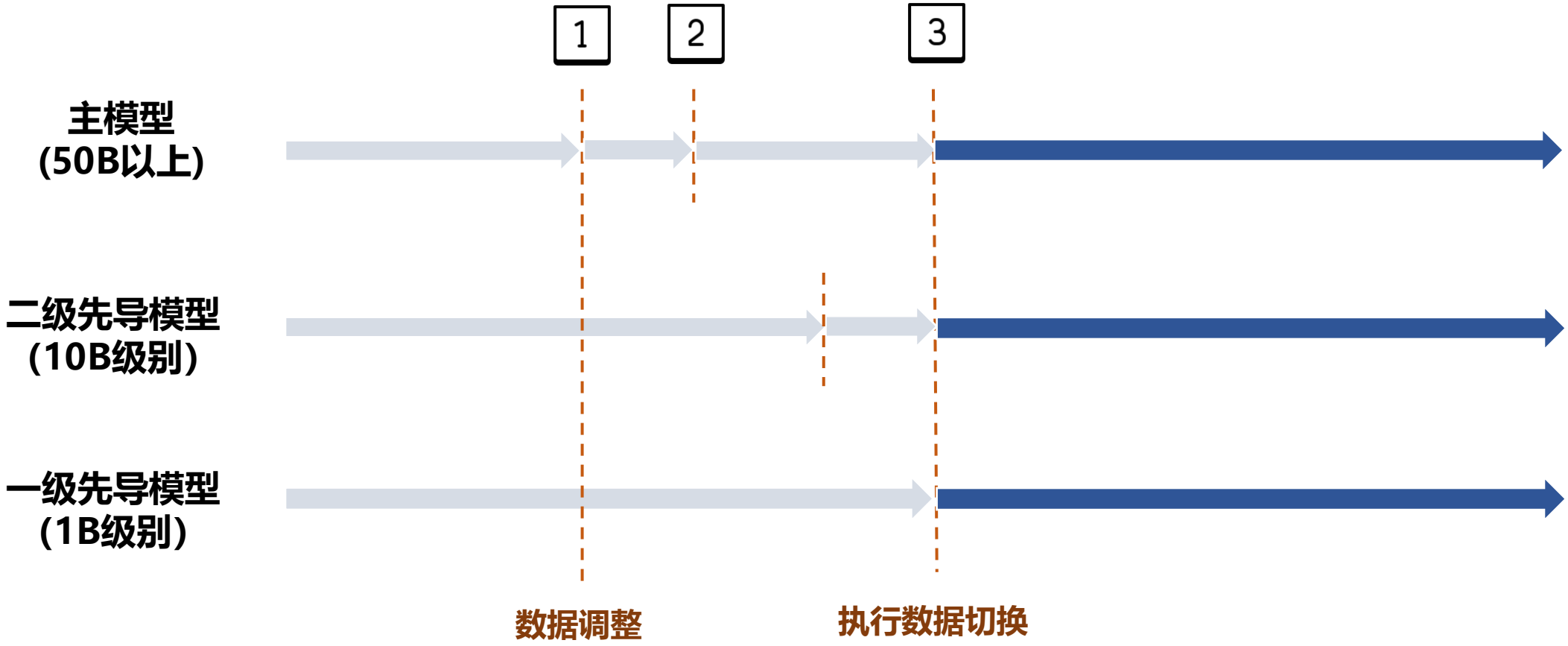
工程化实践

- 多级先导模型



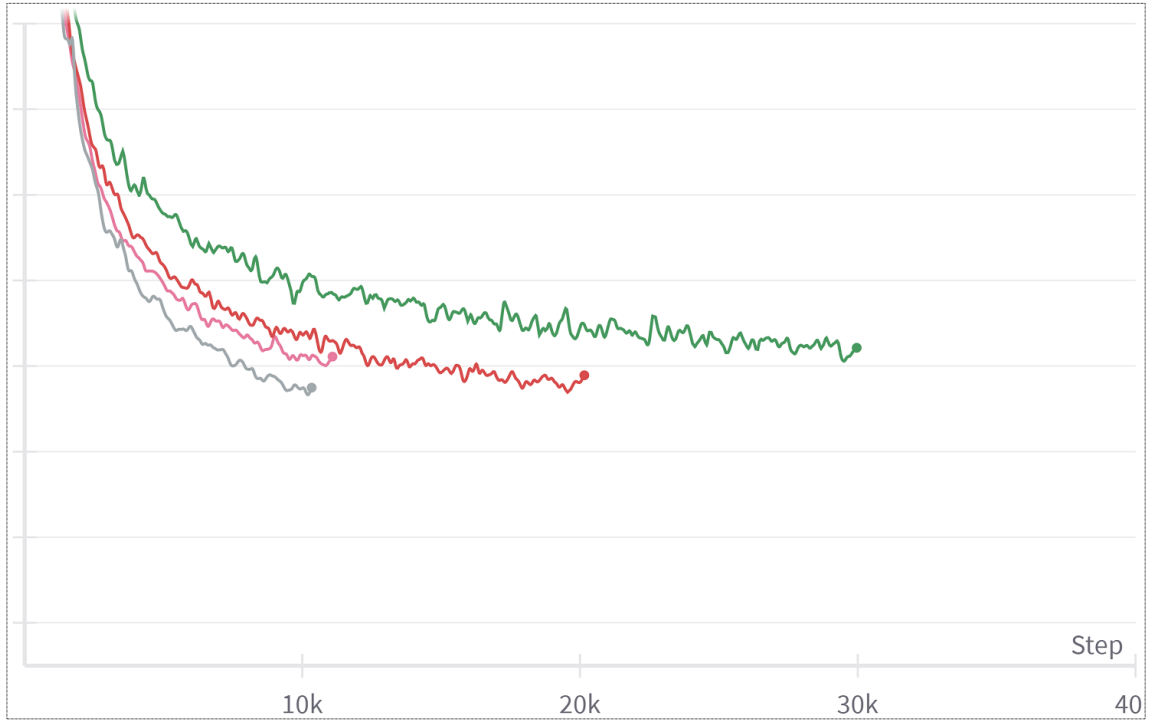
工程化实践

- 多级先导模型

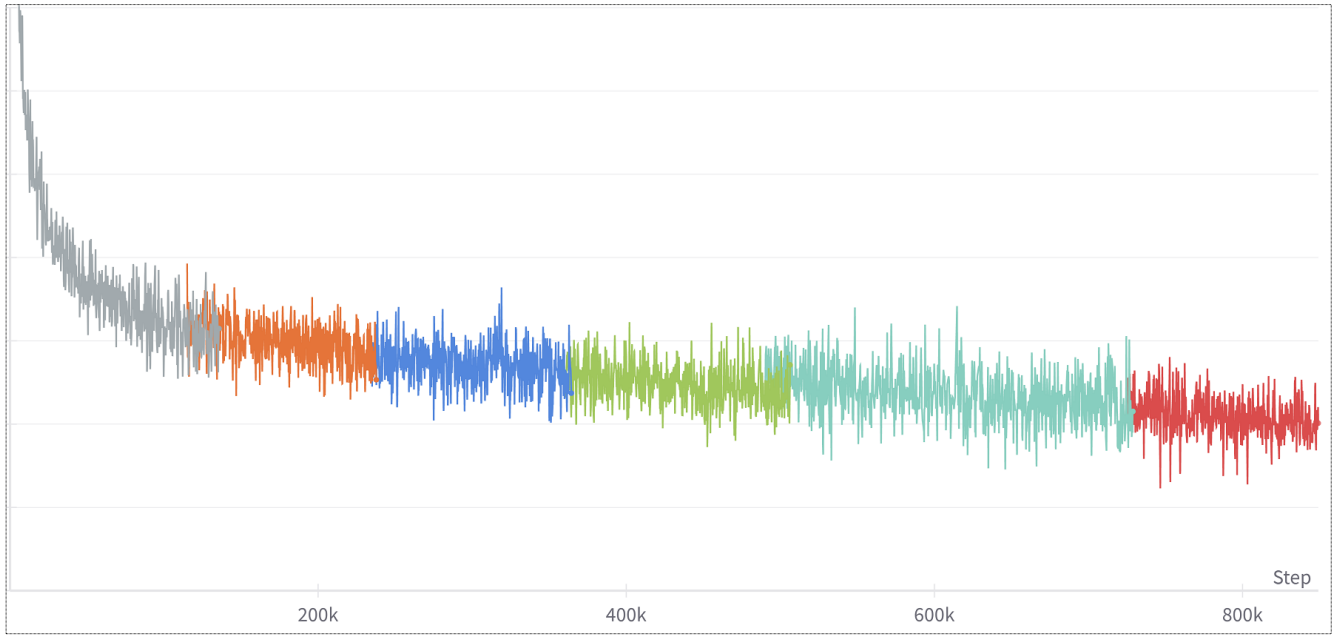


工程化实践

多级先导模型loss曲线



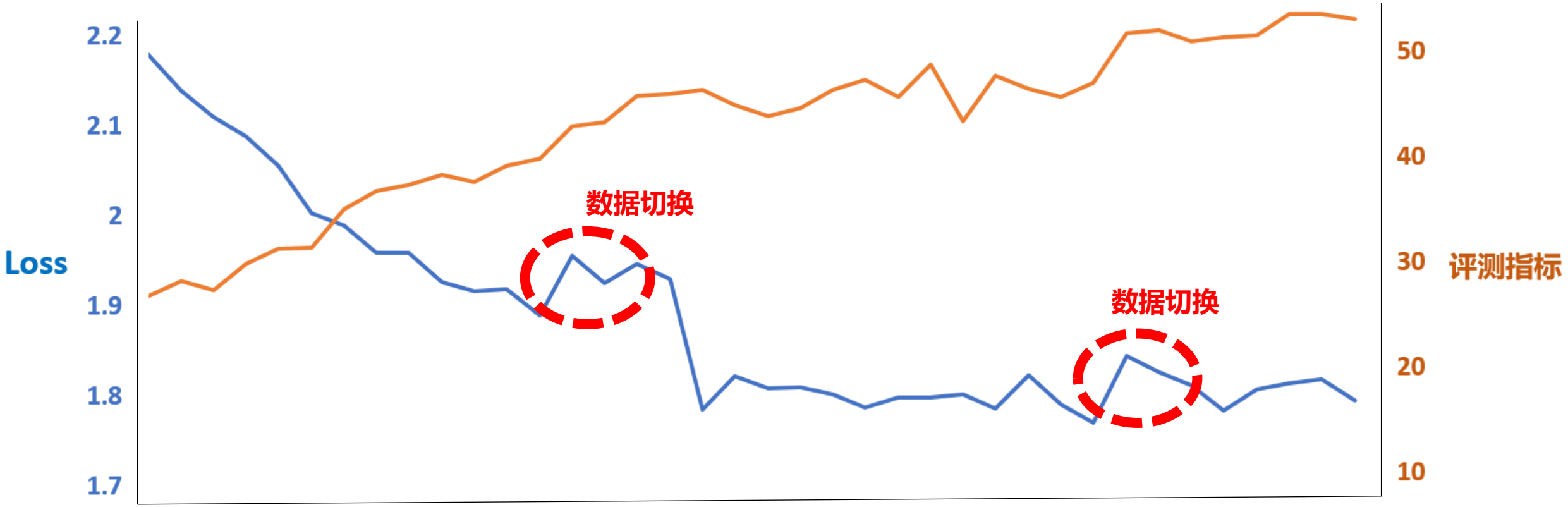
主模型数据块切换loss曲线



即使不做数据调整，分批制作数据块也更符合工程化设计

影响数据版本更新的因素

- 先导模型评测实验效果



先导模型loss曲线

影响数据版本更新的因素

- 不同的数据版本变更方式，触发不同的数据处理流程

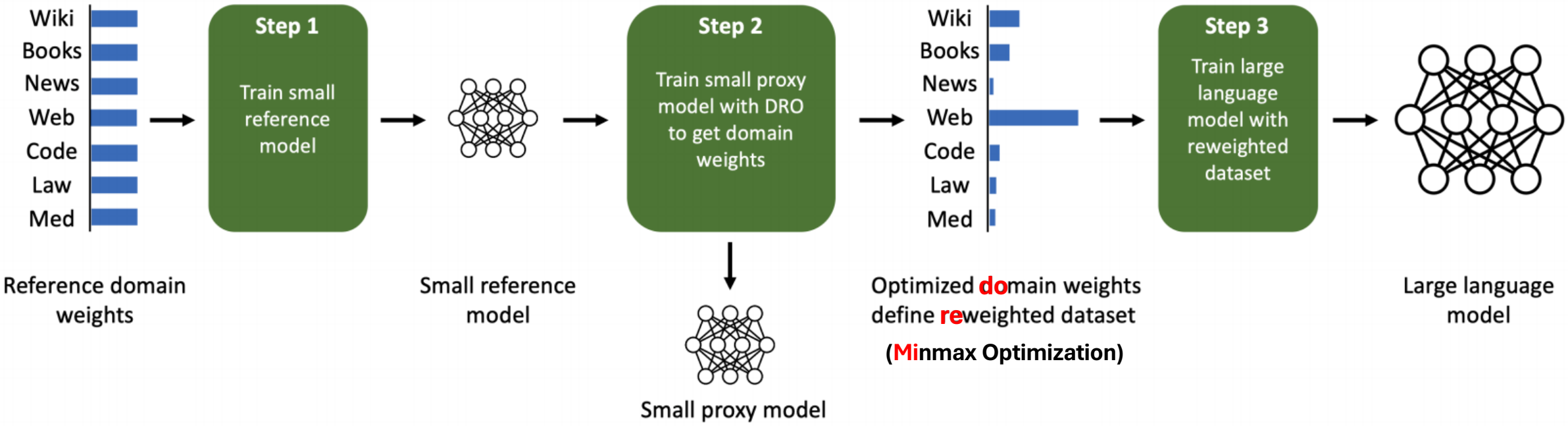
Albalak et al. 2024. A Survey on Data Selection for Language Models

02

预训练数据配比

数据混合 (Data Mixing)

- **DoReMi：不依赖特定下游任务，在小模型上寻找最优数据混合比例**

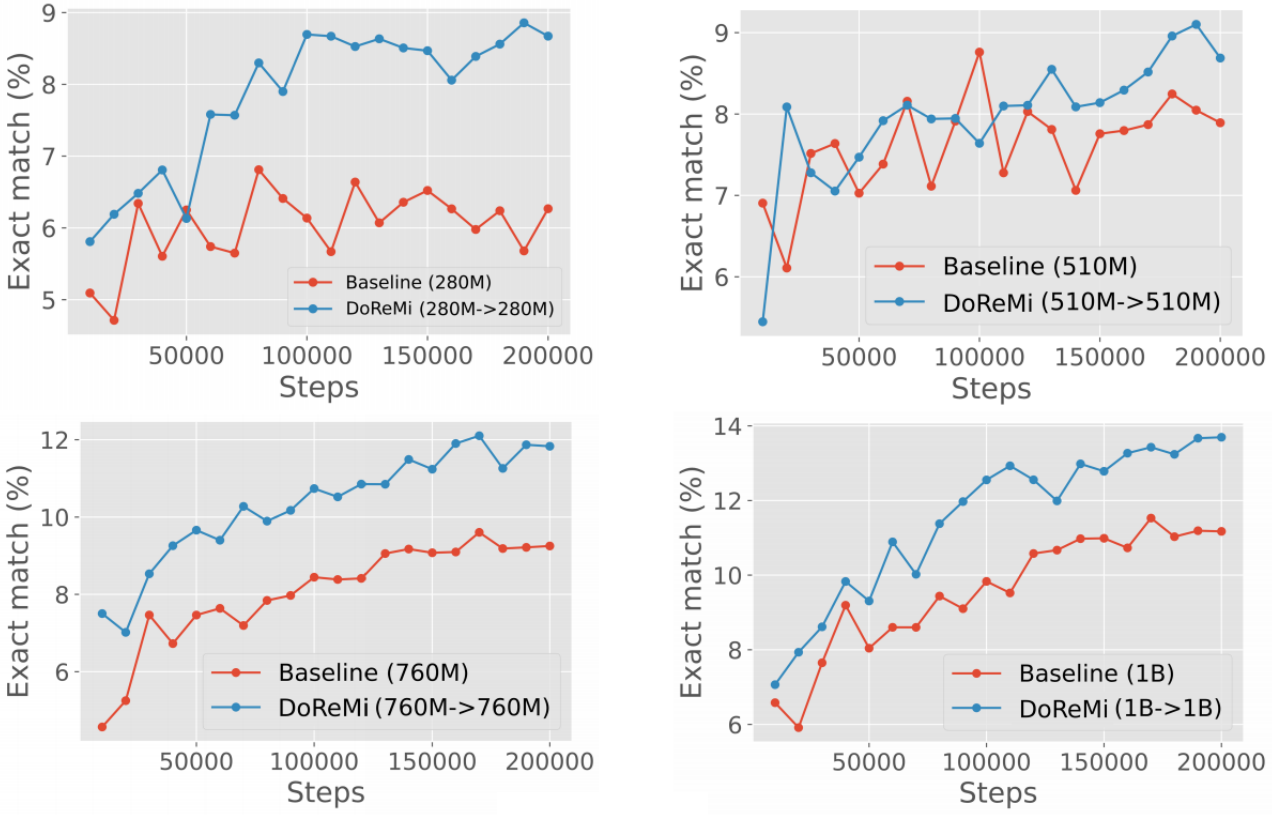


使用 Group DRO 训练“代理模型”，侧重学习与“参考模型” Loss差异最大的域

DoReMi: Optimizing Data Mixtures Speeds Up Language Model Pretraining

数据混合 (Data Mixing)

- DoReMi：相比原始权重，预训练效果提升，且参数规模增长后实验结论一致

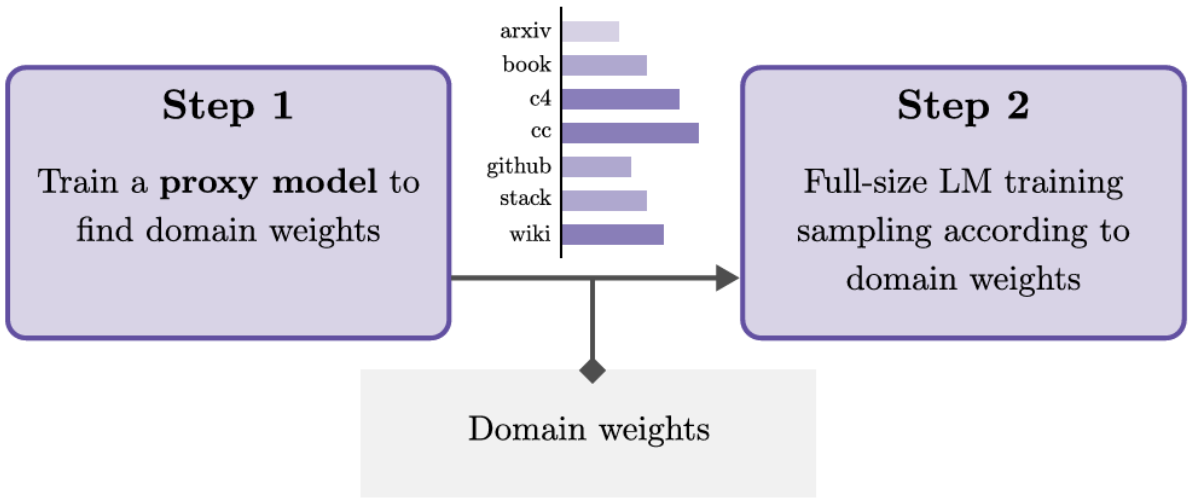


(a) The Pile

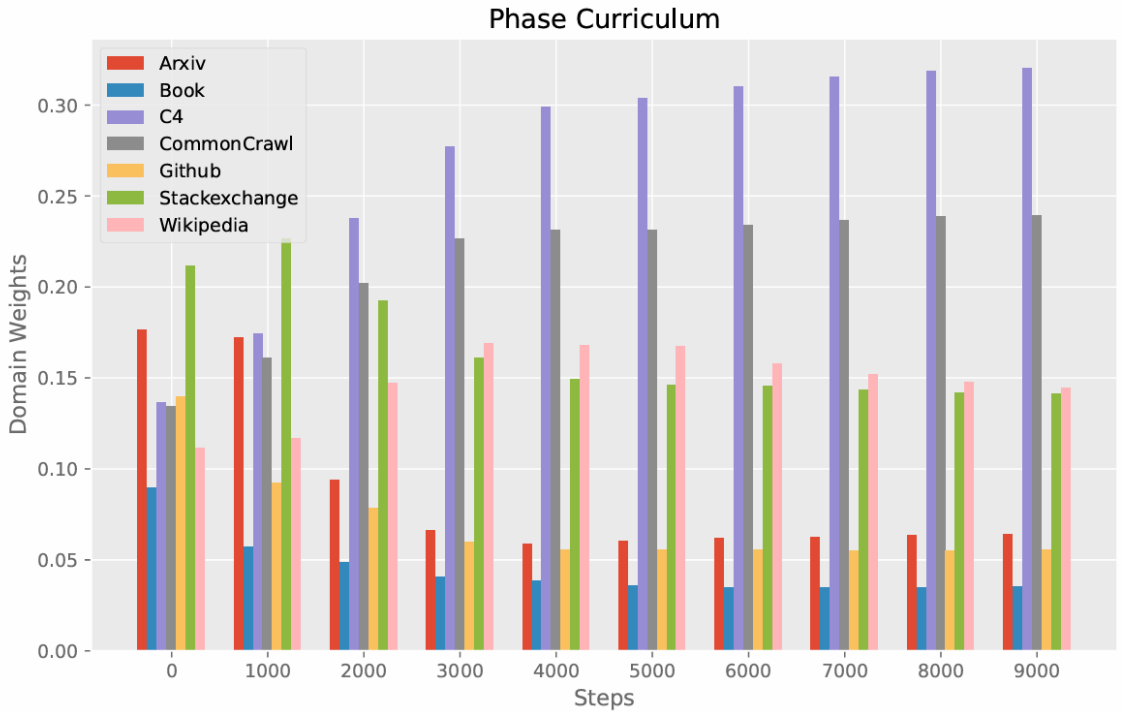
DoReMi: Optimizing Data Mixtures Speeds Up Language Model Pretraining

数据混合 (Data Mixing)

- **DoGE：借助双层优化算法 (BLO)，直接训练代理模型并调整域权重**



无需训练参考模型

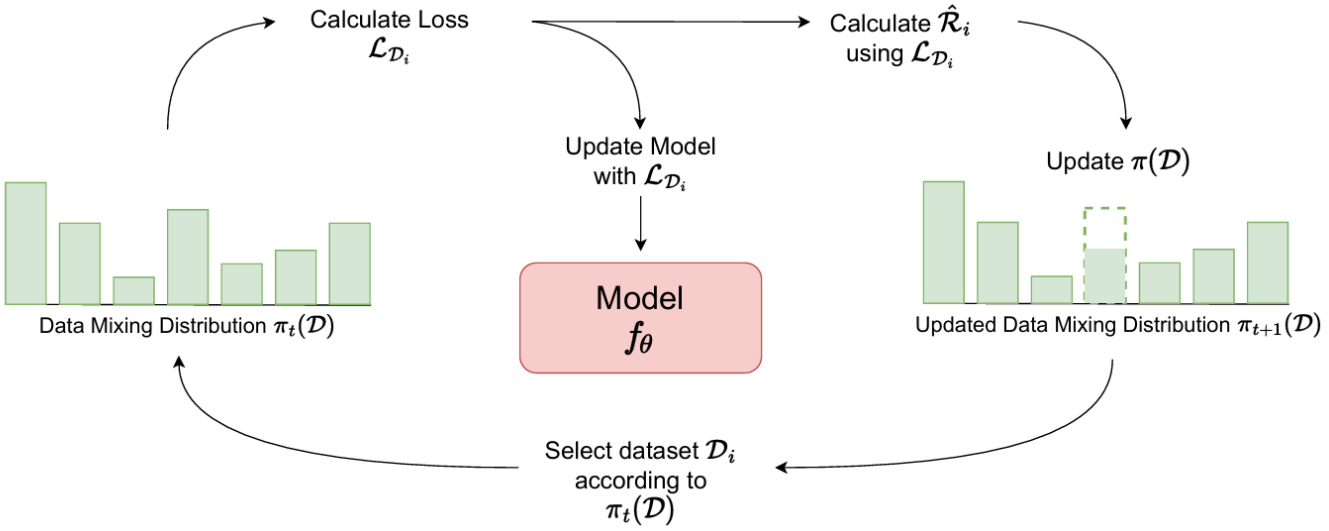


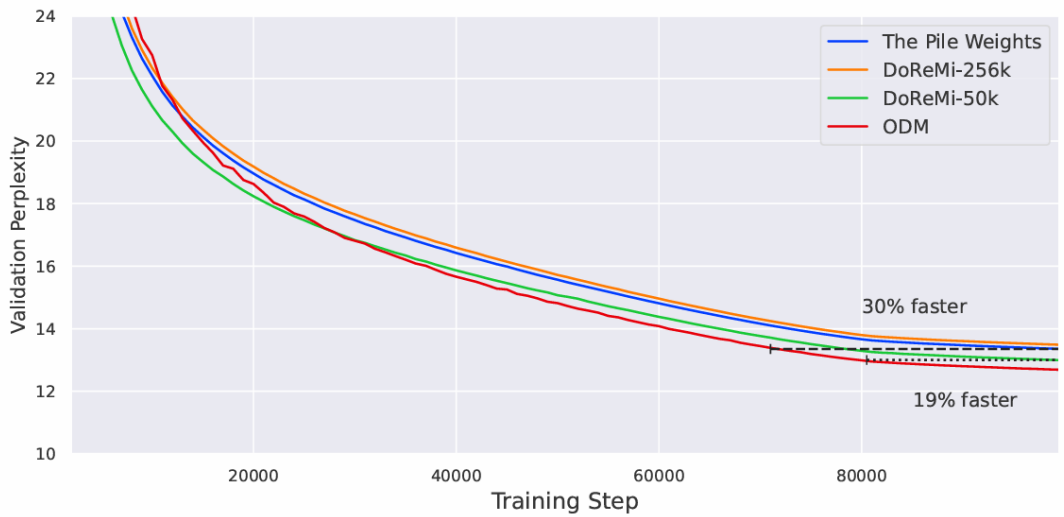
侧重学习对整体梯度更新贡献最大的域

Doge: Domain reweighting with generalization estimation

数据混合 (Data Mixing)

- **ODM：借用多臂老虎机 (MAB) 框架，在训练过程中不断调整领域采样权重**

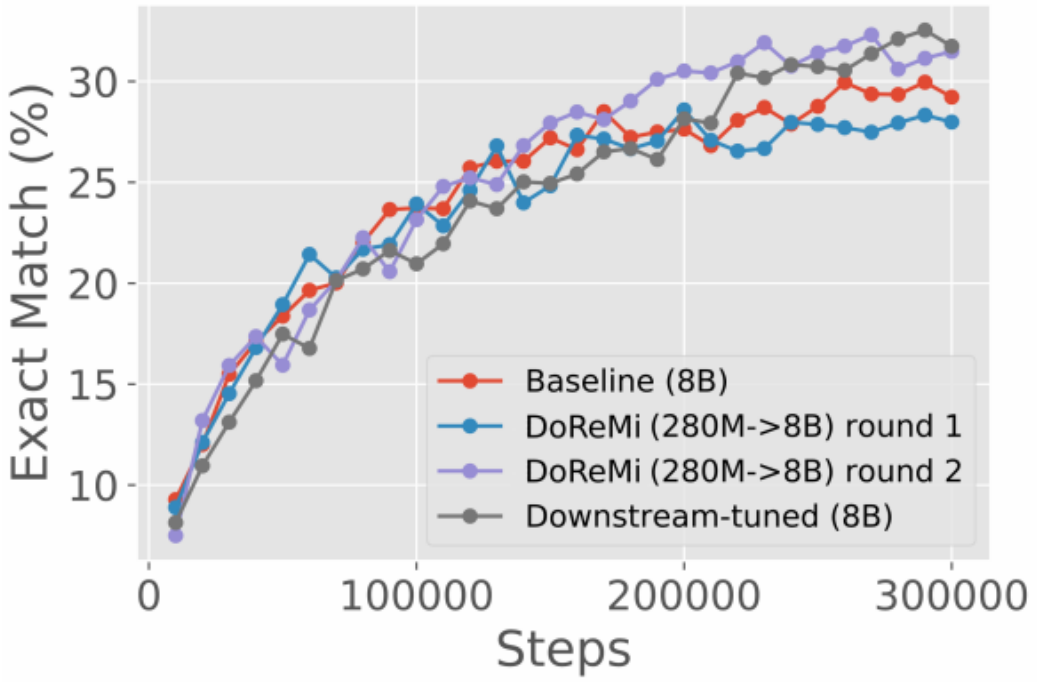




无需训练代理模型

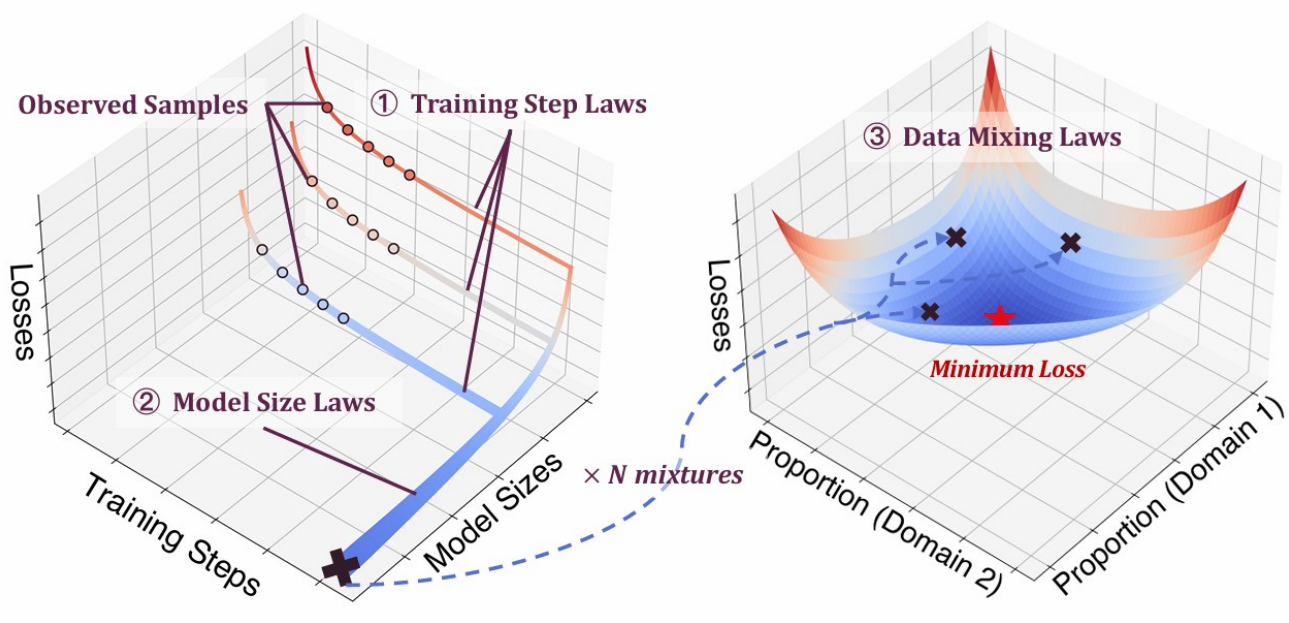
现有方法观察

更换数据集和领域会影响效果



(b) GLaM dataset

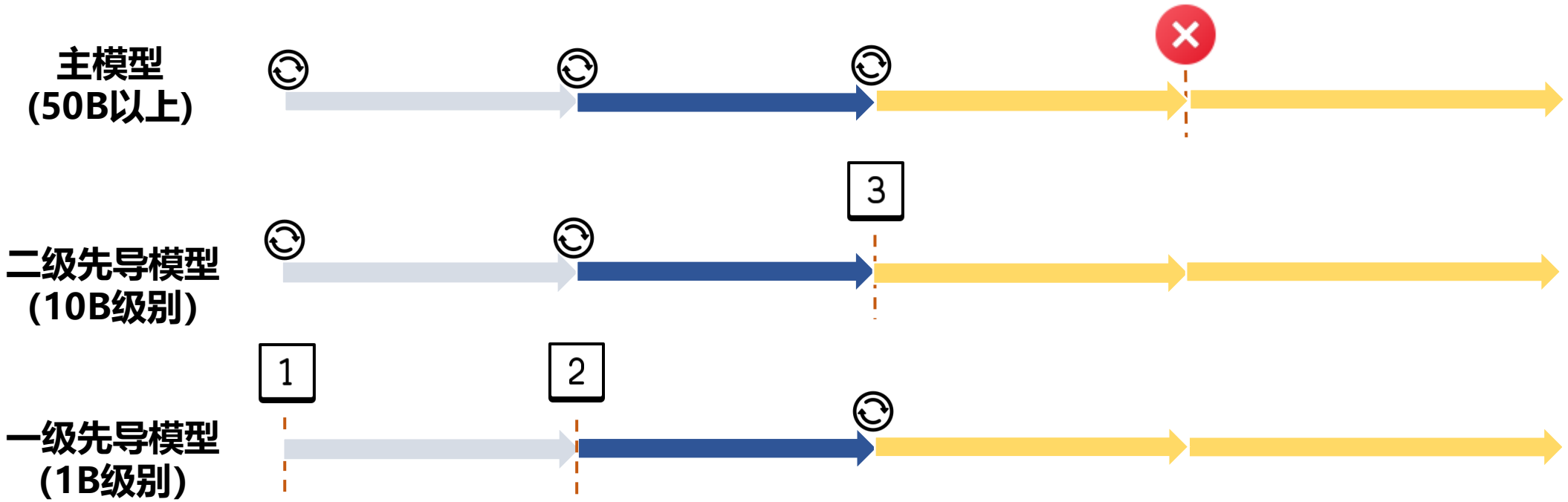
权重变更过程与下游评测无关



DoReMi: Optimizing Data Mixtures Speeds Up Language Model Pretraining
Optimizing Data Mixtures by Predicting Language Modeling Performance

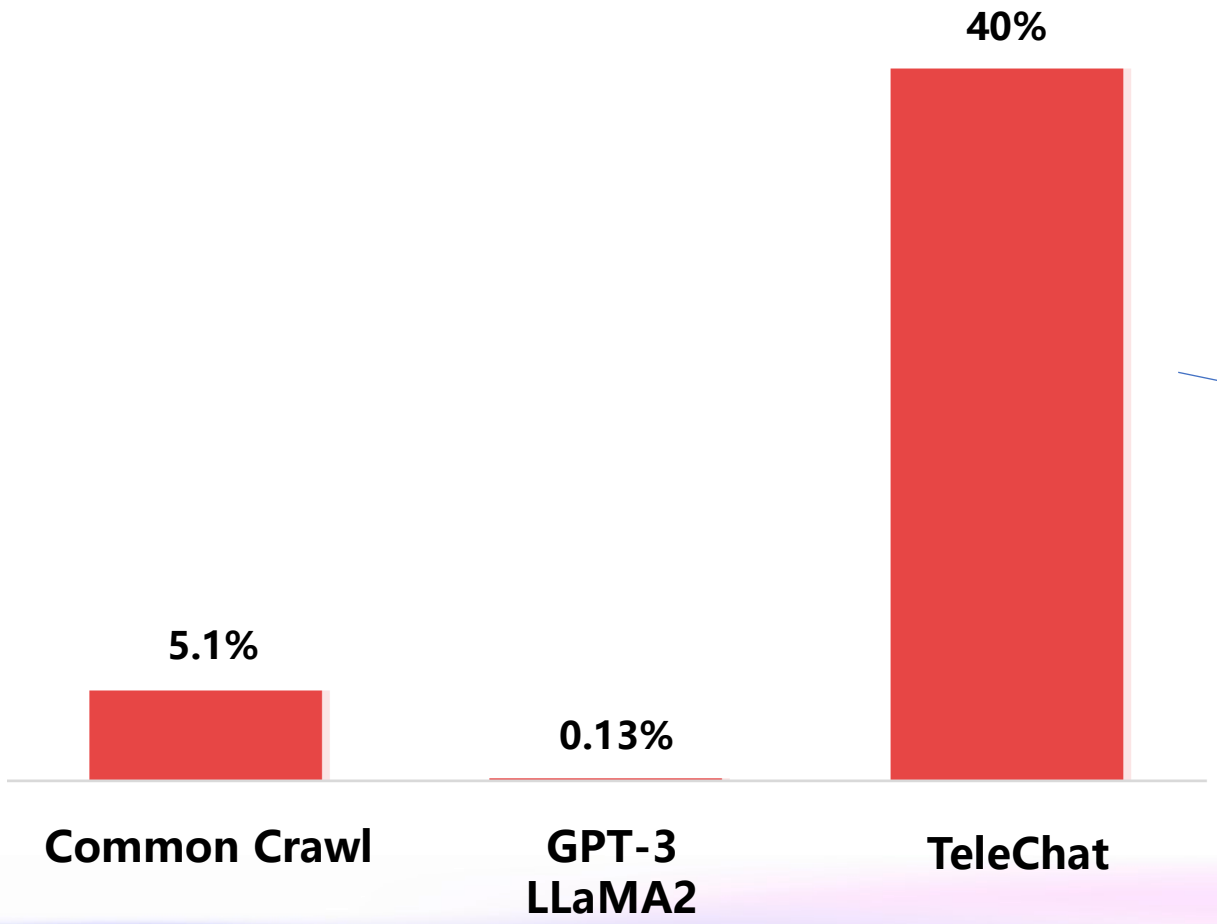
工程化实践

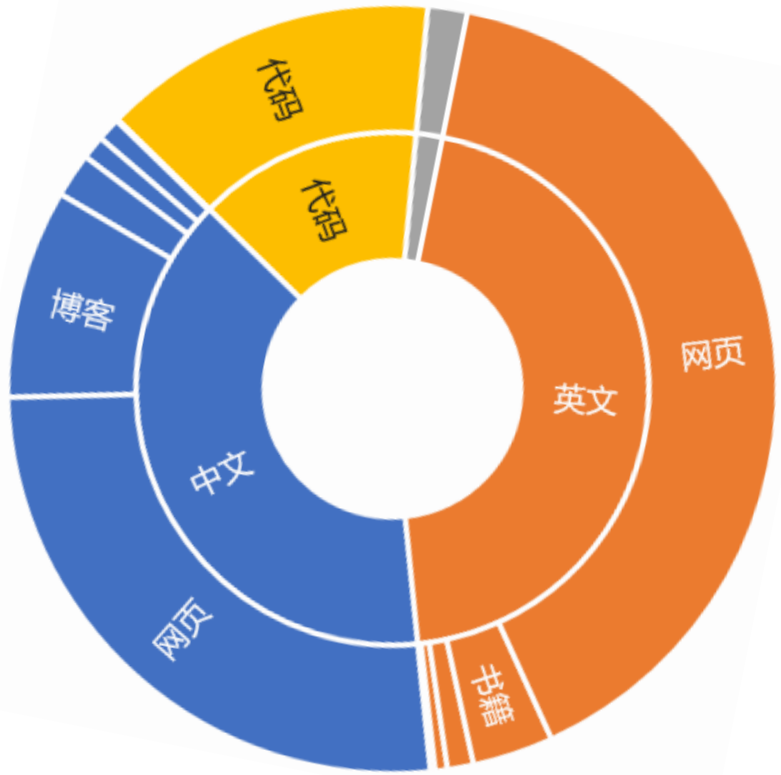
- 每一级先导模型，根据评测指标下降趋势，主动尝试触发配比调整



工程化实践

- 增大中文比例，文本理解、推理、考试能力提升明显，数学能力提升偏小





工程化实践

- 增大数学和题库比例，考试和代码评测指标的“提升速率”加快 (6%~10%)

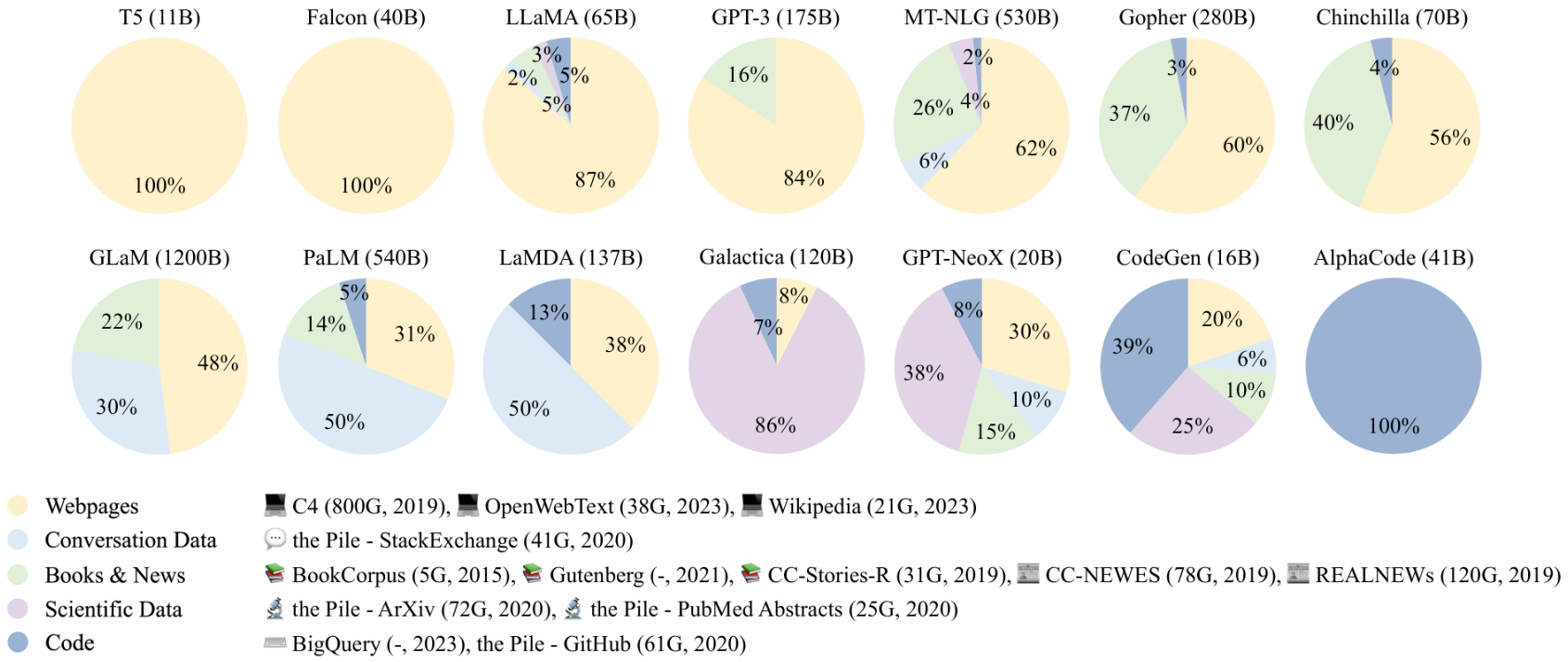
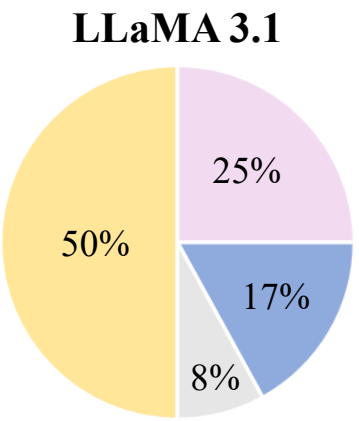
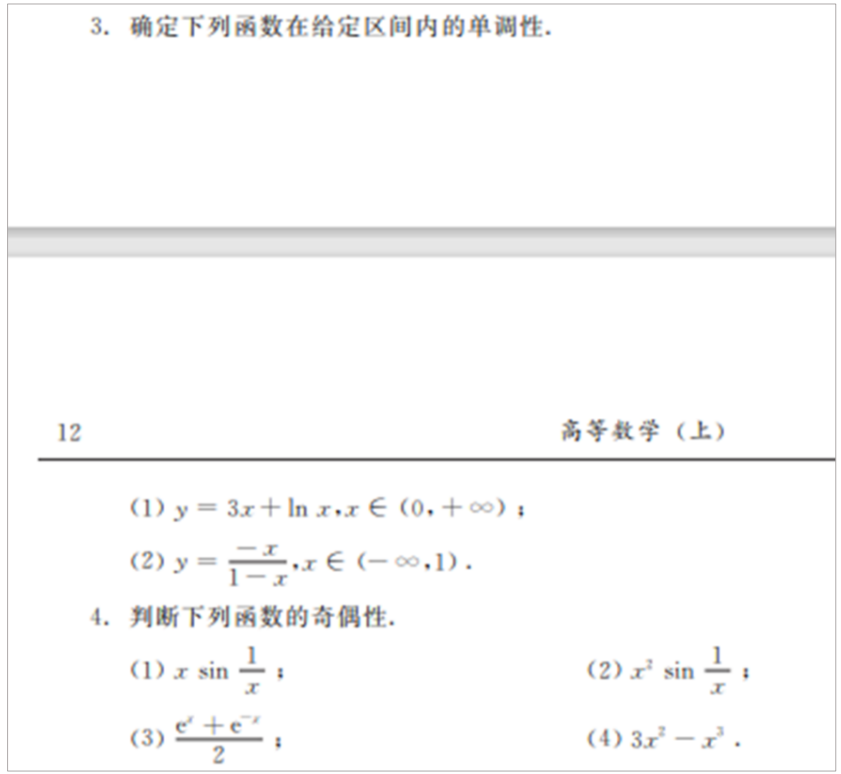


Fig. 6: Ratios of various data sources in the pre-training data for existing LLMs.



工程化实践

- 具备复杂文档解析能力，有助于获取高质量知识类数据



3. 确定下列函数在给定区间内的单调性.

12 高等数学（上）

(1) $y=3x+\ln x, x\in(0,+\infty)$；

(2) $y=\frac{-x}{1-x}, x\in(-\infty,1)$.

4. 判断下列函数的奇偶性.

(1) $x\sin\frac{1}{x}$；(2) $x^2\sin\frac{1}{x}$；

(3) $\frac{e^x+e^{-x}}{2}$；(4) $3x^2-x^3$.

原始PDF文件

```
3. 确定下列函数在给定区间内的单调性.(1) $y=3 x+\ln x, x \in(0,+\infty)$;
(2) $y=\frac{-x}{1-x}, x \in(-\infty, 1)$.
4. 判断下列函数的奇偶性.
(1) $x \sin \frac{1}{x}$;
(2) $x^{2} \sin \frac{1}{x}$;
(3) $\frac{\mathrm{e}^{x}+\mathrm{e}^{-x}}{2}$;
(4) $ 3 x^{2}-x^{3}$.
```

解析结果

(要点：页面排版、公式、表格、阅读顺序、无关内容去除)

3. 确定下列函数在给定区间内的单调性.(1) $y=3x+\ln x, x\in(0,+\infty)$; (2) $y=\frac{-x}{1-x}, x\in(-\infty,1)$.
4. 判断下列函数的奇偶性. (1) $x\sin\frac{1}{x}$; (2) $x^2\sin\frac{1}{x}$; (3) $\frac{e^x+e^{-x}}{2}$; (4) $3x^2-x^3$.

复原效果

工程化实践

- 具备复杂文档解析能力，有助于获取高质量知识类数据

（1）金属氢化物储氢原理

H_2 分子与金属间化合物的反应可以分成 4 个步骤：a） H_2 分子接近金属表面；b） H_2 分子通过范德华力吸附在金属表面（物理吸附状态； c） H_2 分子解离为 H 原子，并在金属表面化学吸附（公式 1-1； d ）占据金属次表面位点并扩散到基体内。吸氢是一个放热的过程，释放的热量称为氢化物的形成焓 ΔH 。

$$H_2 \leftrightarrow 2H$$
$$M + xH \leftrightarrow MH_x(\alpha) \pm \Delta H$$
$$MH_x(\alpha_{\max}) + yH \leftrightarrow MH_{x+y}(\beta) \pm \Delta H$$

原始Word文件

```
\section*{（1）\begin{CJK}{UTF8}{mj}金属氢化物储氢原理\end{CJK}}
\(\mathrm{H}_{2}\) \begin{CJK}{UTF8}{mj}
分子与金属间化合物的反应可以分成\end{CJK} 4 \begin{CJK}{UTF8}{mj}
个步骤\end{CJK}：a） \(\mathrm{H}_{2}\) \begin{CJK}{UTF8}{mj}
分子接近金属表面\end{CJK}；b） \(\mathrm{H}_{2}\) \begin{CJK}{UTF8}{
mj}分子通过范德华力吸附在金属表面\end{CJK}（\begin{CJK}{UTF8}{mj}
物理吸附状态\end{CJK}； c） \(\mathrm{H}_{2}\) \begin{CJK}{UTF8}{mj}
分子解离为\end{CJK} \(\mathrm{H}\) \begin{CJK}{UTF8}{mj}原子\end{CJK}
，\begin{CJK}{UTF8}{mj}并在金属表面化学吸附\end{CJK}（\begin{CJK}{
UTF8}{mj}公式\end{CJK} 1-1； \(\mathrm{d}\) ）\begin{CJK}{UTF8}{mj}
占据金属次表面位点并扩散到基体内\end{CJK}。\begin{CJK}{UTF8}{mj}
吸氢是一个放热的过程\end{CJK}，\begin{CJK}{UTF8}{mj}
释放的热量称为氢化物的形成焓\end{CJK} \(\Delta \mathrm{H}\) 。

\[
\begin{gathered}
H_{2} \leftrightarrow 2 H \\
M+x H \leftrightarrow M H_{x}(\alpha) \pm \Delta H \\
M H_{x}\left(\alpha_{\max }\right)+y H \leftrightarrow M H_{x+y}(
\beta) \pm \Delta H
\end{gathered}
\]
```

解析结果

03

后训练数据筛选

后训练发展趋势

预训练缺失，后训练弥补

多而无序，来源多样

答案文风迥异

数量

几百万条以上

基模与微调维持高度相关

质量

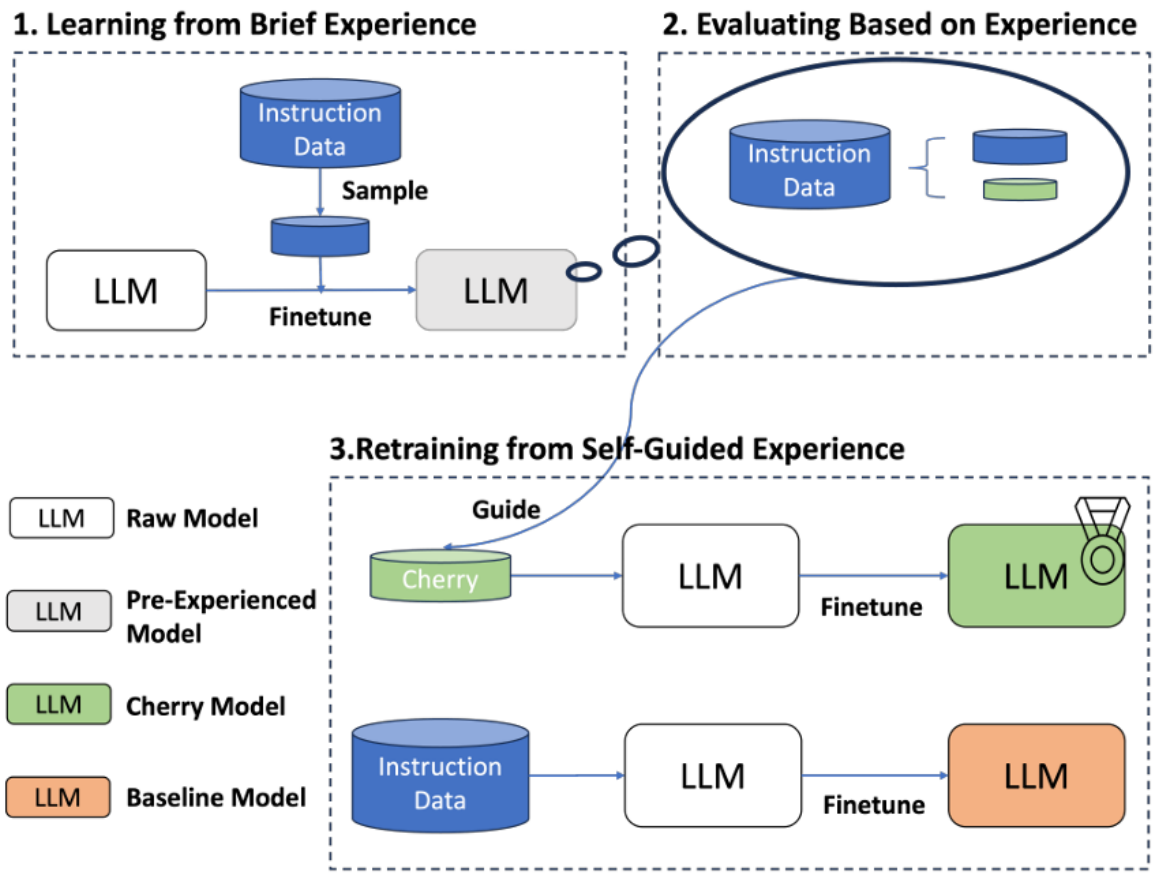
十万条以下

少而全面，指令多样

答案格式规范

数据筛选（Data Filtering）

- CherryLLM：少量数据训练经验模型，根据指令追随难度（IFD）筛选数据

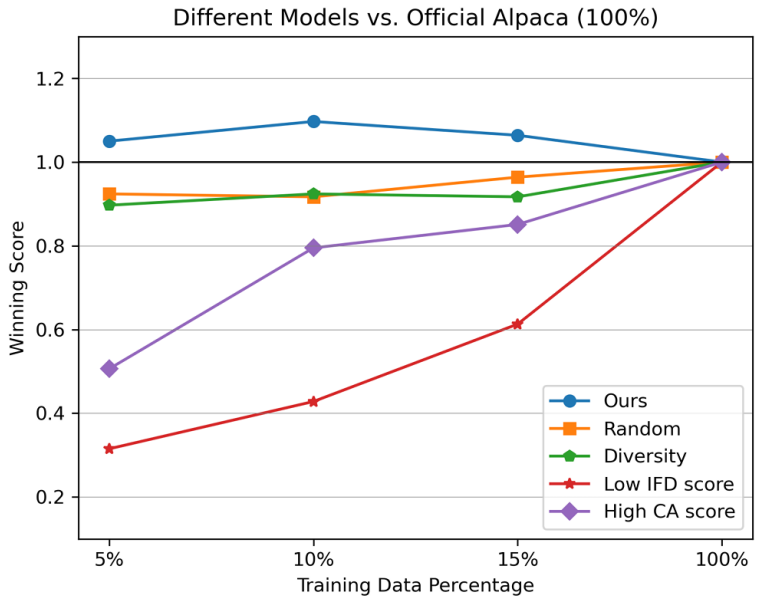


$$\mathrm{IFD}(Q,A) = \frac{\mathrm{CAS}(Q,A)}{\mathrm{DAS}(A)}$$

条件答案得分越大

直接答案得分越小

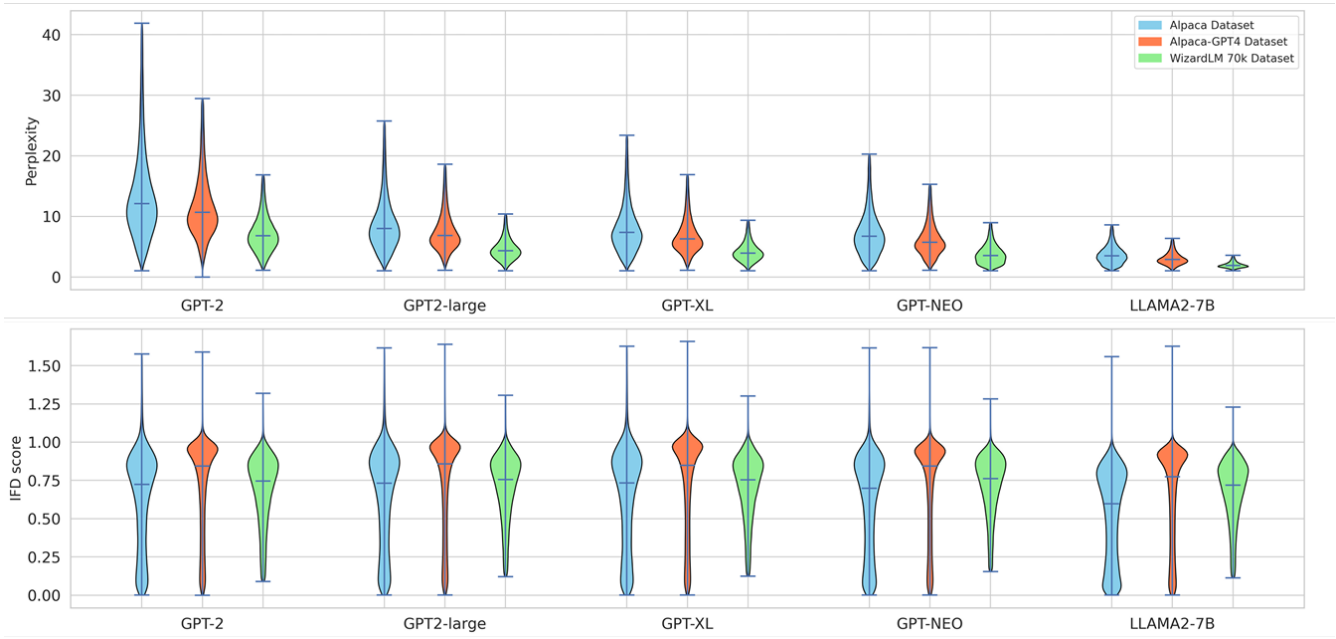
$$= \frac{\log P(a_i|Q,a_1,a_2,\ldots,a_{i-1})}{\log P(a_i|a_1,a_2,\ldots,a_{i-1})}$$



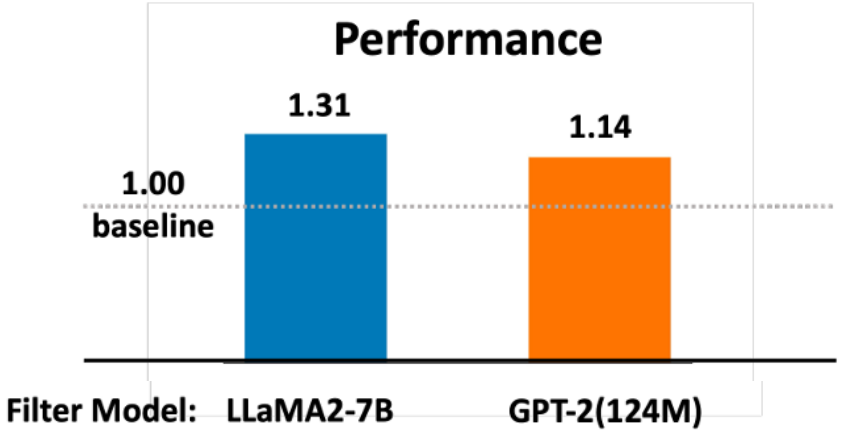
From Quantity to Quality: Boosting LLM Performance with Self-Guided Data Selection for Instruction Tuning

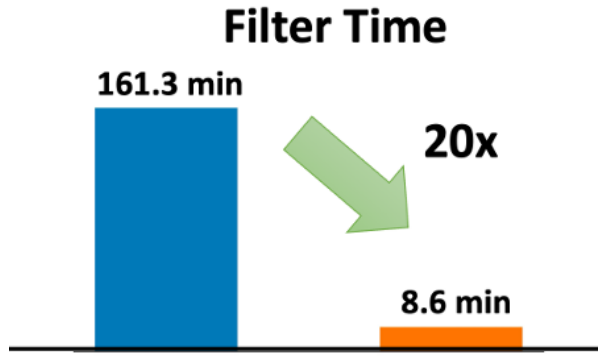
数据筛选（Data Filtering）

- **Superfiltering：以IFD为基础，利用小参数模型执行筛选，减少时间开销**



模型越小，困惑度越大，但IFD分布一致

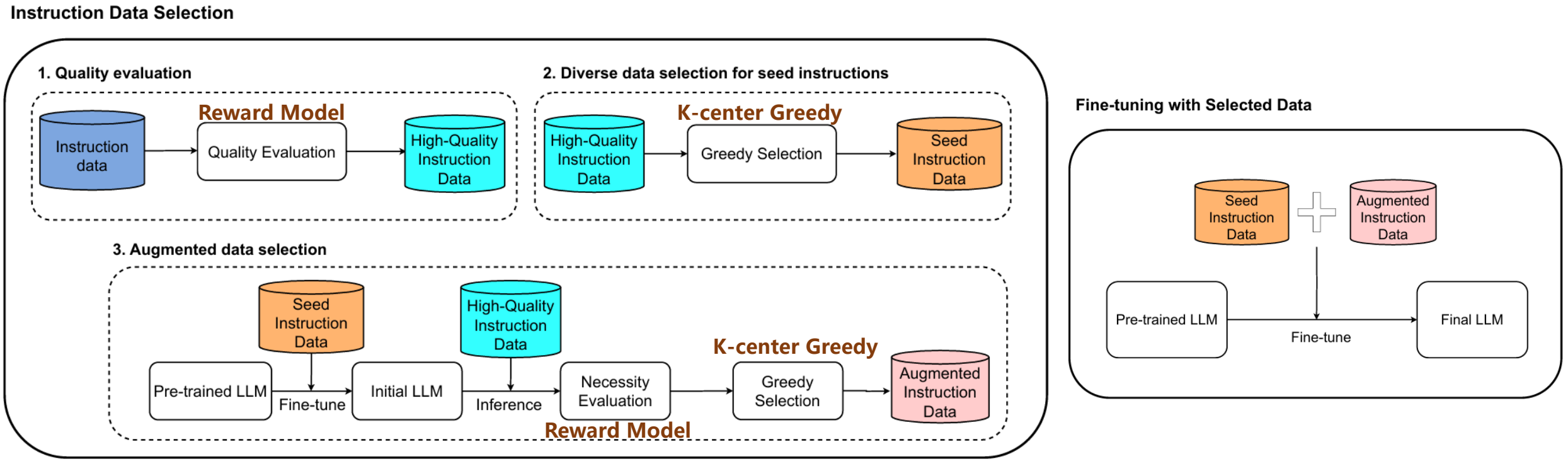




Superfiltering: Weak-to-Strong Data Filtering for Fast Instruction-Tuning

数据筛选（Data Filtering）

- **MoDS：借助奖励模型对数据进行评价和筛选**

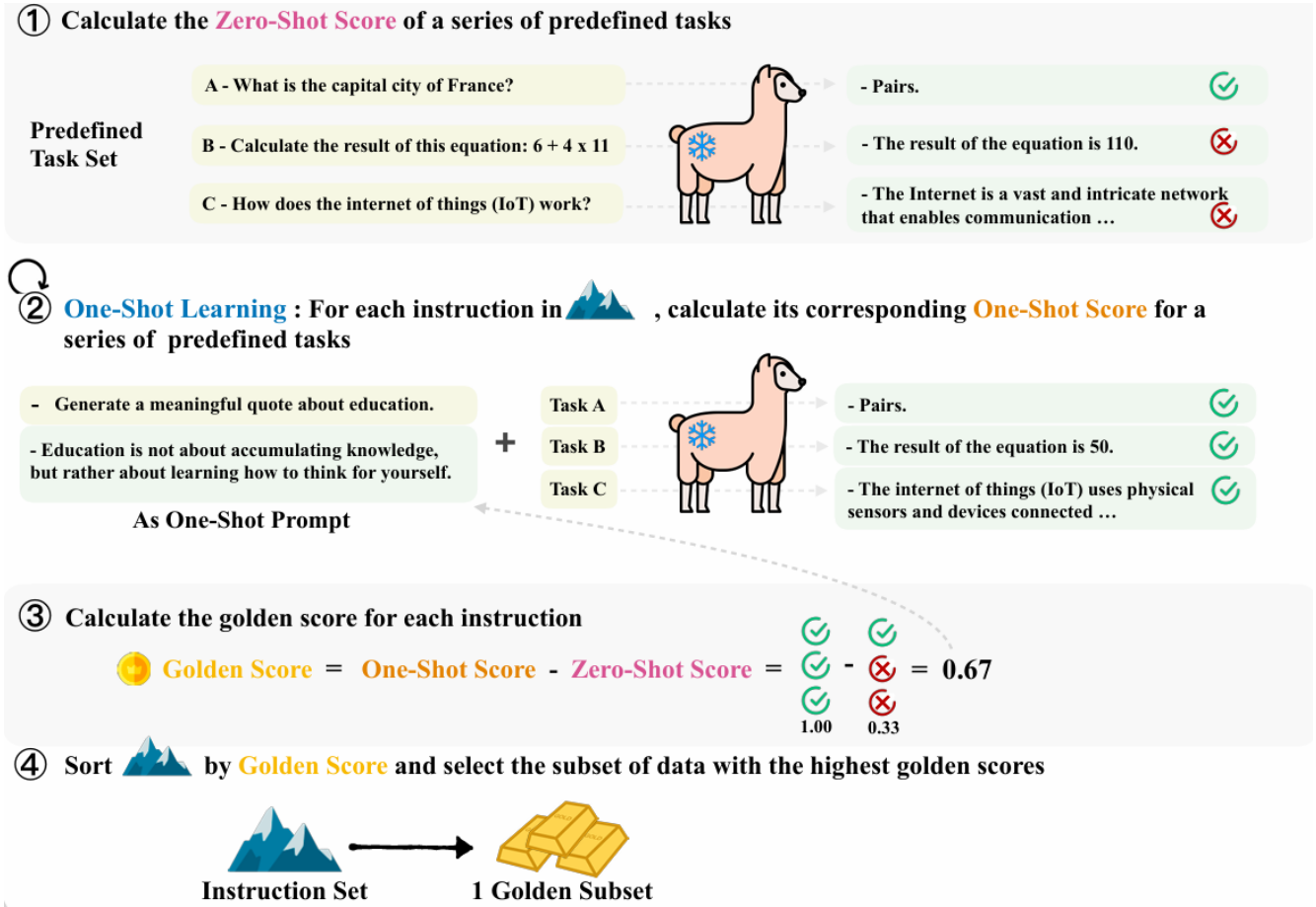


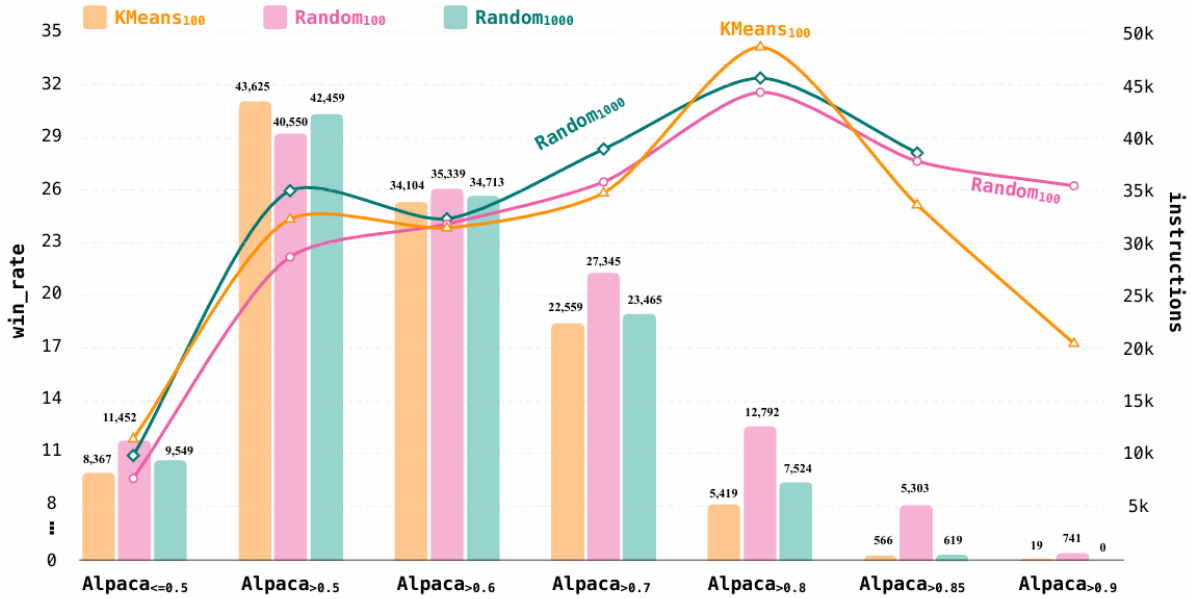
根据预训练模型的偏好，从HQID中再筛选出一些增强数据作为SID的补充

MoDS: Model-oriented Data Selection for Instruction Tuning

数据筛选（Data Filtering）

- **NUGGETS：评估每条样本作为One-Shot样例的价值增益**

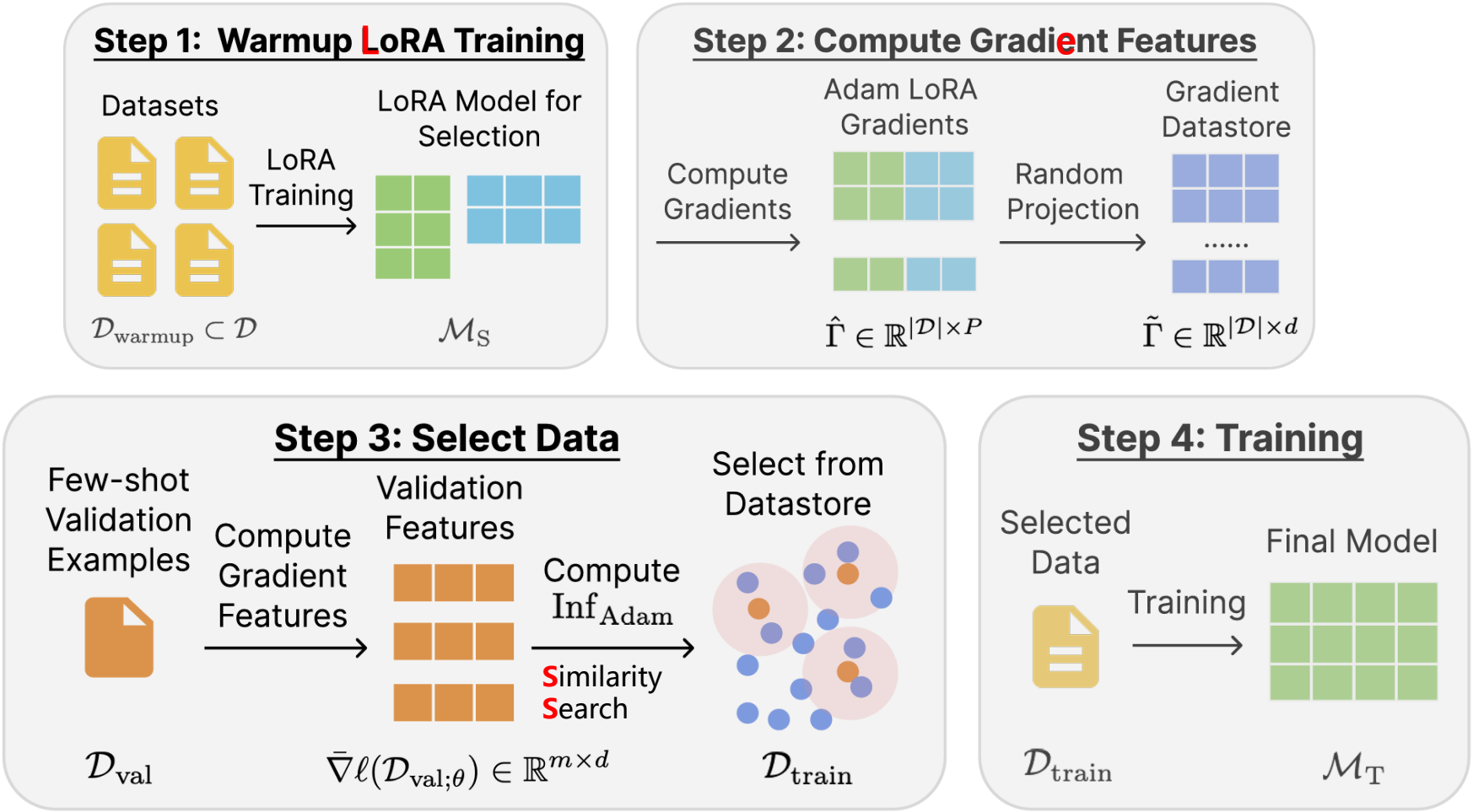




One-Shot Learning as Instruction Data Prospector for Large Language Models

数据筛选（Data Filtering）

- **LESS：根据模型梯度优化方向评估样本对测试集损失的降低程度**



LESS: Selecting Influential Data for Targeted Instruction Tuning

现有方法观察

- “模型类方法”依赖外部模型能力，计算开销大
- “指标类方法”计算效率高，潜在误差大

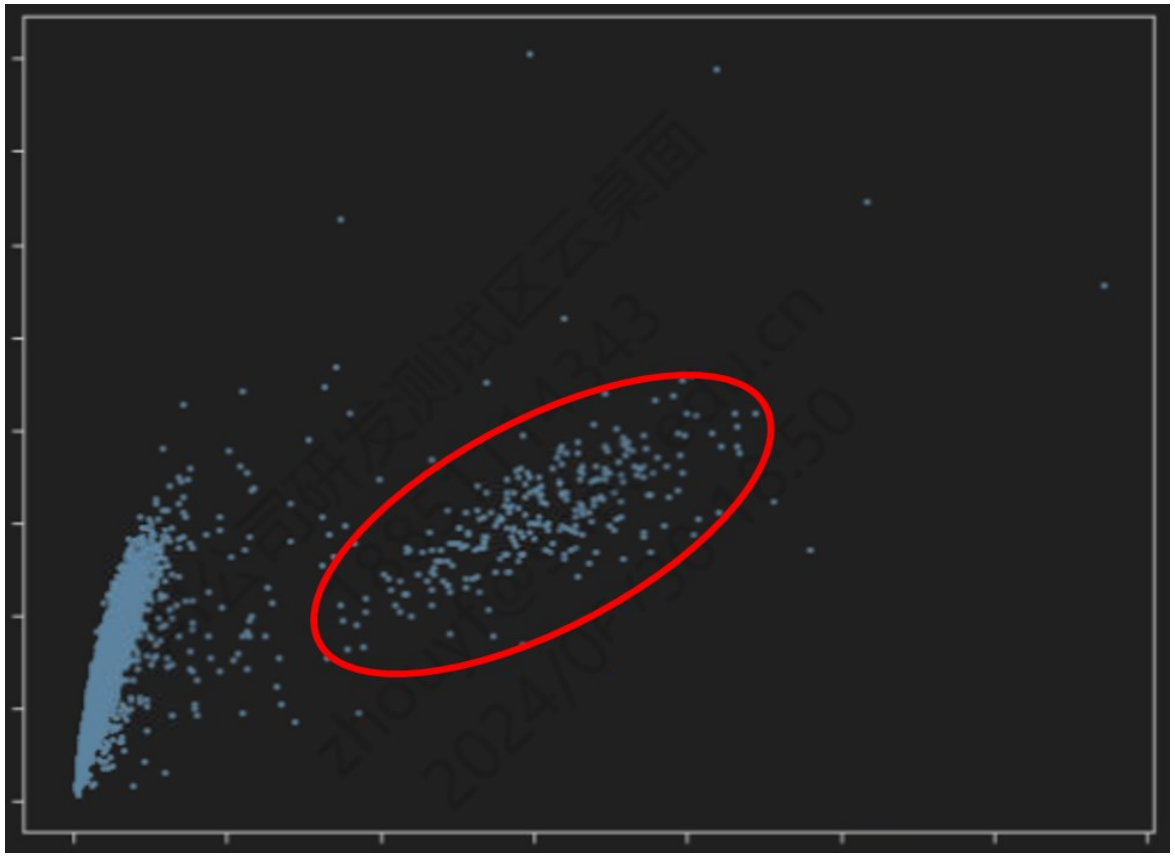
是否存在“相对稳定”的指标

精确性：描述模型输出与标准答案的匹配程度

确定性：描述模型对生成内容的确定程度

理想指标

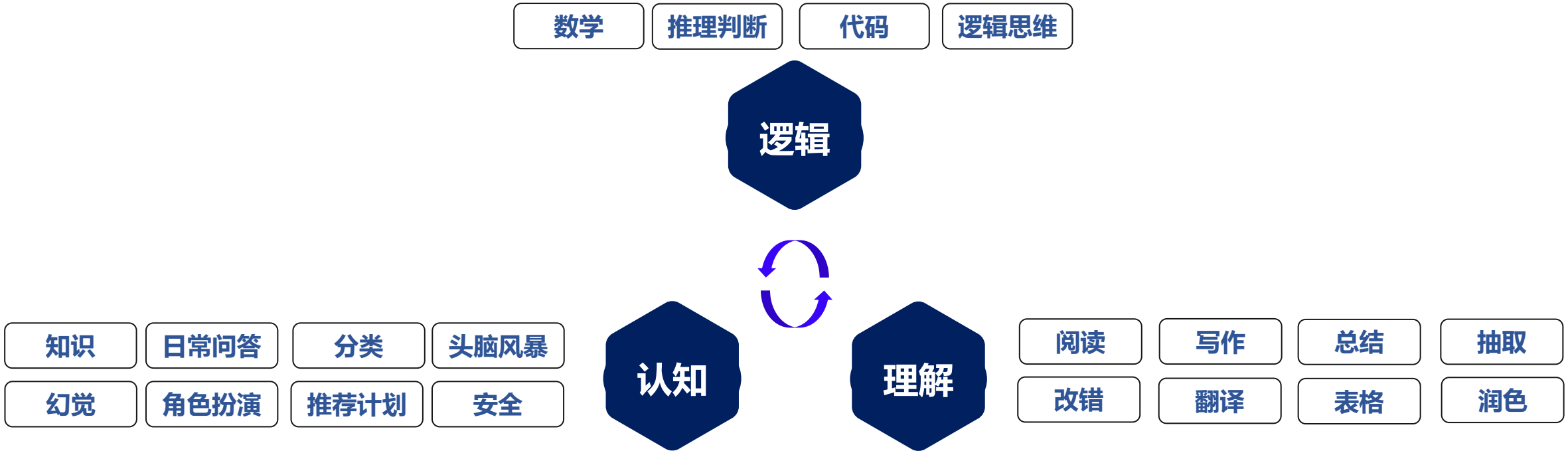
确定性低



模型训练越充分，
符合条件的结果越少

精确性高

工程化实践



重点能力提升

- ★ 结合 PPL 与 RFT，筛选任务更优数据
- ★ 面向标准评测集，定向提升单项重点能力

通用能力融合

- ★ 按认知、逻辑、理解划分能力维度，关联任务互促强化
- ★ 迭代优化数据配比，提升通用能力整体效果

优化效果

通用问答效果 8%

类型	输入问题	优化前	优化后
逻辑推理	烧一根不均匀的绳子，从头烧到尾需要1小时，现有若干条这种绳子，如何记时45分钟	这个脑筋急转弯的答案是"从头烧到尾需要1小时，现在有若干条这种绳子，可以同时烧"。 也就是说，你只需要把绳子首尾相连，然后从一端开始烧，烧到另一端需要45分钟。因为绳子是均匀烧的，所以不管你烧多长的绳子，所需要的时间都是相同的。	要利用不均匀的绳子来准确计时45分钟，我们需要运用一些逻辑分析和策略。首先，我们知道一根绳子从头烧到尾需要1小时。这意味着如果我们同时点燃绳子的两端，它会在半小时内完全烧完，因为两端同时燃烧，速度加倍。 接下来，我们可以按照以下步骤来计时45分钟： 准备两根绳子：我们取两根绳子，分别标记为绳子A和绳子B。 点燃绳子A的两端：由于绳子A的两端同时被点燃，它将在半小时内完全烧完。 点燃绳子B的一端：在点燃绳子A的同时，我们也点燃绳子B的一端。这样，绳子B将开始从一端燃烧。 等待绳子A烧完：当绳子A完全烧完时，已经过去了半小时。此时，绳子B已经燃烧了半小时，但只从一端燃烧，所以它还有半小时才会烧完。 绳子B继续燃烧：点燃B的另一头，继续让绳子B燃烧。由于绳子B还需要半小时才能完全烧完，因此当绳子B完全烧完时，总时间就是1小时的一半再加上半小时，即45分钟。 综上所述，通过同时点燃一根绳子的两端和另一根绳子的一端，并在第一根绳子烧完时，点燃B的另一头继续等待第二根绳子烧完，我们可以准确地计时45分钟。这个策略的关键在于利用绳子两端同时燃烧速度加倍的特性，以及控制不同绳子燃烧的开始和结束时间。

逻辑推理 7%　**幻觉问题** 18%　**数学计算** 40%　**翻译能力** 11%

数学是自然科学的基础

- 仅用3万条数据微调52B模型，效果达到 GPT-4 的91%，GPT-4-1106 的82%
- 其中2.5万条数据与数学相关，其余数据为代码和多轮对话

Table 1: Performance of Tele-FLM-Chat and baselines on Alignbench, rated by CritiqueLLM.

Model	Overall	Reasoning			Language						
		Avg.	Math.	Logi.	Avg.	Fund.	Chi.	Open.	Writ.	Role.	Pro.
gpt-4-1106-preview	7.58	7.11	7.39	6.83	8.05	7.69	7.07	8.66	8.23	8.08	8.55
gpt-4-0613	6.83	6.41	6.49	6.33	7.26	7.16	6.76	7.26	7.31	7.48	7.56
chatglm-turbo	6.36	4.99	4.88	5.09	7.73	7.50	7.03	8.45	8.05	7.67	7.70
Tele-FLM-Chat	6.20	4.61	4.21	5.00	7.79	7.22	7.64	8.53	8.08	7.72	7.59
vs. gpt-4-1106 (%)	82	65	57	73	**97**	94	108	98	98	95	89
vs. gpt-4-0613 (%)	**91**	72	65	79	**107**	101	113	117	111	103	100

52B to 1T: Lessons Learned via Tele-FLM Series

THANKS