AiDD AI+ 软件研发数字峰会
AI+ software Development Digital summit

AI驱动软件研发
全面进入数字化时代

中国·深圳 11.24-25

AI+
software
Development
Digital
summit

大语言模型下的数据及知识管理

彭力 小米AI实验室

科技生态圈峰会 + 深度研习

——1000 + 技术团队的共同选择

KEYLINK ing

K+峰会

K+峰会 深圳站

K+ 全球软件研发行业创新峰会

会议时间：2024.05.24-25

K+峰会 上海站

K+ 全球软件研发行业创新峰会

会议时间：2024.09.20-21

AiDD 峰会

AiDD峰会 深圳站

AI+ 软件研发数字峰会

会议时间：2023.11.24-25

AiDD峰会 北京站

AI+ 软件研发数字峰会

会议时间：2024.07.19-20

AiDD峰会 深圳站

AI+ 软件研发数字峰会

会议时间：2024.11.15-16

演讲嘉宾

彭力

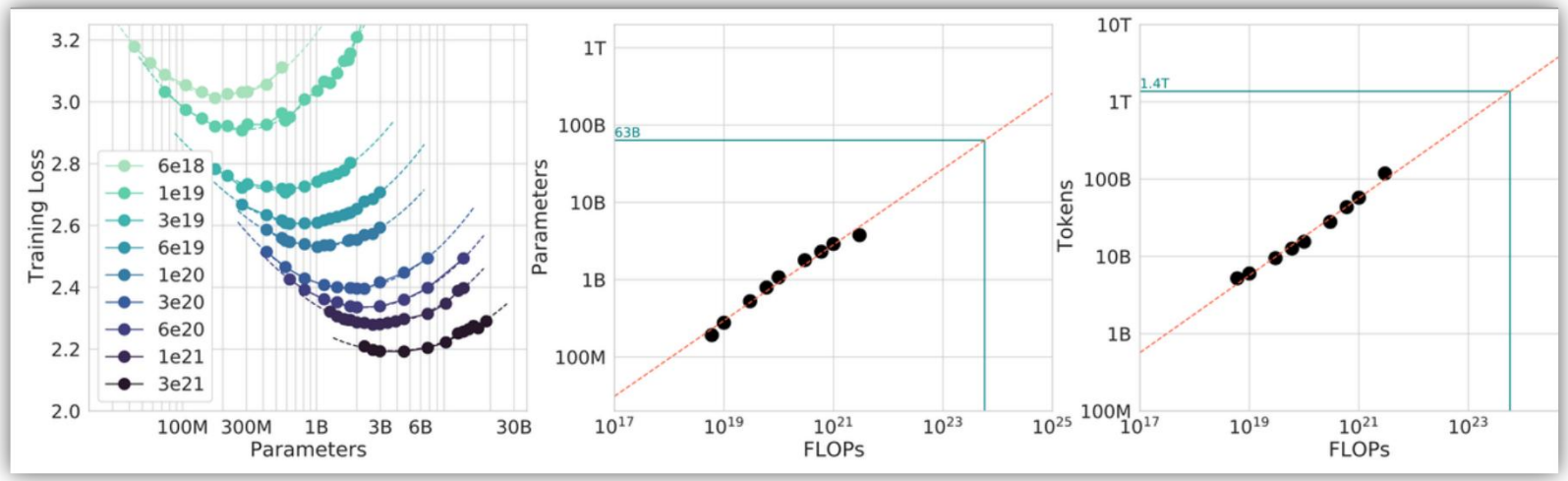
小米集团-AI实验室-大模型数据团队负责人

2012年至2018年曾就职于百度，于2018年5月加入小米。先后负责知识图谱平台及大模型数据团队的技术体系的构建。目前主要负责小米自研大模型的数据及知识构建及自研模型的落地，并在此期间参与知识图谱国家标准的制定。曾在全球人工智能大会、 Data fun talk等多个会议和论坛参加知识图谱相关主题的技术分享

目录
CONTENTS

PART 01
知识和数据管理的重要性

海量数据及知识对大语言模型的影响



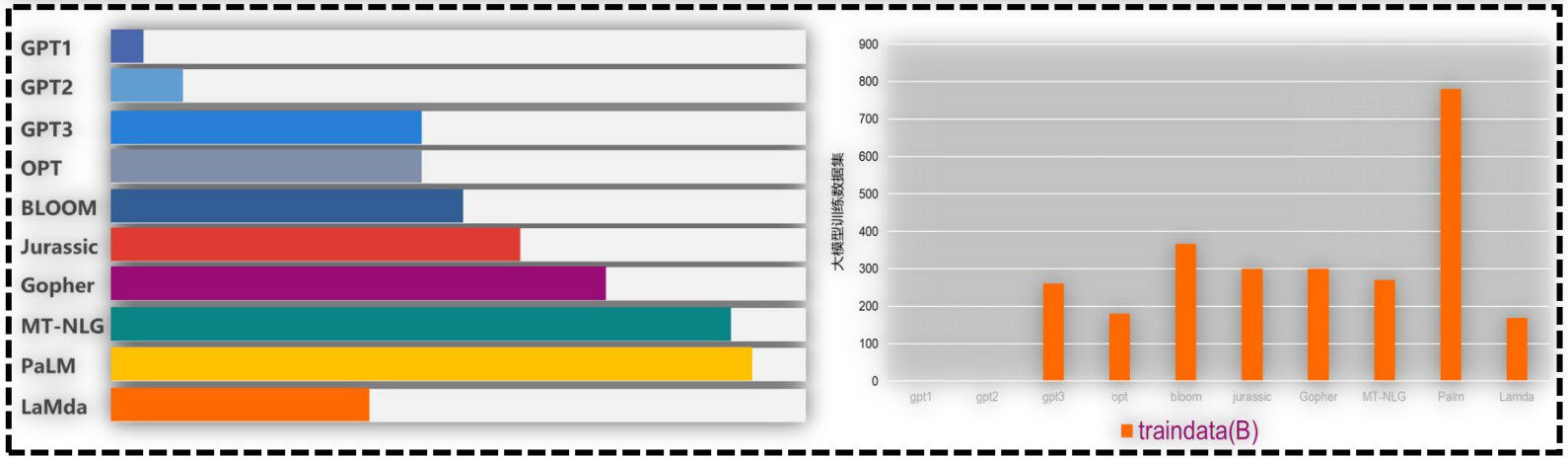


Table 15: Results for several models trained on 1×10^{22} FLOPs. This is with an architecture and data mixture that is different from PaLM 2. When validation loss is roughly equal, downstream performance is approximately equal as well. All evaluations are done in a 1-shot setting. Unless mentioned otherwise, accuracy is used as the evaluation metric. The number of parameters indicates non-embedding parameters.

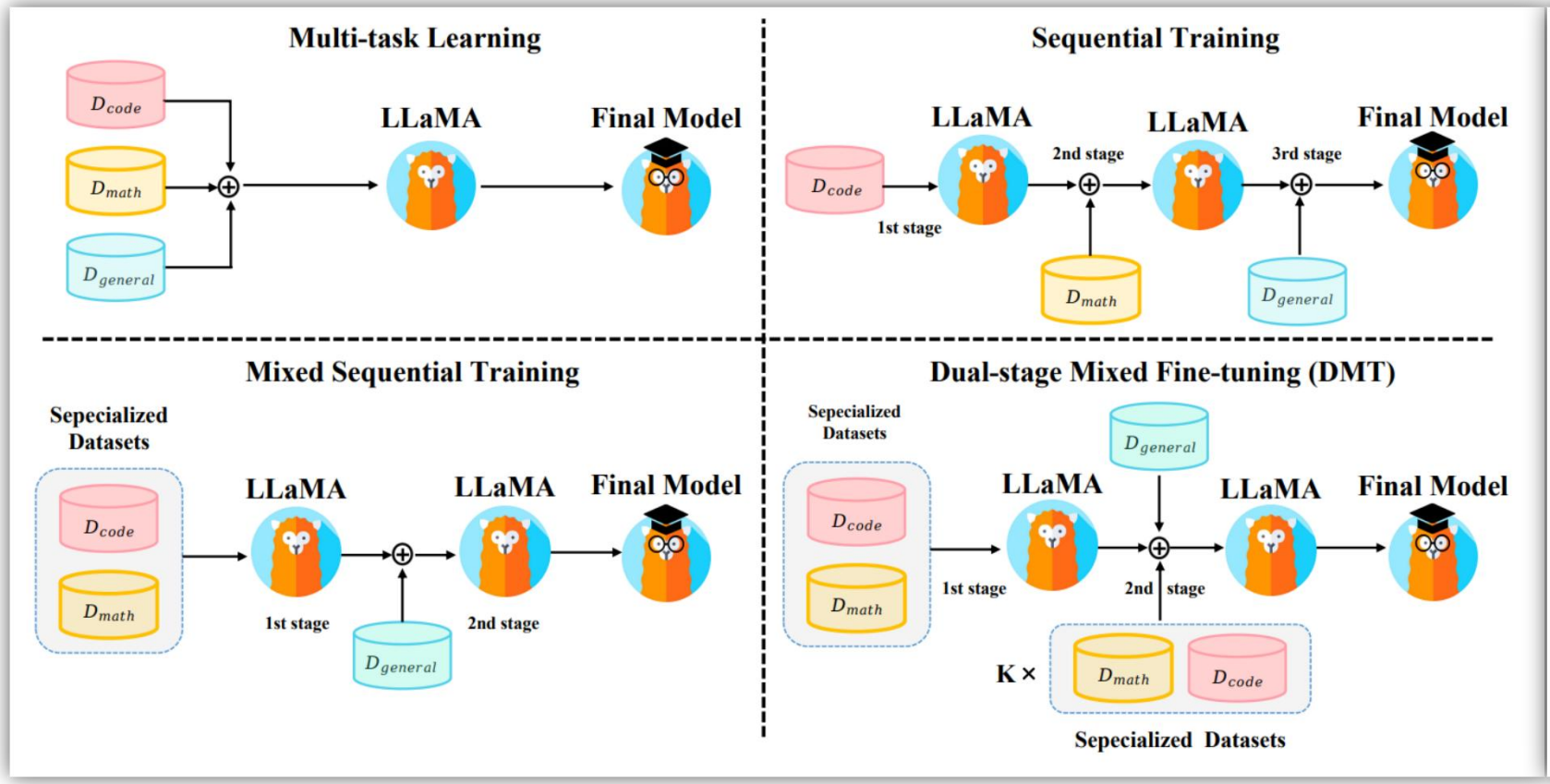
	3.31B	6.08B	8.95B	14.7B
Triviaqa-Wiki (EM)	43.75	**50.57**	49.22	48.74
NaturalQuestions (EM)	10.11	10.97	**12.58**	11.50
WebQuestions (EM)	10.29	11.96	**12.70**	**12.45**
LAMBADA	55.46	59.27	60.97	**63.05**
HellaSwag	69.82	71.84	**72.34**	**72.27**
StoryCloze	80.49	80.97	**81.88**	**81.56**
Winograd	83.15	**85.71**	83.52	84.62
Winogrande	69.14	**70.56**	69.61	70.56
WSC	83.86	**84.91**	**84.56**	82.81
SQuAD v2 (EM)	56.19	**57.66**	55.93	55.74
RACE-H	40.68	**43.80**	**43.51**	42.65
RACE-M	56.96	**59.68**	58.84	58.84
TyDiQA-En (EM)	**43.64**	41.36	**43.86**	**43.64**
PiQA	77.42	77.86	**78.73**	78.02
ARC-C	40.61	**43.60**	42.66	**43.69**
ARC-E	71.55	73.70	**74.28**	71.00
OpenBookQA	50.60	51.60	50.60	**54.00**
BoolQ	68.59	**71.25**	**71.31**	**71.59**
CB	**64.29**	41.07	42.86	51.79
COPA	85.00	87.00	84.00	**90.00**
MultiRC	**63.37**	60.58	60.68	58.91
ReCoRD	88.49	**89.85**	**89.45**	**89.74**
RTE	**63.54**	55.60	**63.90**	**63.90**
WiC	**48.75**	47.49	**48.28**	47.81
ANLI-R1	33.10	34.70	31.70	**35.10**
ANLI-R2	30.70	**33.30**	32.80	31.50
ANLI-R3	32.09	**35.50**	34.00	34.25
Average	57.30	57.61	57.68	**58.26**

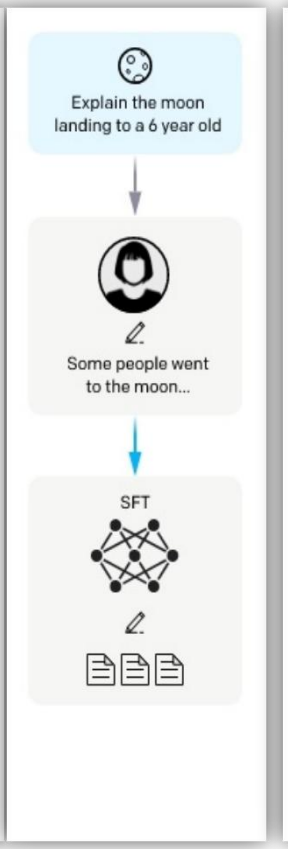
模型参数越大需更越多的知识

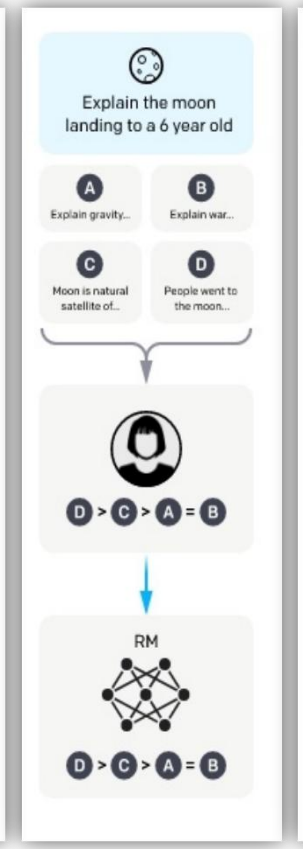
大语言模型基座训练需要更多更丰富的知识

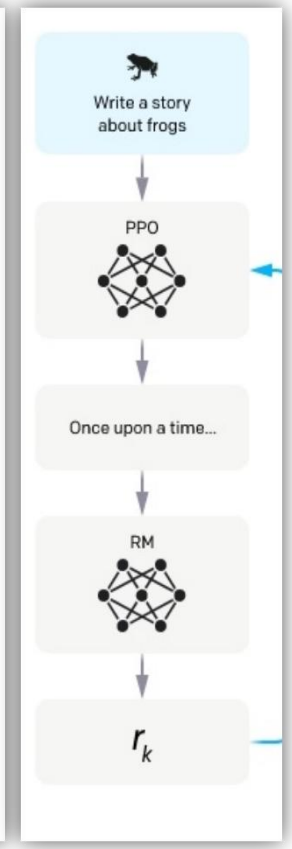
海量数据及知识对大语言模型的影响

除基座训练以外全阶段的性能依赖高质量的知识









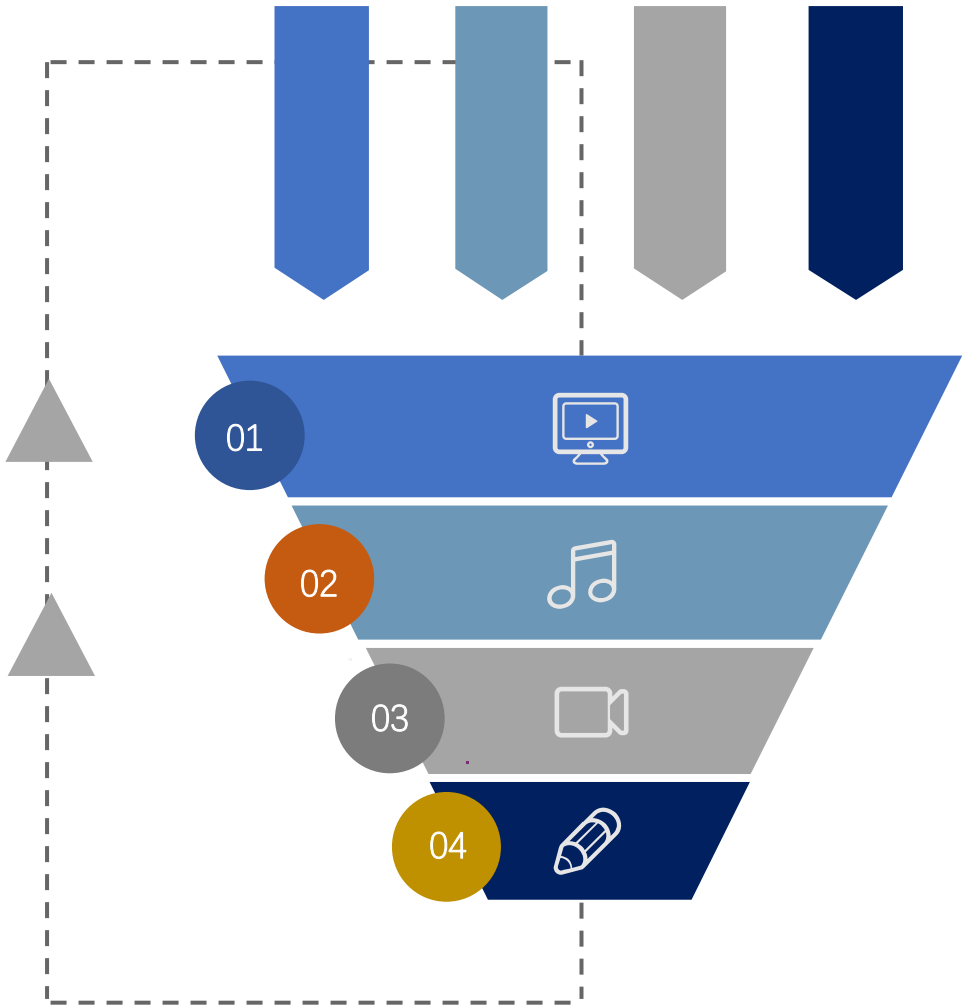
How Abilities in Large Language Models are Affected by Supervised Fine-tuning Data Composition

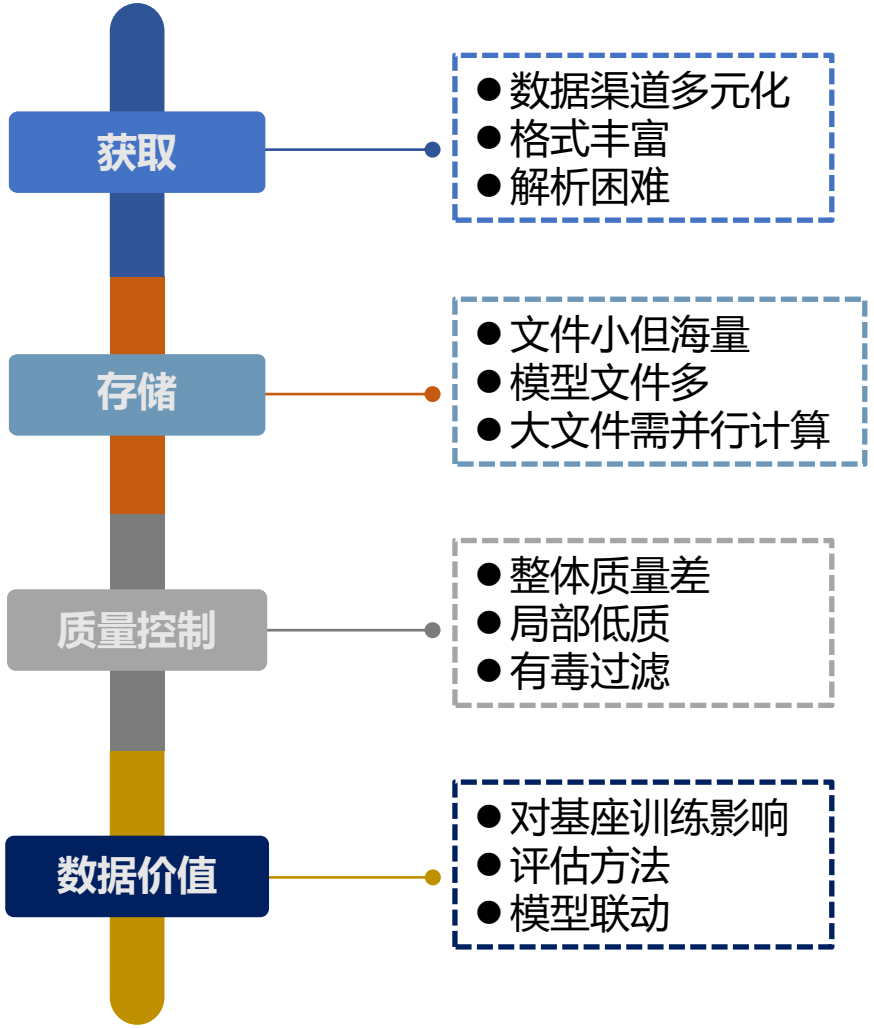
Introducing ChatGPT

PART 02

大模型下数据加工及管理的解决方案

大语言模型中数据及知识管理问题

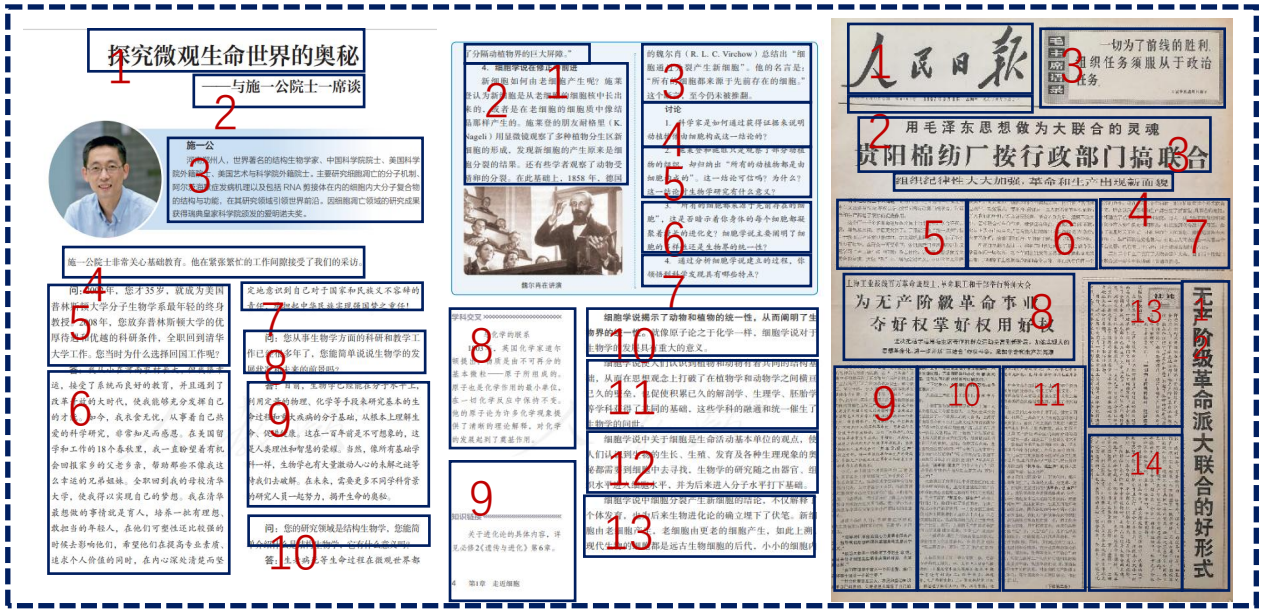




小米大语言模型中数据及知识挖掘

基于文档分析的不同格式的数据和知识提取依然存在挑战

布局识别 → 阅读顺序 → 格式解析
表格/公式/化学式/……



- 布局复杂且包含丰富的元素
- 多栏之间转换影响阅读的顺序
- 多样式的公式及表格等需统一

小米大语言模型中数据及知识挖掘

计算得到主要影响因素，包括驱动能力偏低、活门输出特性分散度大及驱动负载偏大，进一步研究其产生原因，认为是由发动机个体差异或热状态差异导致的，因此在故障现象上表现出一定的不确定性。进一步对上述 3 方面主要因素进行叠加分析，复现了故障现象。

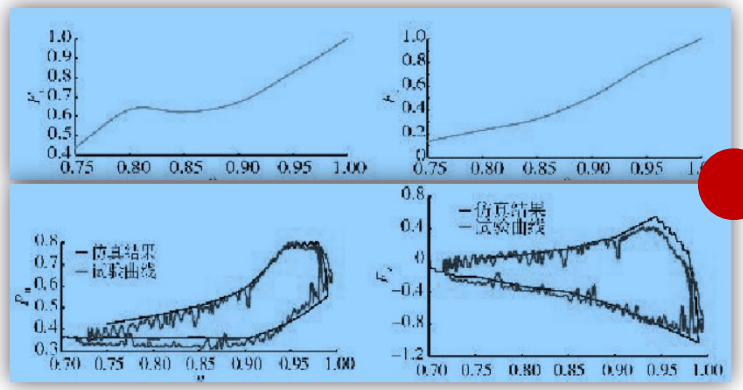


图 8 作动筒 B 腔压力仿真及试验结果（无量纲）　图 9 驱动力仿真及试验结果（无量纲）

3.2 故障模式分析及验证

航空发动机　第45卷

$$F_{dl}=F_n+F_f \quad (1)$$

$$F_{dr}=F_f-F_n \quad (2)$$

$$F_{ql}=P_A\cdot A_A-P_B\cdot(A_B-A_g) \quad (3)$$

$$F_{qr}=P_B\cdot(A_B-A_g)-P_A\cdot A_A \quad (4)$$

表 1 故障因素影响计算

序号	影响因素	VSV 滞后量/(°)
1	泵后压力过低	仿真排除
2	活门泄漏	仿真排除
3	摩擦阻力突变	仿真排除
4	控制器内部机械刚度过低	仿真排除
5	反馈机构位移输入输出偏差	仿真排除
6	作动筒内泄漏	仿真排除
7	驱动能力偏低	0.63
8	活门输出特性分散度大	0.89
9	驱动负载偏大	0.98
10	活门输出特性分散度大＋驱动负载偏大	2.70
11	驱动负载偏大＋驱动力偏低	2.03
12	活门输出特性分散度大＋驱动力偏低	1.79
13	活门输出特性分散度大＋驱动负载偏大＋驱动力偏低	3.35

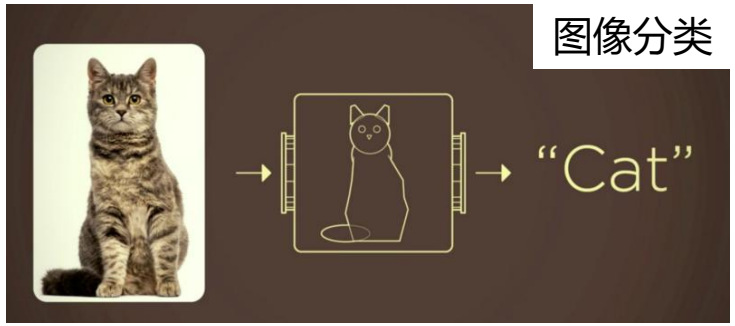
文档元素类别

1 图片　2 图片caption　3 文本段落　4 标题　5 页眉　6 公式　7 表格caption　8 表格

小米大语言模型中数据及知识挖掘

第二阶段:机器学习阶段
2000年

图像分割



图像分类

第三阶段:深度学习阶段
2014年

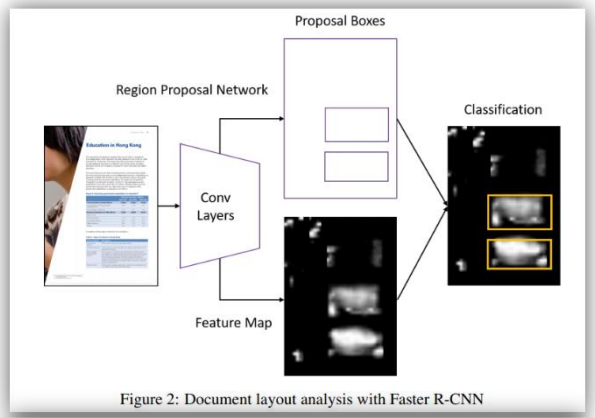


Figure 2: Document layout analysis with Faster R-CNN

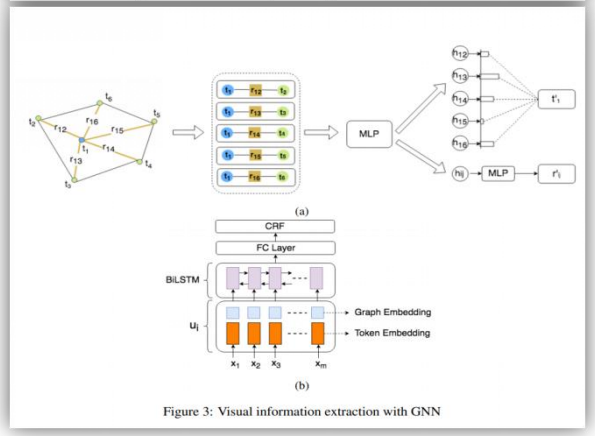


Figure 3: Visual information extraction with GNN

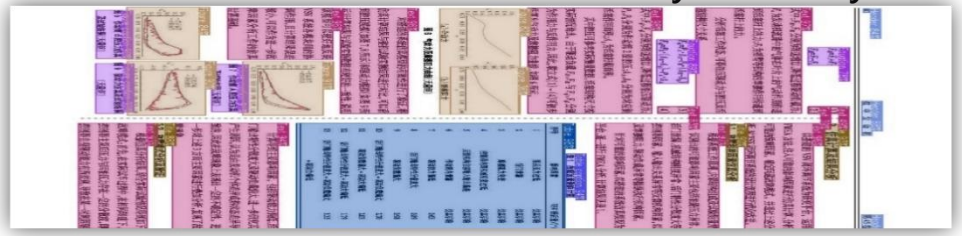
DOCUMENT AI: BENCHMARKS, MODELS AND APPLICATIONS

https://arxiv.org/pdf/2111.08609.pdf

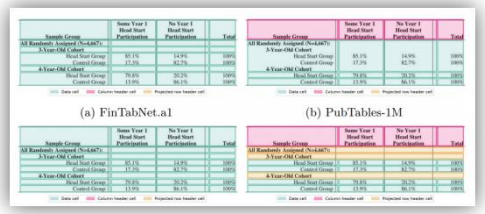
第一阶段: 启发式阶段
1900年(规则)
Docstrum

- 文档图像分类

- 文档布局分析 (Document layout analysis)



- 公式、表格检测结构识别



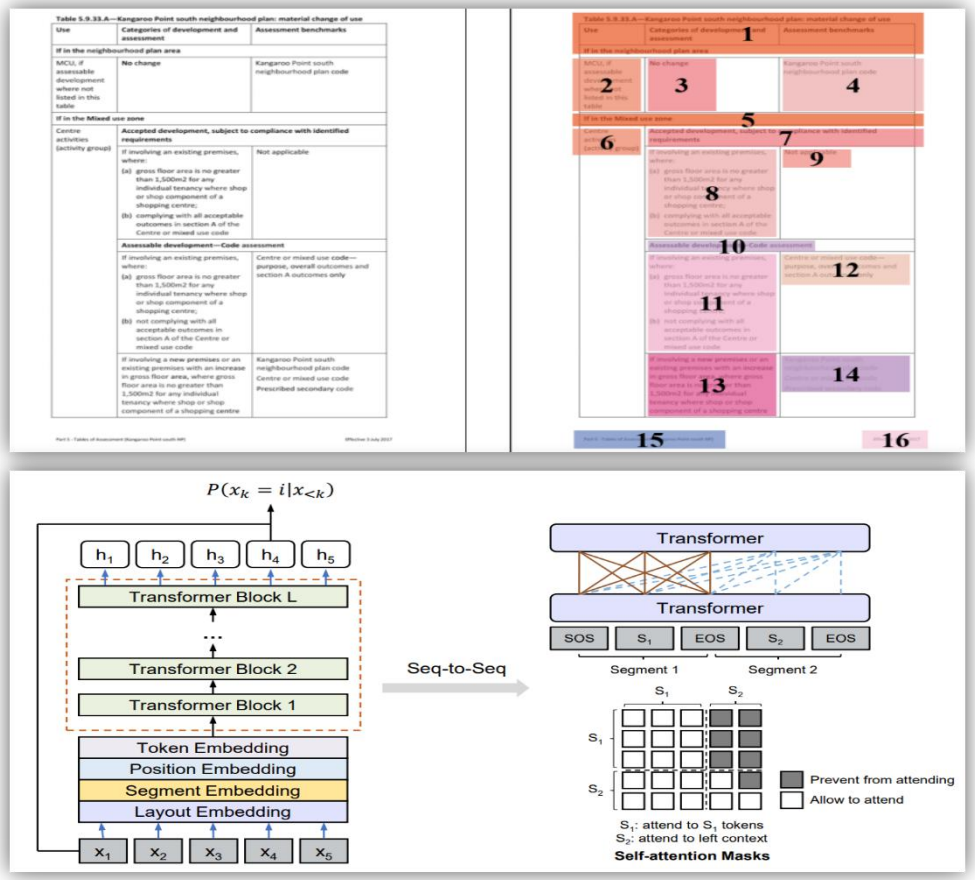
小米大语言模型中数据及知识挖掘

相关布局模型

模型名	模型规模	最佳表现情况	技术方法	发布时间
DiT	-	ON RVL-CDIP 2022 SOTA! Accuracy **92.69** | Param**304M**	-	2022-03
RectiNet-v2	-	ON DocUNet 2021 SOTA! SSIM **0.507434**	Concatenated Skip Connection	2021-02
Faster-RCNN	-	ON Document Layout Recognition Challenge mini-dev 2015 SOTA! Overall **0.95** | Title**0.88** | Text**0.95** | Figure**0.98** | Table**0.98** | List**0.98**	-	2020-10
VisualWordGrid	-	ON RVL-CDIP 2020 SOTA! FAR **28.7** | WAR**18.7**	-	2020-10
RectiNet	-	ON DocUNet 2020 SOTA! MS-SSIM **0.415**	Convolution	2020-07
DewarpNet	-	ON DocUNet 2019 SOTA! LD **8.98**	-	2019-10
DocUNet	-	ON DocUNet 2018 SOTA! LD **14.08**	Concatenated Skip Connection	2018-06
Faster-RCNN	-	ON Document Layout Recognition Challenge test 2015 SOTA! Overall **0.91** | Title**0.82** | Text**0.92** | Figure**0.95** | Table**0.95** | List**0.89**	-	2018-03

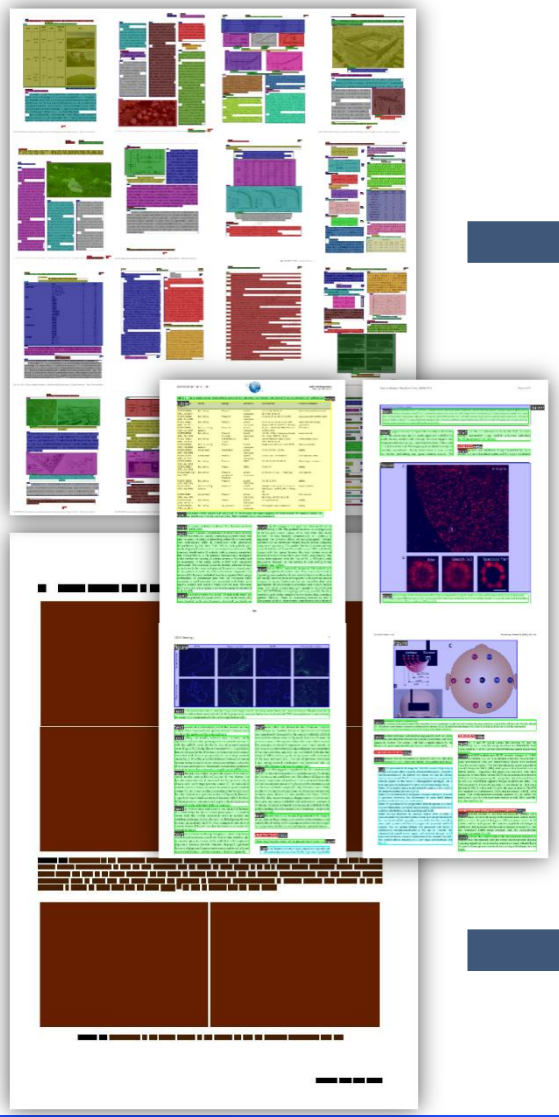
机器之心sota方法排行榜

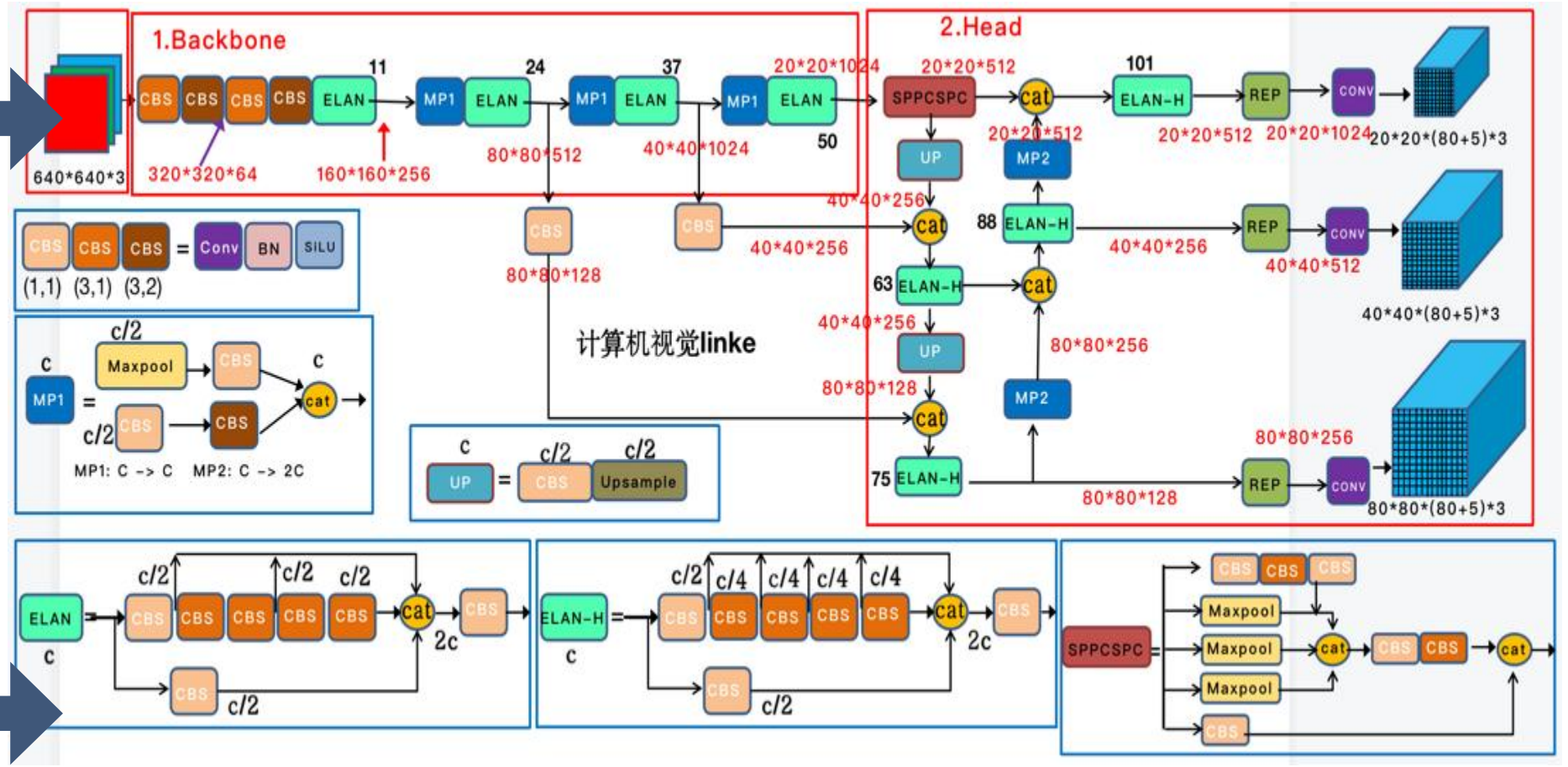
文档块阅读顺序



LayoutReader: Pre-training of Text and Layout for Reading Order Detection

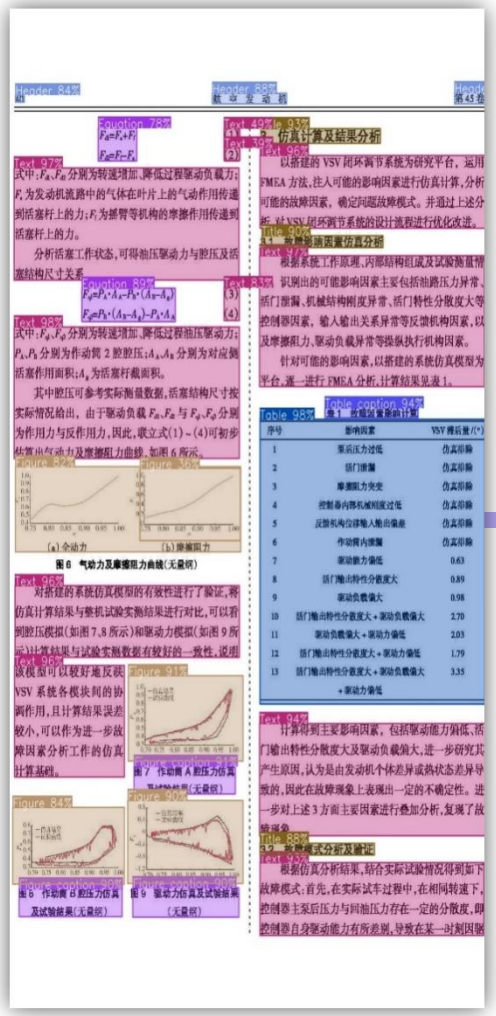
基于目标检测的文档布局识别



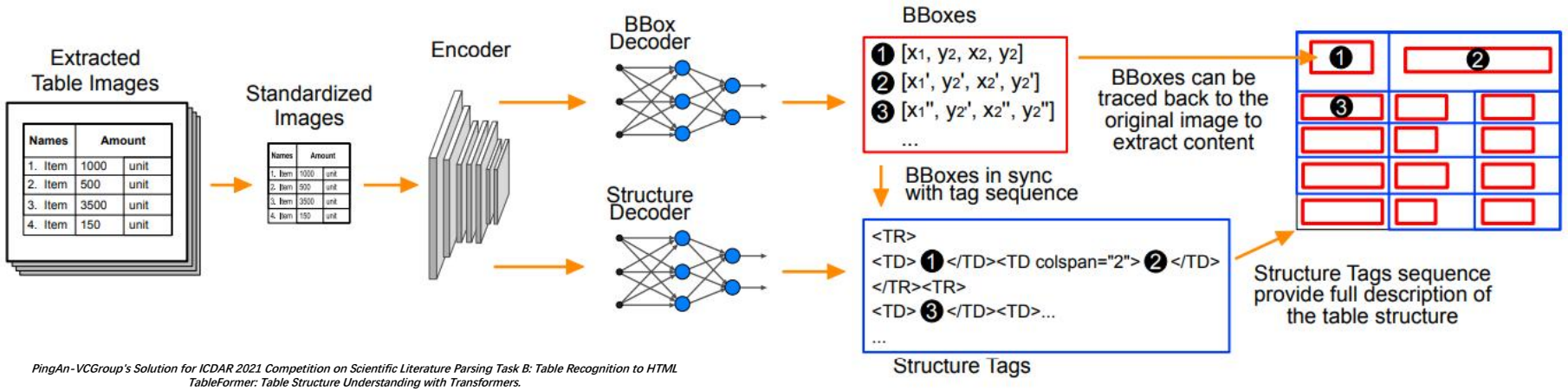


图源自： https://segmentfault.com/a/1190000042194108

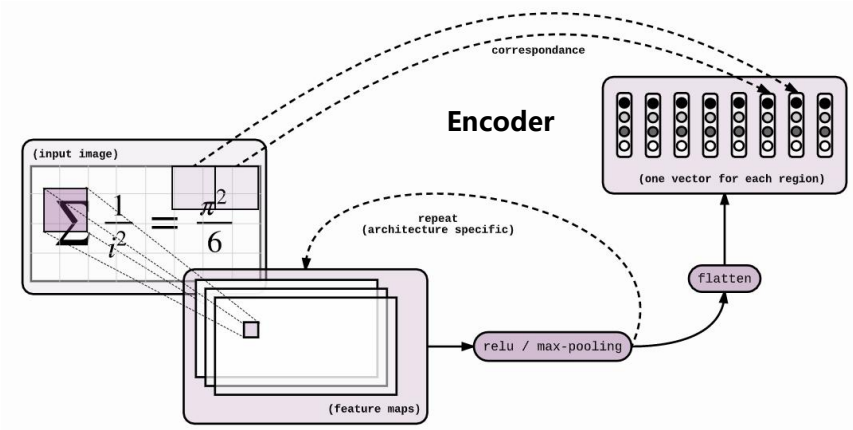
小米大语言模型中数据及知识挖掘

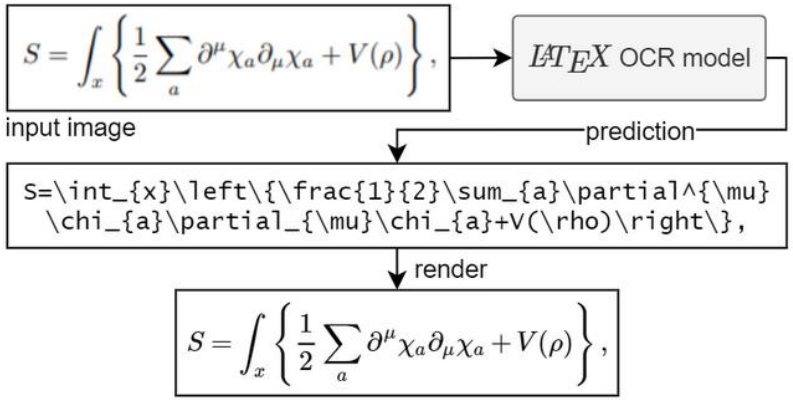


表格解析



PingAn-VCGroup's Solution for ICDAR 2021 Competition on Scientific Literature Parsing Task B: Table Recognition to HTML
TableFormer: Table Structure Understanding with Transformers.





公式解析

小米大语言模型中数据及知识挖掘

word

原生PDF

扫描PDF

图片JPG

Text/Ocr

Latex-ocr

table-ocr

Text/Ocr

40 航空发动机 第45卷

表 1 故障因素影响计算

```
\begin{array}{l}
F_{\mathrm{q} 1}=P_{\mathrm{A}} \cdot A_{\mathrm{A}}-P_{\mathrm{B}} \cdot\left(A_{\mathrm{B}}-A_{\mathrm{g}}\right) \\
F_{\mathrm{q}}=P_{\mathrm{B}} \cdot\left(A_{\mathrm{B}}-A_{\mathrm{g}}\right)-P_{\mathrm{A}} \cdot A_{\mathrm{A}}
\end{array}
```

```
\begin{tabular}{|c|c|c|}
\hline 序号 & 影响因素 & vsv 滥后量 $\left.I^{\circ}\right)$ \\
\hline 1 & 柔后压力过低 & 仿真排除 \\
\hline 2 & 活门泄漏 & 仿真排除 \\
\hline 3 & & 仿真排除 \\
\hline 4 & 控制器内部机械刚度过低 & 仿真排除 \\
\hline 5 & 反钎机构位移输入输出偏差 & 仿真排除 \\
\hline 6 & 作动筒内洽漏 & 仿真排除 \\
\hline 7 & 震动能力偏低 & 0.63 \\
\hline 8 & 活门输出特性分散度大 & 0.89 \\
\hline 9 & 罡动负载偏大 & 0.98 \\
\hline 10 & 活门输出特性分散度大 + 驱动负载偏大 & 2,70 \\
\hline 11 & & 2,03 \\
\hline 12 & 活门输出特性分散度大 + 驱动力偏低 & 1.79 \\
\hline \multirow{2}{*}{13} & 活门输出特性分散度大 + 皎动负裁偏大 & 3.35 \\
\hline & +歌动力偏低 & \\
\hline
\end{tabular}
```

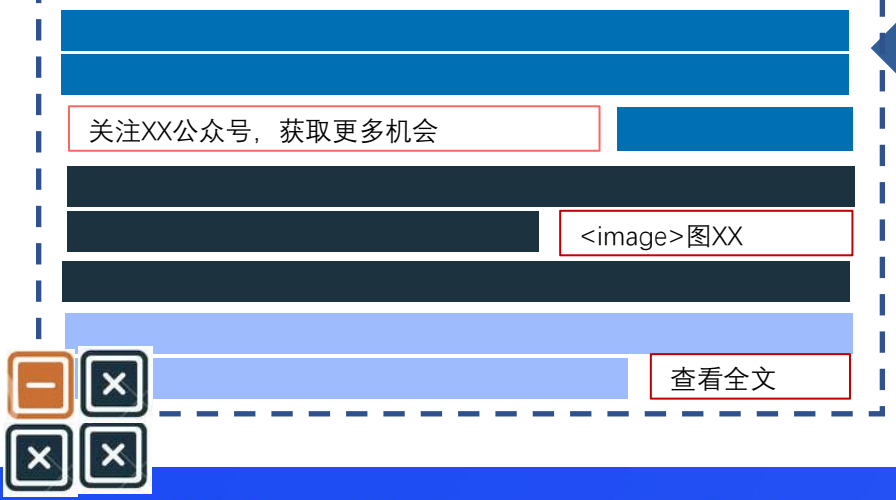
以搭建的 VSV 闭环调节系统为研究平台, 运用 FMEA 方法, 注入可能的影响因素进行仿真计算, 分析可能的故障因素, 确定问题故障模式。并通过上述分析, 对 VSV 闭环调节系统的设计流程进行优化改进。

3.1 故障影响因素仿真分析

```
$$
\begin{array}{l}
F_{\mathrm{q} 1}=P_{\mathrm{A}} \cdot A_{\mathrm{A}}-P_{\mathrm{B}} \cdot\left(A_{\mathrm{B}}-A_{\mathrm{g}}\right) \\
F_{\mathrm{q}}=P_{\mathrm{B}} \cdot\left(A_{\mathrm{B}}-A_{\mathrm{g}}\right)-P_{\mathrm{A}} \cdot A_{\mathrm{A}}
\end{array}
$$

表 1 故障因素影响计算

\begin{tabular}{|c|c|c|}
\hline 序号 & 影响因素 & vsv 滥后量 $\left.I^{\circ}\right)$ \\
\hline 1 & 柔后压力过低 & 仿真排除 \\
\hline 2 & 活门泄漏 & 仿真排除 \\
\hline 3 & & 仿真排除 \\
\hline 4 & 控制器内部机械刚度过低 & 仿真排除 \\
\hline 5 & 反钎机构位移输入输出偏差 & 仿真排除 \\
\hline 6 & 作动筒内洽漏 & 仿真排除 \\
\hline 7 & 震动能力偏低 & 0.63 \\
\hline 8 & 活门输出特性分散度大 & 0.89 \\
\hline 9 & 罡动负载偏大 & 0.98 \\
\hline 10 & 活门输出特性分散度大 + 驱动负载偏大 & 2,70 \\
\hline 11 & & 2,03 \\
\hline 12 & 活门输出特性分散度大 + 驱动力偏低 & 1.79 \\
\hline \multirow{2}{*}{13} & 活门输出特性分散度大 + 皎动负裁偏大 & 3.35 \\
\hline & +歌动力偏低 & \\
\hline
\end{tabular}

根据仿真分析结果，结合实际试验情况得到如下故障模式：
首先，在实际试车过程中，在相同转速下，控制器主原后压力与回油压力存在一定的分散度，即控制器自身驱动能力有所差别，导致在某一时刻因驱
```

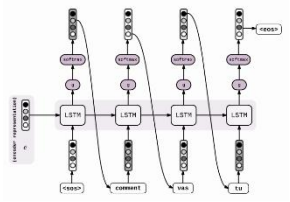
转图片 | 布局元素及阅读顺序预测 | Latex/table ocr / text | 按阅读顺序拼接

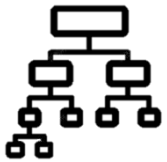
数据及知识挖掘的质量控制



- 宗教信仰
- 身份攻击
- 侮辱
- 威胁
- 亵渎
- 色情

大模型与小模型联动提升模型迭代速度

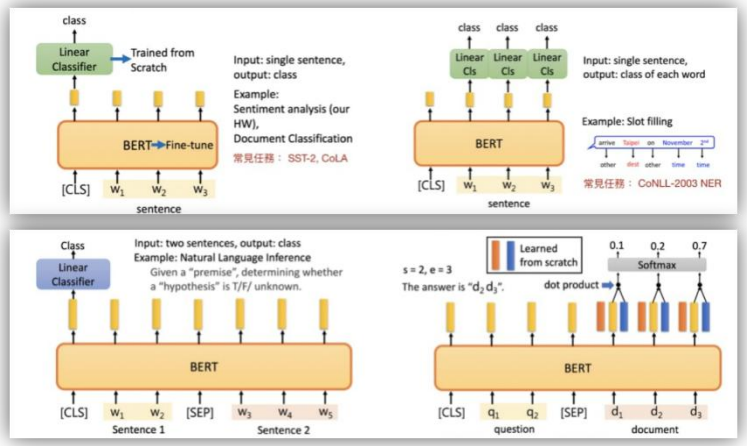




MILM

小模型快速的质量发现及识别清洗

识别　清洗　去重



小模型

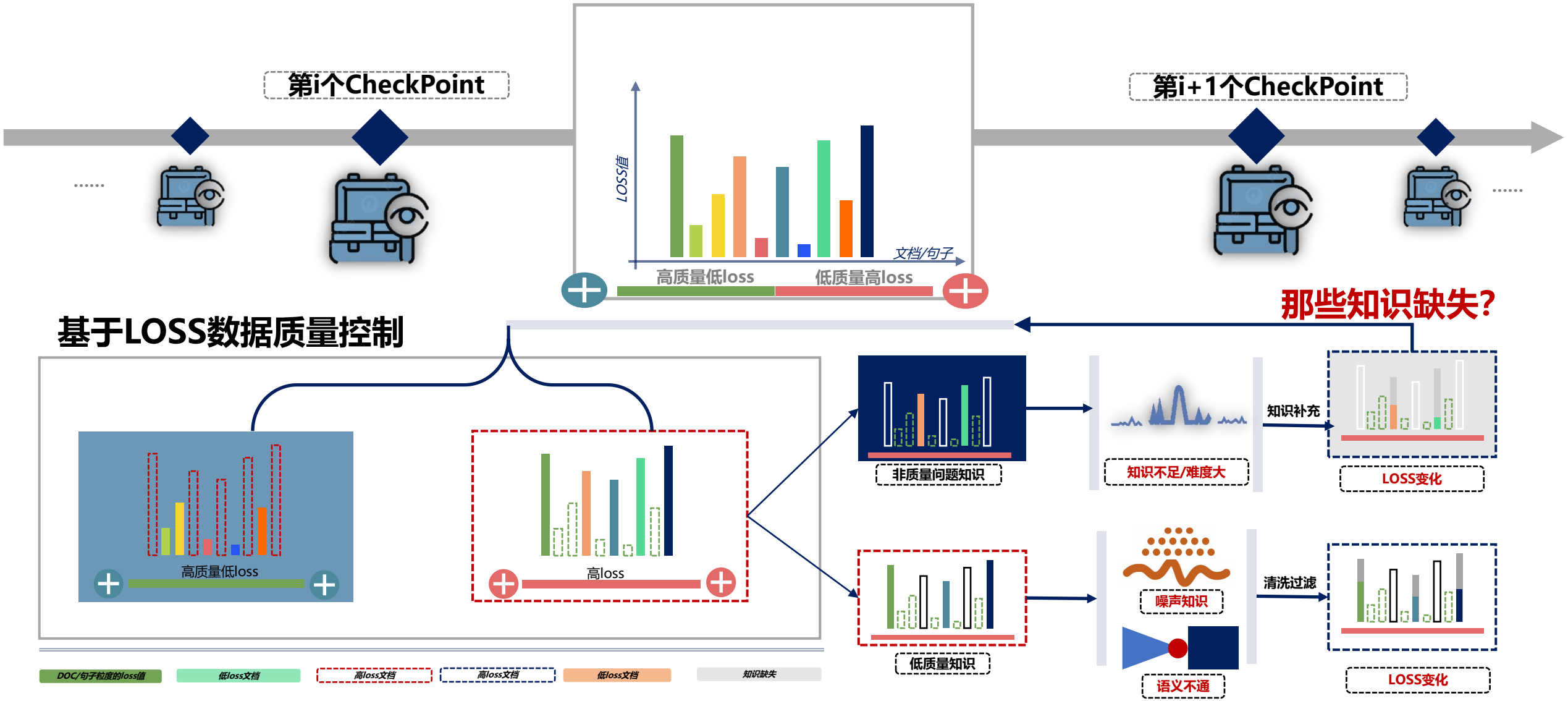
大模型能否发现质量问题?

高质量的数据影响模型的质量

质量　速度　成本

大模型

大模型与小模型联动提升模型迭代速度



丰富知识维度和控制采样提升模型迭代速度

粗粒度数据类别及组成

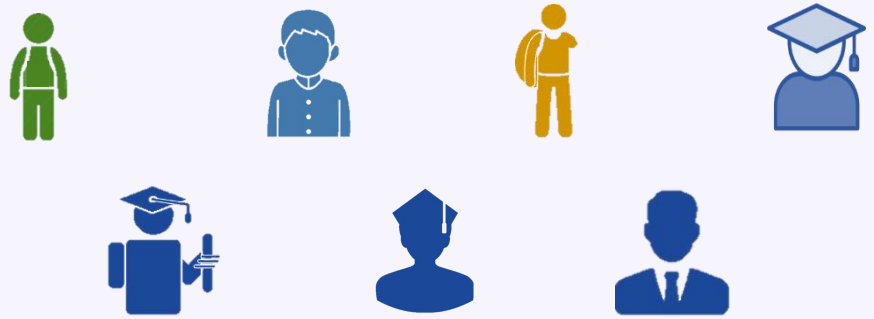
文档格式

数据来源

细化知识分布及统计

行业分类

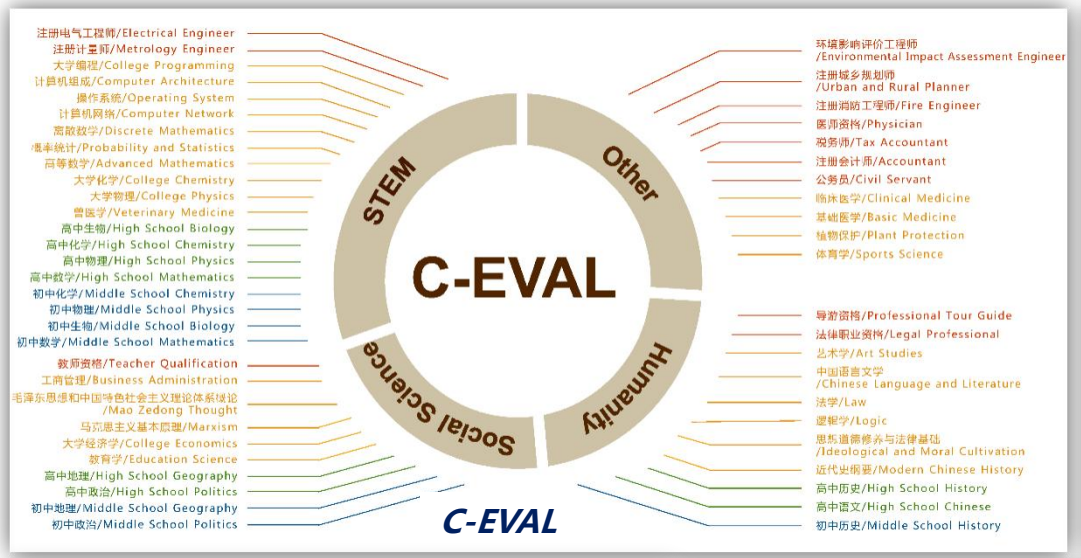
教育分类



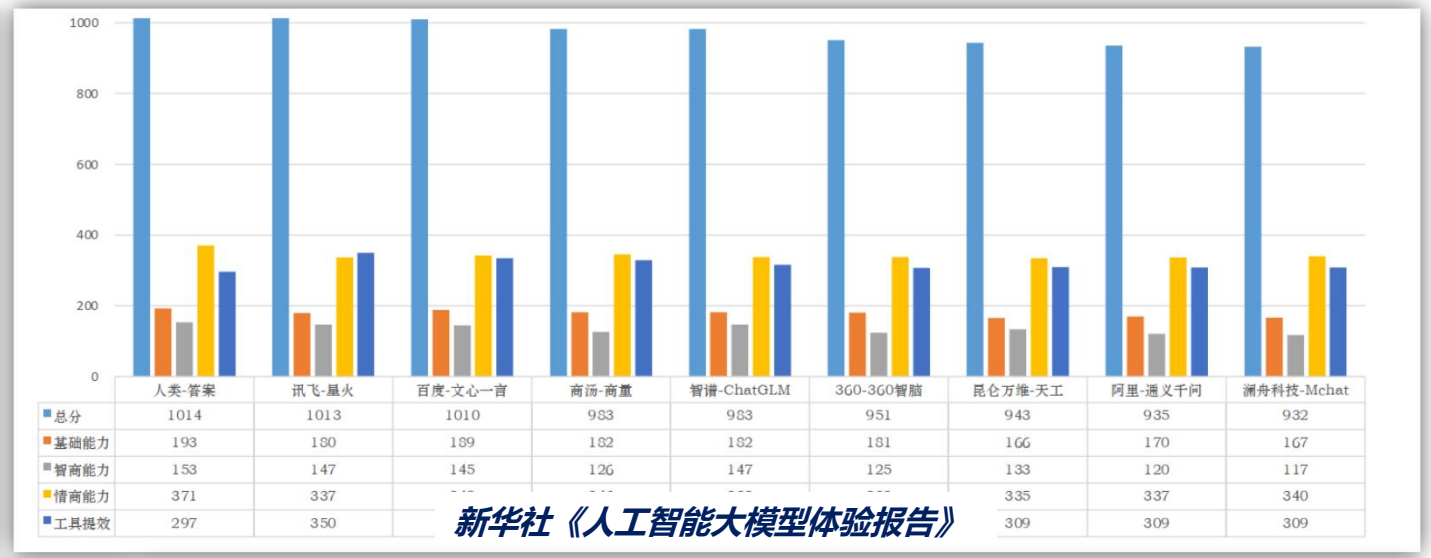
模型及数据质量评估的方法

准确率	完整性	一致性	无偏性	重复率
丰富度	覆盖度	毒性率	满足率	清洁度

模型及数据质量评估的方法

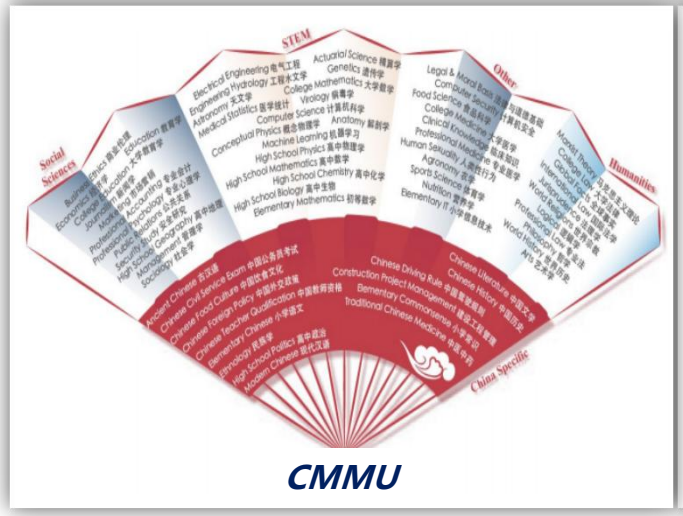


C-EVAL

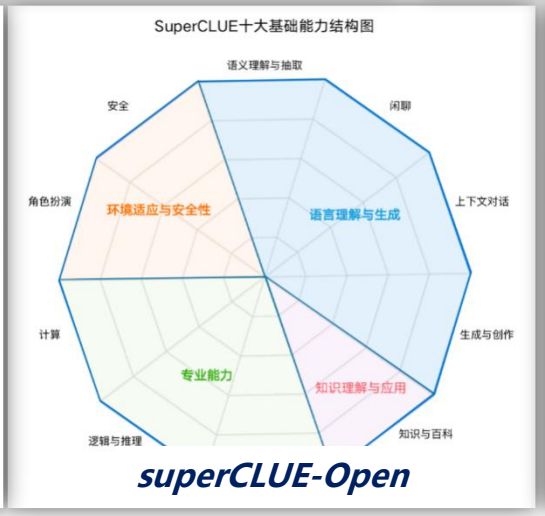


	人类-答案	讯飞-星火	百度-文心一言	商汤-商量	智谱-ChatGLM	360-360智脑	昆仑万维-天工	阿里-通义千问	澜舟科技-Mchat
总分	1014	1013	1010	983	983	951	943	935	932
基础能力	193	180	189	182	182	181	166	170	167
智商能力	153	147	145	126	147	125	133	120	117
情商能力	371	337					335	337	340
工具提效	297	350					309	309	309

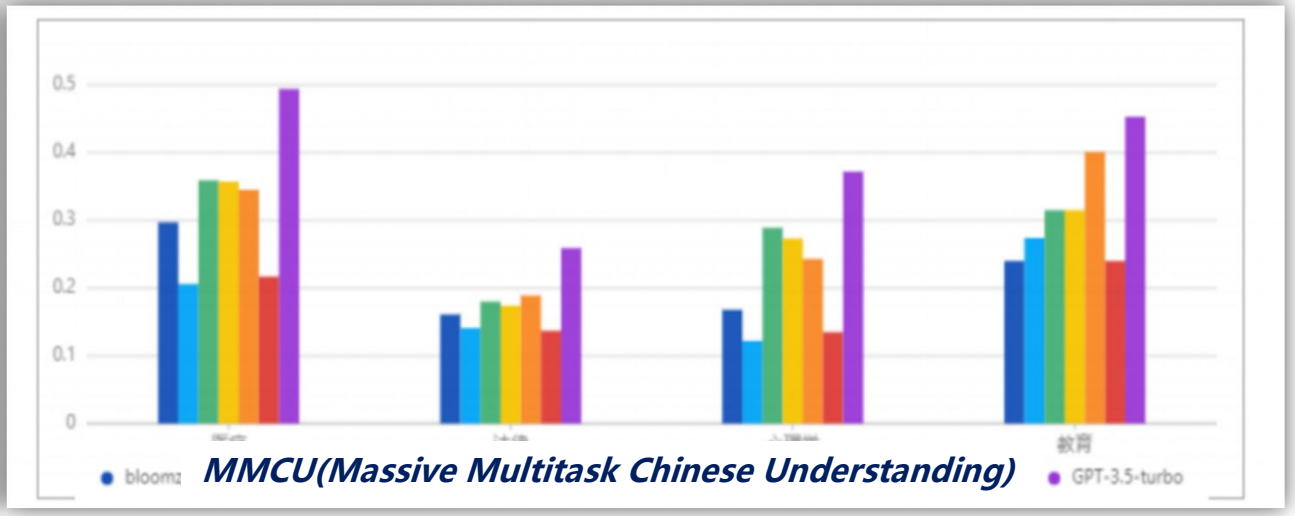
新华社《人工智能大模型体验报告》



CMMU



superCLUE-Open



MMCU(Massive Multitask Chinese Understanding)

小米模型及数据质量评估的方法

信息量　事实性

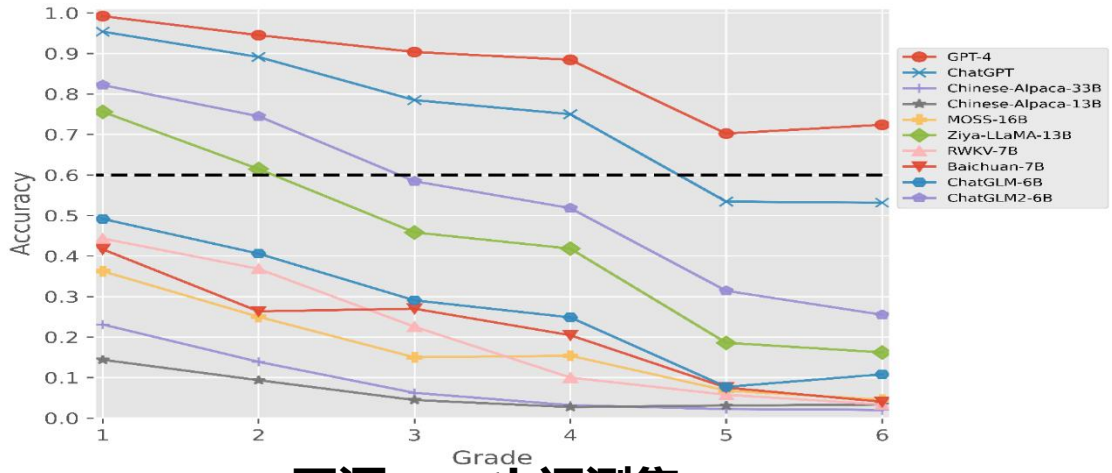
无害性　满足性

评估方法原则及维度

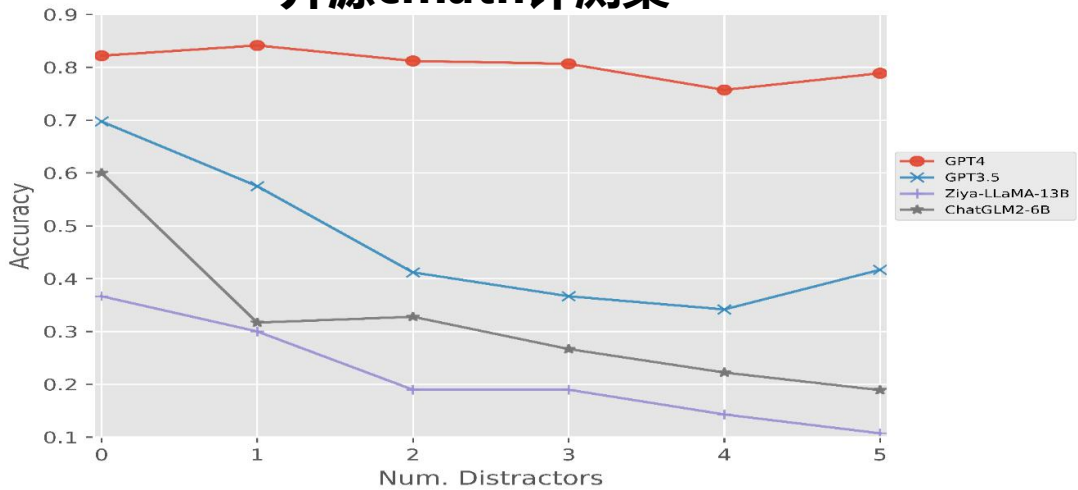
通用能力　**专项能力**

满足程度

0　1　2　3　4



开源cmath评测集



《CMATH: Can Your Language Model Pass Chinese Elementary School Math Test?》

- 对齐开放原则
- 能力多维度评估
- 满足度细化
- 开源专项评测

大模型及数据质量评估依然存在挑战

评估结论置信度低

评估测 + **模型测**

评估人员能力

普通标注人员能力不足，模型能力大于普通标注人员能力

高效评估

生成能力主观性强，无标准答案，自动化评估不客观

评估范围广

通用能力丰富，评测范围广，专业能力分类多

Prompt敏感

提示指令变化对结果影响大，难公平的比较不同的模型

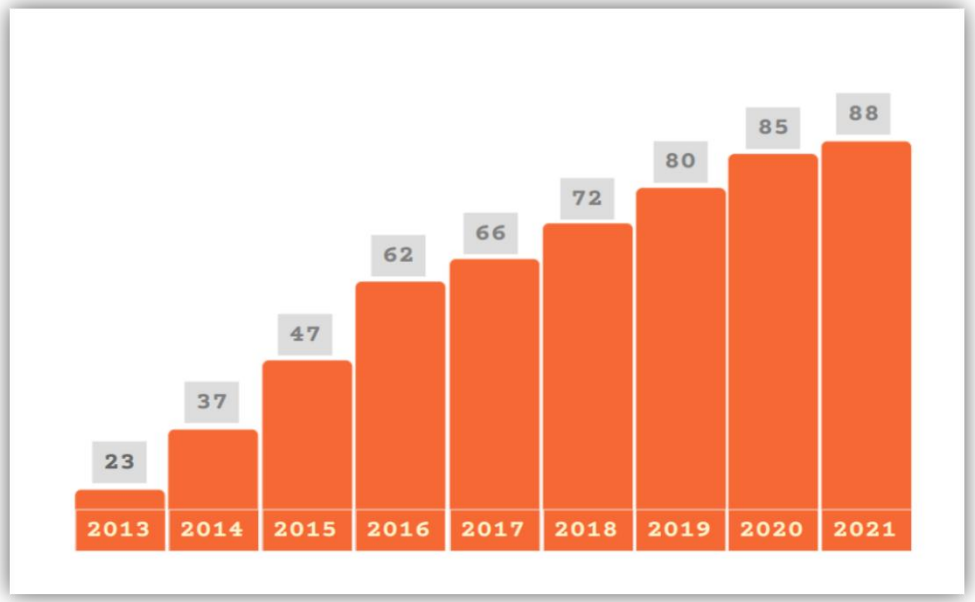
- **低代价**(评价成本低、效率高)
- **可复现性**(相同设置下 (软件、人等) 多次评价有一致性)
- **公平客观** (客观真实的评价模型效果)

大语言模型中数据及知识的存储

88，343，822

100TB数据

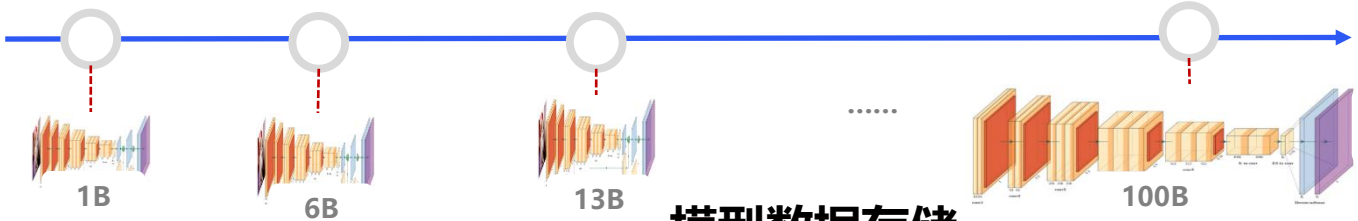
小文件存储



1.海量小文件存储需求，对小文件的预处理及访问和计算提出挑战

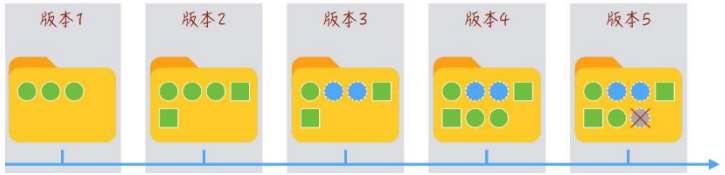
2.不同参数规模的模型文件，相同参数模型下不同超参的每个epoch的存储

3. 数据处理的不同阶段（解析、清洗、去重、分类）等中间态的数据存储

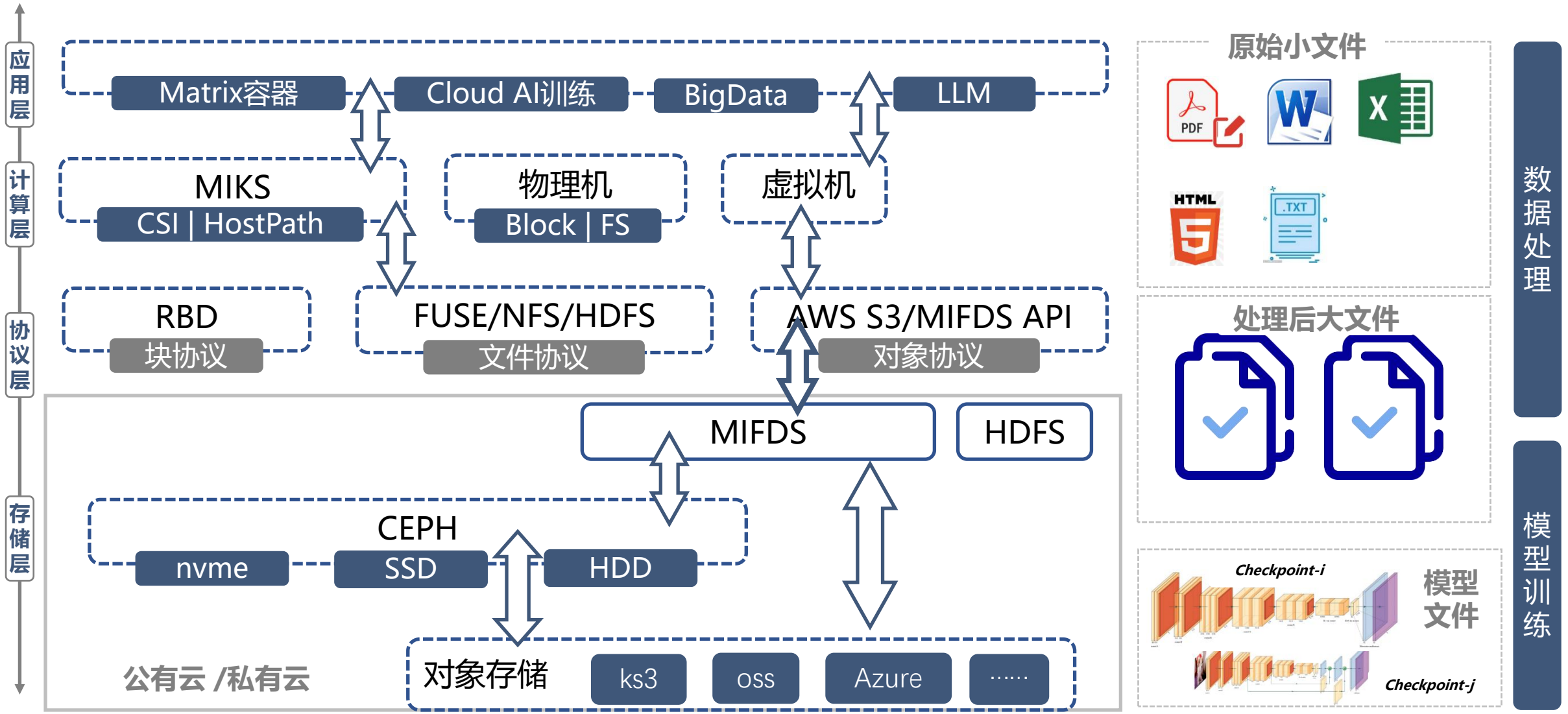


模型数据存储

不同参数规模 X **不同版本**



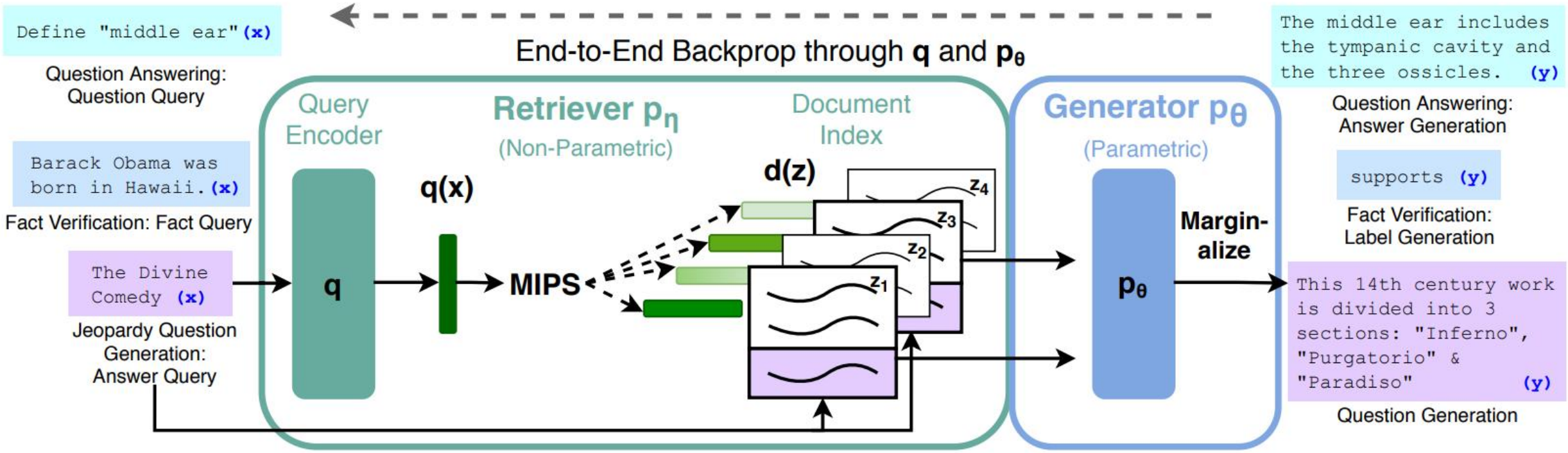
小米大语言模型中数据及知识挖掘



PART 03

大模型在小米领域场景下的应用和方法

LLM+Retrieval-Augmented Generation的领域应用



Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks

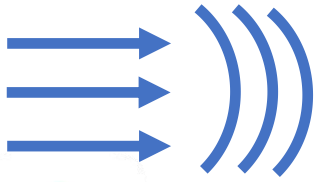
- 模型的知识/数据是隐式的编码在参数中，引入的知识是显式的，优好的扩展性及更可控的时效性。

- 为生成模型引入更丰富的关联和参考信息，可以提升生成模型的性能和丰富度。

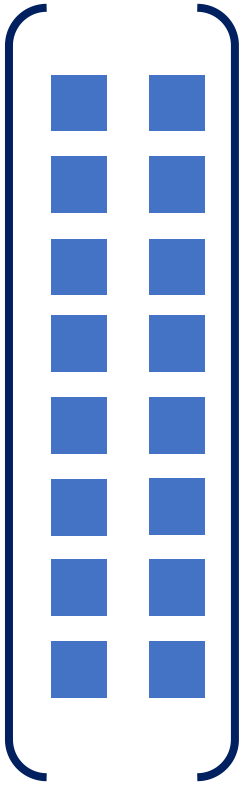
LLM+Retrieval-Augmented Generation的领域应用

query

以色列与巴勒斯坦战争什么状态



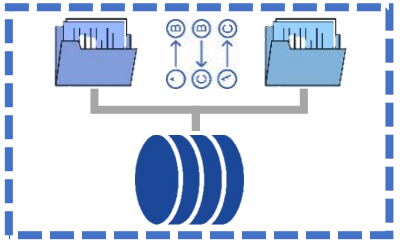
外部知识和数据



向量表示

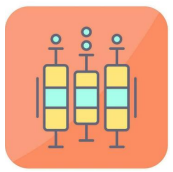
用户问题

以色列与巴勒斯坦战争什么状态



context

最近邻



生成模型/MILM

基座数据

以色列对加沙战争：加沙巴勒斯坦死亡人数超过8000人。

•加沙圣城医院受袭可能性增大
•过去三周巴勒斯坦你儿童遇害人数超近三年全世界遇害总和

最新事件

外部知识 | 隐式向量化 | 向量检索 | Prompt构造 | 模型生成

领域数据高效全参微调的小爱问答场景应用方法

全参微调

高效微调LORA/QLORA

L1 L2 L3 L4

plugin

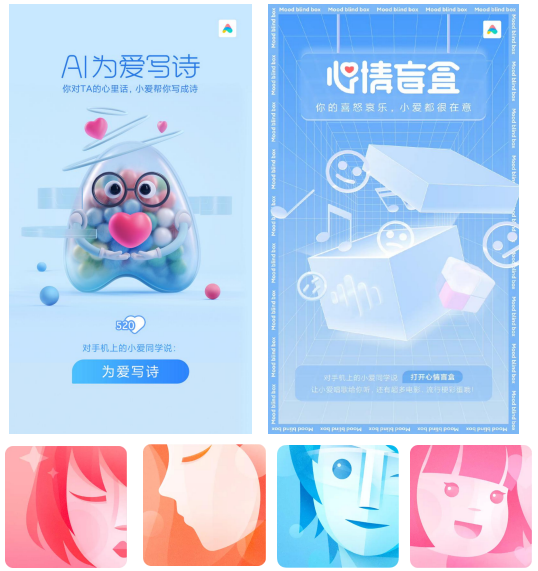
意图理解分发

创作生成

应用助手

设备控制

内容领域



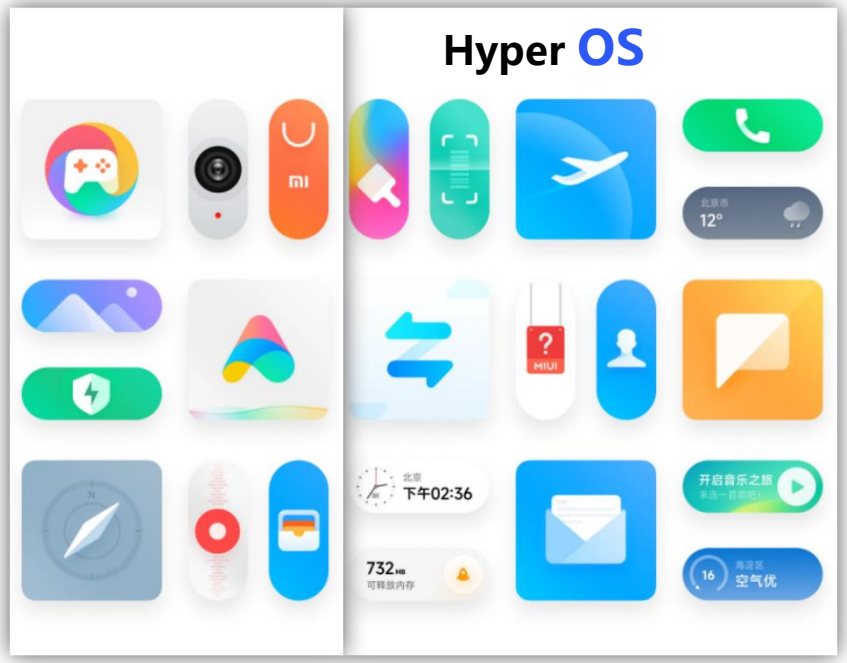
生成创作

应用助手

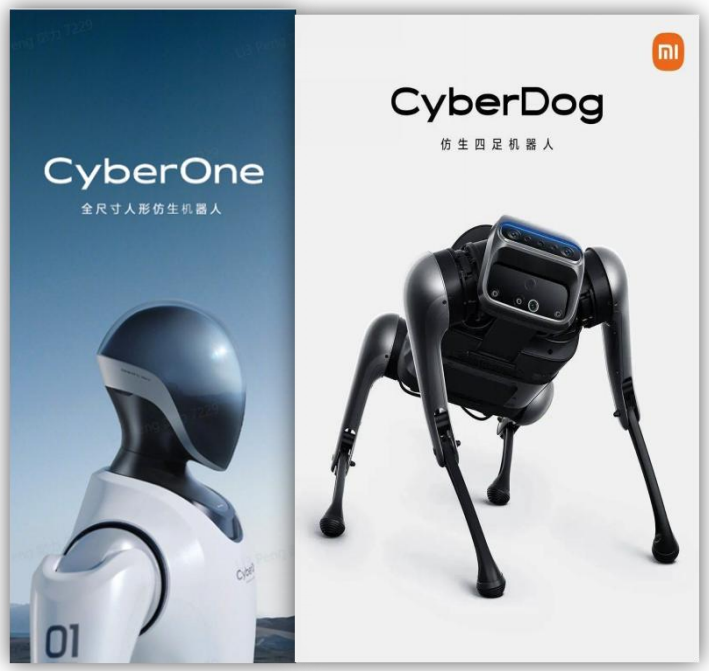
大模型赋能小米应用场景

轻量化 本地部署

智能驾驶



操作系统



仿生机器人

PART 04
总结及展望

总结

01 数据依然是影响LLM性能的决定元素之一

02 知识获取、存储、评估等不断完善

03 大语言模型在小米业务下逐渐成长

04 数据为中心的大语言模型会更强大

AiDD AI+ 软件研发数字峰会
AI+ software Development Digital summit

THANKS