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

大模型在ToB企服领域的技术和应用实践

李翔 WakeData

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

李翔

惟客数据 AI 算法科学家

中山大学人工智能方向博士&博士后，珠海市产业青年优秀人才，在人工智能领域有 11 年的研究与落地经验；熟悉资讯流推荐、画像预测标签、NLP、CV、语音识别等多个 AI 方向，并将对应落地成果发表在国际一流期刊以及申请多项发明技术专利

目录
CONTENTS

PART 01
大模型发展脉络以及对趋势的预判

大语言模型发展概览

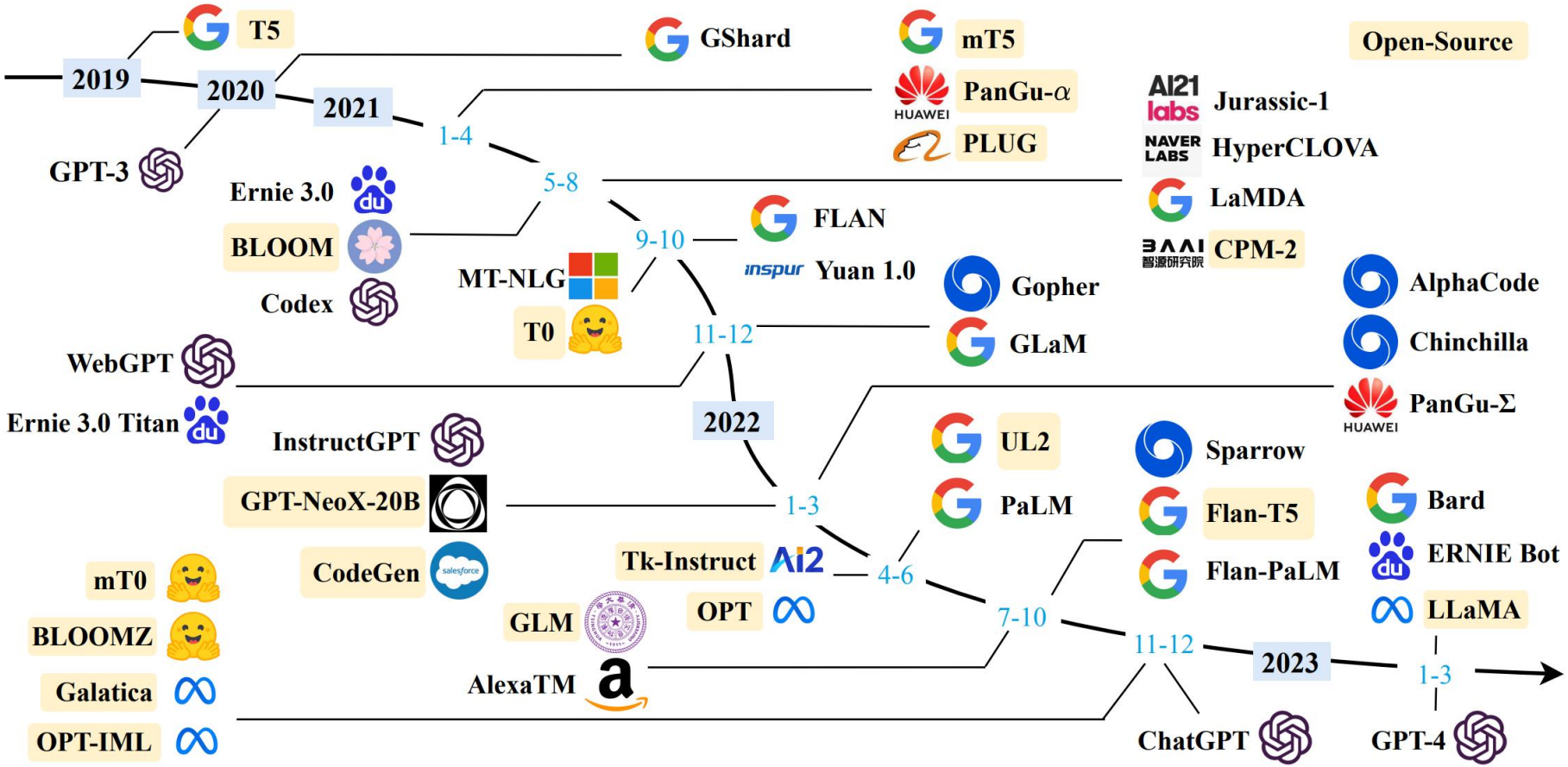
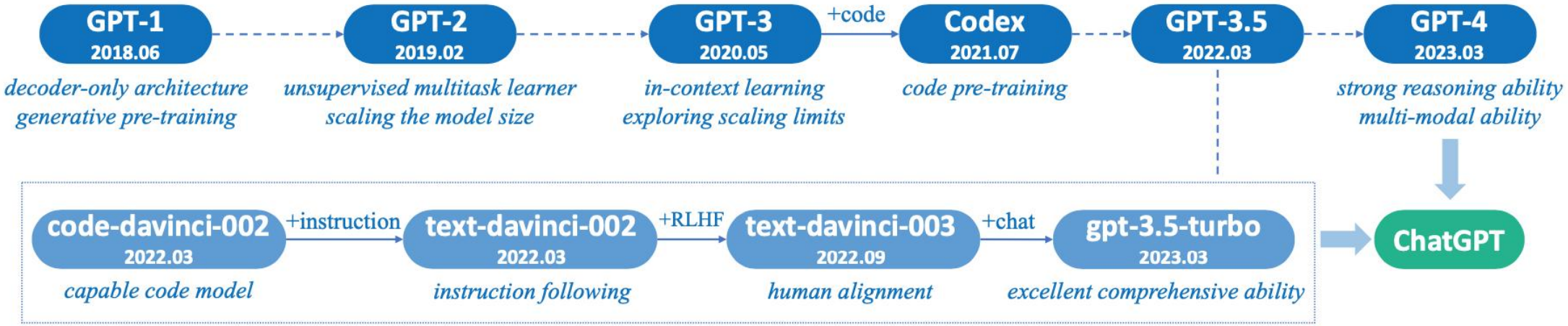


Fig. 1. A timeline of existing large language models (having a size larger than 10B) in recent years. We mark the open-source LLMs in yellow color.

闭源大模型：OpenAI模型的演进

OpenAI



OpenAI DevDay

GPT-4 Turbo

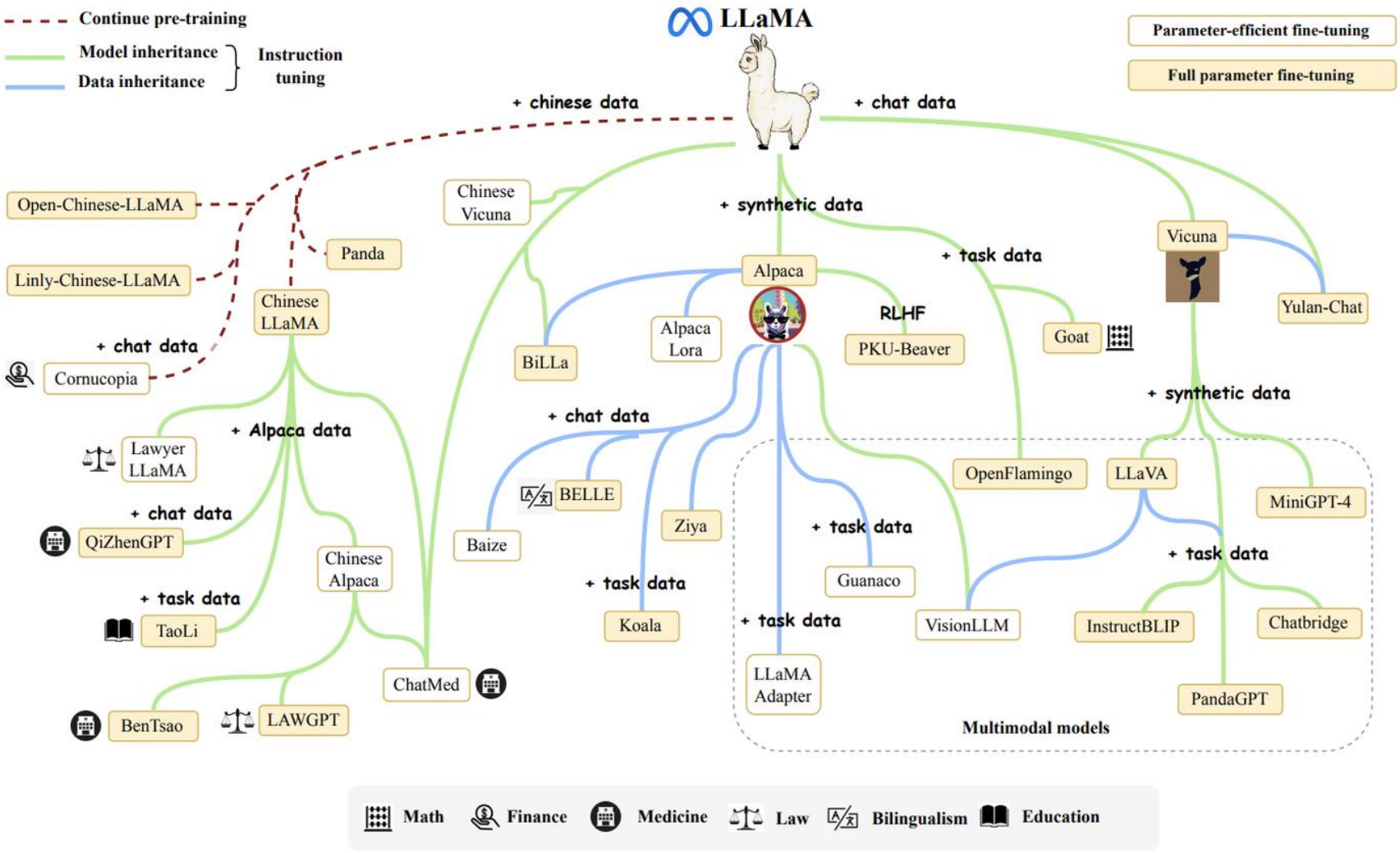
多模态

API降价&提速

Agent工具

GPT商店

开源大模型：LLaMA家族



开源VS闭源：观点和思考

观点一：闭源会一直遥遥领先
来自OpenAI & Anthropic的高层的一个饭局

观点二：开源会无限接近闭源
企业：Meta的LLaMA
高校：研究方向的香饽饽

依据： 壁垒会在哪里？

数据：
- 公开数据，RLHF的标注数据
- 数据积累：OpenAI VS Google
- 相比数据飞轮，数据质量更重要

人才：
- 流通

算力：
- 卡的数量要到什么级别，真的是越多就能形成壁垒吗？

大模型发展方向的预判

1 多模态：GPT4-V、BLIP-2、LLAVA、Qwen-vl、CogVLM-17B

2 对上下文长token的支持：Claude、月之暗面、百川

3 微调技术：LongLora

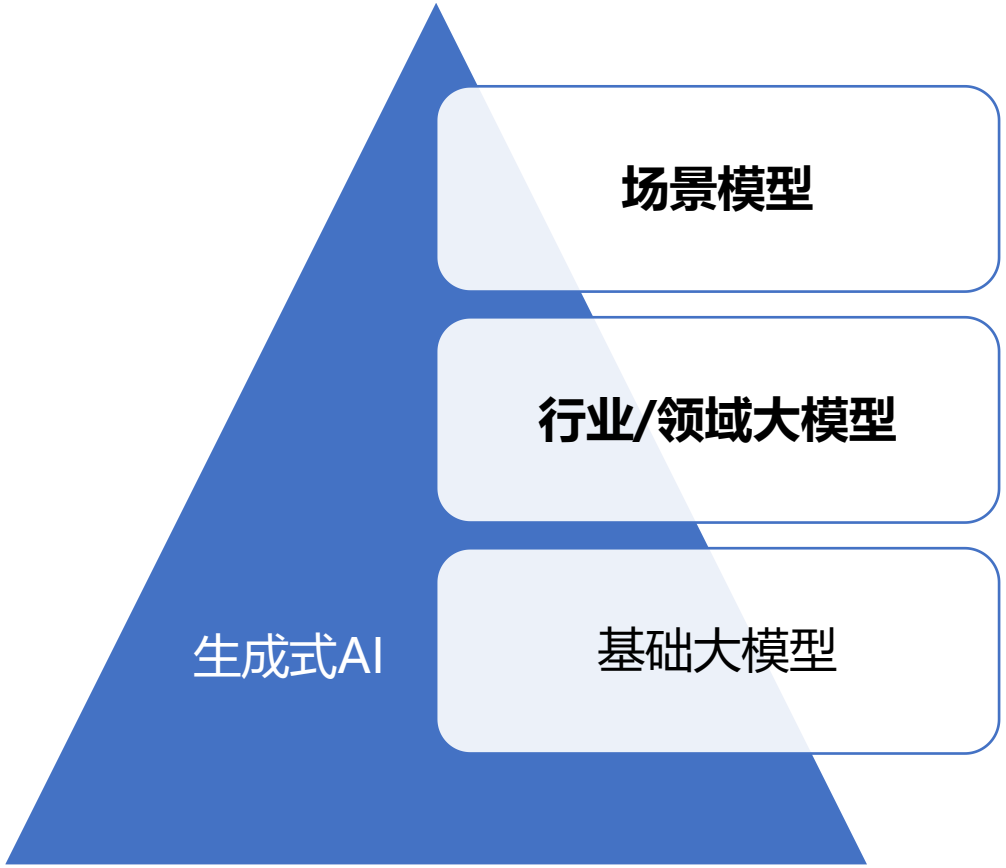
4 Agent：AutoGPT、MetaGPT

5 减少幻觉：长期

PART 02

大模型在ToB企服领域有哪些机会

大模型在ToB领域的产品化和商业化思考



LE	SME
内部价值链	SaaS
外部价值链	+
单点式创新	营销域
系统性创新	服务域
私有化部署	内部效能
独立自建	**行业合作/云端调用**

竞争格局？ 大厂vs.创业？ 闭源vs.开源

“训”vs.“用”：使用大模型的几种方式

01 **提示词模式（Prompt only）**

直接使用提示词调用API

02 **嵌入向量模式（Embedding）**

将知识预处理存入向量数据库，提问时通过相似度查询找到关联知识，然后跟问题一起加入提示词，再调用API

03 **精调模型模式（Fine-tune）**

将知识通过Fine-Tune训练存入大模型

为什么要有领域大模型?

1 一切的一切，都是为了【效果】

2 GPT4 的MoE模式

3 长期过程：很多领域知识不可见

不同的声音：智谱AI不做细分行业的【行业模型】，因为对大模型的通用性有信心

为什么要做私有化大模型?

1 私有数据让大模型的**效果**更好

2 数据**隐私**和安全

3 降低大模型使用**成本**

相关公司：

- MosaicML ：Databricks以13亿美金收购，在上一轮的融资中，其估值为2.2亿美元，估值提升6倍
- Reka：5800万美金A轮
- Mistral AI：1.13亿美元种子轮
- 智谱AI
- MiniMax
- ……

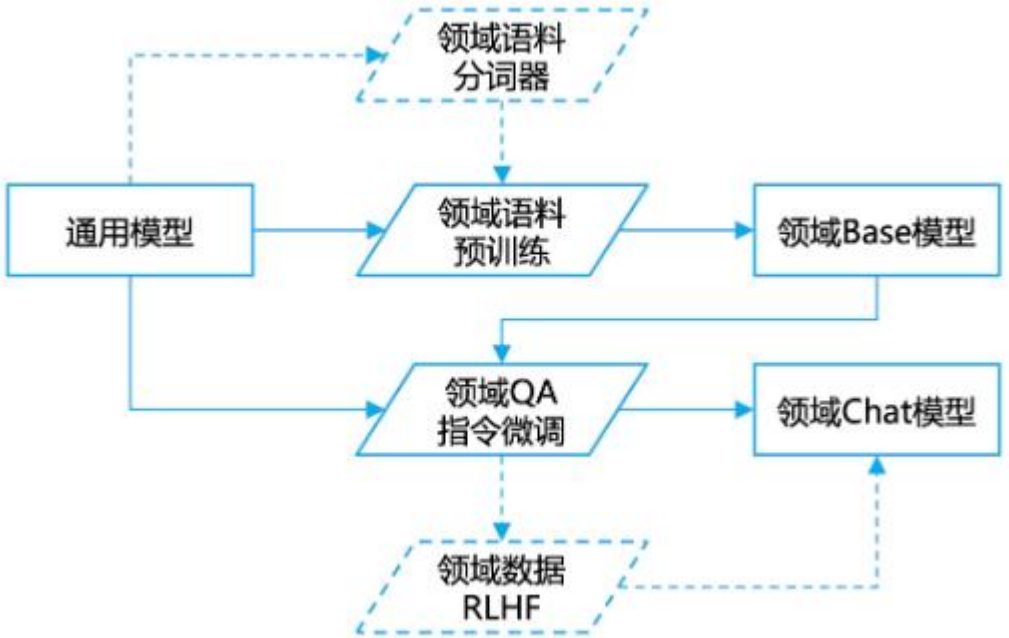
Microsoft paper claims ChatGPT 3.5 has ~20 billion parameters arxiv.org/abs/2310.17680

System description			Python (	
System	**Model**	**#P**	**top-1**	**t**
T5	t5-large	770M	80.4	
CodeT5	codet5-large	770M	80.5	
GPT-3	text-davinci-003	175B	**82.5**	
ChatGPT	gpt-3.5-turbo	20B	80.6	
StarCoder	starcoder	15.5B	79.2	
CodeT5+	codet5p-16b	16B	79.6	
CodeGen	codegen-350m	350M	80.1	
Diffusion-LM	Custom	50M	70.4	
GENIE	Custom	93M	73.2	
CODEFUSION	**Custom**	75M	80.7	

PART 03
私有化大模型的一些技术细节

如何挑选模型?

什么是BaseModel，什么是ChatModel



如何挑选模型?

测试数据集	范围	形式	规模
C-Eval	人文、社科、理工等52个学科	13948道单选题，涉及52个学科，4类不同难度（初中、高中、大学、专业）	学科知识，难度跨度合适，缺乏对生成表达能力的考察
CMMLU	常识类、人文、社科、理工等共67个主题	11,528道单选题，其中67个主题每个主题至少105道题	学科知识，选择题适合快速评测，缺乏对生成表达能力的考察
Gaokao-Bench	2010-2022年高考试卷，包括文科和理科	2811道题目，包括选择、填空、解答	数据质量高，范围窄、跨度小，主观评测需要人工参与，成本较高
仅供参考--在榜单上得到高分的方式：从GPT-4的预测结果蒸馏，找人工标注然后蒸馏；在网上找到原题加入训练集中微调模型 。然而这样得到的分数是没有意义的			

如何挑选模型?

公司	模型	Token	规模
Meta	LLaMA		7、13、70
TII	Falcon	1T-3.5T	7、40、180
智谱	chatGLM		6
百川	Baichuan	2.6T	7、13
上海AI lab	InternLM	2.3T	7、20
阿里	Qwen	2.4T-3T	7、14
零一万物	Yi	3T	6、34
元象	Xverse	2.6-3.2T	7、13、65
…	…	…	…

#	Model	Creator	Access	Submission Date	Avg ▼
0	Yi-34B	零一万物	Weight	2023/11/2	81.4
1	BlueLM-7B	vivo	Weight	2023/11/7	73.3
2	Qwen-14B	Alibaba Cloud	Weight	2023/9/22	72.1
3	Yi-6B	零一万物	Weight	2023/11/2	72
4	XuanYuan-70B	度小满AI-Lab	Weight	2023/9/21	71.9
5	ChatGLM3-6B-base	Tsinghua & Zhipu.AI	Weight	2023/10/26	69
6	GPT-4*	OpenAI	API, Web	2023/5/15	68.7
7	XVERSE-65B	XVERSE Technology	Weight	2023/11/5	68.6
8	Nanbeige-16B-Base	Nanbeige LLM Lab	Weight	2023/11/8	63.8
9	LingoWhale-8B	深言科技(DeepLangAI)	Weight	2023/11/3	63.6
10	Qwen-7B v1.1	Alibaba Cloud	Weight	2023/9/12	63.5

如何低成本部署?

性能参数	V100 PCIe	A100 80GB PCIe	A800 80GB PCIe	H100 80GB PCIe
微架构	Volta	Ampere		Hopper
FP64	7TFLOPS	9.7TFLOPS		26 TFLOPS
FP32	14TFLOPS	19.5TFLOPS		51 TFLOPS
FP16 Tensor Core		312TFLOPS		756.5 TFLOPS
INT8 Tensor Core	62 TOPS	624 TOPS		1513 TOPS
GPU 显存	32/16GB HBM2	80GB HBM2e		80GB
GPU 显存带宽	900 GB/s	1935GB/s		2TB/s
最大热设计功耗 (TDP)	250 瓦	300 瓦		300-350W
多实例 GPU		最多 7 个 MIG 每个 10GB		
外形规格		PCIe 双插槽风冷式 或单插槽液冷式		PCIe 双插槽风冷式
互连技术	NVLink：300 GB/s PCIE：32 GB/s	搭载 2 个 GPU 的 NVIDIA" NVLink" 桥接器：600GB/s PCIe 4.0：64GB/s	搭载 2 个 GPU 的 NVIDIA" NVLink" 桥 接器：400GB/s PCIe 4.0：64GB/s	NVLink：600GB/s PCIe 5.0：128GB/s
服务器选项		搭载 1 至 8 个 GPU 的合作伙伴认证系统和 NVIDIA 认证系统		

GeForce RTX 3090：40TFLOPs
GeForce RTX 4090：83TFLOPs~100TFLOPs

瓶颈是显存和通信、licence

GPU 利用率不高，原因有2个维度：
1 故障
2 显存、通信限制

如何低成本部署?

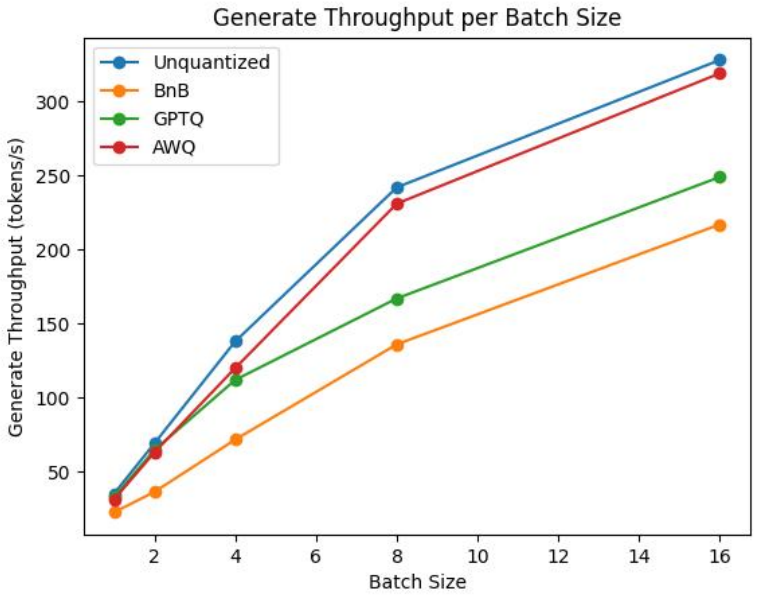
1 量化—性能退化很小

LLaMA 65B INT4 VS LLaMA 30B INT8

INT8、INT4降低显存消耗，但会**降低推理速度**

建议基于官方提供的量化模型

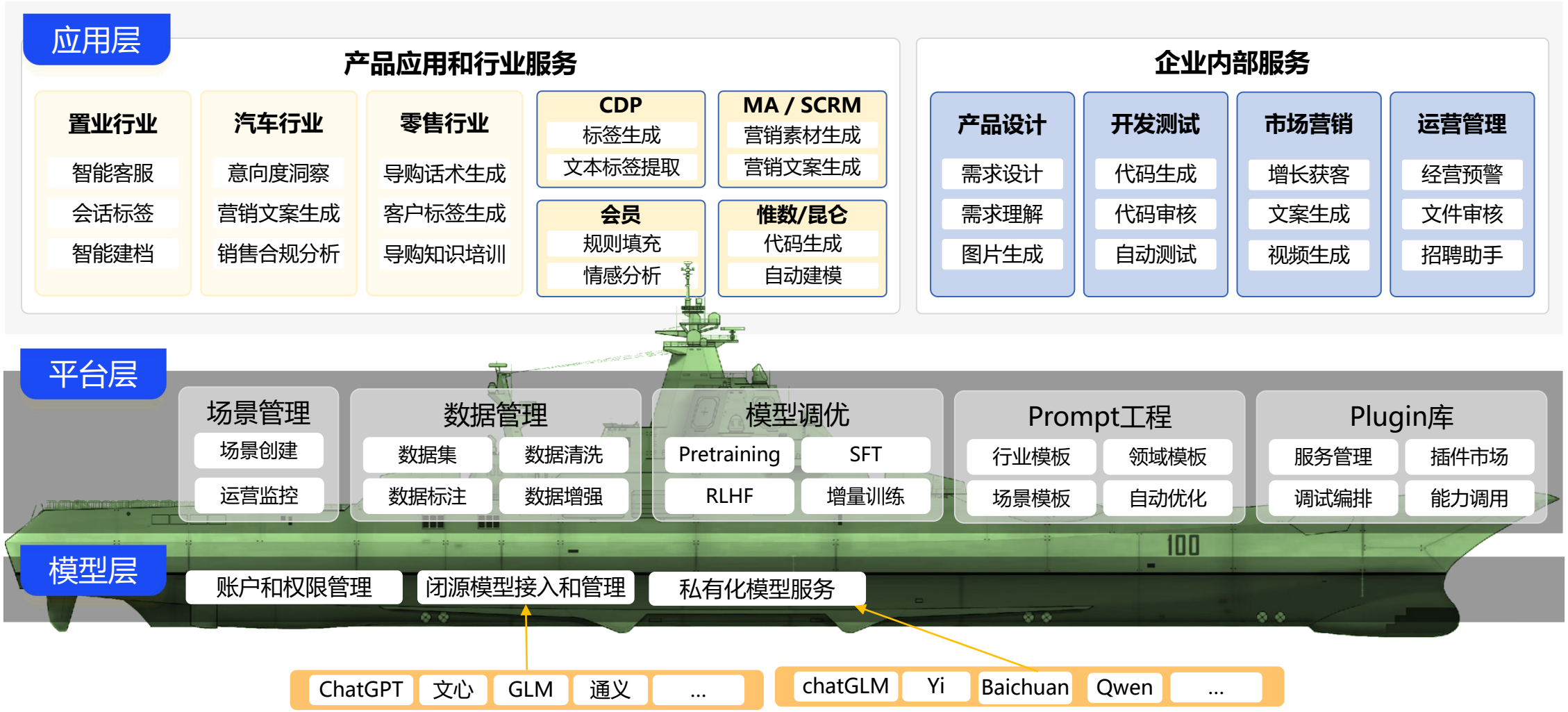
模型	显卡	FP16	INT8	INT4
chatGLM3	T4	18 token	3 token	6 token
LLaMA-7B	A4000	28 token	10 token	17 token
Baichuan2-13B	A800	35 token	10 token	



2 TensorRT-LLM、TVM

PART 04
WakeData的解决思路和实践

新一代领域大模型平台：WakeMind



独立自建私有化大模型的落地方法

场景梳理

盘点业务系统、识别关键痛点、整理潜在场景

方案设计

将场景需求与AI能力匹配
评估策略和方案

模型微调

整理预训练数据和标注
大模型训练

效果评测

准备用例和测试数据
评测效果并分析优化

落地跟踪

Prompt调整、业务流程上线，
效果跟踪和问题反馈

领域大模型落地经验

不动产领域

1 数据集：

57M tokens：专业文章、国家GB标准文档、书籍、论坛圈子；
问答对：56k
600M tokens：通用数据

2 三种训练方式：

重头训：贵、需要海量数据（先基础后专业）
基于Basemodel或者Chatmodel直接instruction-tuning
基于Basemodel先post-training再instruction-tuning

3 模型：

Baichuan2 13B

4 结论：

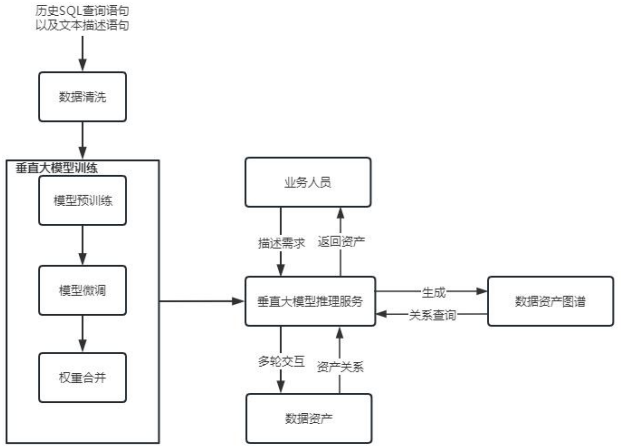
1 Basemodel直接instruction-tuning
2 Chatmodel 直接instruction-tuning
3 Basemodel post-training（尽量只用领域数据，不要融合数据），再instruction-tuning
4 instruction-tuning 可以做融合数据

产品和行业服务场景

场景1

数据资产建设

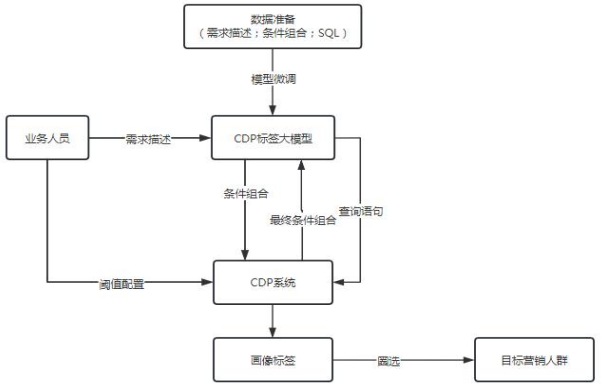
适用人员：业务
适用场景：快速低成本的搭建数据资产图谱，提高数据资产的复用率



场景2

自动生成画像标签

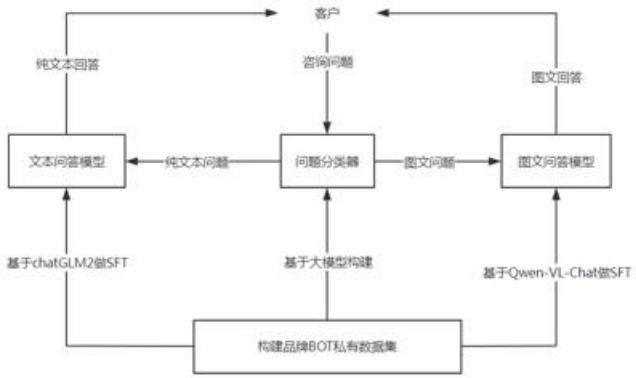
适用人员：业务
适用场景：CDP中自动生成画像标签，提升标签构建效率



场景3

多模态智能对话

适用人员：业务
适用场景：基于地产行业垂直场景数据，服务地产营销和服务对话场景

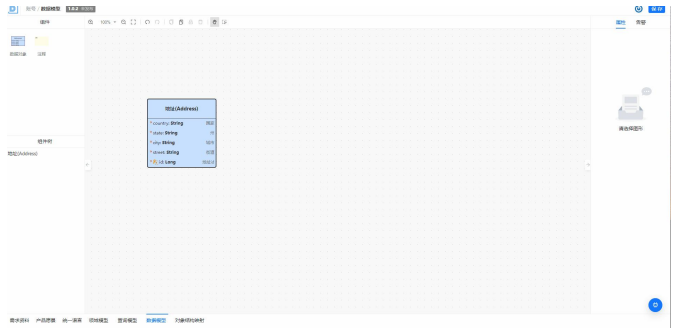


企业内部服务场景（1/2）

场景1

数据建模

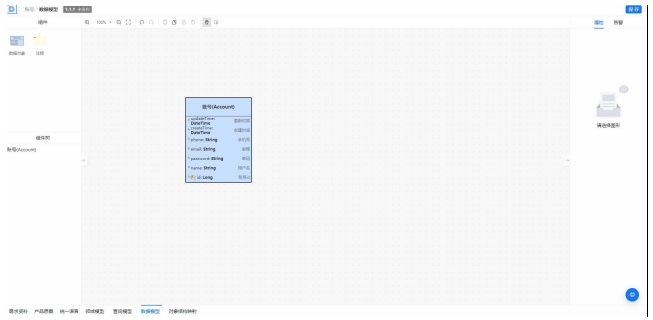
适用人员：架构师、工程师
适用场景：架构师通过ER图来表示数据的业务流转与存储逻辑



场景2

SQL编写/优化/诊断

适用人员：架构师、工程师、运维
适用场景：对SQL进行诊断和优化，提升编写效率



场景3

统一语言自动补充

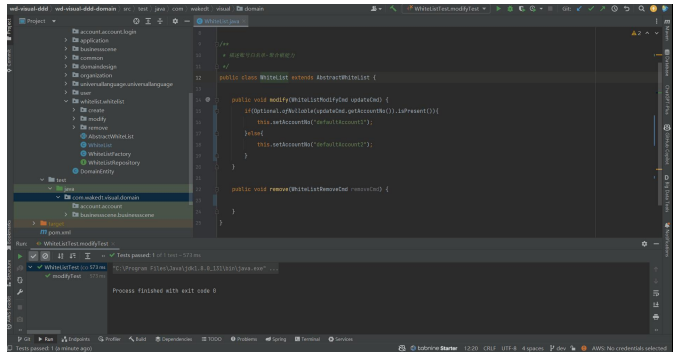
适用人员：产品、架构师、开发测试
适用场景：在软件开发的任意节点都可以添加统一语言，快速预生成统一语言的英文、定义和举例，提高开发效率

企业内部服务场景(2/2)

场景4

单元测试

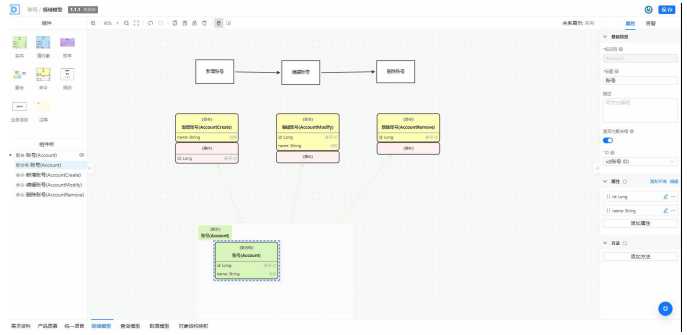
适用人员：开发人员
适用场景：在代码写完之后，通过大模型自动生成单元测试



场景5

DDD知识辅助

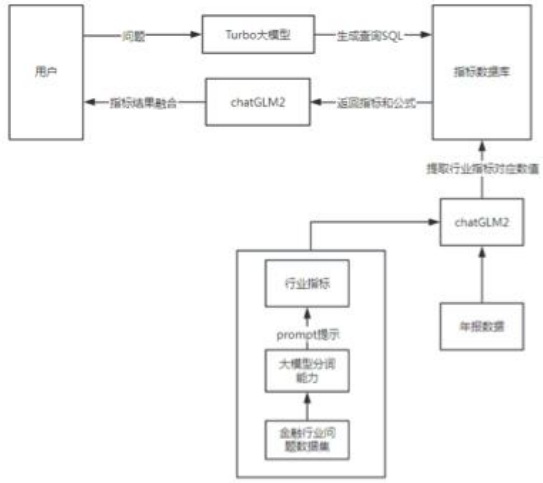
适用人员：架构师、产品、开发测试
适用场景：在DDD领域知识方面，辅助引导正确构建领域模型



场景6

数值指标精准回答

适用人员：业务
适用场景：基于企业自有数据，确保大模型回答指标数据的准确性



最后的话
AGI即将发生 坚守长期主义

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

THANKS