对大模型训练数据安全治理的思考

王峥

阿里研究院
数据经济研究中心

目录
CONTENTS

01 大模型训练所需的数据类型

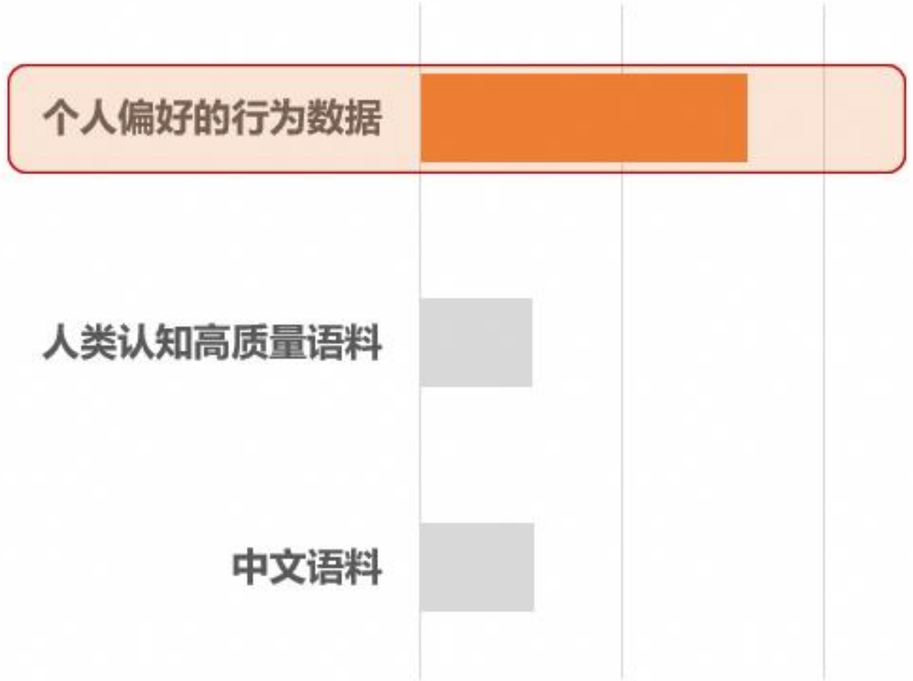
大模型训练需要哪些数据?

	训练阶段			
	基础模型			行业模型
	1、预训练	2、监督微调	3、强化学习 (RLHF)	
需求数据	世界海量知识	人类认知	人类认知	领域知识
数据内容	• 互联网多年沉淀 • 各类公开网页 • 书籍期刊 • 百科 • **代码** • 专业问答	• 人类编写的问答示例 *问：什么是大模型?* *答：大模型(Large Language Model)是一种大规模的自然语言处理模型，具有以下特征:* *1、参数数量巨大.....*	• 人类对模型答案打分排序 *问：什么是大模型?* 答案1 👍👍👍 答案2 👍 答案3 👍👍 答案4 -	• 行业积累的行业经验和专业知识 ***法律**：法律法规、裁判文书、案例分析、仲裁文书、法学论文等* ***医疗**：包括药品说明书、诊断报告、医学论文等.....*
	"广"	"齐"		"专"

必须澄清的误解：模型训练并不依赖个人信息

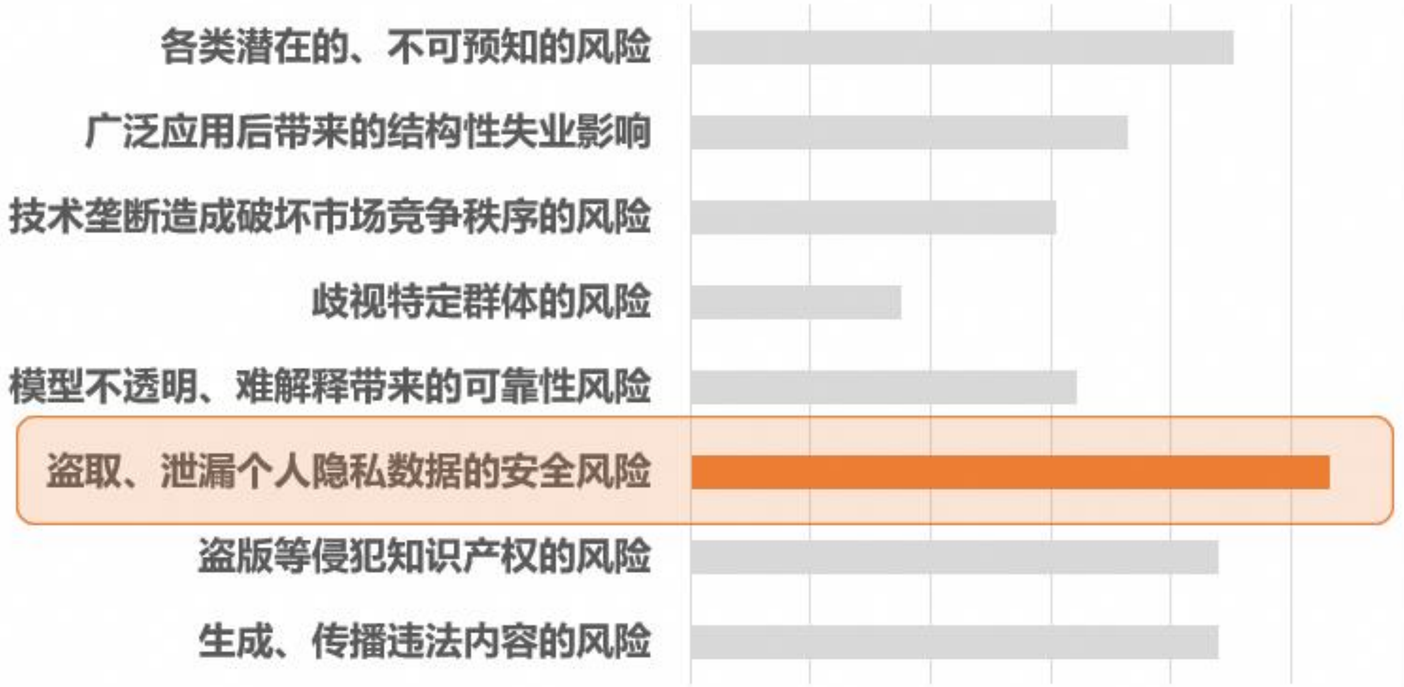
误认为大模型训练需要个人数据...

您认为影响我国大模型发展是因为缺乏什么数据?



...也误认为大模型会侵犯隐私

您认为生成式人工智能给人类社会带来哪些风险? (n=455)



02

训练语料的安全机制

各时代人工智能风险的演进和迭代：AI时代的风险识别

IT时代的风险
强调技术系统的安全

DT时代的风险
管控AI分配决策的合理性

AI时代的风险
引导AI创造知识、对齐价值

对齐价值：内容安全、知识产权、隐私安全

补充能力：去个性化，合成数据风险

控生态：开源治理、用户管控、国际互操作

强调限制：个信保护、算法黑箱、公平公正

护平台：全周期安全、抗攻击能力

保企业：数据安全、系统安全

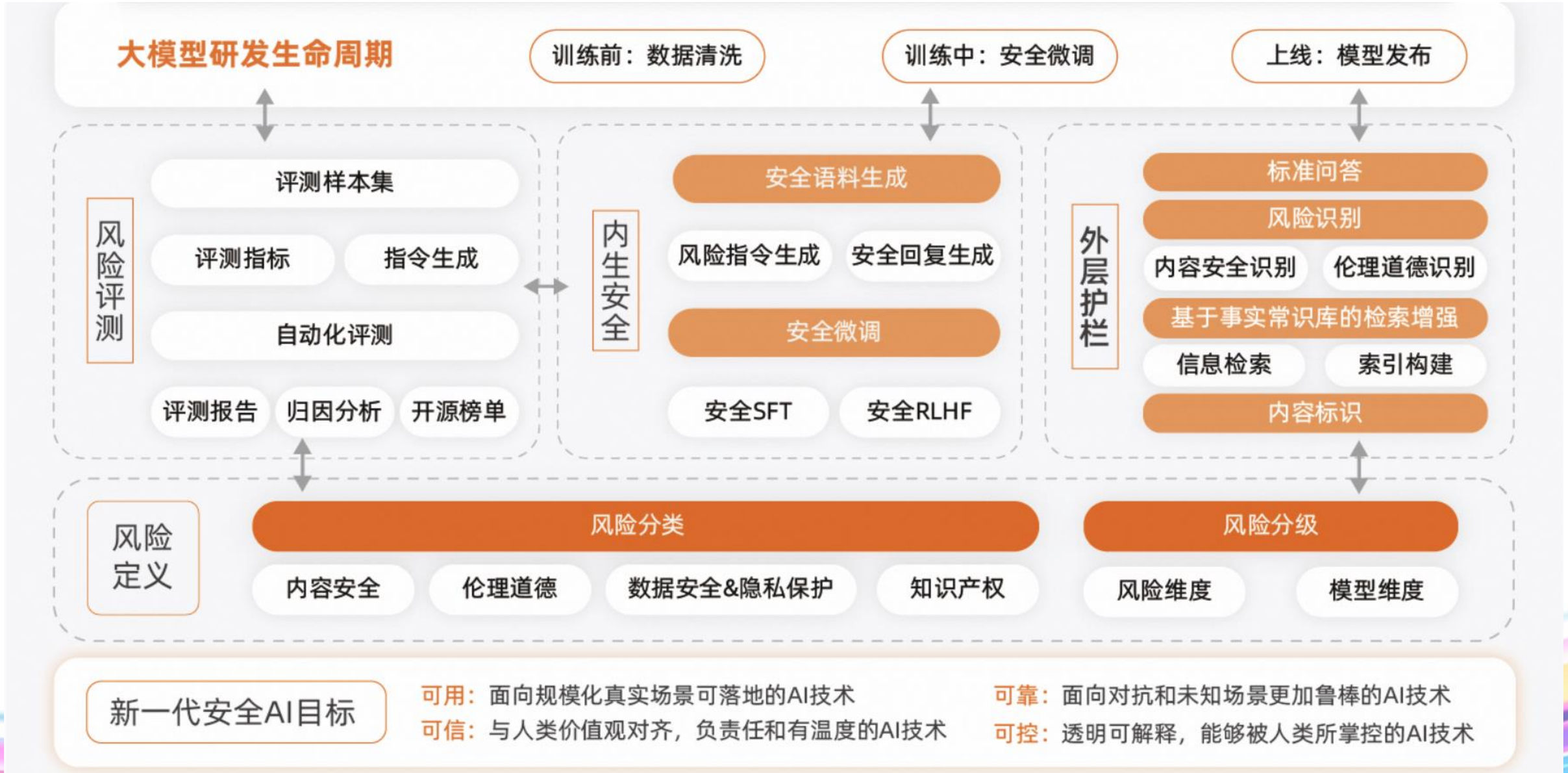
精神文明

物质文明

技术文明

security → safety

安全是大模型的核心竞争力，阿里巴巴治理的思路与核心技术



03 对数据安全治理新模式的思考

从模型视角看训练数据的安全治理

训练阶段对数据的使用特点：

- **个人信息**：模型训练阶段不依赖个信，对公开个信属于合理使用
- **版权数据**：对版权类语料属于转换性使用，是为了掌握客观规律，构建模型的基础能力，并不是复制式拷贝，属于**合理使用**

治理思路的变迁：

- **重视数据的可及性**：输入端的前置使用限制 → 输出端的管控和事后救济
- **提升数据的供给**：鼓励安全类数据集的开放共享
- **新技术的应用提升安全性**：如合成数据对个信的保护

谢 谢