



大模型 训练数据 白皮书

LARGE LANGUAGE
MODEL'S TRAINING DATA
WHITE PAPER

大模型训练数据白皮书

专家委员会

安筱鹏 阿里云智能副总裁

袁 媛 阿里研究院执行副院长

宋志刚 数字中国研究院（福建）院长

编写组主要成员

傅宏宇 王 峥 赵志远 郑达真

张 荣 陈岳峰 李天宇 徐 强

编写单位

阿里巴巴集团

数字中国研究院（福建）

阿里云智能集团



欢迎关注“阿里研究院”公众号



AliResearch
阿里研究院

关于我们

阿里研究院是阿里巴巴集团的内设智库机构，多年来扎根于阿里巴巴丰富的数字科技商业生态，依托前沿的产业实践和大量的创新案例，围绕集团“用户为先，AI 驱动”的战略重心，聚焦于科技创新、数据和算法治理等领域的研究。

联系我们

aliresearch2023@service.alibaba.com

目录

CONTENTS

01	训练数据对大模型发展的重要性	02
02	模型训练所需的数据类型	03
	2.1 训练大语言模型的数据	03
	2.2 训练多模态模型的数据	04
	2.3 训练数据的常见疑问和误解	04
	2.3.1 大模型训练并不依赖用户个人信息	04
	2.3.2 中文语料短缺不是制约我国大模型发展的重要因素	05
03	科学理解高质量数据的含义与作用	06
	3.1 高质量数据的重要性	06
	3.2 高质量数据的标准	07
	3.2.1 高质量数据类型的三重不确定性	07
	3.2.2 同类数据的评估标准并不完全一致	08

04

合成数据作为解决训练数据供给不足的新方案 09

- 4.1 训练数据供给不足带来的思考 09
- 4.2 合成数据的定义 10
- 4.3 合成数据的必要性 10
- 4.4 合成数据的生成方法及分类 11
- 4.5 合成数据在模型训练中的作用 12
 - 4.5.1 预训练语料的新物种 12
 - 4.5.2 提升对齐语料获取效率的加速器 13
- 4.6 解决训练数据供给不足的新方案 14
- 4.7 在发展中治理的合成数据 16

05

对大模型训练数据治理的思考 17

- 5.1 大模型对训练数据的使用特点 17
- 5.2 大模型训练数据合规的治理之智 18

06

政府与社会力量协同的训练数据生态 19

- 6.1 美国的现状 19
- 6.2 中国的现状 21

07

阿里巴巴集团在大模型训练与应用的探索 23

08

以更开放和务实的方式解决高质量训练数据供给 24

