Gartner
Peer Community™

用于合成数据的生成式AI

当前，随着ChatGPT的发布，生成式AI成为了万众瞩目的焦点，但其早已应用于合成数据并对数据分析（D&A）领域做出了重大的贡献。生成式AI可以填补现实世界数据源的空白，甚至优化模型结果。那么，数据分析专业人士目前是如何使用合成数据的？他们又面临着哪些挑战？

一分钟洞察：

- 企业机构之所以采用AI生成的合成数据，是因为现实世界数据存在可访问性、复杂性和可用性等方面的挑战
- 部分合成的数据最为常见，基于文本合成的数据是最常用的合成数据类型
- 领导者已经了解到，合成数据能够提高模型的准确性和效率
- 现实世界缺少源数据，往往会给合成数据带来挑战
- 为确保合成数据的质量，大多数企业机构已经实施了最佳实践

关于热点话题的一分钟洞察仅面向Gartner Peer Community成员提供。注册会员后即可立即获得超100多个话题的洞察，且每周均可获得全新洞察。

数据收集时间：2023年4月1日至14日

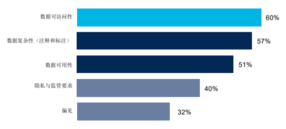
受访者：150名与AI生成型合成数据有关的IT和数据分析领导者。

现实世界数据存在可访问性、复杂性和可用性方面的挑战，促使企业机构采用AI生成的合成数据

大多数参与调研的**IT和数据分析领导者**表示，其企业机构之所以**采用AI生成的合成数据**，是因为现实世界数据存在可访问性（**60%**）、复杂性（**57%**）和可用性（**51%**）方面的挑战。

3%的受访者表示，其企业机构在**现实世界数据方面没有面临任何挑战**。

贵企业机构是因为现实世界数据方面的哪些挑战，才决定采用AI生成型合成数据的？请选择所有适用项。



非平衡数据**31%** | 我们在现实世界数据方面没有遇到过任何挑战**3%** | 其他**0%**

n = 150

"模型必须不断被训练，合成数据对我们的帮助很大。"

——专业服务行业的C级高管，员工<1,000名

"这是AI真正有所助益的一个领域。"

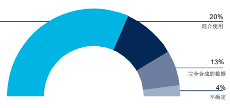
——消费品行业副总裁，员工10,000名以上

问题：您如何看待AI生成的合成数据？

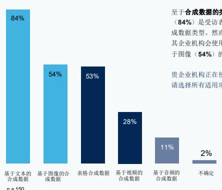
与完全合成的数据相比，部分合成的数据最为常见，基于文本合成的数据是最常用的合成数据类型

大多数受访者表示，其企业机构使用部分合成的数据（**63%**）或混合使用部分合成的数据与完全合成的数据（**20%**）。

贵企业机构是否使用完全或部分合成的数据？



n = 150



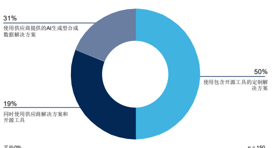
n = 150

至于**合成数据的类型**，基于文本合成的数据（**84%**）是受访者所在企业机构最常使用的合成数据类型。然而，**超过一半**的受访者表示，其企业机构会使用表格（**53%**）合成数据或基于图像（**54%**）的合成数据。

贵企业机构正在使用哪种类型的合成数据？请选择所有适用项。

50% 的受访者表示，其企业机构使用包含**开源工具**的定制解决方案来**生成合成数据**，而**31%**的受访者使用**供应商解决方案**来生成合成数据。

贵企业机构是如何生成合成数据的？



其他0%

n = 150

"用于合成数据的生成式AI仍处于早期阶段，要在整个企业机构内采用该技术十分困难，而且投资回报率也无法计算。监管问题不容忽视。"

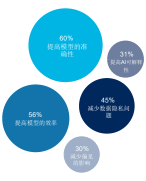
——金融行业C级高管，员工10,000名以上

"生成式AI（技术）具有高度的短视偏差，为数据选择合适的供应商仍然是一个挑战。"

——金融行业经理，员工1,000-5,000名

问题：您如何看待AI生成的合成数据？

合成数据能够提高模型的准确性和效率，但许多企业机构在使用合成数据时都面临着缺乏现实世界源数据或现实世界源数据质量低下的挑战



在受访者的企业机构中，合成数据最常见的优势是提高模型的准确性（**60%**），提高模型的效率（**56%**）和减少数据隐私问题（**45%**）。

合成数据对贵企业机构有何益处？请选择所有适用项。

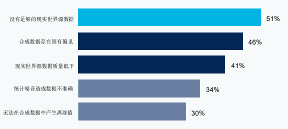
提高数据团队的效率25% |
重新平衡数据集23% |
降低数据泄露风险19% |
减少过拟合14% |
上述选项均不适用3% | 其他0%

n = 150

约有一半（**51%**）的受访者表示，其企业机构使用合成数据时面临着**现实世界缺乏源数据**的问题。超过三分之一的受访者遇到过合成数据存在固有偏见（**46%**）、现实世界源数据质量低下（**41%**）或统计噪音导致数据不准确（**34%**）等挑战。

只有**2%**的受访者在其企业机构中**没有经历过任何**合成数据的**挑战**。

贵企业机构在使用合成数据时遇到过哪些挑战？请选择所有适用项。



缺乏专业知识**25%** | 资源不足**24%** | 准确度下降**23%** |
与现有数据系统的整合**14%** | 算力成本**12%** |
选择合适的供应商**11%** | 确定适当的效用指标**11%** |
法律/道德问题（例如，重识别风险）**10%** |
我们在合成数据方面没有遇到过任何挑战**2%** | 其他**0%**

n = 150

"提高医疗数据准确性的同时减少偏见是十分困难的。到目前为止，唯一的方法是将现实世界数据进行标记化处理，从而在确保数据的准确性和质量的同时减少风险。"

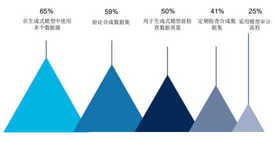
——金融行业总监，员工10,000名以上

问题：您如何看待AI生成的合成数据？

为确保合成数据的质量，大多数企业机构已经实施了最佳实践

65% 的受访者在**生成式模型中使用多个数据源**，以确保获得**高质量**的合成数据。验证合成数据集（**59%**）和用于生成式模型前检查数据质量（**50%**）也是受访者中常见的最佳实践。

为确保获得高质量的合成数据，您实施了哪些最佳实践？请选择所有适用项。



评估合成数据质量**24%** |
我们还没有实施任何最佳实践**7%** | 其他**0%**

n = 150

"AI生成的合成数据相当敏感，需要以安全的方式进行处理。"

——教育行业经理，员工10,000名以上

"AI生成的合成数据具备潜在的优势，但使用这类数据必须考虑道德方面的问题以及准确性和实用性的限制。"

——金融行业经理，员工5,000 - 10,000名

"必须将人力资源洞察与AI生成的合成数据结合起来，最大化提高数据的效率。"

——专业服务行业经理，员工5,000 - 10,000名

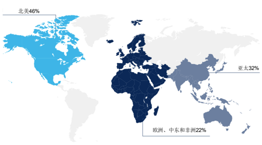
问题：您如何看待AI生成的合成数据？

想要从同行那里获取更多类似的洞察？

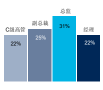
点击此处，发挥同行的力量，建立联系并推动对话，从我们强大的同行社区获得可执行的专业建议。

受访者分类

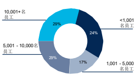
按地区划分



按职级划分



按公司规模划分



注：由于四舍五入，百分比总和可能不等于100%

Gartner

Gartner IT 领导者　　成为Gartner客户