

行业研报

人工智能行业系列（一）：智能数据标注技术三大实现路径

Industry Research Report: Artificial Intelligence Industry Series (1) - Three Realization Paths of Intelligent Data Labeling Technology

業界調査レポート：人工知能業界シリーズ（1）——インテリジェントデータアノテーションテクノロジーの三つの実現パス

报告标签：人工智能、数据标注、数据采集、监督学习

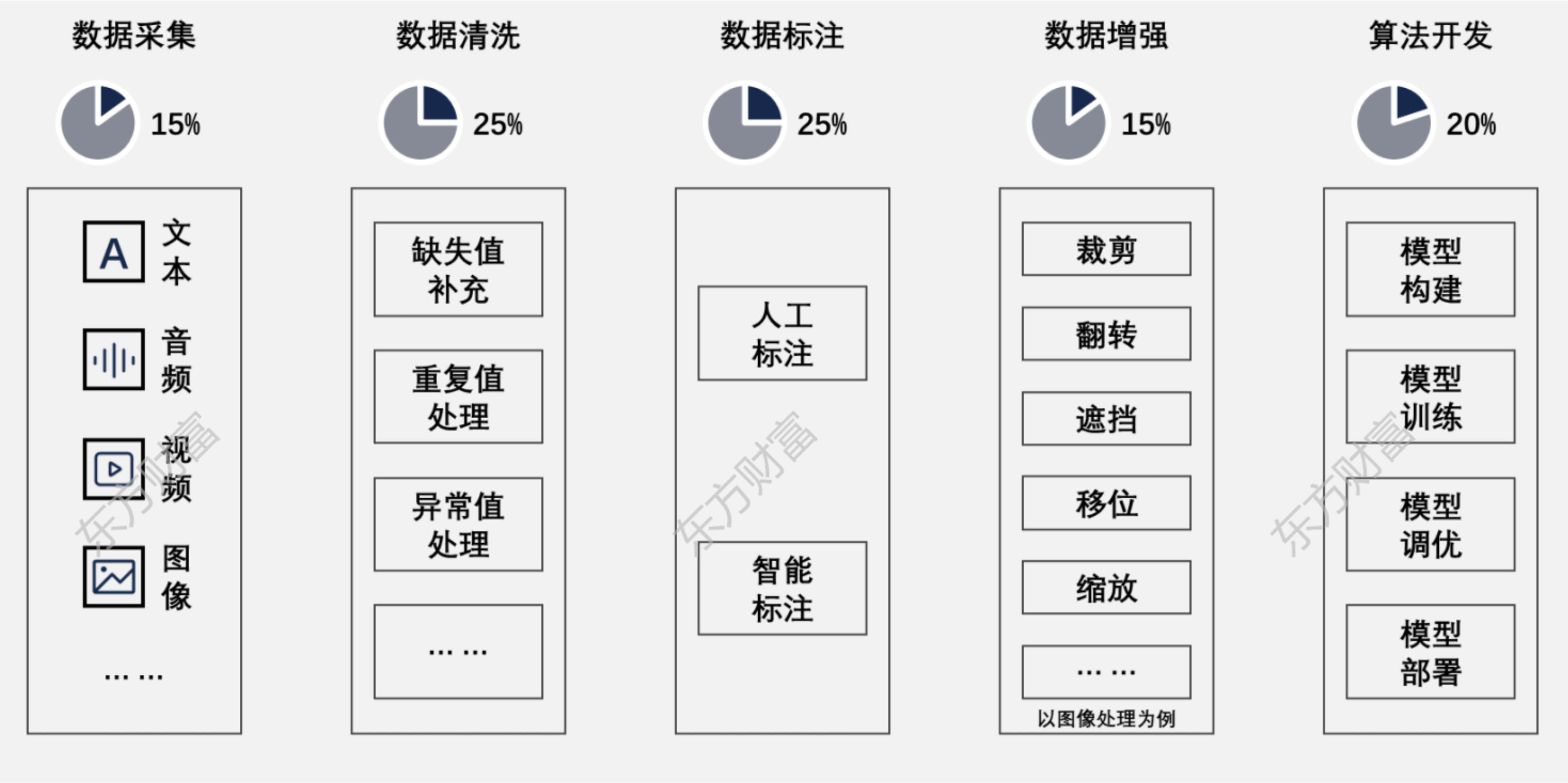
主笔人：谢子博

Q1：数据标注是什么？为什么业界需要推进数据标注的智能化？

■ 数据标注是人工智能项目开发的核心工作流程之一，成本压力推动智能数据标注技术发展

人工智能项目的开发可分为以数据采集、清洗、标注、增强等流程为主体的数据准备工作，以模型构建、训练、调优、部署为主体的算法开发工作。其中，由于数据准备工作可用的自动化辅助工具较为匮乏，数据准备工作对人力劳动的依赖程度较高，整体所需的工作时间占人工智能项目开发流程的80%。数据标注工作指对模型训练所需的数据样本中的特征要素进行标识与区分。由于目前人工智能技术的发展仍处于监督学习阶段，以深度学习为代表的AI算法模型在训练过程中对数据内涵信息及数据之间逻辑的学习及验证基于数据的特征标识实现，数据的标注工作具有必要性，是数据准备乃至人工智能项目开发的核心工作之一。与数据准备其余工作流程类似，数据的标注工作高度依赖人力劳动。冗长的工作周期及庞大的人力成本成为制约人工智能行业发展的主要因素之一。数据标注服务供给侧的痛点催生市场对自动化工具的需求，推动智能数据标注技术的发展与规模化应用。

图表1：人工智能应用开发流程及时间占比



来源：头豹研究院

Q2：人工智能数据集服务行业目前较为成熟的商业模式有哪些？

■ 数据集服务行业商业模式逐渐固化，定制化服务市场需求占比较高

数据标注行业较为成熟的商业模式可分为定制化服务及标准化服务。定制化服务根据客户需求可分为全定制化服务及标注定制化服务。全定制化服务指客户仅针对AI项目开发需求提出数据集需求，由数据厂商完成从数据集构建、数据采集到数据处理、质检等多个流程。标注定制化服务指厂商对客户提供的未标注数据集进行标注。标准化服务包括数据集服务与智能标注服务。数据集服务指厂商根据市场需求，自主生产具备较高通用性的数据集，并统一定价出售。智能标注服务搭载于AI开发平台，为AI开发者提供基于平台云计算能力的智能标注服务。

从市场需求层面来看，相对于定制化服务，标准化服务的市场需求较小。造成该需求差异的主要原因来自于AI开发的重心逐渐由数据通用性较高的互联网等领域逐渐倚向数据通用性较低的新兴领域或专业领域。从短期来看，定制化服务的市场需求占比将保持持续增长的趋势。

图表2：人工智能行业数据服务商业模式



来源：头豹研究院

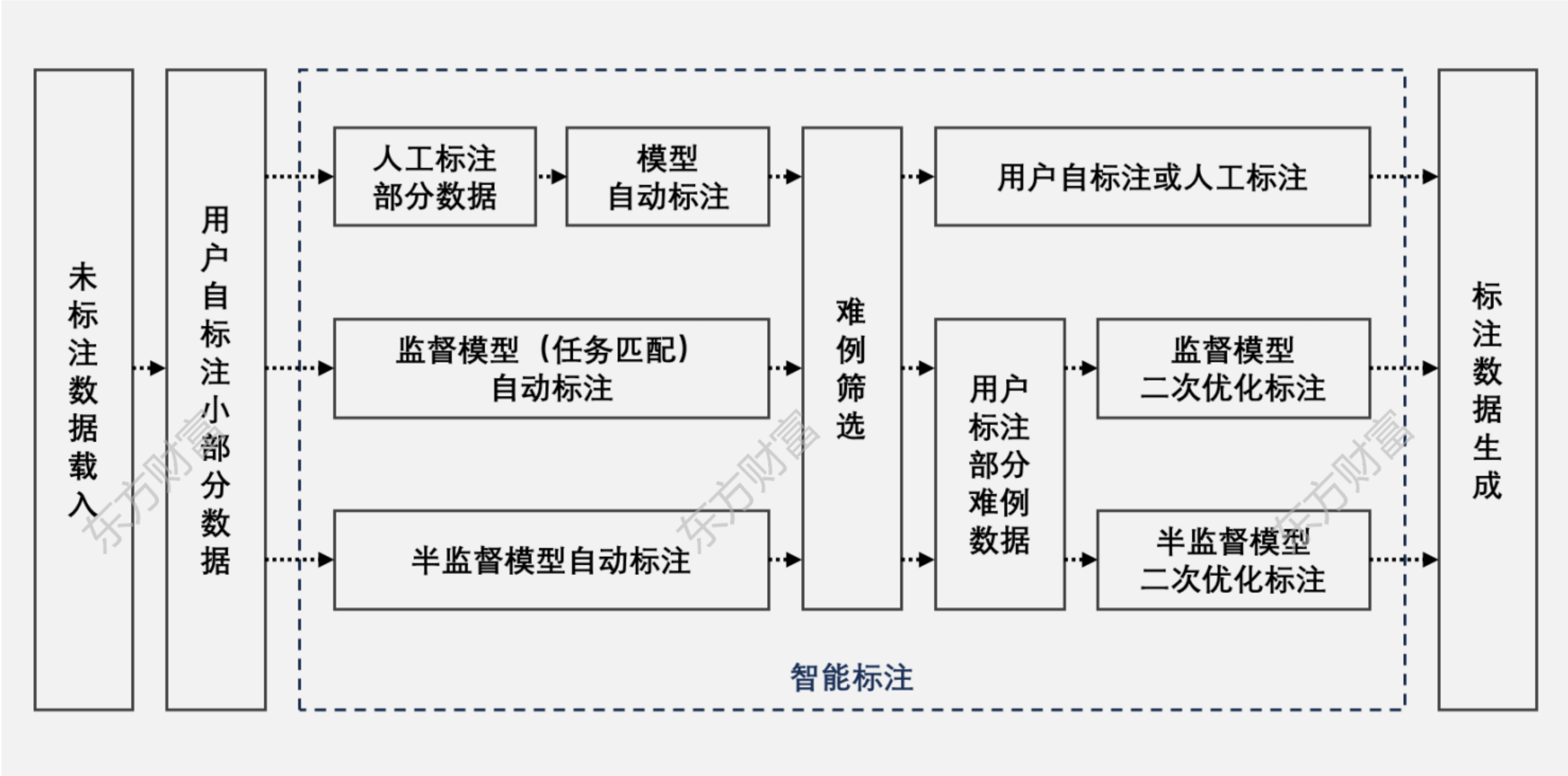
Q4：目前用实现智能数据标注的主要技术路径有哪些？

■ 智能数据标注服务基于成熟AI算法模型实现，尚未实现对人工标注的完全剥离

智能数据标注指利用成熟的AI算法模型对数据进行自动化标注，从而降低人工标注在数据标注服务中的占比。智能数据标注实现路径的差异主要来自于对选用模型的区别及人工标注参与的占比：

- 首次标注：智能标注服务用户对部分未标注数据进行标注，以其作为智能标注工作的指引。在需求方完成部分标注工作后，智能标注服务可采用进一步人工标注，增加智能标注样本参考量以提高AI算法模型标注准确率；或是直接采用监督模型或半监督模型对剩余未标注数据进行标注。AI算法模型在完成第一次标注的过程中会对难以标注的数据样本进行筛选；
- 二次标注：难例数据样本交由用户或标注团队完成剩余标注工作，或进行人工优化再交由AI算法模型进行二次智能标注。该过程将循环数次，直到数据的标注满足用户需求。

图表3：智能标注服务实现路径



来源：头豹研究院

Q4：从短期来看，人工智能技术的发展将为数据集服务行业带来怎样的机遇与挑战

■ 小样本学习

小样本学习（Few-shot Learning）指模仿人类认知机制，通过对少量实例即可构建对新概念的认知，实现人工智能技术应用的机器学习技术。小样本学习通过对原始数据进行分解等多种手段，实现在数据样本量较小的场景中机器学习的应用。相较于传统的机器学习，由于小样本学习对训练数据的需求量较小，应用小样本学习的人工智能开发项目对包含数据标注在内的数据准备服务需求降低。在单个项目对数据需求量的减少的同时，小样本学习推动垂直应用领域人工智能的规模化应用。**从总体上看，小样本学习对数据集服务市场需求的增长具有正向作用。**

■ 预训练模型

预训练模型（Pre-trained Models）指从大规模数据中提取并学习数据特征信息，获取与具体任务目标不存在显著关系的算法模型。预训练模型具有模型层数复杂、模型参数量大等特点，在训练过程中需使用的数据样本较多，数据标注需求较高。

预训练模型通过对数据、特征、模型、参数等的迁移，在完成模型开发与训练的同时降低模型对数据样本、硬件算力的使用。与小样本学习类似，预训练模型的应用在降低单个项目开发对数据标注需求的同时，推动人工智能技术在新兴领域及专业领域的规模化应用。**从总体上看，预训练模型对数据集服务市场需求的增长具有积极意义。**

■ 半监督学习与无监督学习

人工智能技术的发展目前仍处于监督学习阶段，以深度学习为基础的人工智能项目的开发与落地对标注数据依赖程度较高。半监督学习与无监督学习可从未标注数据中发展内涵的信息及逻辑，实现计算机对未标注数据的自学习。理论上，半监督学习与无监督学习技术可切实降低对标识数据的依赖，从而降低对数据标识服务的市场需求。然而，现阶段半监督学习仍处于初步应用阶段，无监督学习仍处于理论阶段。**从短期来看，人工智能技术的实现仍将以监督学习技术为主。**随着人工智能技术应用的进一步深化，包括数据标注服务在内的数据集服务市场需求仍将保持高速增长，数据集服务行业将迈入高速发展阶段。

来源：头豹研究院

方法论

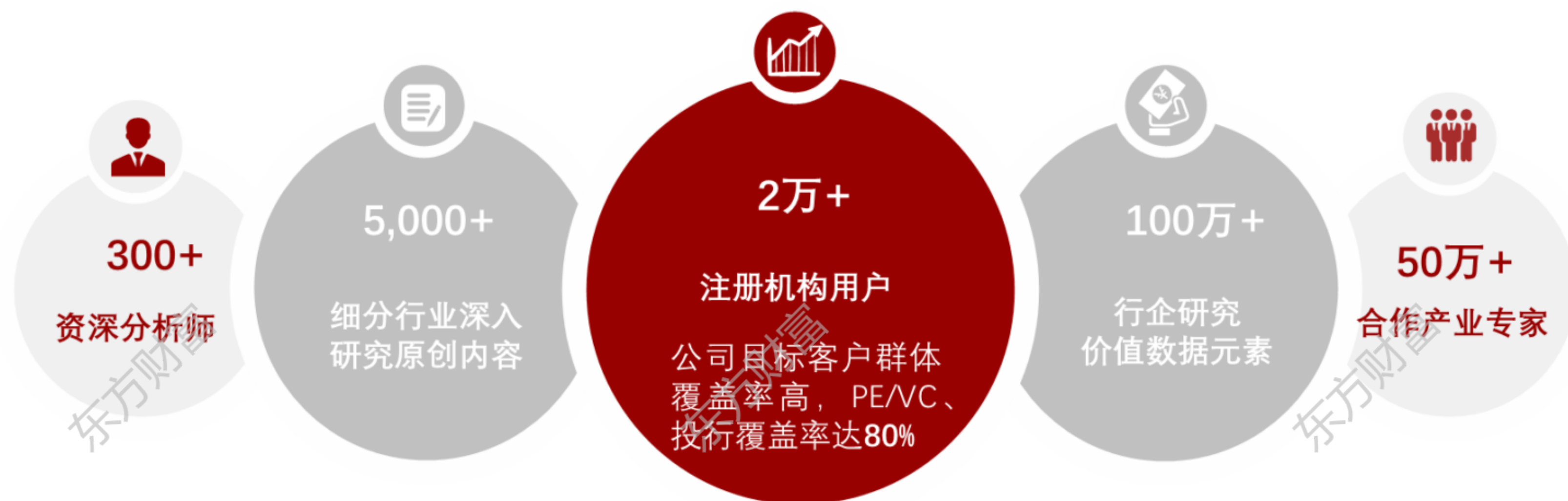
- ◆ 头豹研究院布局中国市场，深入研究10大行业，54个垂直行业的市场变化，已经积累了近50万行业研究样本，完成近10,000多个独立的研究咨询项目。
- ◆ 研究院依托中国活跃的经济环境，从人工智能、数据标注、智能标注等领域着手，研究内容覆盖整个行业的发展周期，伴随着行业中企业的创立，发展，扩张，到企业走向上市及上市后的成熟期，研究院的各行业研究员探索和评估行业中多变的产业模式，企业的商业模式和运营模式，以专业的视野解读行业的沿革。
- ◆ 研究院融合传统与新型的研究方法，采用自主研发的算法，结合行业交叉的大数据，以多元化的调研方法，挖掘定量数据背后的逻辑，分析定性内容背后的观点，客观和真实地阐述行业的现状，前瞻性地预测行业未来的发展趋势，在研究院的每一份研究报告中，完整地呈现行业的过去，现在和未来。
- ◆ 研究院密切关注行业发展最新动向，报告内容及数据会随着行业发展、技术革新、竞争格局变化、政策法规颁布、市场调研深入，保持不断更新与优化。
- ◆ 研究院秉承匠心研究，砥砺前行的宗旨，从战略的角度分析行业，从执行的层面阅读行业，为每一个行业的报告阅读者提供值得品鉴的研究报告。

法律声明

- ◆ 本报告著作权归头豹所有，未经书面许可，任何机构或个人不得以任何形式翻版、复刻、发表或引用。若征得头豹同意进行引用、刊发的，需在允许的范围内使用，并注明出处为“头豹研究院”，且不得对本报告进行任何有悖原意的引用、删节或修改。
- ◆ 本报告分析师具有专业研究能力，保证报告数据均来自合法合规渠道，观点产出及数据分析基于分析师对行业的客观理解，本报告不受任何第三方授意或影响。
- ◆ 本报告所涉及的观点或信息仅供参考，不构成任何证券或基金投资建议。本报告仅在相关法律许可的情况下发放，并仅为提供信息而发放，概不构成任何广告或证券研究报告。在法律许可的情况下，头豹可能会为报告中提及的企业提供或争取提供投融资或咨询等相关服务。
- ◆ 本报告的部分信息来源于公开资料，头豹对该等信息的准确性、完整性或可靠性不做任何保证。本报告所载的资料、意见及推测仅反映头豹于发布本报告当日的判断，过往报告中的描述不应作为日后的表现依据。在不同时期，头豹可发出与本报告所载资料、意见及推测不一致的报告或文章。头豹均不保证本报告所含信息保持在最新状态。同时，头豹对本报告所含信息可在不发出通知的情形下做出修改，读者应当自行关注相应的更新或修改。任何机构或个人应对其利用本报告的数据、分析、研究、部分或者全部内容所进行的一切活动负责并承担该等活动所导致的任何损失或伤害。

头豹研究院简介

- ◆ 头豹是中国领先的原创行企研究内容平台和新型企业服务提供商。围绕“协助企业加速资本价值的挖掘、提升、传播”这一核心目标，头豹打造了一系列产品及解决方案，包括：**报告/数据库服务、行企研报服务、微估值及微尽调自动化产品、财务顾问服务、PR及IR服务**，以及其他企业为基础，利用大数据、区块链和人工智能等技术，围绕产业焦点、热点问题，基于丰富案例和海量数据，通过开放合作的增长咨询服务等
- ◆ 头豹致力于以优质商业资源共享研究平台，汇集各界智慧，推动产业健康、有序、可持续发展



四大核心服务

研究咨询服务

为企业定制**定制化报告服务、管理咨询、战略调整**等服务

行业排名、展会宣传

行业峰会策划、**奖项评选**、行业白皮书等服务

企业价值增长服务

为处于不同发展阶段的企业，提供与之推广需求相对应的“**内容+渠道投放**”一站式服务

园区规划、产业规划

地方**产业规划**，园区企业孵化服务

报告阅读渠道

头豹官网 —— www.leadleo.com 阅读更多报告

头豹小程序 —— 微信小程序搜索“头豹”、手机扫上方二维码阅读研报



添加右侧头豹分析师微信，身份认证后邀您进入行研报告分享交流微信群



详情咨询



客服电话

400-072-5588



上海

王先生： 13611634866

李女士： 13061967127



深圳

李先生： 18916233114

李女士： 18049912451



南京

杨先生： 13120628075

唐先生： 18014813521

头豹 Project Navigator 领航者计划介绍



备注：活动解释权均归头豹所有，活动细则将根据实际情况作出调整。

头豹 Project Navigator 领航者计划与商业服务

- 头豹以**研报服务**为切入点，根据企业不同发展阶段的资本价值需求，以**传播服务**、**FA服务**、**资源对接**、**IPO服务**、**市值管理**为基础，提供适合的**商业管家服务解决方案**



扫描上方二维码
联系客服报名加入

备注：活动解释权均归头豹所有，活动细则将根据实际情况作出调整。

读完报告有问题？

快，问头豹！你的智能随身专家



扫描二维码
即刻联系你的智能随身专家

千元预算的
高效率轻咨询服务



STEP03 解答方案生成

大数据×定制调研
迅速生成解答方案

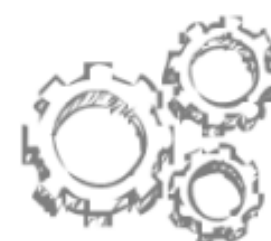


STEP01 智能拆解提问

人工智能NLP技术
精准拆解用户提问

STEP04 专业高效解答

书面反馈、分析师专访、
专家专访等多元化反馈方式



STEP02 云研究院后援

云研究院7×24待命
随时评估解答方案

