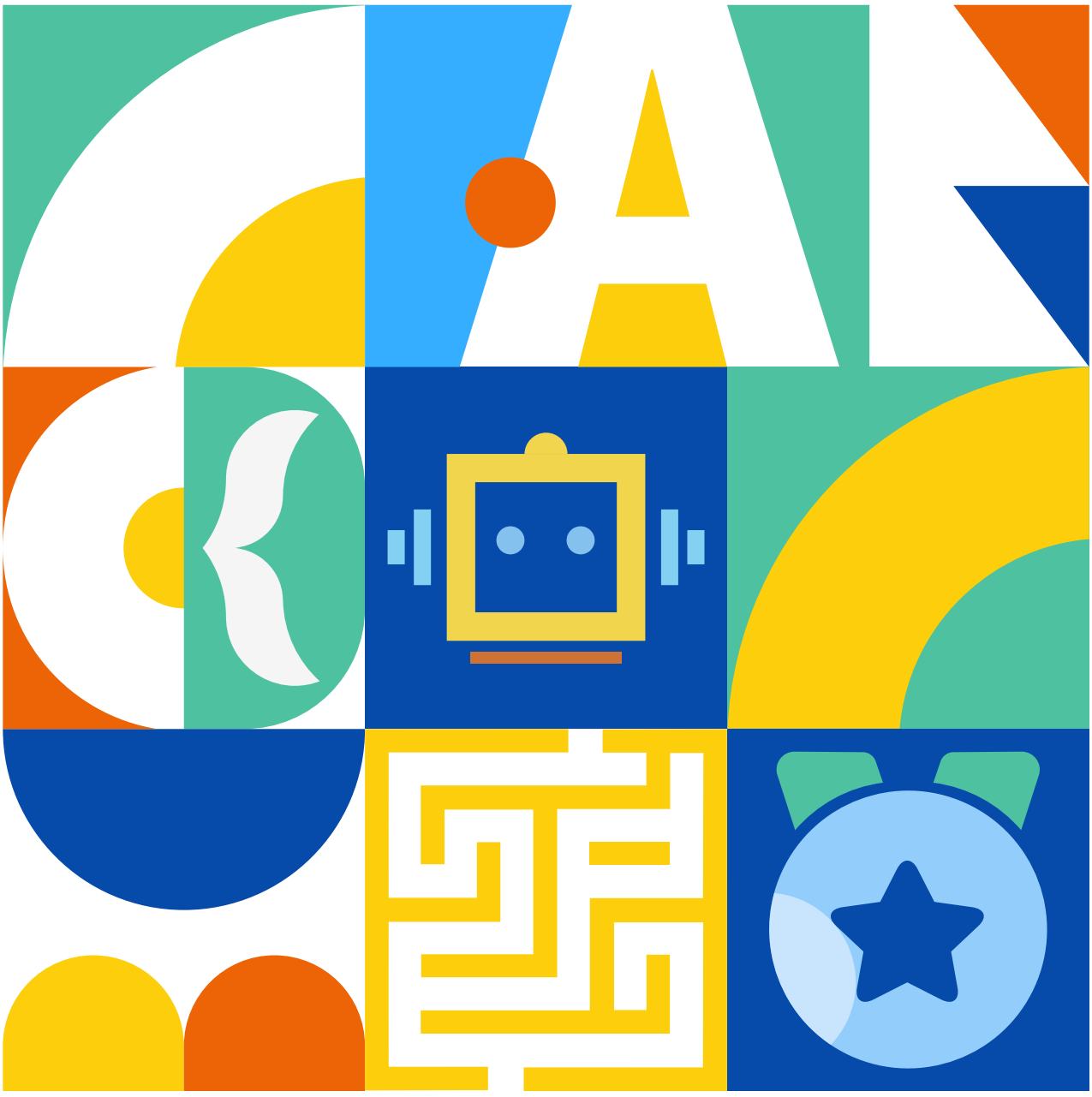
White Paper of
2021 AI Competition

2021人工智能竞赛白皮书

1000场竞赛的深度分析



中国人工智能产业发展联盟
Artificial Intelligence Industry Alliance

Heywhale 和鲸
数 据 科 学 协 同 平 台

CAICT
中国信息通信研究院

北京郵電大學
Beijing University of Posts and Telecommunications
人工智能学院

Datawhale

目录
CONTENTS

ARTIFICIAL INTELLIGENCE

前言

数字科技正逐渐渗透到经济、商业、社会生活方式等方方面面，人们对于数据智能时代的来临充满了期待。然而，从现阶段的发展来看，我们距离真正的大数据时代仍有距离，各行各业的数字化转型进程未能如期落地。数据的力量究竟该如何激发？带着这样的疑惑，我们把视角对准了人工智能竞赛这一载体，以期为行业提供启发。通过全面的调研、分析，我们得到了寻找创新路径和探索最佳实践的总结，形成这份《人工智能竞赛白皮书》分享给大家。

全球人工智能竞赛已经历了 20 余年的探索，其发展起点可追溯至 1997 年由国际顶级学术会 议 KDD（Conference on Knowledge Discovery and Data Mining，知识发现和数据挖掘会议） 推出的 KDD Cup。此后，一系列国际顶尖学术会议也陆续牵头组织人工智能竞赛。

国际顶会人工智能竞赛往往会积极兼顾市场应用的需求，通过与企业、政府机构合作，聚集当下 技术与数据应用中的难点寻求解决思路，不断缩短科研成果与落地应用之间的距离。创立于 2010 年的 Kaggle 作为目前全球最大的数据科学技术分享社区，更是为人工智能竞赛的平台化发 展奠定了模式化基础，其合作伙伴除了 Google、Facebook、Airbnb、Yelp 等互联网时代诞生的巨头，亦不乏 Walmart、Airbus、Genentech 等声名日久的传统行业领头羊。

放眼中国市场，人工智能竞赛自 2014 年萌芽以来，数量每年以翻倍之势增加，主导者不乏权威的 科研机构与高校，亦有来自各行各业的领军企业。身担城市管理与民生重任的政务机构也在“大众创业、万众创新”的政策驱动下，开始意识到数据的力量，牵头组织了多场大型人工智能竞赛。

人工智能竞赛绝不仅仅只是“主办方发布赛题、传播品牌美誉，参赛者参与开发、收获丰厚奖金” 的公关活动。对双方而言，人工智能竞赛已经成为了数字化转型的“云端实验室”，很多领先机构已经把人工智能竞赛当作是他们最佳的“数字化练兵场”、“数字化人才挖掘器”和“数字化生态构建器”。主办方紧跟数据智能的发展趋势，通过挖掘数据价值的场景和树立“数据 + 算力 + 人才 + 算法”的价值闭环，积累数据科学研发的核心能力；参赛选手能够积累如何在真实业务场景中应用先进技术的实际经验，在与众多高手的交流切磋中不断成长。

基于这一现状意义，在中国人工智能产业发展联盟（AIIA）的指导下，和鲸科技携旗下中国领先的第三方人工智能竞赛平台“和鲸社区 (Heywhale. com)”，联合中国信息通信研究院数据研究中心、北京邮电大学人工智能学院、人工智能开源社区“DataWhale.club”，共同发布《2021 人工智能竞赛白皮书》，分析全球市场人工智能竞赛发展趋势，提出对人工智能竞赛推动前沿技术落地的创新价值思考，并在业界开源优质人工智能竞赛运营方法论与最佳实践。

第一章

人工智能竞赛是数字经济创新发展的重要推动方式

一、人工智能竞赛是人工智能技术创新、应用发展的重要推动方式，受到各国政府高度重视

二、人工智能竞赛是各地人工智能产业创新发展的重要实现方式

三、人工智能竞赛为各行业数字化转型创新持续输出动能

ARTIFICIAL INTELLIGENCE

一、人工智能竞赛是人工智能技术创新、应用发展的重要推动方式，受到各国政府高度重视

在国家层面，人工智能竞赛助力数字经济创新发展，突破重点技术瓶颈，推动人工智能产业高质量发展。2017 年 7 月，国务院印发《新一代人工智能发展规划的通知》，人工智能正式上升至国家战略层面，此后各类关于加快推进数据应用的政策相继出台。可以看到，在国家宏观政策的大力推进和坚实支持下，大数据、云计算以及人工智技术的不断迭代和成熟正在催生日渐规模化、网络化和生态化的内生驱动的生态经济体，其基础都在于借助算法和算力实现对数据价值的洞察。人工智能竞赛活动是人工智能技术创新、产业应用实践、赋能数字经济改革创新的重要方法。通过组织各类人工智能竞赛活动可汇聚各方技术力量，共同研究突破关键技术瓶颈，解决人工智能技术应用落地难点，并提升公众对人工智能技术的认知。

当前，全球多个国家发布人工智能国家战略，并在战略中指明应加强人工智能竞赛、合作研讨，实现本国人工智能技术的重大突破，并通过竞赛方式促进人工智能解决方案的开发，恢复本国经济增长。部分国家政府部门牵头组织竞赛，成果解决人工智能相关的科学和工程问题。

可以看到，以人工智能竞赛作为载体，以问题为导向，以人才为抓手，汇聚人工智能领域的各界科技创新力量和智慧，在共同推动人工智能 + 产业发展，加快算法产业化进程中起到重要作用。人工智能竞赛作为人工智能赋能数字经济改革创新的重要手段和窗口，也越来越受到各国的高度重视。

二、人工智能竞赛是各地人工智能产业创新发展的重要实现方式

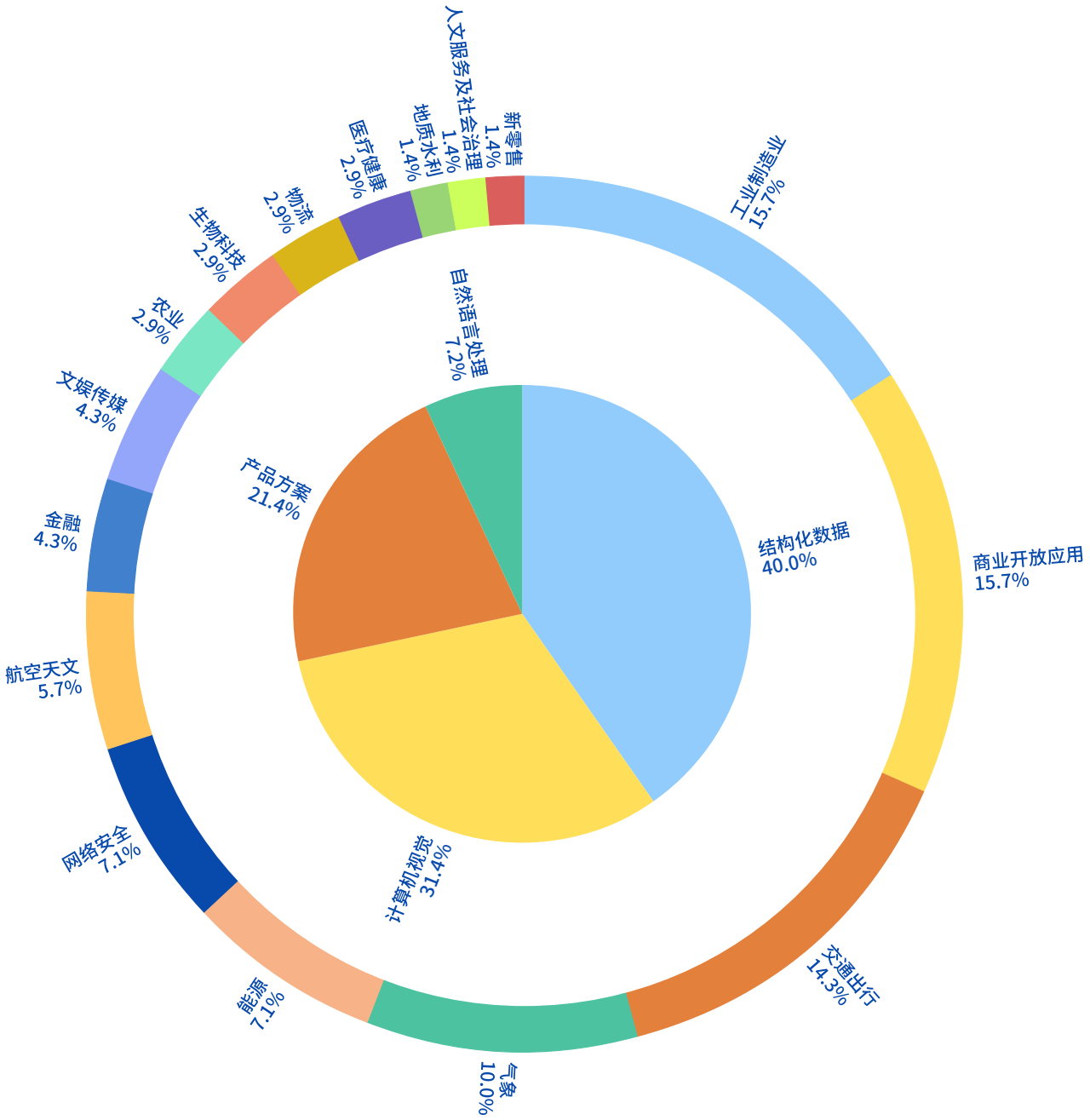
我国各省、市贯彻国家创新驱动战略，结合各地人工智能产业特色，举办各类人工智能创新、创业大赛，推动人工智能技术在各行各业的应用创新及成果转化，汇聚人工智能专业人才，促进地方人工智能产业发展。多地举办面向不同人群的人工智能竞赛，覆盖从学生教育、专业技术人员，到行业专家的不同类型竞赛活动，既促进人工智能技术的普及，又面向具有应用价值的场景，探寻技术解决方案。

助力政府实践产业落地和建设数据创新生态

在“大众创业、万众创新”的号召下，在政府资金、政策和转型决心的支持下，政务部门在大力倡导政务管理数据化、智能化的同时，也成为了有序开放优质公共数据资源、释放公共数据红利的引领者，为建设数据创新生态建设提供着政策支撑和资源支持。

如何借助数据科学建设智慧城市，是政务机构面对的重要问题。政务部门的数据积累具有数量更大、来源更多、质量更好、纬度更广等特征，其应用领域遍布城市管理的方方面面。为了探索数据科学在城市管理中的应用可能性，政务机构较为偏好开放性的产品方案赛题。具有明显的行业标签，如教育局、交通局、气象局等，赛题应用场景多与其自身行业标签一致；而没有明显行业标签的政务机构，如人民政府，其赛题场景则非常广泛。

图 1-1 政务部门类赛事主办方的赛题场景分布和技术类型分布



数据来源：和鲸科技（heywhale.com）经调研统计得出。引用请注明出处。

表 1-1 人工智能竞赛对科研机构的价值总结

	人工智能竞赛的亮点价值
前沿科研进展	融合丰富的资源支持 开辟创新的科研价值探索路径
产研交流	与工业界进行紧密合作 促进了跨学科领域产研交流
科研标杆	释放前沿科研的实用价值 树立起前沿科研成果创新应用探索的标杆

数据来源：和鲸科技（heywhale.com）经调研统计得出。引用请注明出处。

案例特写：医疗大数据 Datathon——产研协作树立医疗数据创新应用的标杆

赛事主办方：中国医院协会医学工程专业委员会 /《中国医疗设备》杂志社

2019 年第四届 PLAGH-MIT 医疗大数据 Datathon 于 11 月在北京召开。这一数据实践活动最初由麻省理工大学团队和哈佛医学院团队发起，由中国人民解放军总医院率先将该模式引入国内，旨在融合海量医疗数据物基础上，汇集医疗专家、数据科学家、统计分析师等不同知识背景和专攻技术领域的参与者，利用真实的医疗数据，通过组队协作以项目竞赛的形式解决临床医学面临的众多问题和未满足的医疗需求。

通过跨学科合作竞赛的创新模式，Datathon 激活了海量医疗大数据的实用价值，成为数字化创新趋势下高效、敏捷、低耗的练兵场——前期的训练营阶段集合了医疗、数据、计算机等不同学科背景的人士，组织进行专业知识的互通交流和学习，快速打通专业壁垒；在正式竞赛阶段，不同背景的参与者通过优势互补的组队构成了数字化创新的基础组织单元，在竞赛环境中协作挖掘医疗数据的应用潜力、探索解答临床实践中遇到的问题，为如何开展医疗大数据研究提供了新思路。

迄今为止，医疗大数据 Datathon 已在全球范围内持续组织了 40 余场，历年赛题方向广泛，涉及了影像报告的自动语意分析与信息提取、基于急诊大数据的急性高乳酸血症短期治疗效果预测评估模型构建的研究、以及消化道出血发病风险与气象条件之初步探究等丰富的内容，有效推动了医疗大数据的实践应用落地。

> Datathon 为医疗研究提供了一个突破时空、地域、文化和学科限制的创新协作平台，项目成果的开源共享和不同观点的激烈碰撞促进了数据和技术在临床医疗领域发了更广泛、更持续的创新价值。
>
> ——《科学·转化医学》期刊评论

扫描上方二维码
了解更多赛事详情

案例特写：“莱斯杯”全国第二届“军事智能机器阅读”挑战赛——军事领域的前沿科研探索

赛事主办方：中电莱斯信息系统有限公司

中电莱斯作为我国军事科研的重要力量通过人工智能竞赛积极探索 NLP（自然语言处理）的前沿发展，赛题内容将 NLP 技术应用于对真实中文文本段落的阅读理解，要求参赛选手训练的算法模型能够基于对多篇文章的内容定位和深度理解，为测试集中给定的问题找到最优答案。

人工智能竞赛将语料库数据、NLP 技术和人才加以串联，并提供了云计算资源和协作开发工具等基础设施，以阅读理解与推理问答的竞赛成果为基础，创造性地探索了 AI 技术在军事场景的应用价值，将有望颠覆以往情报整编工作以人工分析提炼为主的模式，逐步实现由机器替代繁琐、重复性情报整编业务流程。作为中国在军事领域进行 AI 创新探索的标杆案例，该人工智能竞赛也将加强中电莱斯对于中文语料库建设的信心。

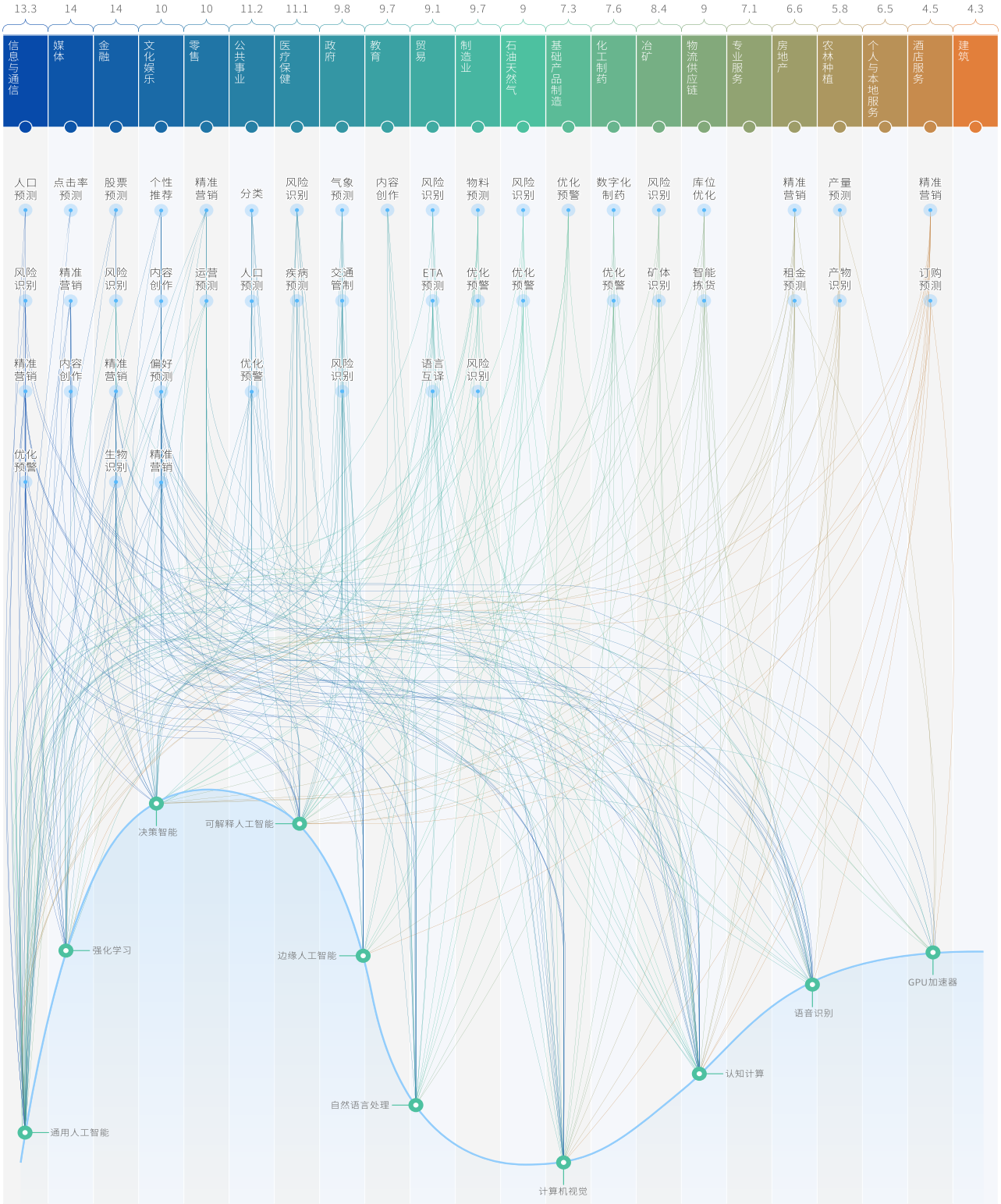
扫描上方二维码
了解更多赛事详情

三、人工智能竞赛为各行业数字化转型创新持续输出动能

在各类人工智能竞赛活动中，赛事汇聚产学研各领域专家，实现以问题为导向的产学研协同创新。通过举办选拔各行业优秀的人工智能创新应用方案，加速形成各细分领域的人工智能应用共识和标准。

图 1-2 不同发展周期的前沿技术在各行业场景中的应用分布

备注：以下数值为各行业应用 AI 前沿技术的总指数得分，从左到右依次排名



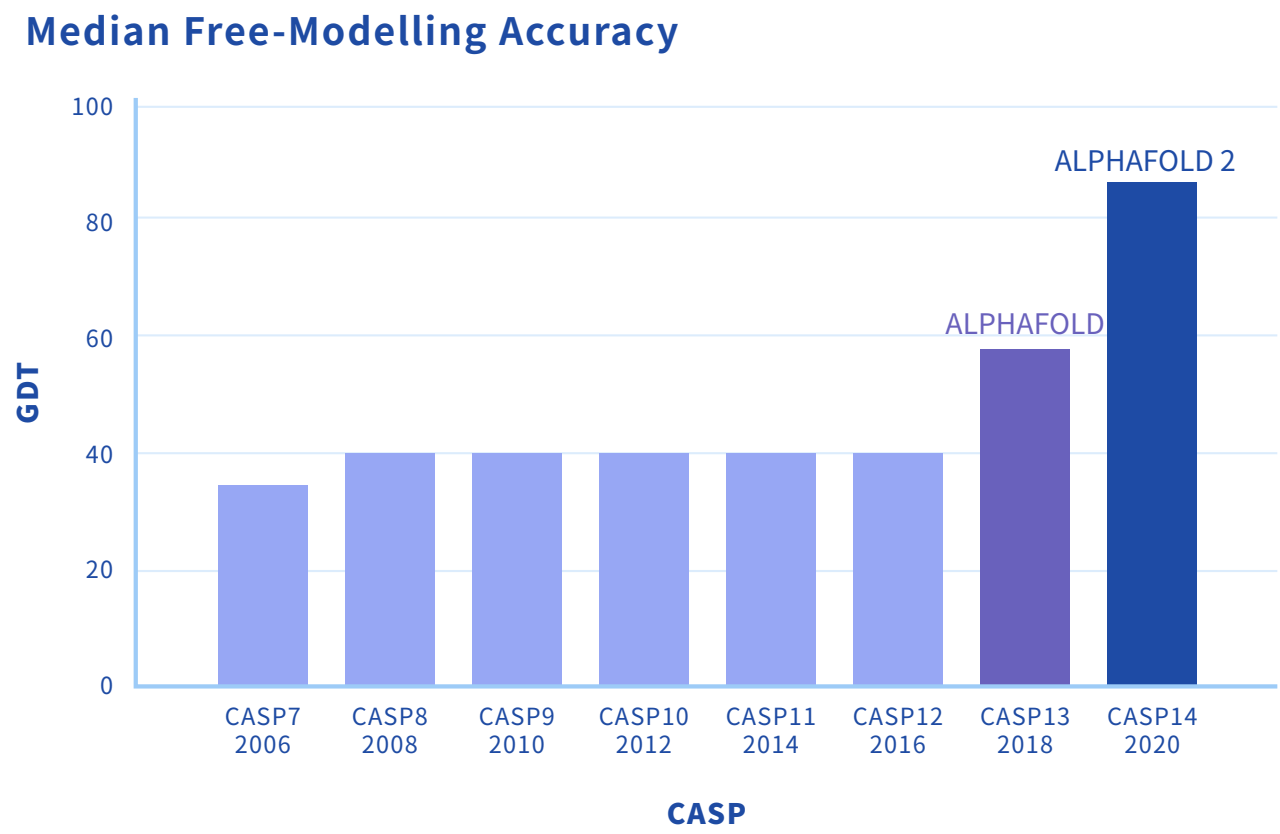
数据来源：和鲸科技（heywhale.com）制作。引用请注明出处。

伴随着技术水平的日趋成熟，越来越多的行业开始尝试融入信息技术，技术落地也成为数字化创新的首要目标。组织机构希望加入数字化转型的趋势性浪潮，然而，数字化却面临着诸如抗拒改变的文化氛围、有限的分享和协作、业务尚未准备就绪、人才短缺、现有实践不适宜进行数字化、以及成本投入巨大等多方面的挑战。

对绝大多数组织机构而言，推进数据科学项目耗时漫长、所费不赀。从集思广益提出创意，到通过审慎评估形成切实可行的方案，继而投入大量尖端人力进行算法开发，成果新鲜出炉后投入现实场 景加以应用验证，再到确定其效果与回报后推行全面部署——每 一个环节都会牵涉到大量资源的协调投放与各方利益诉求的平衡，更少不了富有预瞻性的战略判断、科学专业的运营管理、以及发现误差后能够进行敏捷的优化调整。

因此，人工智能竞赛为数据驱动的技术创新落地和数字化转型升级提供了充沛的生产力。例如，近年来知名度提升的国际蛋白质结构预测竞赛（CASP）由马里兰大学的 John Moult 教授和加州大学戴维斯分校的 Krzysztof Fidelis 教授联合创建的国际性比赛，旨在评估、促进和确认最佳的蛋白质结构预测手段。2018 年，DeepMind 开发的第一代 AlphaFold 首次参加 CASP 并且拔得头筹。而 2020 年，新一代的 AlphaFold2 在 CASP 中的表现更为惊艳，其评估结果比 2018 年提升了，被媒体誉为“解决生物学 50 年来的重大挑战”。

图 1-3 2006-2020 年 CASP 比赛中最佳蛋白折叠预测系统的评分表现



图片来源：DeepMind Blog

进而可以看出，在行业层面，人工智能竞赛在实现以问题为导向的产学研协同创新，选拔各行业优秀的人工智能创新应用方案，以及加速形成各细分领域的人工智能应用共识和标准方面都能够发挥重要作用。

第二章

人工智能竞赛是敏捷的数据探索与应用创新方式

一、数字化转型成本高昂，暗藏诸多风险

二、整合关键创新要素，快速实验创新应用

三、低耗高效可持续，生态化创新动能不断

ARTIFICIAL INTELLIGENCE

作为赛事主办方，通过人工智能竞赛可以提升数据资产的利用价值、挖掘高质量数据人才、征选优质的人工智能应用方案、以及构建自身的开放创新的数字化生态，在这些方面人工智能竞赛都有着得天独厚的优势，可以把它看作为是一种敏捷的价值探索与应用创新的手段。

一、数字化转型成本高昂，暗藏诸多风险

数字化创新需要不同的技能、工作实践、组织模式甚至是文化。要让一个为结构化的、有序的、面向流程而设计的组织，变成为生态系统、适应、学习和 实验设计的组织，是非常困难的。 ——Gartner 研究副总裁 Marcus Blosch

推进数据科学项目每一环节都面临着诸多难题

数据科学项目的推动需要经历一系列艰难环节，但凡有一环节出现迟滞或脱节，都有可能导致重大损失，或在瞬息万变的数字化时代丧失创新发展的先机。

二、整合关键创新要素，快速实验创新应用

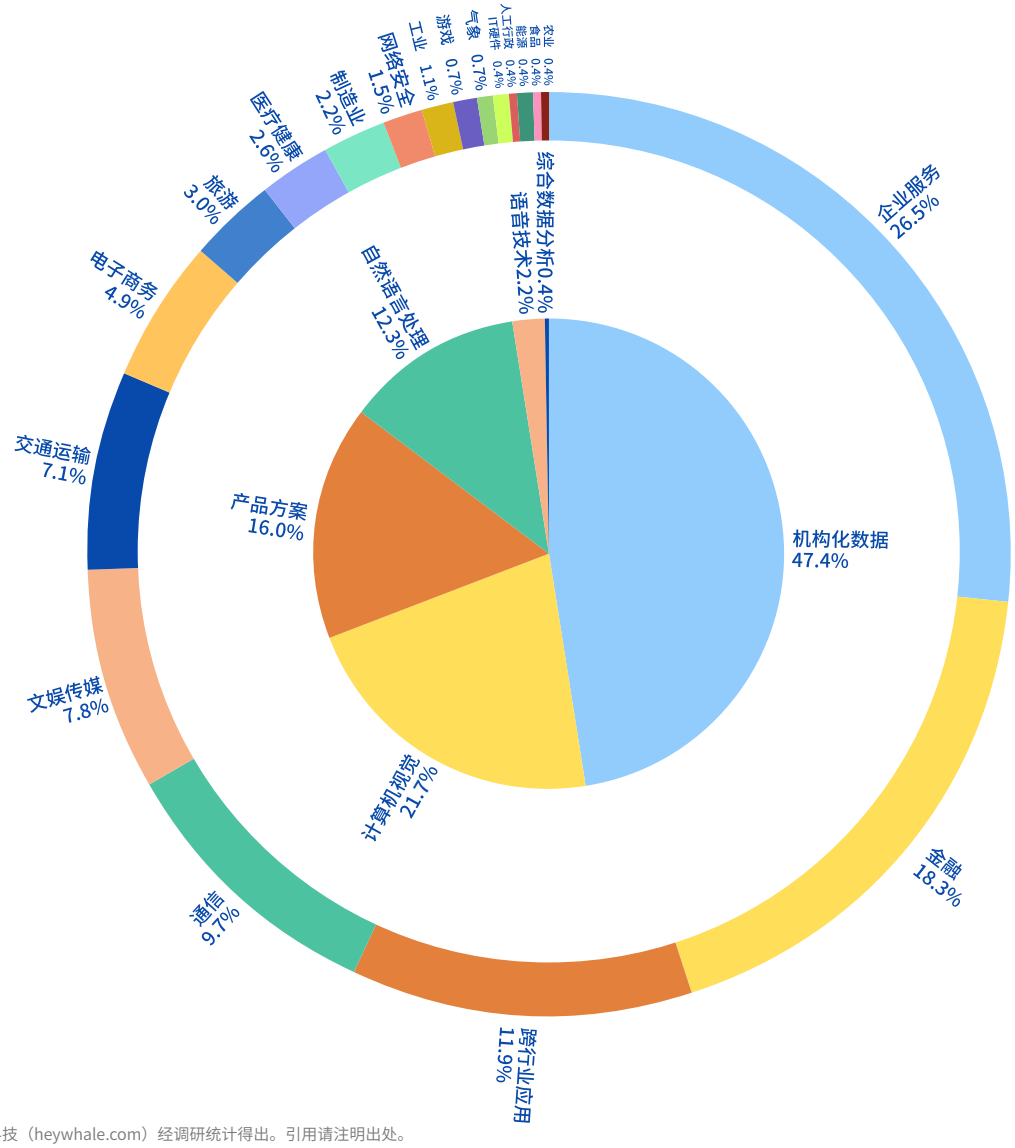
引领企业进行技术落地场景的创新探索

企业举办人工智能竞赛的核心诉求，是挖掘潜在的创新方向和应用场景。通过人工智能竞赛的举办，企业打造出了充分融合数据、算法、人才、算力等各项创新要素的标杆案例，以赛题要求切入，让创新探索过程更加具象、创新探索结果更加明确，为企业的数字化投入建立信心。

同时，人工智能竞赛**作为数据研发的缩影，也成为了企业积累数据科学能力的实验室。**在贴近真实业务场景的竞赛环境下，为可复用的数据科学方法论指导赛事设计与运营，厘清数字化转型的目标与路径，把握数据科学项目的生命周期，更加有的放矢地制定策略。

分析 2014 年至今国内的 400 场人工智能竞赛发现，企业在所有人工智能竞赛主办方中占比高达 67%，金融、通信、文娱传媒、交通运输和电子商务是发起人工智能竞赛数量最多的五大行业领域，合计占比高达 75%。企业最为偏好结构化数据挖掘类赛题（47.4%），其次是计算机视觉（21.6%）。具体来看，金融行业赛事主办方的赛题场景相对发散，除金融关联领域外，亦涉及电子商务、交通出行、农林养殖业、新零售等“金融 +”场景。通信行业在 5G 技术的商用趋势与产业互联的影响下，还会关注安防、交通、旅游、人文、社会治理、及文娱传媒等“万物互联”和“智慧城市”等相关领域。

图 2-1 企业类赛事主办方的赛题场景分布和技术类型分布



数据来源：和鲸科技（heywhale.com）经调研统计得出。引用请注明出处。

人工智能竞赛连接起的每一项要素都在数字化创新中扮演着不可或缺的重要角色：云计算资源和数据科学协同开发工具是基础设施，人才和 AI 技术是生产力，数据是基础生产资料，环环相扣的创新效果通过以赛题承载的应用场景得以直观呈现。以人工智能竞赛为中轴，汇聚起更多的人才、组织起更频繁的跨学科领域交流、涉足更广泛的技术面，积累更全面更优质的数据，从而带来更加敏捷高效的创新探索和性能突出的创新成果。

与收益不断高企相伴的是成本的逐步下跌：

· 由竞赛催动的规模经济促使基础设施的研发成本不断摊薄；

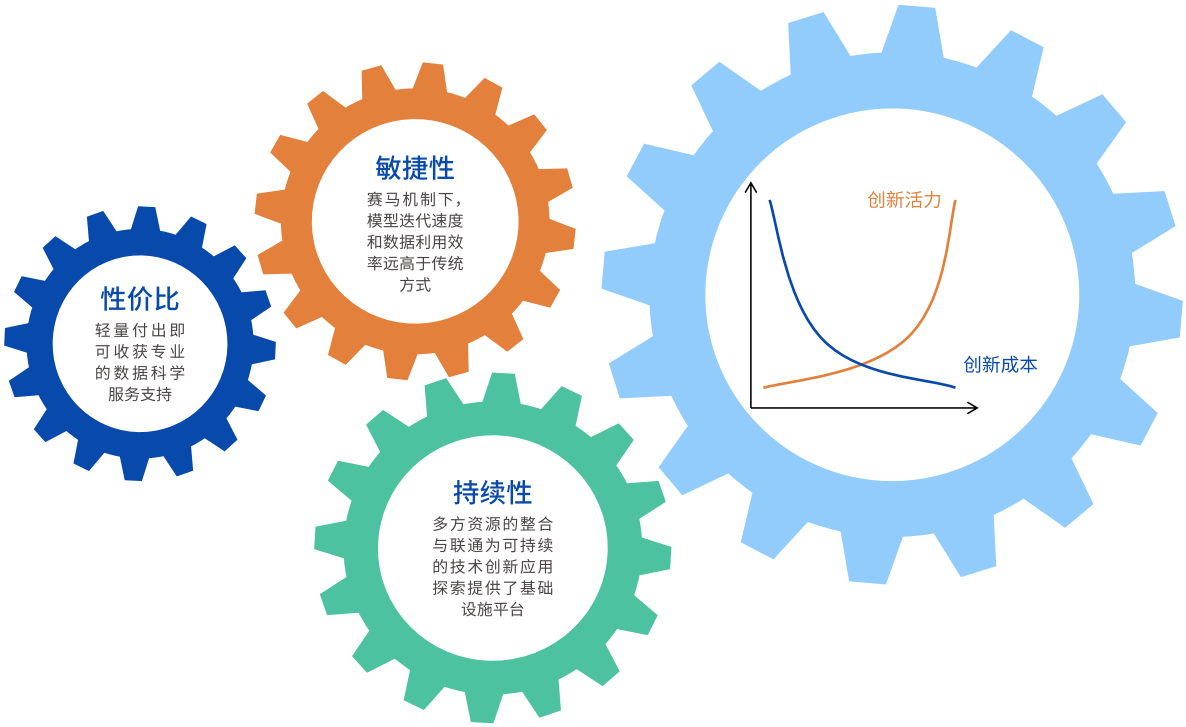
· 通过竞赛培养和积累的核心能力也让获取数据科学研发经验和制定竞赛制度与标准的边际成本不断减少；

· 规模日益巨大的数据科学人才在向竞赛输入其专业能力和创造力的同时，也在获取持续的实践训练和成长，人才聚集和培养成本得以压缩；

· 社区为交流互动提供了发声地，工具为团队合作提供了功能保障，协作不再成为难点。

· 创新的巨轮一旦转起，就会越来越快，最终开拓出更多有效果有价值的应用场景，实现令数字化价值指数级增长的飞轮效应。

兼具性价比、敏捷性、持续性的人工智能竞赛带来数字化价值指数级增长的飞轮效应



三、低耗高效可持续，生态化创新动能不断

竞赛低耗高效可持续，创新探索敏捷化

各类组织机构不懈探索更简单、更便捷的数字化战略，根据 Gartner 在 2017 年发布的数字化战略调研报告，在众多数字化创新手段中，人工智能竞赛依然具有相对较高的成功率。因为在确保有 100% 的信心推动数据科学项目成功着陆之前，**人工智能竞赛在性价比、敏捷性和持续性等方面表现突出，不仅有利于低成本的发展策略试验和排错，亦能充分积累数据科学实战经验。**

数字化创新方法投入对比

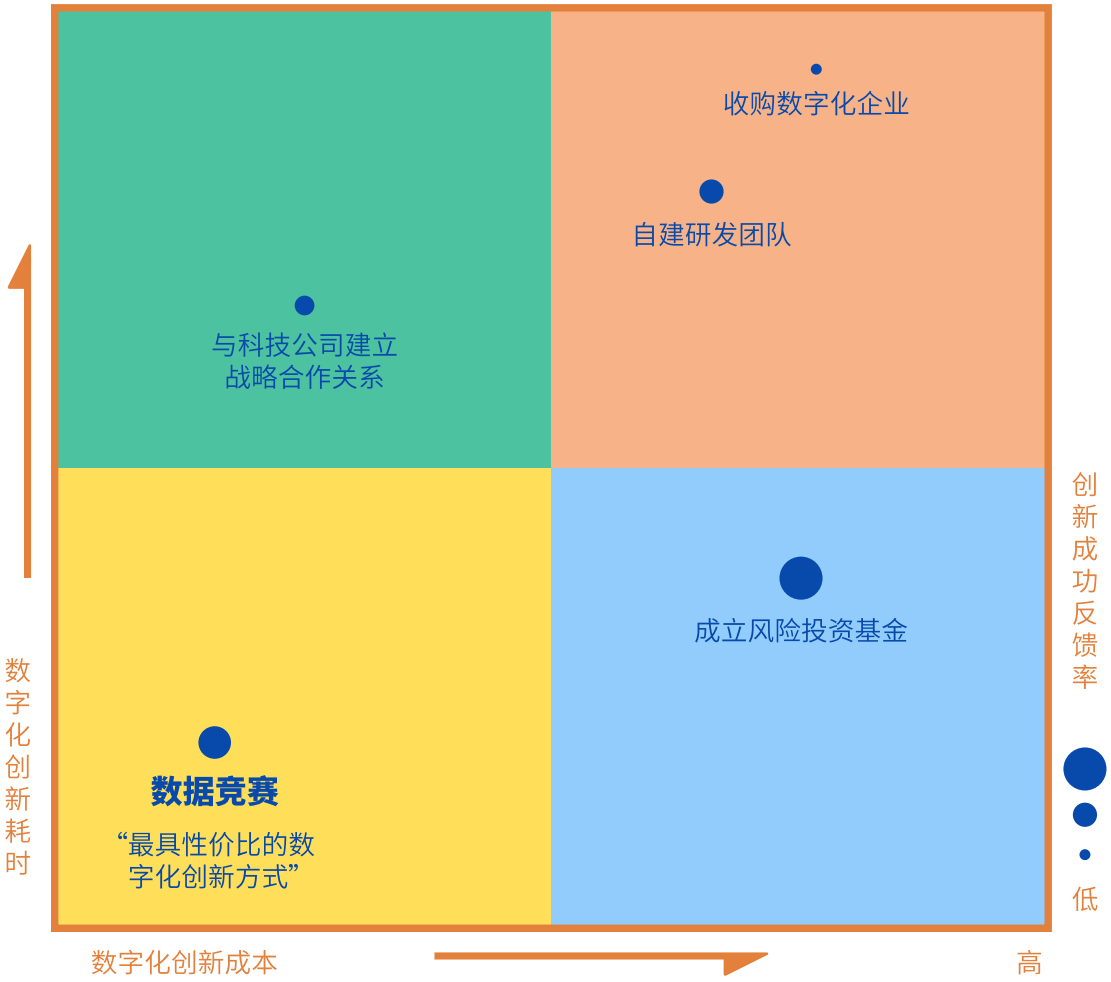
	自建团队	风险投资	外包	数据竞赛
人力成本	需要10人的管理、对接和QC团队	需要10人的团队进行投前研究与投后管理	需要1-3人的对接团队	**需要1-3人的对接团队**
时间成本	一般不少于半年	评估在一个月左右 项目周期不少于半年	技术外包周期不小于1年	**2-3个月左右**
资金成本	数百万人民币	数百万人民币	数十万至数百万人民币不等	**20-50万人民币不等**
其他资源	服务器、电脑、计算资源等	投后管理资源支出	办公场地等	**几乎不需要其他资源投入**
风险抵御	企业内部资源周转的成本风险	投资失败风险	外包后成果的延展性和可实现性	**较低的风险**
效果评估	可能会获得较好的解决方案	可能会成功捕获优秀的创新方向	可能会获得良好的外包成果	**创新探索效果可检验；更开阔以及创新性更强的解决方案；人才触达以及品牌号召力提升**

· **性价比：**人工智能竞赛创造了化繁为简的环境条件，在更加紧凑的周期内，能够直接感受推动数据科学创新和技术落地的步骤与方法。无论是自行成立数字化研发中心，还是与产业同行进行战略合作，往往首先需要在 IT 基础设施上有巨大投入，而在竞赛环境里，赛事主办方往往只需要承担一个业务负责人的角色，与竞赛运营平台沟通其创新诉求，配合轻量级的付出即能够获得赛事运营团队和数据科学研发人员的专业服务与支持。

· **敏捷性：**传统的数据科学项目受限于团队人数和团队精力，信息反馈也往往需要较长的周期。而人工智能竞赛由于其特殊的赛马机制，模型迭代的速度和对于数据的利用效率往往远高于常规的数据科学项目。作为缺乏前车之鉴的创新之举，推进数据科学项目总少不了反复的试验和排错，传统方式下的每一次试验往往因投入巨大可能导致决策者投鼠忌器。但籍由人工智能竞赛以更经济的投入换取更快速的反馈，便为多轮连续验证创造了空间，实现了敏捷的创新。

· **持续性：**以人工智能竞赛为契机，加强企业、高校和科研机构、政务部门之间广泛深入的合作，促进多方资源的整合与联通，既是促进产教融合的数据科学人才培养模式变革的创新之举，也为研协同打通了渠道、树立了标杆，为可持续的技术创新应用探索提供了基础设施平台，市场价值必将迎来指数型增长。

图 2-2 不同数字化创新方式的成本、耗时和成功率对比



数据来源：和鲸科技（heywhale.com）制作。引用请注明出处。

表 2-1 人工智能竞赛对企业的价值总结

	人工智能竞赛的亮点价值
创新探索	技术落地应用场景探索，挖掘潜力项目 检验 AI 投入与建设效果，优化创新战略 压缩试错成本，敏捷迭代创新思路和方向
技术应用标杆	打造 AI 技术和云计算资源的应用标杆案例 加强数据资产的积累、治理与应用 提升技术投入信心
数据科学能力	积累推动数据科学自研的核心能力
产研交流	建立跨学科背景人才的交流学习机制 多元思维碰撞，保持与最新技术趋势的同步性
企业形象	将数字化建设成果与社会进行及时的分享 树立起科技驱动、AI 持续投入的前沿形象

数据来源：和鲸科技（heywhale.com）经调研统计得出。引用请注明出处。

各类赛事主办方的数字化创新需求与难点通过人工智能竞赛得到了满足

	企业	政府	高校	科研机构
核心诉求	挖掘潜在技术创新方向 和数据应用场景，积累数 据科学核心能力	政务管理数据化、智能化 利用数据建设智慧城市 构建开放共赢的数据生态体系	需要进行数据科学学 科建设，培养优秀的数 据科学人才	科研成果创新落地 保证科研方向的领先性
目标汇总	提升数据科学创新应用的核心能力		完善教学培养体系 实现“教学练习实践”	提升科研水平，进行科研技 术成果的创新落地
困难挑战	人才不足、研发落地成本高、创新动力不足、成果反馈周 期长、数据纬度广探索方向发散亟需收敛……		数据储备不足、质量差，缺乏功能完备的工具平台和贴近实 战的场景来进行实践，教学及研发方向容易与市场割裂	
最佳策略	聚集多方创意、人才、技术，在一个小的场景中进行成果 验证，再进行全面的部署		学生能在真实的商业数据和应 用场景下进行实践训练	增加与企业的互通合作，共同 进行科研探索和成果孵化

人工智能竞赛

第三章

人工智能竞赛帮助促进高校的学科交叉与产研融合

一、促进高校加强学科建设与学科交叉

二、推动科研机构加速产研融合与技术发展

ARTIFICIAL INTELLIGENCE

一、促进高校加强学科建设与学科交叉

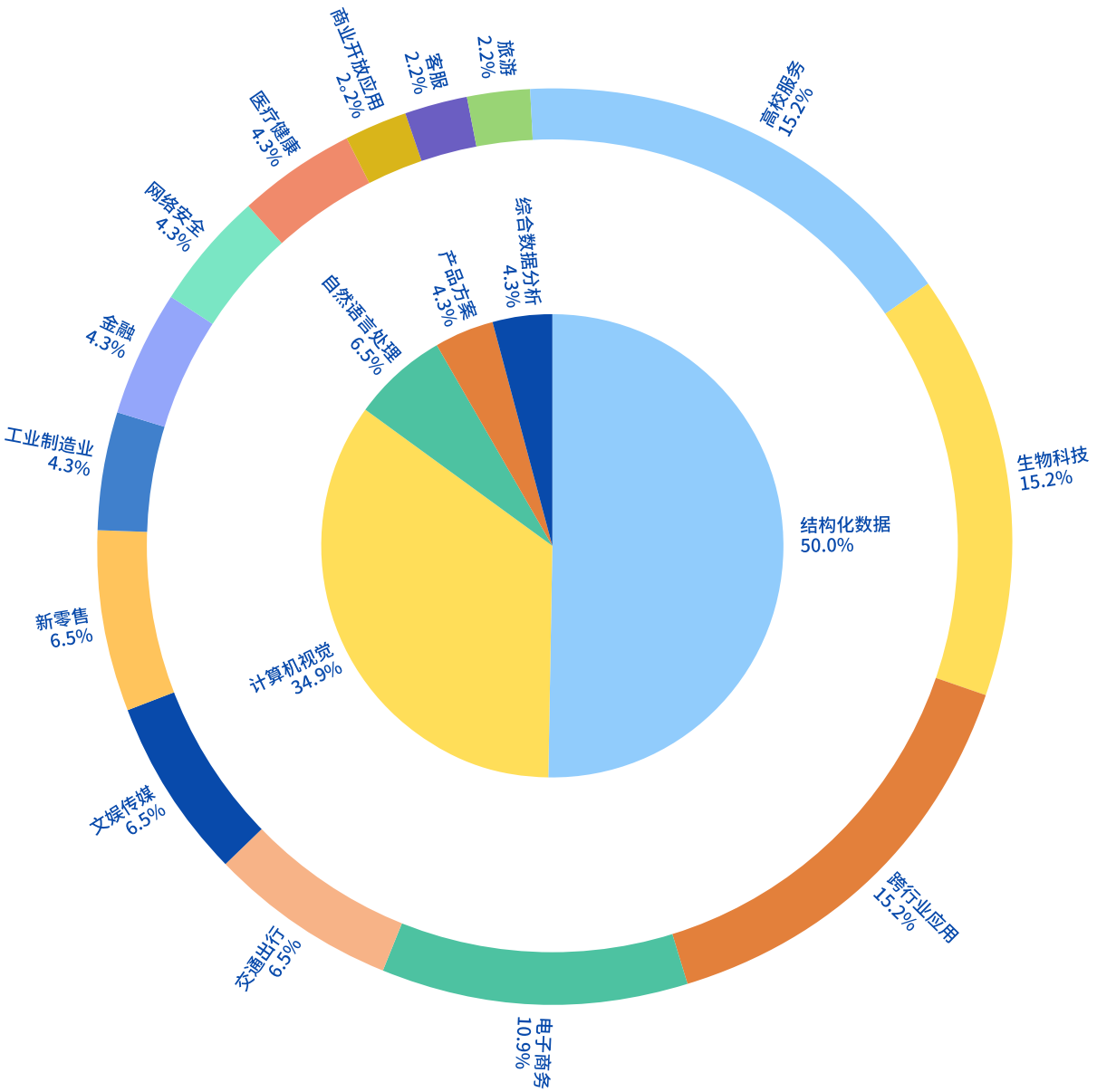
促进高校加强学科建设与人才培养

在国家相关政策的鼓励和推动下，越来越多的高校开始进行数据科学的学科建设和人才培养。高校学科建设需要持续的资金投入和资源配套，数据相关专业仍缺乏成熟的学科建设经验，学生则需要充足的优质数据、功能完备的工具平台和贴近实战的场景来进行实践——这一系列问题都是困扰着高校的现实困难。

人工智能竞赛作为高校推进大数据相关学科建设的重要手段，将教学、科研、实训深度结合，以赛题的形式发布，广泛吸引校内的人才参与竞争和交流学习，探索将理论知识转化为实践应用，帮助人才在竞赛场景中以赛促学，帮助人才认识理论学习、基础知识的意义与重要性。

在此背景下，高校围绕技术热点、学科建设以及创新应用三个方向进行赛题的策划，广泛涉及社会服务、工业制造、旅游、生物科技、网络安全和医疗健康等多个行业，以加强学生的理论和技能在多元场景下的应用能力。

图 3-1 高校类赛事主办方的赛题场景分布和技术类型分布



数据来源：和鲸科技（heywhale.com）经调研统计得出。引用请注明出处。

由于人工智能赛事涉及的赛道非常广泛，除了计算机相关的专业外，还可能需要材料科学、电气工程、自动化工程、通信工程、数学等学科的专业知识，甚至还有部分赛道需要航天工程、测绘与地理、物理学、医学、人文与社会科学等与计算机科学不太相关的学科背景。对于拥有多领域人才的高校和科研机构而言，涉及到多领域的人工智能赛事对于不同学科的交流以及交叉学科融合研究与教学都有极大的促进作用。

表 3-1 人工智能竞赛对高校的价值总结

	人工智能竞赛的亮点价值
实践学习	理论学习与演练实践相结合 加强学生锻炼实效，检验人才培育效果 优化教学和人才培养方式
学科建设	探索如何将理论知识转化为实践应用 打造学科建设的标杆案例
数据资源	接触到更真实、更优质、更具时效性的产业数据
高校形象	展示高校的学科资源和教学成果 树立重视学科建设与人才培养的形象

数据来源：和鲸科技（heywhale.com）经调研统计得出。引用请注明出处。

二、推动科研机构加速产研融合与技术发展

推动科研机构加速产研融合与技术发展

科研机构是技术研究与探索的引领者，也是数字化转型的推动者。基础研究、应用研究和开发研究是科研系统中三个紧密相连的重要环节，三者协调统一的发展才能够真正实现科研成果的价值转化。

在人工智能领域，人工智能竞赛为这个过程的进化创造了条件，集中了数据、人才等必要的基础要素，提供了跨学科领域的交流环境。同时，前沿科研的实用价值通过竞赛被释放、被感知，也获得了持续的社会关注和资源支持。

科研机构的赛题技术方向和应用场景富有前瞻性，集中于和自身行业背景属性相同的场景中，地质、水利、气象、传媒是较为常见的赛题场景。

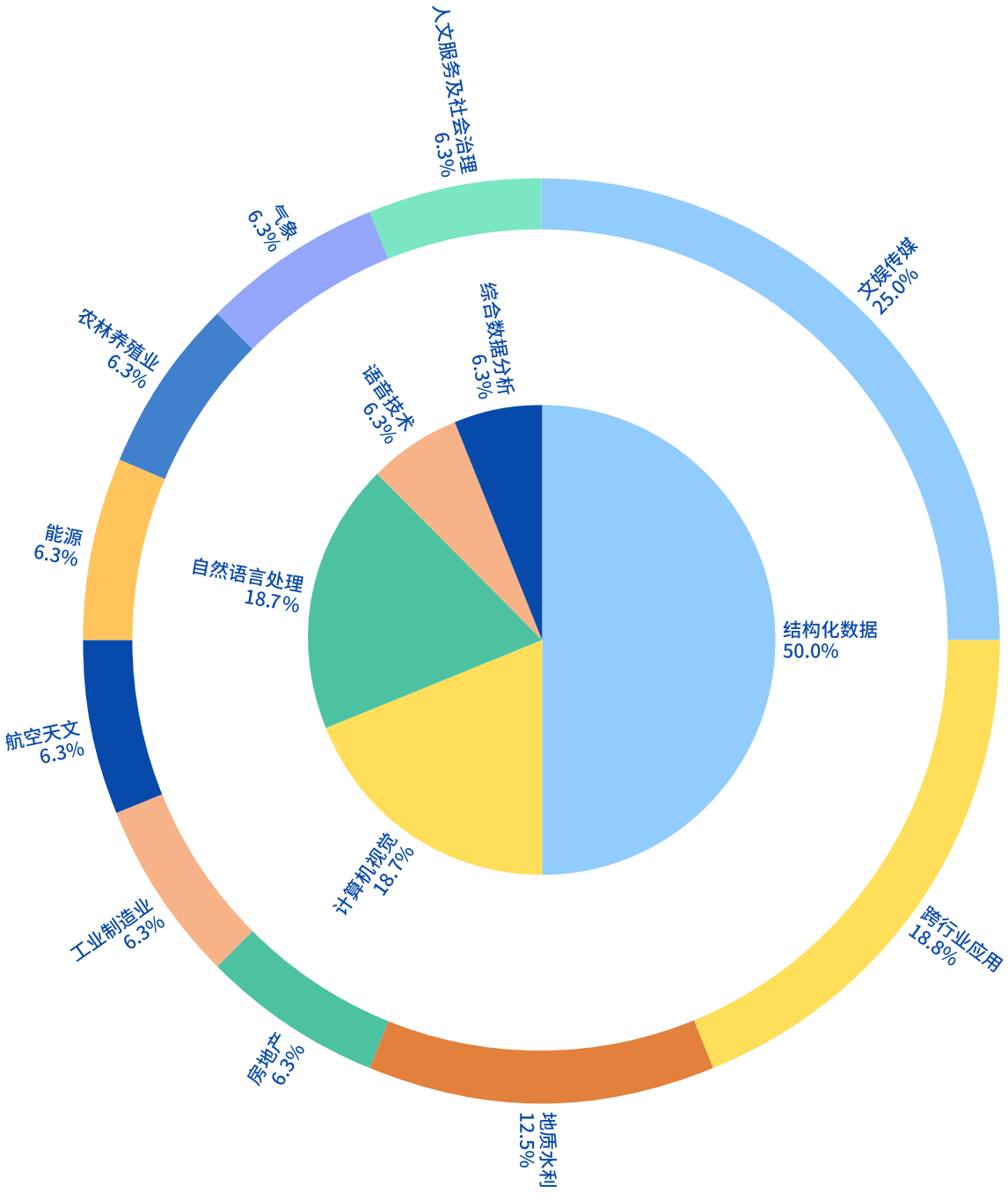
- 基础研究：为技术发明创造理论前提的
- 应用研究：在理论基础上针对特定目标进行研究
- 开发研究：将基础研究和应用研究成果投入生产实践

表 3-2 人工智能竞赛对科研机构的价值总结

	人工智能竞赛的亮点价值
前沿科研进展	融合丰富的资源支持 开辟创新的科研价值探索路径
产研交流	与工业界进行紧密合作 促进了跨学科领域产研交流
科研标杆	释放前沿科研的实用价值 树立起前沿科研成果创新应用探索的标杆

数据来源：和鲸科技（heywhale.com）经调研统计得出。引用请注明出处。

图 3-2 科研机构类赛事主办方的赛题场景分布和技术类型分布



数据来源：和鲸科技（heywhale.com）经调研统计得出。引用请注明出处。

第四章

选手眼中的人工智能竞赛

一、为什么参加人工智能竞赛

二、希望通过参赛收获

三、认为竞赛对自己的意义

ARTIFICIAL INTELLIGENCE

一、为什么参加人工智能竞赛

竞赛离不开人才的参与，那么人工智能竞赛对于参赛选手有哪些价值，他们参与竞赛的动机和意愿是又什么？通过面向逾千名选手的调研中发现，选手报名参与竞赛的动机主要有如下几方面：

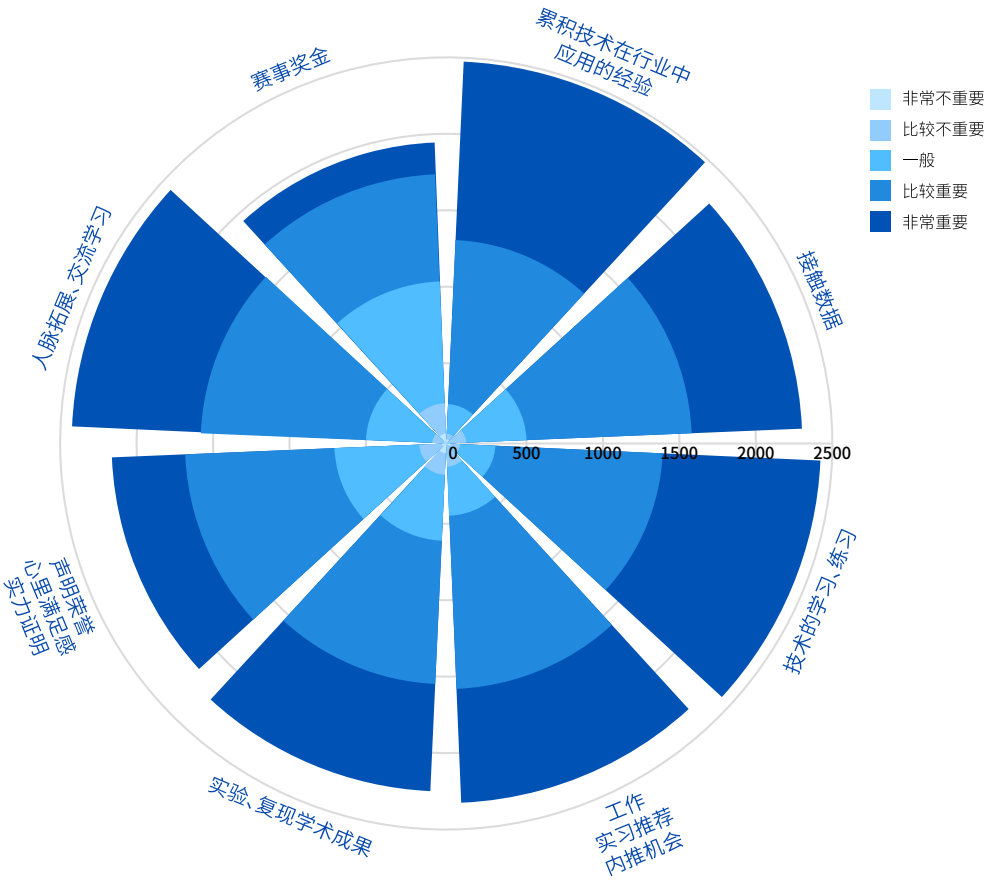
1. 实战经验：参与竞赛可以接触业界实际的业务场景与真实数据，助力人才累积相关技术在具体行业场景下的应用经验。

2. 技能提升：竞赛过程中能极大锻炼工程实践，复杂分析，团队协作等能力，培养技术与跨领域认知策略，提升人才综合素质。

3. 交流学习：由具体的赛题和充满探讨交流氛围的竞赛学习环境，助力人才拓展人脉，建构知识与人际关系的连接。

4. 报酬奖励：一般竞赛都有不菲的赛事奖金和并提供工作机会，在人工智能竞赛中取得的成绩也日益成为证明个人能力的权威背书，对促进人才职业发展与财富积累颇有裨益。

图 4-1 调研对象对于各项人工智能竞赛收获诉求的重要性评价



数据来源：和鲸科技（heywhale.com）经调研统计得出。引用请注明出处。

二、希望通过参赛收获

图 4-2 不同水平调研对象对于各项人工智能竞赛收获诉求的重要性评价

根据在人工智能竞赛中的排名表现，调研对象被划分了黄金玩家、白银玩家和青铜玩家三大类。大家对人工智能竞赛的诉求在呈现出一定程度上的一致性之外，也在某些具体内容上有细微的差异：

- 黄金玩家依然把赛事奖金列为最大诉求，也远比其他人更加倚重通过竞赛拓展人脉和加强交流学习；
- 相对入门的青铜玩家则更加看重能通过竞赛接触数据和获取工作机会。

数据来源：和鲸科技（heywhale.com）经调研统计得出。引用请注明出处。

三、认为竞赛对自己的意义

人工智能竞赛有重要意义和丰富价值，吸引着越来越多的人才投身人工智能竞赛，为更好的了解人工智能竞赛对于参赛者的意义？我们设计了自我效能感、内部满足感、外在报酬感等维度去探究选手们眼中人工智能竞赛对于他们的意义。

量表维度	量表指标
自我效能感	即使竞赛再困难，我也能努力创新去解决
	我如果努力去做，能够在竞赛中取得好名次
	我有所需的专业技能去完成竞赛
	我能比参与竞赛的其他人提供更有价值的方案
内部满足感	参与竞赛是我的兴趣，很享受竞赛的过程
	竞赛给我一个机会去做我擅长的事
	我喜欢竞赛的整体氛围
	竞赛排名前列，让我很有成就感
	我有机会通过参与竞赛交流学习知识
外部报酬感	竞赛对应的物质奖励对我很有吸引力
	竞赛能够有助于我发展事业的机会
	竞赛可以增强我在就业市场的竞争力

1. 自我效能感

自我效能感是个体结合自身的实力和任务难度，对于完成任务目标的自信程度。故在参与人工智能竞赛的人才中，其自我效能感会随着持续参与的结果不断进行调整。具有高水平自我效能感的选手，在竞赛中更可能相信付出努力，可以获得满意的绩效，也即高自我效能感的人才容易形成较高的期望估值并持续的付出和留存。

根据深度访谈选手反馈，人工智能竞赛是锻炼技能实践验证能力的机会，在学习与就业的不同阶段都会参与竞赛去不断提升自己的能力，培养自我效能最大的困难是刚接触竞赛时，知识储备不足，缺少相应的路径和指导。相比竞赛中名列前茅，更看重在竞赛中多次累积的经验，底层的数据分析能力训练与跨领域认知策略的培养，更让自己有信心并不断参与竞赛。

2. 内部满足感

内部满足感不是为了一些可感知可分离的结果做事，而是满足内心的需求。拥有诸如乐趣、挑战感、归属感等内部动机的使用者更能够持续参与和使用。本年鉴中持续参与动机中的内部满足感动机分为学习成长、身份构建、虚拟社区感三个方面。

（1）学习成长——学习成长是指人才在参与人工智能竞赛中，能够学习或提高的相关技能，继而能力提高会引发参与者的持续与意向，当个体在参与的过程中得到学习和提升，也容易被社群接纳建立归属感，也就会更频繁的参与。

（2）身份构建——身份构建是指通过在竞赛中，释放自己的创造力和想象力，来证明自己能力并进一步构建自己的身份角色，也称自我肯定。身份构建和自我肯定能够使得竞赛选手产生沉浸的体验，保持持续的参与意愿和行为。选手希望通过持续的参与

赛和任务，能够证明自己的能力并且构建自己的身份。

（3）虚拟社区感——虚拟社区感会使其对社区负有责任感，从而在竞赛平台中想要持续的参与。竞赛选手在社区中互相支持建立自己的角色并且认同别人的身份以及产生信任，会产生社区归属感，进而有持续参与的意愿。

根据深度访谈选手反馈，在人工智能竞赛中和志同道合的小伙伴一同参与竞赛，通过学习成长，从竞赛小白一路成长为竞赛大神，通过社交不断深化关系链，并在社群与社区中渐渐构建起影响力和社会地位，对于自己很有满足感和社区归属感，愿意分享自己的竞赛经验，自发维护竞赛氛围，愿意为人工智能竞赛生态建设不断贡献力量。

3. 外部报酬

外部报酬能够间接地得到内心满足的需求，是一种可感知可分离的事务，例如竞赛奖金，工作机会等，越高的奖励的设置，越能增强持续参与意愿。人工智能竞赛平台具备将知识进行变现的有偿属性，人才对于获取报酬的预期是激励和吸引他们参与的十分重要的因素。

根据深度访谈选手反馈，人工智能竞赛比起奖金更看重的是工作 / 实习机会，人工智能竞赛带来的就业竞争力的提升，和工作机会要比一次性的奖金更为吸引人，一次完整参与竞赛的项目经历可以为简历增添不少分数，获取 Top 名次就更有竞争优势，目前也有不少公司或机构通过人工智能竞赛遴选人才，在如今数字经济时代，算法相关岗位竞争相对激烈，人才通过人工智能竞赛获取更多报酬机会，具备相应的意愿，但因为竞赛的名次有限，竞争依旧激烈，故意愿相对自我效能感与内部满足感低一些。

人才是人工智能竞赛的核心源泉，竞赛的价值体现与人才内在的动机满足相互促进，共同发展。随着时代的发展，面对社会对人工智能人才的泛在需求，人才应跟紧技术发展与产业升级的步伐，意识到人工智能竞赛对于个人成长与发展的价值，广泛积极参与竞赛，不断夯实基础知识，积极学习跨学科、多领域技能，以解决实际问题为导向，持续积累项目经验，全方位提高综合素质能力。

第五章

组织人工智能竞赛的挑战

一、对竞赛复杂性的调研结果

二、人工智能竞赛的运营复杂度详解

ARTIFICIAL INTELLIGENCE

一、对竞赛复杂性的调研结果

调研发现，在多数的参赛选手心目中，人工智能竞赛的质量取决于三个要素：主办方的品牌知名度、赛事激励的丰厚程度、竞赛是否提供真实的数据。这些是吸引优质参赛选手，促进产出更高水平成果的关键要素。

但是，这三点仍不足以保障优质竞赛的完成。人工智能竞赛牵扯到**应用场景、数据、算力、算法和人才**等方方面面的复杂要素，包含了**数据科学研发过程中的主要环节**，从赛前筹备到赛中运营，再到赛后跟进，从应用场景挖掘到配套数据处理，从计算环境搭建到算法应用，从参赛选手管理到成果测评，**每个环节都需要专业能力和精细运营的支撑**。

2018 Kinetics 视频行为分类比赛——数据失误导致重大竞赛事故

2018 年，Kinetics 视频行为分类比赛遭遇了赛事组织的多次事故。起因在于赛事运营方发布的训练集数据中错误地混入了测试集数据。一周后，赛事运营方重新发布的数据集再度出现重大疏忽——所有数据标签被泄漏，意味着能够验证模型效果的答案遭到提前揭示，选手无须花费精力训练可靠的模型。参赛选手向赛事运营方发送邮件反应数据集的问题，也未得到及时的回应和解决。赛事的公平性和参赛团队的积极性遭到了破坏，同时引起了舆论争议，赛事被迫延期。

扫描上方二维码
了解更多赛事详情

1. 主办方对竞赛有高预期

不同的赛事主办方举办人工智能竞赛的初衷各有侧重，其共同的诉求是赛事进程平稳顺畅、舆论评价正面积极、参赛选手能够准确理解赛题、参赛选手技术能力出色、赛事成果具有优越表现。这需要不同环节的专业服务能力紧密配合，促成赛事实现最好的效果。正是意识到赛事筹办和组织的复杂性，主办方存在多方面的顾虑，担心对于人工智能竞赛的投入无法获取预期回报，甚至损伤品牌形象。

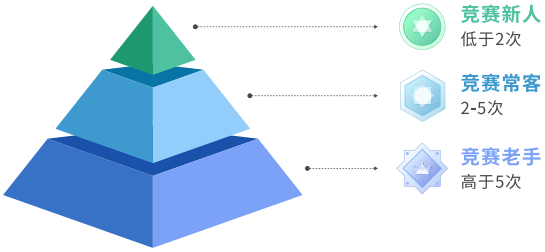
> 人工智能竞赛还是需要依赖专业的办赛平台，因为人工智能竞赛筹办区别于普通的活动组织，企业缺乏专业的的赛事运营人员和赛题设计人员。
>
> ——上海电信 互联网部产品技术研发中心副经理 葛正荣

> 术业有专攻，优秀的企业更懂得互补 + 合作 = 共赢的价值。同盾除了不断锻造自身技术实力外，也积极推行在产学研领域的企业价值，创办专业的赛事、选拔潜在的专业科技人才，这需要专业的赛事平台参与其中，从学术性与商业性相融合的赛题设计、到赛制中的数据质量与敏感把控，以及兼顾数据安全与赛题可解性，都需要更加专业的、垂直的竞赛平台来合作共建。同时，专业的竞赛平台拥有过硬的技术实力，可以勘验参赛选手的成果。在整个赛事的宣传节奏把控上，也能够创造出丰富的内容及足够触达到目标群体的有效渠道。
>
> ——同盾大学运营人员 娜娜

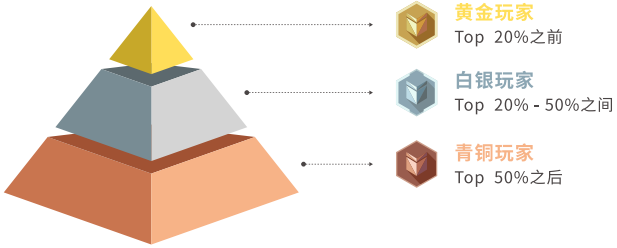
2. 参赛者对竞赛有严要求

选手的技术潜力的释放程度和技术能力的发挥水平，高度依赖于赛事全生命周期的技术支撑和运营管理。人工智能竞赛参赛者调研的结果显示：**组织“优质的人工智能竞赛”需要兼备专业的技术储备和丰富的项目管理经验。**

按2018年至今参与全渠道数据竞赛的频率



按2018年至今参与全渠道数据竞赛平均成绩排名表现



从参赛频率的维度来看：

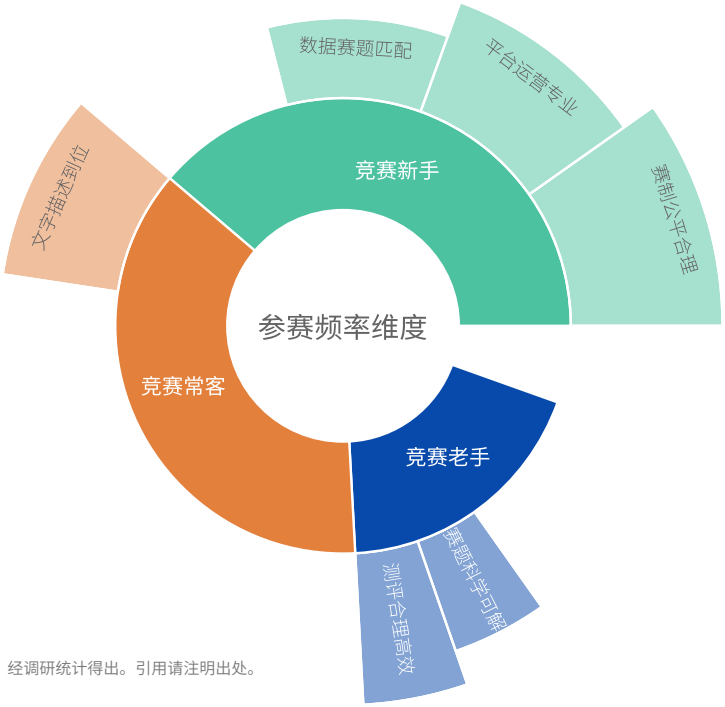
- 竞赛新手：关心自己是否值得将精力投入到竞赛中，获得经验与成长，认为优质的人工智能竞赛首先应该具备科学合理的赛题、数据和测评办法，竞赛页面的文字描述需要清晰、专业、合理。
- 竞赛常客：在赛题与数据相匹配的基准要求被满足后，则更关心完善的赛制、自动测评等功能点能够保障良好的参赛体验。
- 竞赛老手：奖金或工作机会等实质性激励是他们首要考虑因素，且激励也需要以合理、可解的赛题为基础；此外，他们也会倾向于投入精力加入前沿性难题的探索。

图 5-1 不同参赛频率的调研对象判断人工智能竞赛是否优质的考虑因素排序

	竞赛新人	竞赛常客	竞赛老手
赛题、数据、测评的文字描述充分、到位、正确	1.0	6.0	4.0
赛题课题科学正确、合理可解	2.0	7.0	2.0
数据与赛题目标匹配	3.0	1.0	6.0
自动测评方法科学合理	4.0	3.0	9.0
竞赛平台/主办方运营专业、响应与解答及时	5.0	11.0	8.0
具有公平的赛制	6.0	2.0	10.0
自动测评反馈报错信息清晰明了	7.0	5.0	7.0
自动测评即时出分、运行较快	8.0	4.0	3.0
配备好用的在线AI开发工具、够用的云资源算力与在线储存	9.0	8.0	12.0
有领域权威专家的主观评审与点评	10.0	14.0	15.0
赛题难度属技术前沿或行业应用前沿	11.0	9.0	5.0
数据质量良好	12.0	10.0	7.0
仅需提交模型文件而非答案文件	13.0	15.0	13.0
激励（包括奖金、工作机会等各类激励）丰厚	14.0	12.0	1.0
数据真实	15.0	13.0	14.0
主办方知名度大	16.0	16.0	11.0

数据来源：和鲸科技（heywhale.com）经调研统计得出。引用请注明出处。

图 5-2 不同参赛频率的调研对象认为影响人工智能竞赛质量最重要的因素



数据来源：和鲸科技（heywhale.com）经调研统计得出。引用请注明出处。

从参赛成绩的维度来看：

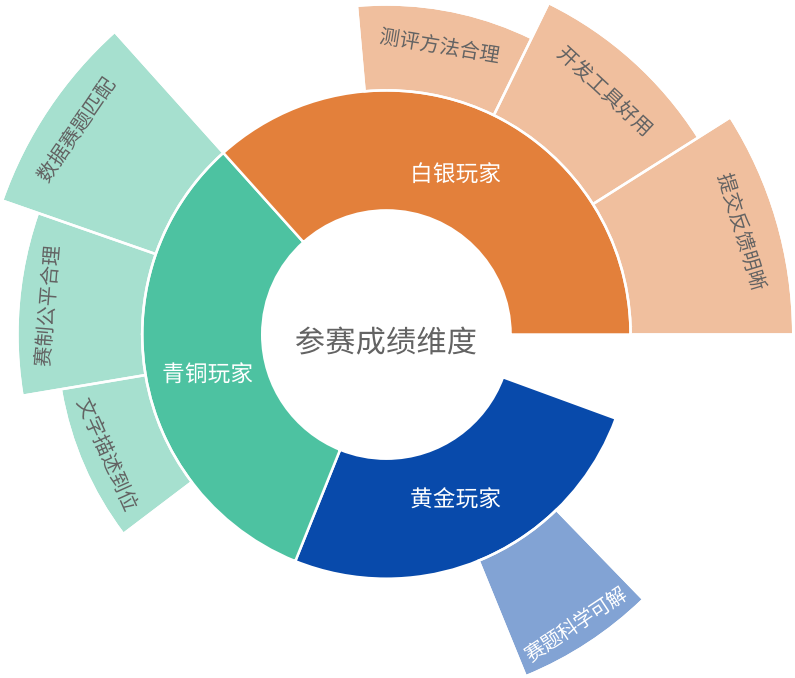
- 青铜玩家：除了要求题意表达通畅清晰外，希望公平的赛制和完善的测评功能能够保障竞赛的锻炼价值。
- 白银玩家：最关心赛题与数据匹配程度；进而关心有科学的测评算法合理评估选手的表现；而为了更有效率地展现实力，白银玩家也要求优质的人工智能竞赛配套功能全面、性能优越的算法开发工具和计算资源。
- 黄金玩家：对自己的能力充满自信，因此对于优质人工智能竞赛的定义简单直接：不浪费时间、不枉费付出，即赛题、数据、评审办法科学合理，且有敏捷完善的产品功能予以支撑，从而保证其精力最大程度投入解题本身；同时要有与自己的开发成果相般配的丰厚回报；水平一流的黄金玩家比起其他人也更愿意投身前沿课题的研发。

图 5-3 不同参赛成绩水平的调研对象判断人工智能竞赛是否优质的考虑因素排序

	青铜玩家	白银玩家	黄金玩家
赛题、数据、测评的文字描述充分、到位、正确	1.0	5.0	3.0
自动测评即时出分、运行较快	2.0	9.0	4.0
具有公平的赛制	3.0	7.0	8.0
赛题课题科学正确、合理可解	4.0	10.0	1.0
自动测评方法科学合理	5.0	2.0	10.0
自动测评反馈报错信息清晰明了	6.0	4.0	9.0
数据与赛题目标匹配	7.0	1.0	5.0
竞赛平台/主办方运营专业、响应与解答及时	8.0	8.0	11.0
仅需提交模型文件而非答案文件	9.0	15.0	15.0
有领域权威专家的主观评审与点评	10.0	13.0	16.0
配备好用的在线AI开发工具、够用的云资源算力与在线储存	11.0	3.0	13.0
赛题难度属技术前沿或行业应用前沿	12.0	6.0	6.0
数据质量良好	13.0	11.0	7.0
激励（包括奖金、工作机会等各类激励）丰厚	14.0	12.0	2.0
数据真实	15.0	14.0	14.0
主办方知名度大	16.0	16.0	12.0

数据来源：和鲸科技（heywhale.com）经调研统计得出。引用请注明出处。

图 5-4 不同参赛成绩水平的调研对象认为影响人工智能竞赛质量最重要的因素



数据来源：和鲸科技（heywhale.com）经调研统计得出。引用请注明出处。

可以看出，调研对象虽然拥有不同程度的参赛经验和成绩表现，但对于人工智能竞赛从赛题、评审到工具、运营均拥有复杂而严苛的要求——竞赛内容富有实际意义、竞赛服务专业周到。选手需求的复杂性和多元性，对于赛事的主办方和赛事平台，也提出了严格而专业的要求。

二、人工智能竞赛的运营复杂度详解

优质的人工智能竞赛需要把控住赛事的全流程，每一环节的成果产出都将直接影响其它环节的工作，影响整场赛事的质量。这对赛事组织运营者的业务解读能力、技术能力和沟通能力提出了持续的考验。

1. 数据科学项目管理，支撑赛事成果有效性

通过需求分析和赛题设计，赋予人工智能竞赛创新意义和业务价值，进而借由数据和计算资源管理保障该创新价值落地，有着综合性的挑战。

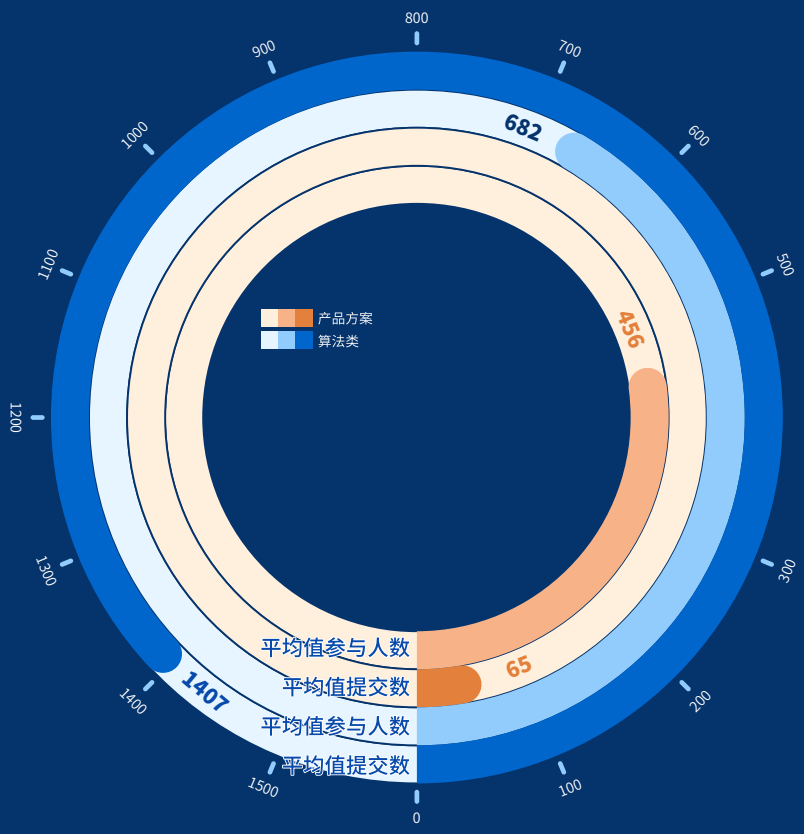
a. 需求分析和赛题设计

需求分析和赛题设计的首要任务是能够充分理解、拆解、定义竞赛主办方的真实诉求，使其可被人工智能竞赛承载和实现。赛事组织方需要对业务场景进行赛题边界的确定，使得赛题方向与主办方的行业特征、业务特点、数据储备相匹配。

面向大众广泛收集开放性方案的竞赛形式也很常见，因为许多行业的数字化转型正处于起步阶段，需要创新方向的借鉴与启发。**通过对国内数百场人工智能竞赛的分析，我们发现开放性方案题在报名热情上只稍逊于算法题，但最终成果的提交数量却远低于算法题。**

算法赛题需要对应用场景进行高度收敛，开放性方案题则不设边界，能让选手充分发挥创意，往往可收获意想不到的、但实则可行的数据与技术在场景中的应用方案，从而可为技术进一步落地指明方向。也是由于边界发散，导致考核标准难以界定，成果表现从而也往往不够稳定。

图 5-5 开放性方案题与算法题的参与规模与提交数量对比



数据来源：和鲸科技（heywhale.com）经调研统计得出。引用请注明出处。

算法模型类赛题相比开放性方案赛，有更加稳定的表现，其设计过程也存在着诸多考验：在需求分析结果和赛事数据条件的基础上，在成本限制、计算资源、数据安全的约束条件下，最终的赛题需要平衡可解性、前沿性和创新性等多方面的要求。

算法赛题需要以清晰明确的表达方式，帮助参赛选手充分理解其考察重点和开发目标，同时配套科学的测评方法对赛事效果进行合理评价。

b. 数据管理

赛题设计需要兼顾是否可解、是否具备业务应用价值或前沿技术探索价值等多方面的因素，而这皆以“数据”为基础。**数据层面有三个重点：数据质量、数据集切分、数据安全**，任何一点的疏漏都会导致潜力巨大的赛题难以兑现价值。

（1）数据质量

数据质量问题可能发生在数据收集、治理、加工和采样的各个阶段，数据质量低下表现在字段缺失、信息过时或信息不准确等方面。在人工智能竞赛的时间要求下，**数据质量低下容易导致选手分散过多精力进行数据的二次处理，限制了算法所能达到效果的天花板**。

质量低下的数据难以反映实际生产生活中的真实情况，最终影响模型成果在业务场景的表现。另外一种常见的数据质量问题是信息交互泄漏（Data Leakage），如数据的某些属性、特征、字段已潜在暴露了需要通过算法模型进行预测的信息结果。

一个易懂的例子是，一道赛题要求参赛者识别图片中的动物是否为鸟类，而在进行数据采样时，所有鸟类的图片 ID 均以 1 开头，而所有非鸟类的图片 ID 均以 2 开头，参赛选手通过图片 ID 字段即能判断结果，通过取巧的模型也有不俗的分数表现，但没有业务价值。利用信息交互泄漏漏洞而取得虚高分数的行为，也会破坏竞赛的专业性、公平性，激发负面的评价。

（2）数据集切分

数据集切分是模型构建的重要前置工作。在人工智能竞赛中，数据需要被切分为训练集和测试集两部分。其中训练集供参赛者用以搭建、训练模型，测试集则用以检验其模型的性能水平。举例来说，在数据量总体有限的情况下，**若测试集的数据量过少，则无法客观反应数据全貌。**

如果在切分数据集时未能控制好数据分布，也会影响模型实际表现的上限，容易出现模型在某个数据集上过拟合的情况，进而影

响其泛化性能。

（3）数据安全

出于对数据所涉敏感信息或商业机密的保护需要，数据安全是赛事主办方重点关心的问题，数据脱敏是这一问题的常见手段。优质的脱敏算法要充分保障数据的隐秘性，确保参赛选手无法通过技术手段逆向解码，获取脱敏前的信息。

如果数据的脱敏处理欠缺科学性，一方面可能会反而脱敏失效，另一方面也会导致参赛者在特征工程阶段被误导。

图 5-6 数据管理需要兼顾三个层面的重点

数据来源：和鲸科技（heywhale.com）经调研统计得出。引用请注明出处。

c. 成果管理

成果管理是影响数据科学项目项目效果的“最后一公里”，成果版本管理混乱、成果不可复现都是常见的赛事项目管理问题。

混乱的成果版本管理会导致团队协作效率低下、数据科学项目进度缓慢，并且最终在验证算法效果时难以做到对应版本的跟进和回溯。成果的不可复现，会导致算法成果在实现应用时**达不到预期效果，投入产出比模糊。**

d. 计算资源管理

充沛的算力可以满足 AI 模型的优化需求，也容易导致对于算法创新的忽视，算力供给也会带来较大的经济负荷。因此，搭建统一的云端竞赛环境便成为必然趋势，不仅能够使竞赛突破物理空间的限制，也能够鼓励选手在一定的算力限制条件下侧重算法层面的设计与创新。

对此相对的情况是，计算资源供应不足，则会出现模型运行时间过长、内存溢出、资源互相占用等现象，**导致参赛者难以在既定时间内完成对模型的迭代优化**。

因此，对竞赛全程的计算资源消耗量做精确的预估，充分平衡成本限制、赛题难度和赛事规模等多方因素，是赛事筹备的重要任务。

2. 保障赛事相关方体验需要面面俱到

a. 赛制

公平公正的人工智能竞赛环境和合理的赛制设计，是保障赛事主办方和参赛选手体验的基础，其中牵涉了科学的成果测评、反作弊机制等方面。

在自动测评方面，测评代码要严格契合赛题数据的要求，确保竞赛开发成果的性能表现和参赛选手的能力水平能够得到合理的评估，即具备可解释性和结果的可复现性。同时，测评代码还需要保证多人并行大量提交下的稳定性，确保参赛体验流畅。

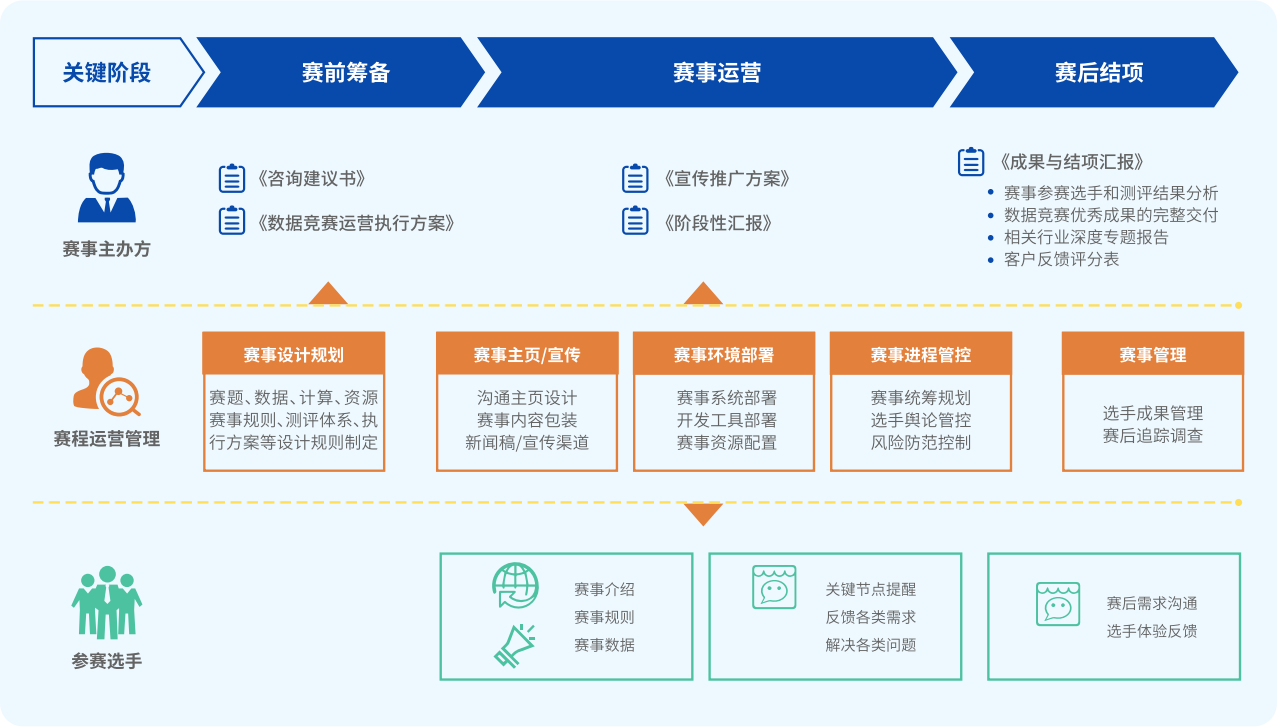
在常见的作弊手法中，包括利用排行榜和测评代码的漏洞，通过有规律的提交去猜测正确答案的排行榜探查（Leaderboard Probing）行为，通过高频提交粗暴堆分或人工标注答案的暴力刷分（Boosting Attack）行为，为了追求分数表现刻意构建过拟合模型的行为。

b. 赛程运营管理环节

一场人工智能竞赛牵涉赛事主办方、赛事运营者、参赛选手、媒体等方方面面的利益相关集体，彼此之间背景、分工、诉求不尽相同，兼顾各方体验的协作具有较大难度。而赛程运营管理则以串联者的角色整合了贯穿人工智能竞赛前期筹备、赛事运营、赛后结项等全生命周期的各类赛事相关要素，工作范围涉及资源统筹调配、赛事流程推进、选手社区运维、阶段进展汇报、风险防范控制、成果管理移交等方方面面。

缺乏成熟专业的赛程运营管理，轻则可能延滞人工智能竞赛进度的顺利推进，重则可能导致前期投入无法产出效益。

图 5-7 赛程运营管理具有相当的复杂性



数据来源：和鲸科技（heywhale.com）经调研统计得出。引用请注明出处。

c. 工具支撑

人工智能竞赛的组织具有较高的技术难度，从最初要求顺畅地跑通竞赛全流程，到能够进行云端协同开发，再到保障日益敏感的数据安全，越来越多的问题有赖于产品工具的功能保障。

（1）竞赛工具 1.0 时代——人工智能竞赛系统

人工智能竞赛系统是人工智能竞赛最基础的工具系统，为了实现对人工智能竞赛全生命周期的管理。作为连接参赛选手、赛事运营管理者和赛事主办方的线上桥梁，其功能覆盖竞赛页面信息配置、竞赛任务发布、评审系统设置和参赛成果统计等方面，目标是提高赛事运营管理者工作效率，便于赛事主办方了解赛事进程。

人工智能竞赛的公平性在很大程度上有赖于评审环节是否妥当，而评审又涉及复杂的算法和规则：**既要保障不同复杂程度的客观评审规则的自动化执行，又要对主观评审的分工权重与规则做出合理呈现，也能对多次提交和赛程各阶段的结果实现高效处理和统计，及时反馈报错信息或得分情况，可以说评审系统是整个竞赛系统开发的难点和重点。**

图 5-8 竞赛工具 1.0 的功能要求

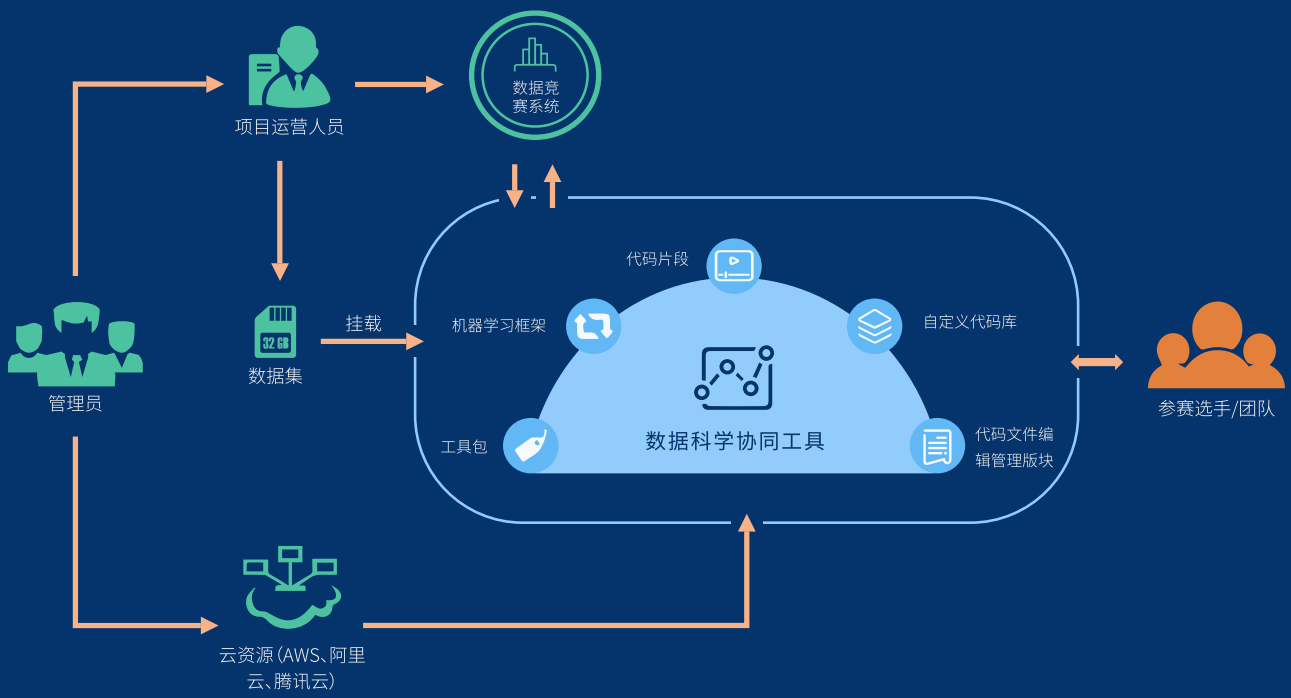
数据来源：和鲸科技（heywhale.com）经调研统计得出。引用请注明出处。

（2）竞赛工具 2.0 时代——人工智能竞赛系统 + 数据科学协同平台 + 云计算

在企业上云和协同开发的趋势下，竞赛工具 2.0 时代迎来了数据科学协同平台的整合。通过实现在云端协作进行数据分析和算法建模等工作，让人工智能竞赛真正贴近业界实际的数据科学研发场景。

目前全球范围内接受度较高的工具包括国外的 Jupyter Notebook、Kaggle Kernels、Google Colab，以及国内的 K-Lab。在人工智能竞赛环境中需要至少满足三个方面的要求：**提升效率，令参赛选手更加专注于开发本身；加强协同，为参赛团队打造交互式体验；弹性调度云计算资源，降低算力应用门槛与成本的同时，也让比赛更公平更有效。**

图 5-9 竞赛工具 2.0 的功能要求

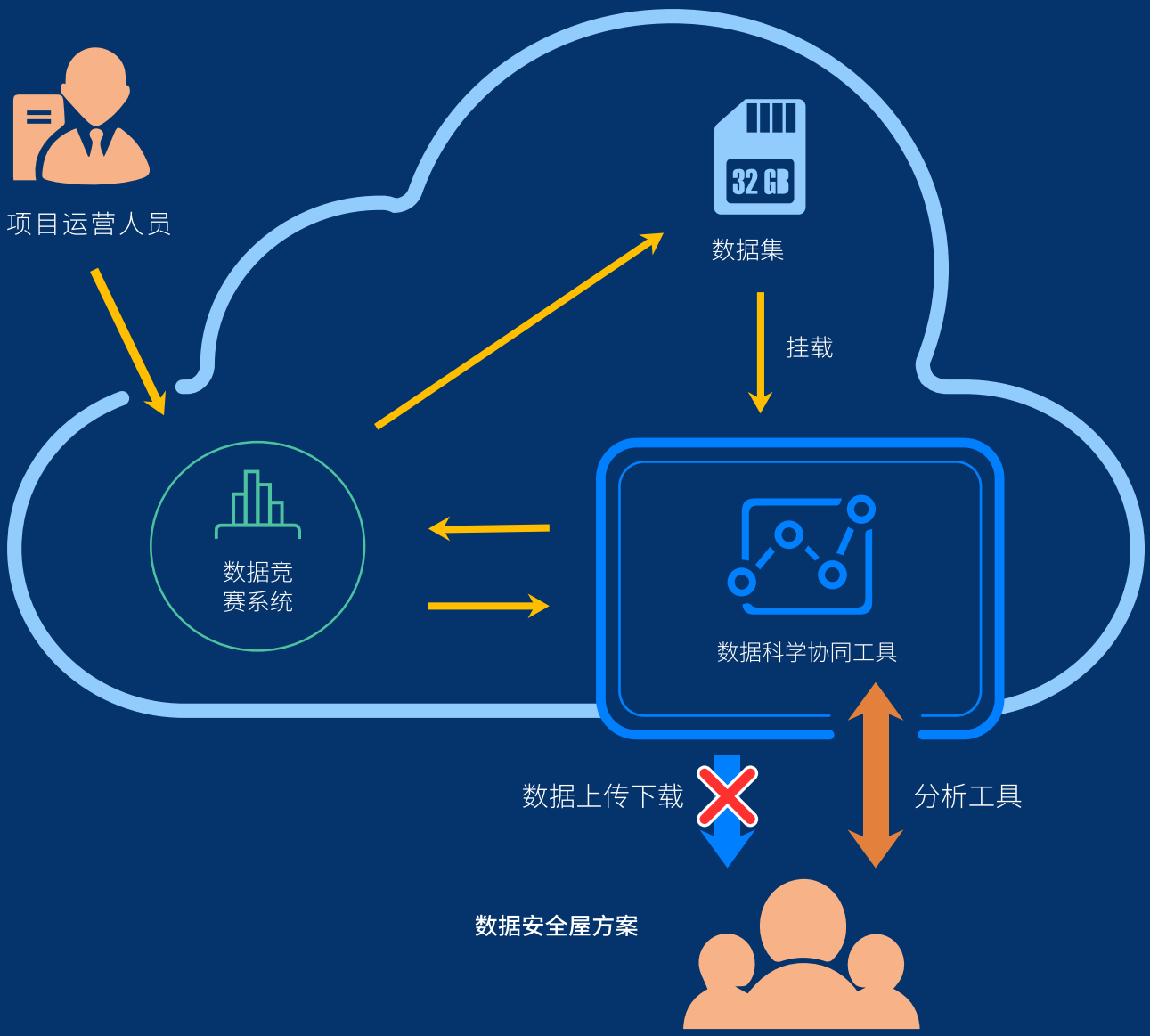


数据来源：和鲸科技（heywhale.com）经调研统计得出。引用请注明出处。

（3）竞赛工具 3.0 时代——人工智能竞赛系统 + 数据科学协同平台 + 云计算 + 数据安全

在大数据产生巨大社会价值的同时，一次次的数据安全事故也引发了全社会的担忧。人工智能竞赛作为数据公开应用的大规模尝试，数据安全的保障更是重中之重。除了采用数据脱敏等传统手段外，通过竞赛工具的技术功能来保障数据安全也日益受到重视。

图 5-10 竞赛工具 3.0 的功能要求



数据来源：和鲸科技（heywhale.com）经调研统计得出。引用请注明出处。

人工智能竞赛各环节的工作内容，展示了数据科学研发的项目项目所需具备的基础能力。**通过数据、云计算、人工智能技术、数据科学人才的连接，人工智能竞赛在帮助赛事主办方探索创新方向、树立技术驱动的品牌形象同时，也在为其在内部积累数据、治理数据、应用数据提供了全流程的借鉴。**

在这个过程中，赛事主办方可以充分利用外部优质资源加速实现技术落地的创新探索，同时积累扎实的数据科学核心能力。

第六章

如何组织一场人工智能竞赛

一、人工智能竞赛的关键要素

二、人工智能竞赛的必备条件

三、人工智能竞赛的八大保障

ARTIFICIAL INTELLIGENCE

一、人工智能竞赛的关键要素

作为整合 AI 技术和应用场景的枢纽，连接数据人才和业务问题的桥梁，优质的人工智能竞赛既要能够激发数据人才的能动性和生产力， 又要满足赛事主办方的主要诉求。前者需要打造良好的参赛体验， 后者又须培养扎实的数据科学运营能力，即优质的人工智能竞赛应同时满足竞赛体验的顺畅度和竞赛成果的有效性。

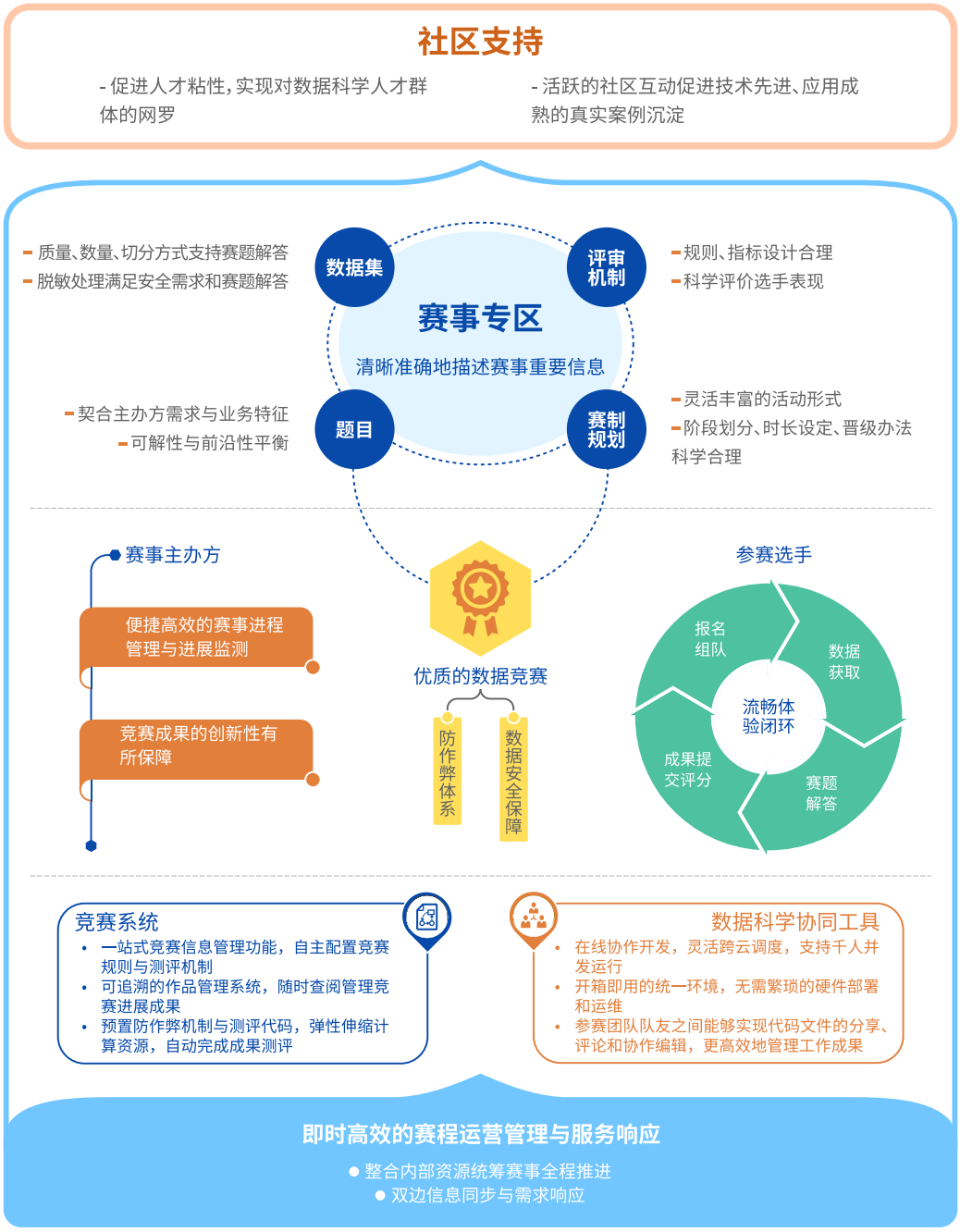
● 竞赛体验的顺畅度

竞赛体验的顺畅度以系统化、专业化的运营手段严格把控赛事质量，以易用的工具、合理的计算资源支撑赛事功能，保障各个环节的平稳推进和过渡，激励参赛选手在公平、公正的竞赛环境下充分发挥能动性和专业实力，高效创造出性能优越的竞赛成果。

● 竞赛成果的有效性

竞赛成果的有效性在充分理解赛事主办方业务特征与办赛诉求的基础上，通过赛题最大化呈现技术在具体应用场景下的价值，将人工智能竞赛打造成为高度整合“数据 + 算力 + 算法 + 应用场景 + 人才”的标杆， 为数字化创新提供借鉴与思路。

■ 图 6-1 优质数据科学竞赛的关键要素



数据来源：和鲸科技（heywhale.com）经调研统计得出。引用请注明出处。

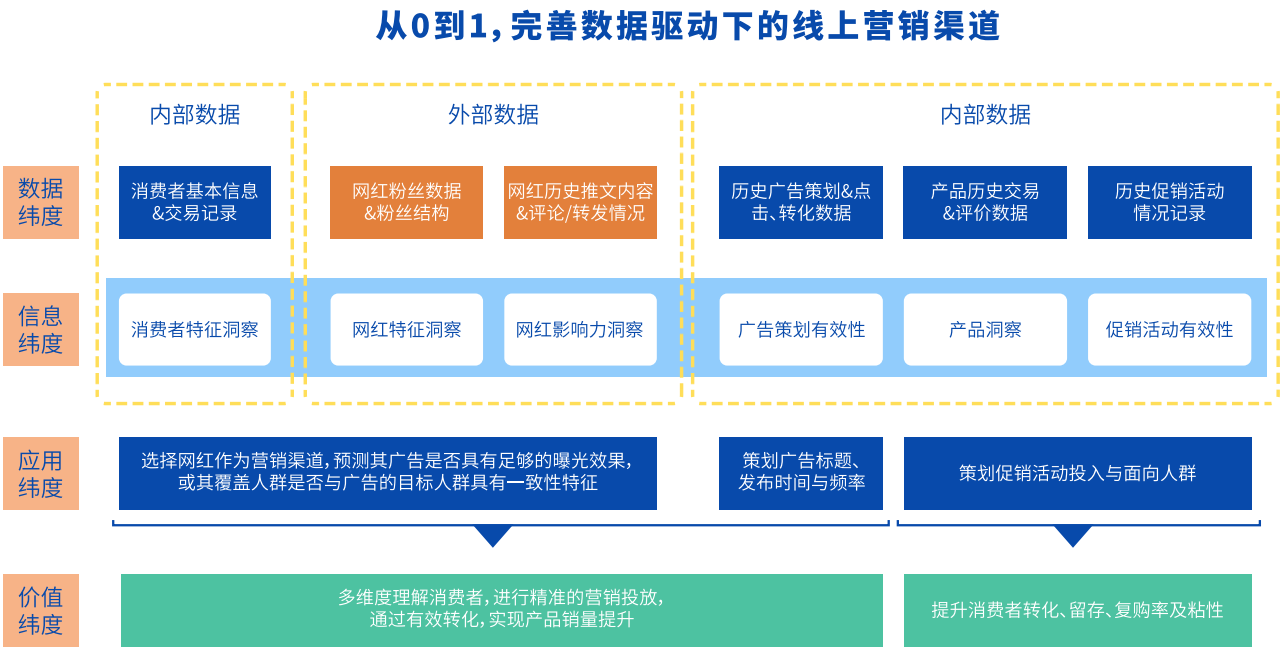
二、人工智能竞赛的必备条件

1. 贴合实际场景的需求分析和赛题设计

为了快捷、高效、准确地寻求数据创新的切入点和赛题方向，竞赛运营平台在日常工作中需要注重对各行业、各领域的洞察积累和研究分析，挖掘各行业的业务特征、发展趋势和价值生态。

在面对具体的人工智能竞赛项目时，结合赛事主办方的实际业务价值和实现风险，快速敏捷地构建数据创新解决建议。

图 6-2 和鲸数据创新解决建议，以快消行业的线上营销网红渠道场景为例



数据来源：和鲸科技（heywhale.com）经调研统计得出。引用请注明出处。

在初步厘清具有较高实现把握的赛题应用场景方向之后，则进入了具体的赛题设计阶段。赛题设计需要考虑如下原则：

- 赛题的未来应用价值在紧凑的竞赛周期内能够基于现行技术实现；
- 赛题与现有数据条件充分契合；
- 赛题具有充分的可解性，又能够保有一定的难度和挑战；
- 赛题以清晰、明确的表述方式向参赛选手予以呈现；
- 赛题配套的测评方法需要综合考量学术指标、生产指标与竞赛效果；
- 引入行业专家共同参与赛题的打磨并对赛题质量提出意见。

案例特写：中国银联高校极客挑战赛——赛题拟定充分贴合实际业务价值

赛事主办方：中国银联

中国银联于 2019 年 6-9 月举办了高校极客挑战赛，希望为 2020 年的校园招聘提前网罗优秀的数据科学人才，同时与高校进行技术碰撞。因此，赛题的拟定需要与银联的业务特征高度契合。最终，竞赛运营平台设计的赛题要求参赛选手基于脱敏和采样后的约 40,000 条用户消费行为数据，预测在未来的一段时间内，用户对于某产品是否会有购买和收藏的行为。相较于大量结构化数据支撑的赛题，这场竞赛着重考察了选手在业务冷启动时对于小数据量的适应和调整能力，而这也符合银联大量新生业务在数据量积累不足的条件下仍赖于通过数据建模提供决策支撑的真实情况。

2. 审慎的数据筹备与质量管控

（1）数据甄选和评估

数据甄选不仅要做到能够从数据源头上支撑赛题的可解性，配合赛事主办方积累数据采集和管理方面的能力。竞赛运营平台需要积累国内外的开放数据资源，形成丰富的数据案例库，以此为基准帮助赛事主办方进行数据采集和管理。

在实际竞赛中，要第一时间面向赛事主办方收集数据样例，深入了解数据采样过程、方法和数据的业务特色，结合赛题的应用场景和技术类型，从字段的完整性、关联性等多个方面系统性地检查数据集质量。

过小的数据量可能不足以支撑赛题的解答，最终开发出的算法模型性能有限，而过大的数据量需要占用巨大的储存空间、消耗海量的计算资源，同时也会使得参赛选手的分析开发工作耗时日久，从而违背了人工智能竞赛经济、高效、敏捷的初衷。

竞赛成果投入真实场景的应用，会面临种种现实约束条件，需要结合数据实际情况预判可能出现的质量风险，并采取合理的手段甄选质量优良、数量合理的数据，提出改善数据质量、或对数据进行加工的方法建议。

案例特写：同盾科技声纹识别建模大赛——数据处理充分考量竞赛成果未来应用条件

赛事主办方：同盾科技

同盾科技希望籍由人工智能竞赛引领语音智能识别领域的工业级算法创新，要求参赛选手开发算法模型来判断任意两段语音出自同一说话人的概率，从而加速声纹识别在金融风控方向的技术落地。考虑到声纹识别模型未来在目标场景中的应用存在各类约束条件，如模型要能在各性别、口音中都能发挥平衡效果，模型对长句和短句均具有较快的识别效果等，因此在为赛题筹备数据时就应予以提前考虑，并对数据做出相应的加工处理。

声纹识别模型未来应用时的约束条件	
场景噪音	在部分音频上人工叠加背景场景噪声
说话人的方言、性别影响	在测试集中等分了方言、性别的语音数据对
模型在说话人长短不一、内容不一的语音对上都能有较好表现	1. 做文本无关（text-independent） 的说话人比对 2. 将有限的音频在兼顾等分的条件下进行合理的人工拼剪，形成数千对长短不一的数据，并与短句音频进行比对
通常深度学习模型在越大规模的数据集上的效果会明显优于经典模型，因此需通过更大规模的数据来再训练，以增强模型	额外补充公开数据来扩充数据量级

在优质数据的充分支撑下，260 支参赛团队最终呈交出了出人意料的优质成果。其中，冠军团队实现了一种端到端的声纹特征学习模型，EER 得分高达 0.0011111 分（越接近零分意味着模型精确度越高），已经达到甚至超越了声纹识别的工业级水平，引发了广泛的行业关注。

扫描上方二维码
了解更多赛事详情

（2）数据集切分

赛事的竞赛数据集由四个部分组成：作为模型开发输入的训练集数据、作为模型开发输出的训练集标签、作为模型测试输入的测试集数据（无标签样本）、作为评估模型的真实测试集标签。

目前普遍的数据集切分方法是按一定比例进行随机切分，这样的切分方式表面看似公允，但没有考虑数据分布特征具有的不确定性。举例来说，如果训练集和测试集过于相似，模型在训练集上的过拟合表现则无法在测试集上被验证发现，最终得分虚高；而如果训练集和测试集的相似度低，则基于训练集数据特征训练出的模型在测试集上则可能有不同的表现。

为了避免随机切分带来的潜在问题，对抗验证（Adversarial Validation）不失为行之有效的办法：

· 首先，采用一系列分析手段来量化判断在不同的切分方式下，训练集和测试集之间的相似度变化；

· 在此基础上，结合实际赛题的业务价值、难度定位、效果预估等多方面因素找到最合适的相似度指标；

· 最后，依据该相似度指标切分数据集，邀请专业人士进行测试，确保切分结果能够保证比赛质量的同时考察模型的泛化性能。

图 6-3 数据集切分示意

数据来源：和鲸科技（heywhale.com）经调研统计得出。引用请注明出处。

3. 科学有效的竞赛成果评价机制

设计科学的评价机制，能够有效评估参赛选手的真实水平和竞赛成果的价值，帮助赛事主办方对于人工智能竞赛的意义有更加直观的感受，也能为数据科学研发的迭代指明方向。从参赛选手的视角出发，在其付出和能力得到客观评价的同时，更重要的是能够在清晰合理的评价反馈引导下更加高效地优化模型性能。

（1）评审机制

评审机制包括评审类型、评审次数和评审反馈。

其中，评审类型包括客观测评和主观评审两大类：

· 客观测评是通过测评脚本进行的自动化评审，评价的是竞赛成果模型在一些既定维度上的客观表现效果；

· 主观评审则会兼顾考查竞赛成果模型的实现成本、在时间和空间上的运行复杂程度、算法思路的创新性等，甚至包括参赛选手最终在决赛路演中是否清晰地传达了自己的解题思路等。

为了保障评审机制的平衡性，需要对评审次数和评审反馈内容进行精细化设计。具体来说，过多的评审次数可能引发刷分现象，过少的评审次数据又限制了参赛选手通过获得反馈来进行成果优化的机会，因此在两相之间需要有所平衡。

评审反馈是指参赛选手提交成果，接受评审后所能获得的反馈信息。比如在客观测评后能够看到评审算法所反馈的分数，或若提交失败，收到报错原因的具体反馈等。其中，报错反馈的设计尤为需要周全与具体，其设计要能尽可能同时给到选手全部的提交报错原因，能提供建设性提示。

（2）客观评审

客观评审是指通过设置考察指标，对参赛选手开发的算法模型的推理效果进行评估。由于客观评审指标不存在主观干预，因此可直接通过编写测评算法脚本，自动完成对参赛成果的评审过程，并输出评审结果。

合理的测评算法需要从**有效性、稳定性和安全性**三个维度出发予以保障。首先保证测评指标的有效性，在此基础上尽可能提高其稳定性和安全性，优先观察安全性指标，一旦低于某个阈值，则停用该指标。

有效性需要考虑测评指标能够充分反应算法模型投入实际应用时的业务侧重点，并能够如实考察选手的模型在该侧重点上的表现效果。

> 示例：
>
> 通过摄像头实时画面识别工人进入施工现场是否佩戴安全帽的算法更着重考察的应是查全率（Recall），即确保尽可能所有未佩戴安全帽的目标都能够被检测到并发出告警，宁肯因为错误识别佩戴了安全帽的目标而发出冗余告警，也要杜绝因为遗漏而导致的安全隐患。
>
> 无人超市的顾客在前往自助柜台结账时，机器会打开摄像头检测顾客是否为数据库内记录的有过逃单行为的顾客，如发现疑似历史逃单客户则会给工作人员发起提示。在这个过程中，算法需更加着重考察查准率（Precision），即尽量不错怪顾客，否则会给顾客带来不必要的麻烦甚至伤及其名誉。

有效性又分为排名有效性和数值有效性。前者衡量的是该指标能否正确地对选手表现进行排名，后者衡量的是选手的排名差距是否合理地反映了选手模型的性能差距。一般而言，测评指标的排名有效性与其数值有效性之间不存在绝对的关系，因此在指定该测评算法的时候需要在满足排名有效性的前提下尽可能地满足数值的有效性。

指标的稳定性是指，当被机器学习中的随机因素干扰时，排行榜体现的分值与排名也不会发生剧烈波动，依然能够准确地反映选手的能力和模型的效果。

通常容易被所忽略的是测评指标的安全性，是指该指标能够尽最大可能确保数据信息不会被泄露，即避免参赛选手在提交次数限制范围内不断提交，从而通过和测评分数比对来推断测试数据的信息，进而实施作弊来提升分数。

（3）主观评审

主观评审需要与客观测评算法有一定的区分度，从而能对参赛选手的表现做出更综合的评价，主要是从竞赛成果的**有效性、创新性和性能**三个维度进行考虑。

有效性是指选手提交的成果模型是完整的，且代码能够完成复现。创新性则考虑了成果模型是否采用了与现有方法不同的创新思路，也是判断其是否具有更多潜在价值的重要依据。

性能是在客观测评算法之外进一步考查其在运行过程中消耗的计算资源、时间，以及技术的普适性和复用性。举例来说，不少选手在竞赛中为了追求客观测评算法的分数表现，会采用大模型的堆叠，从而计算速度有所折扣。

比如一道要求开发模型能实时识别并定位视频画面中人脸五官的赛题，如果仅追求识别的精准度而堆叠复杂的模型，实际运行时的识别和定位过程则需要耗费十余秒而非实时完成，那么其未来的商用价值必将大幅缩水，也难以取得好的成绩。

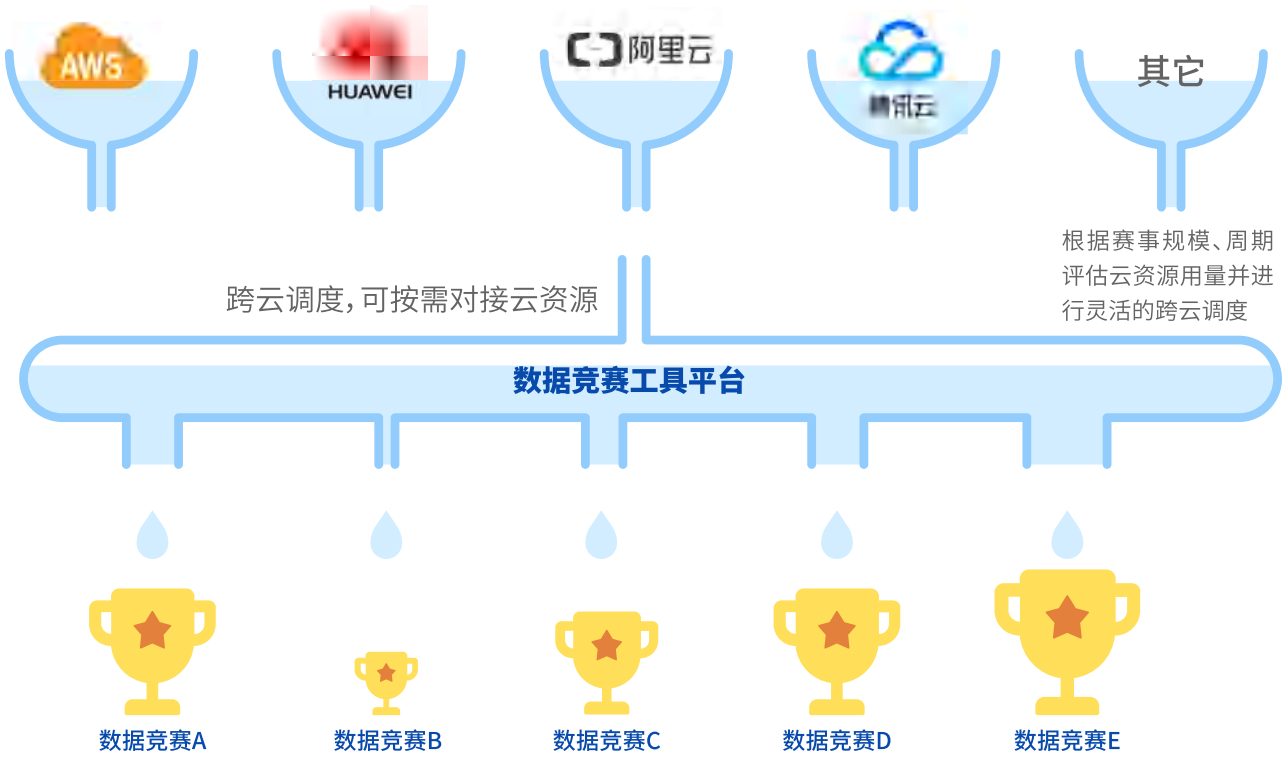
考虑到竞赛成果需要兼具实际应用性和技术先进性，需要在业务背景和技术背景上跨学科领域的专家共同完成主观评审。科学的方法论需要综合考虑不同背景的专家应从哪些指标出发进行评审、以及不同指标的计分规则和权重如何设计。

4. 平衡成本和效率的计算资源管理与配给

云端竞赛环境下需要为参赛选手配给相应的计算资源，在充分平衡成本投入和建模效率的原则基础上，计算资源的预估需要注重从过往办赛经验中总结计算资源使用规律。同时，竞赛运营平台在日常工作中要对各类具体算法问题的最新技术进展、人才发展水平和计算资源要求的变化保持高度关注，力图构建起业界领先的计算资源用量预测模型。在面临具体赛事时，结合赛事主办方对参赛规模、赛事周期的要求，以及赛题的技术类型和难度，对计算资源用量进行合理预估。

考虑到复杂的竞赛环境和赛题内容要求，在实际赛程中，计算资源也需要实现对 AWS、阿里云、腾讯云等国内外各大主流云厂商进行灵活的跨云调度，以防止单一云厂商的算力资源用量无法满足同时段内多场竞赛的算力资源使用需求。在竞赛运行时，会对相应的计算资源进行弹性调度，当用户使用结束后及时自动释放运算资源，从而降低计算资源的浪费率。

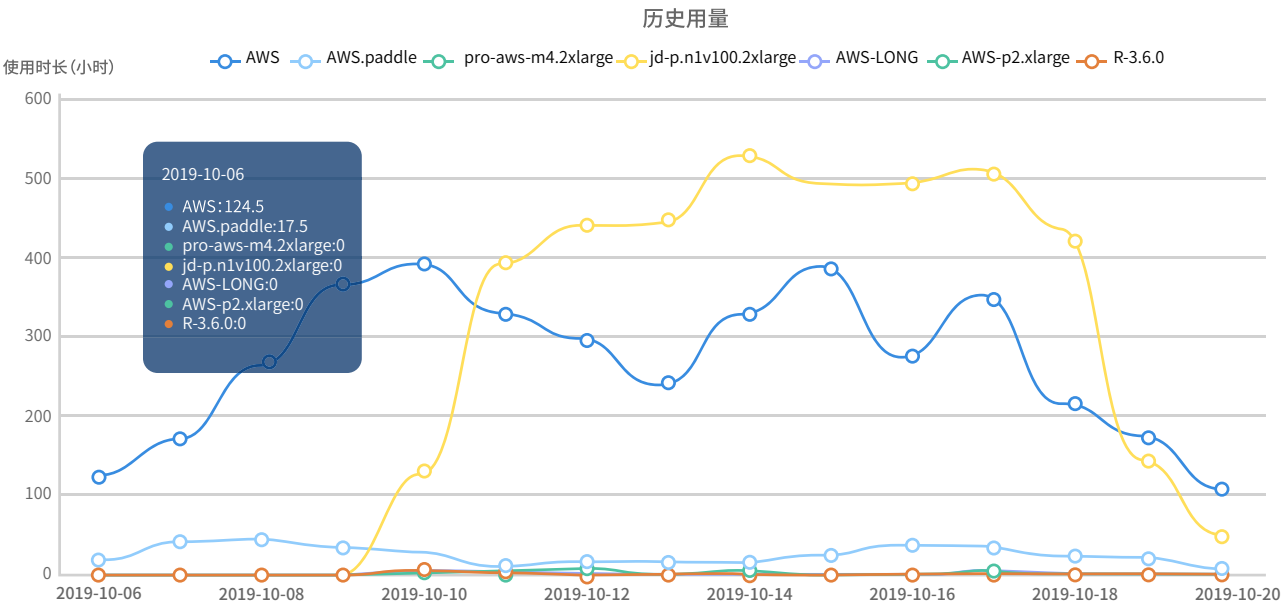
图 6-4 根据赛事规模、周期评估云资源用量并进行灵活的跨云调度



数据来源：和鲸科技（heywhale.com）经调研统计得出。引用请注明出处。

此外，赛事主办方和竞赛运营平台在人工智能竞赛的举办过程中，都能够对算力用量按个人、分时段进行实时监控，对异常使用行为进行及时干预，而历史用量统计数据也能够进一步服务于计算资源用量预测模型的优化。

图 6-5 在赛程中实时监控计算资源用量



数据来源：和鲸科技（heywhale.com）经调研统计得出。引用请注明出处。

在科学的方法论指导下完成以上赛事筹备工作后，建议对全流程进行严格的完整测试，除了内部测试外，联系数据科学领域经验丰富的头部选手和产业专家进行外部测试也不失为理性选择。通过模拟参赛选手视角，在测试环节发现可能出现的问题并予以提前解决，为赛事的顺利进行和成果价值实现打上更牢靠的保险。

三、人工智能竞赛的八大保障

1. 数据脱敏、防泄漏等安全保障

解决赛事主办方后顾之忧的数据安全保障

（1）数据脱敏

在通过数据脱敏来解决数据安全的问题时，需要严格评估在赛题要求下敏感数据是否具备使用价值，对于无建模价值的敏感数据会直接在数据采集阶段予以剔除。

之后的脱敏处理要遵循两个原则：一是脱敏前后数据格式类型不变，二是脱敏不能改变数据的分布。这样才能够保证脱敏后的数据在实际比赛当中仍能保留足够的建模价值、在实际场景也能发挥应用潜能。而在此基础上还要考虑，能在尽可能少的随机性前提下完全隐藏相关的信息，规避参赛选手逆向还原脱敏数据的现象。

（2）数据泄露

部署了数据科学协同平台和计算资源的云端竞赛环境将数据挂载在工具内，竞赛一旦结束，数据自动下线，具有数据安全的保障性。为了进一步严控数据被下载至本地、避免数据泄露，数据科学协同平台可以考虑两种不同级别的方案来克服多数机构和企业既想投身于人工智能竞赛的创新探索、却又担心数据泄漏的顾虑。

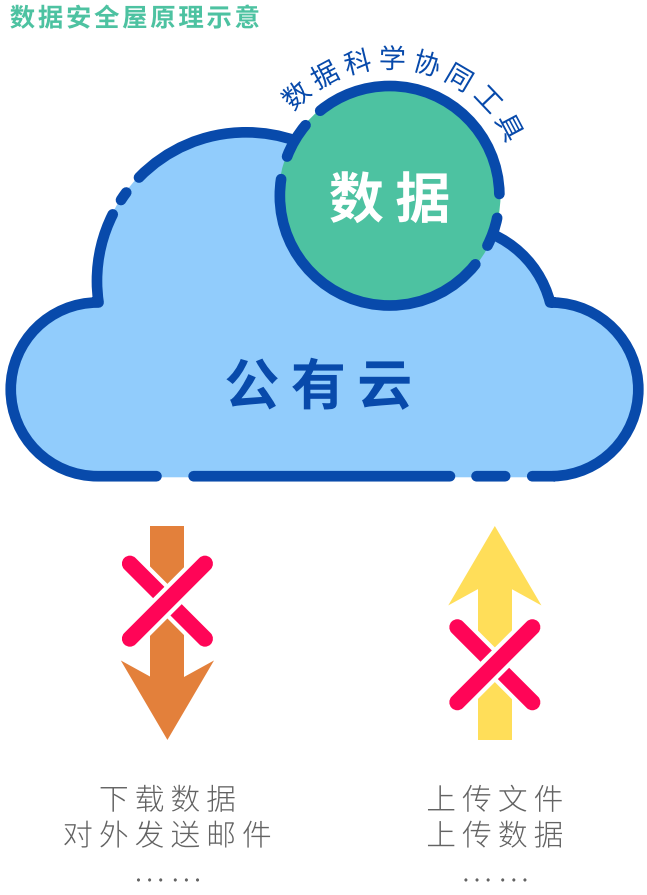
· 监测限制下载流量

在限制开发工具的下载流量并对其进行实时监测的条件下，要将开发工具中挂载的数据集下载至本地需要耗费参赛选手大量时间，在数月的竞赛全周期内最多只能完成不足 5% 的竞赛数据下载，对于选手而言成本高昂。与此同时，竞赛选手依然能够顺畅、自由地向开发工具中安装所需工具包，正常的竞赛体验和开发工作不会受到影响。

· 断网安全屋

断网安全屋的方案对于数据安全的保护更为严格——通过网络控制功能限制所有对外对内的访问流量，如此一来，选手便完全丧失了下载数据的可能性。而为了保障正常的参赛体验，该方案又可以在断网条件下提供特殊访问通道的白名单功能，配置域名或 IP 地址，通过固定出入口的放行使得参赛选手能够安装必备工具。

■ 图 6-6 数据安全屋原理示意



数据来源：和鲸科技（heywhale.com）经调研统计得出。引用请注明出处。

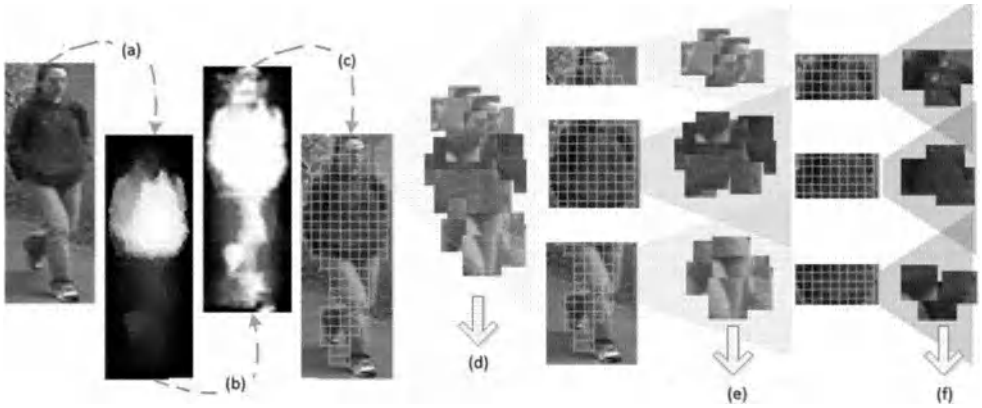
案例特写：快手 - 用户兴趣建模大赛——数据脱敏保障数据安全的同时缓解数据存储压力

赛事主办方：快手

内容生产和分发是短视频发展的两大关键要素。在快手，每天都有数以亿计的视频被播放、被上传、被转发，服务器需要在转瞬间处理海量信息与内容，过多的违规内容会挤压内容生产者的流量，而匹配不精准的内容推荐也是对资源的巨大浪费。在有限的时间内，仅仅依靠人力显然无法高效完成对视频内容的识别、理解，以及对海量用户的精准推荐。可以说，人工智能算法的应用对于快手保持互联网领先身位具有非常重要的作用，因此，快手努力推动着人工智能的创新探索与数据科学的商业应用。

快手在 2018 年举办的用户兴趣建模大赛中，要求参赛选手开发基于视频内容理解的推荐算法，预测用户点击某视频内容的概率。然而，规模巨大的视频数据在竞赛环境中带来了较大的数据存储和算力压力。而利用常用的预训练视觉模型，将中间层的参数输出作为原图像的脱敏数据，则能够有效解决这一问题。

举例来说，下图中的 (a) 为视频某一帧图像的原始状态，能够清晰地看到是一个行走中的人。而 (b) 是提取了图像中行人的身形轮廓特征， (d) (e) 则提取了图像中行人的诸如眼睛、鼻子、下巴等局部特征。通过提取预训练模型的前几层输出，实现了对原始视频的规模降级和脱敏处理，但依然能够判断出图像的主体是一个人。



而在更高的数据安全要求下，还可再进一步作数据脱敏。可将预训练模型的倒数第二层输出作为原图像的脱敏数据，则数据文件大小在进一步降低的同时，也完全无法复原，从而使参赛选手无法知道图像的直观内容，以此实现对图像内容的尽可能保密，既缓解了数据压力，提升了参赛选手的开发效率，还在一定程度上保护了原始视频的私密性。

扫描上方二维码
了解更多赛事详情

2. 公平公正严防作弊的竞赛环境

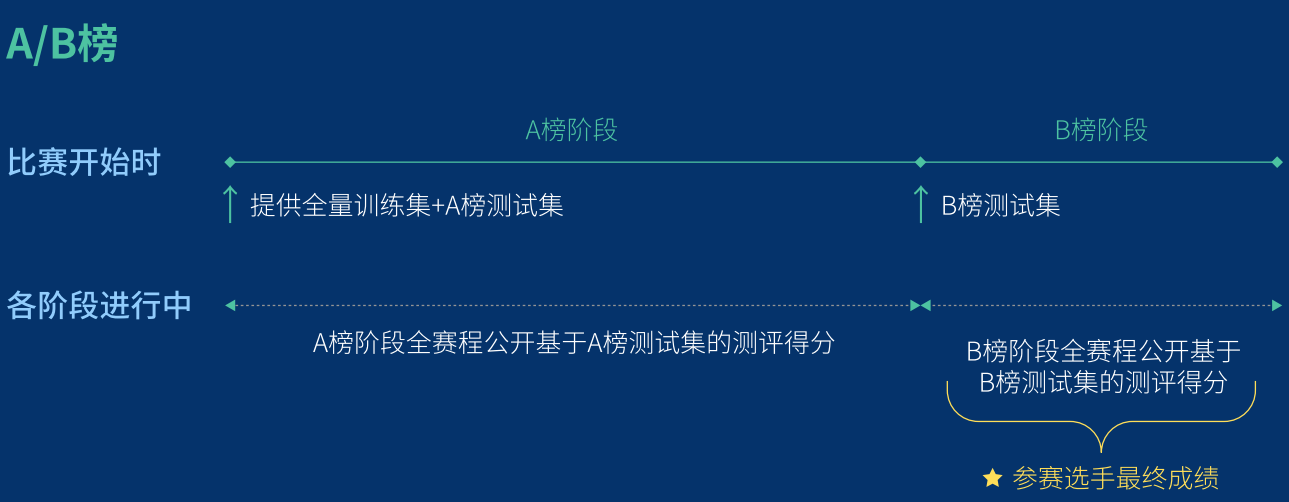
（1）采用多元机制打击作弊行为

投机的作弊行为和不合理的赛制规划也会损伤赛事的公平性和选手的积极性，进而破坏竞赛成果的有效性和真实价值，优质的人工智能竞赛需要在反作弊机制方面进行有效的设计。

通过限制参赛选手的提交次数，同时追踪选手的提交行为可以有效解决刷分的问题。具体来说，可在竞赛之初为每支参赛团队的答案提交动作配置专属 token（计算机身份认证令牌），从而方便追溯提交文件的来源。如某份答案文件的提交选手 ID 与其所属团队的提交 token 不一致，则有理由怀疑出现了窃取答案的情况，需要进行深入调查。

双榜机制在 Kaggle 等多个平台和国际性赛事中得到了不断应用和验证，除了能应对刷分现象外，还对排行榜探查、过拟合等行为具有显著的打击效果。

图 6-7 双榜反作弊机制



数据来源：和鲸科技（heywhale.com）经调研统计得出。引用请注明出处。

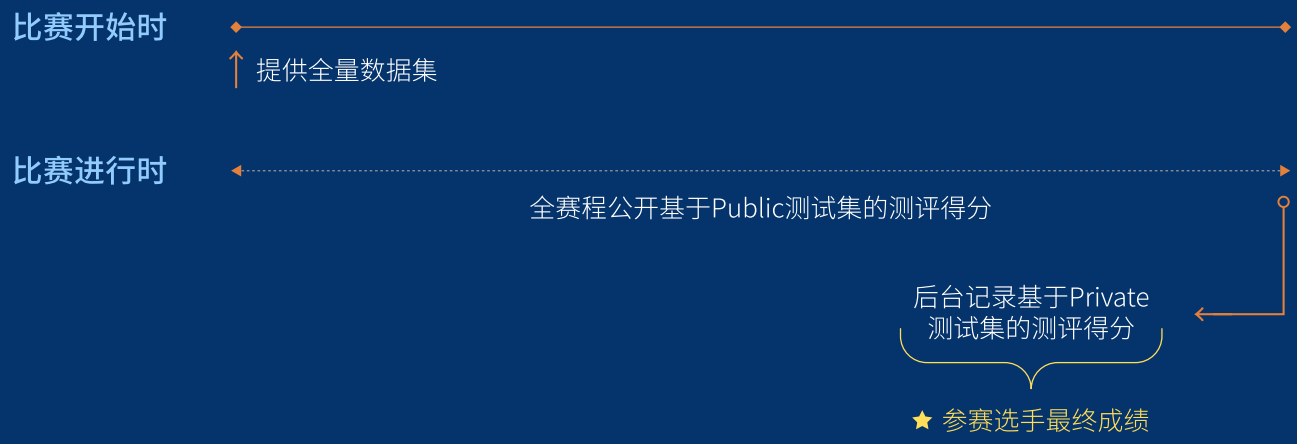
· A/B 榜 - 规避利用过拟合模型上分

目前，A/B 榜是业内公认较为有效的防作弊手段之一，其借鉴了数据科学当中对模型进行交叉验证的思路，为了防止模型在某个数据集上过拟合而失去必要的泛化性能。

运用在比赛中，就是将数据集分别切分为对应 A 榜和 B 榜的两份，在不同阶段陆续提供给参赛选手，以 B 榜的结果作为验证和衡量模型实际表现的标尺。如果选手在先期的 A 榜阶段过分追求分数表现而采用了与 A 榜数据集过拟合的模型，甚至人为手标答案，则会有在 B 榜阶段遭遇成绩下滑。

6-8 双榜反作弊机制

Pulic/Private榜



说明：竞赛平台将测试集切分为Public、Private两份，但并不对外公开切分情况。
选手仅能看到基于Public测试集的测评得分，但并不知道这部分数据集的具体内容。

数据来源：和鲸科技（heywhale.com）经调研统计得出。引用请注明出处。

· Public/Private 榜 - 升级机制，模型效果更有保障

Public/Private 榜机制源来已久，经过多年的迭代和比赛的验证成为了海内外赛事圈选手共同认可的评审机制——参赛选手在比赛伊始便被授权使用全量数据集进行模型训练，但实际上全量测试集已经被竞赛运营平台私下分割为了 Public 和 Private 两部分，但分割方式对外保密。后续在比赛进程中，选手能够看到的公开成绩仅仅是基于 Public 部分测试集测评得出的结果，而属于 Private 部分测试集的得分则会被记录在后台，台，并作为最终排名的依据。

Public/Private 榜机制对 A/B 榜存在的一些问题进行了修正，在此机制下，全量数据集在比赛伊始就被公布，选手有充裕的时间研究和分析数据集的规律和数据特征的性质，从而构建出性能更加优越的算法模型。此外，由于选手并不知道 Public/Private 榜各自对应的测试集是如何分割的，为了追求分数表现而过拟合的情况也得以遏制，有更强大的动机来增强自己模型的泛化性能，使其更加贴近真实应用场景。

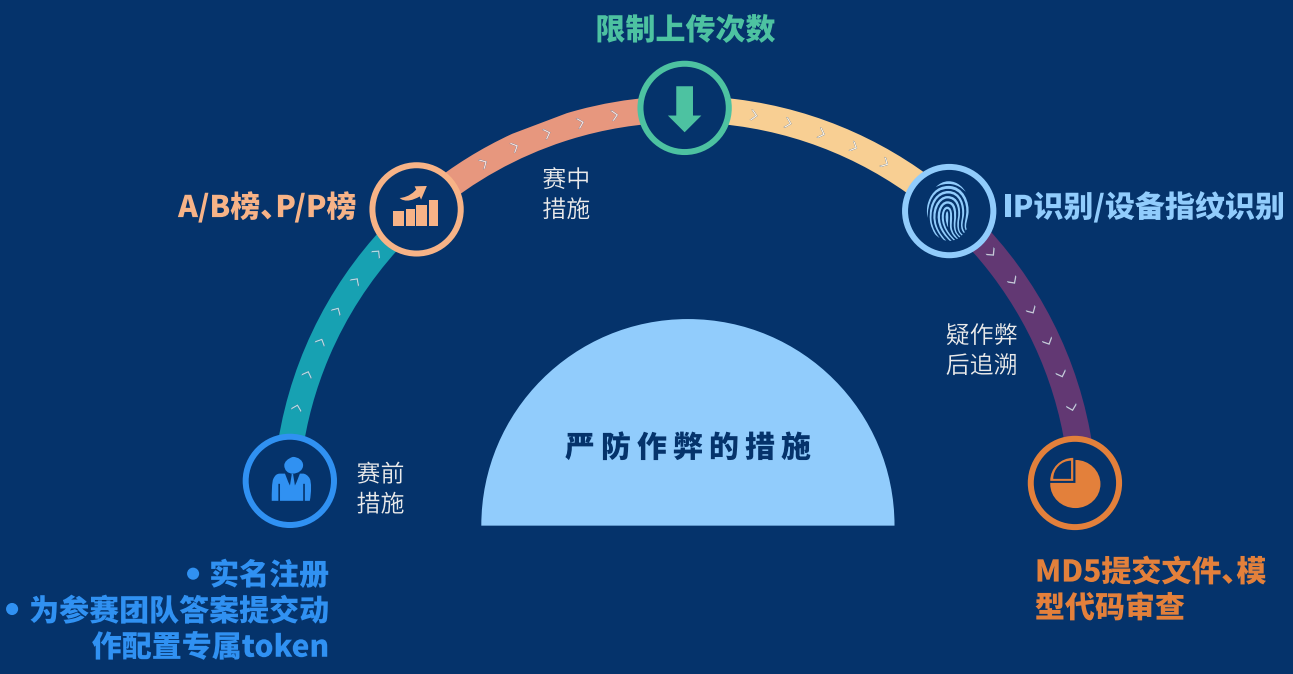
> A/B 榜存在的问题：
>
> B 榜测试数据在比赛靠后阶段才提供→选手难以了解数据全貌，易在 A 榜上过拟合
>
> 最终排名以 B 榜为准，为避免针对性调优、手标答案，B 榜阶段窗口期只有 24-48 小时→选手的参赛体验遭到破坏

（2）以严谨的成果审核作为双重保险

高度相似的成果文件背后可能是参赛选手间共享答案的不端行为，而这一作弊现象在提交答案文件 MD5（每份成果文件独有的 128 位散列值，用于确保信息传输的完整一致）相似度分析下则会被溯源。

在答案文件之外的另外一重保障是要求参赛选手提交模型代码。一方面便于从包含数据特征处理、特征构造、训练策略在内的多个方面进行代码相似度对比，考察选手是否不正当地使用了他人的代码，另一方面也可人工审核代码并复现。

图 6-9 全流程保障赛事环境的公平性



数据来源：和鲸科技（heywhale.com）经调研统计得出。引用请注明出处。

3. 参赛选手享有流畅的闭环体验

在人工智能竞赛中，参赛选手需要先后或多次经历报名组队、数据获取、赛题解决、成果提交与评分几个重要环节。优质的人工智能竞赛在制定比赛规则时应综合考虑初赛、复赛、决赛等每个阶段的里程碑目标、持续时长和晋级要求应如何设定。

竞赛运营平台自身亦需要充分发挥数据的价值，结合过往办赛经验，针对不同的赛题类型与配套技术的发展程度、参赛规模、成本限制等多方面因素，从而对比赛规则设计思路进行持续的迭代，以此满足在多元竞赛需求下的赛事公平性。在此基础上，通过设计流畅的选手体验闭环使得选手的生产力得到更高效、更专注的释放。

（1）报名组队

· 竞赛信息获取：竞赛页面对赛事背景、赛程规划、赛题要求、竞赛数据、测评机制、奖励安排、参赛须知、参赛者权益说明等竞赛基本信息具有清晰的结构呈现、准确的语言表述；

· 报名组队：具有明确、快捷的报名组队操作功能，同时能够快速索引潜在的组队成员，并与之直接在竞赛平台上进行即时交流或发送组队邀约。

（2）数据获取

· 竞赛数据与赛题内容高度契合，能够支撑赛题的可解性；

· 竞赛数据处理完备、字段完整、信息准确；

· 训练集和测试集的切分科学合理；

· 竞赛页面对数据情况有准确到位的描述，并提供数据及字段样例供选手参考；

· 具有专属的竞赛数据分发方式，便于选手高效获取。一般数据规模较大的竞赛都会直接将数据挂载在统一的云端竞赛环境中，选手无需费时下载，可通过在线且体验良好的数据科学协同平台直接读取并使用。

（3）赛题解决

· 赛题解决有赖于团队成员间的密切配合和高效协作，配置能够突破本地环境限制的数据科学协同平台和合理的计算资源将可以大幅提升参赛选手的解题效率，从而为优质成果的诞生奠定基础；

· 以在人工智能竞赛环境中的 Kaggle Kernels 为例，这个可在浏览器中直接运行的交互式编程笔记本，能够快速线上调用 Kaggle 平台上数千个数据集，通过适配谷歌云服务和预装必要的工具包，保证所有参赛选手无需受限于本地配置环境即可在统一的环境中进行成果开发；

· 优质的人工智能竞赛对于在线数据科学协同平台与计算资源的具体要求将在下一小节中展开详述。

（4）成果提交与评分

· 提交页面配备了清晰的提交指引、详细的评审算法与评审机制说明；

· 成果文件能够快速上传至竞赛系统，且能够即时得到清晰明确的报错信息反馈，如在竞赛系统中内置格式检查工具，对选手提交的成果文件进行自动化的格式检查并给出反馈；

· 竞赛涉及多种类型的技术问题，因此也需要支持多种格式和大小的成果文件，如对于较大的文件，最好可提供特定的提交工具支持断点续传，从而不会发生网络环境不稳定导致的多次提交不成功的挫败感；

· 得分反馈：竞赛系统配备科学合理的自动测评算法，能够尽快运行、即时出分，同时提供清晰合理的反馈说明。

图 6-10 参赛选手的重要体验环节

数据来源：和鲸科技（heywhale.com）经调研统计得出。引用请注明出处。

4. 云端竞赛环境突破本地配置局限

提供云端竞赛环境突破本地配置限制

针对机器学习问题，提供充足的数据和算力能带来比较可观的模型效果提升。但这并非长久之计，在大多数业务情况下，其数据积累和治理程度不足，无法形成海量可用数据，大规模的算力也是高昂的成本。因此，在一定算力消耗的条件下鼓励敏捷的算法创新，能够带来更大的经济效益，更具可行性和持续性。

拥有大规模算力条件的参赛选手，可能更容易取得优异的排名成绩，但这与“创新”和“公平”的初衷相违背。最终的比拼变成了谁掌握了更多的计算资源，而不是谁在算法层面诞生了更多突破性想法，而且，这样得到的算法模型通常在实际生产环境中可用性也较低。因此，在云端的竞赛环境、研发环境成为了重要的支撑。

云端统一竞赛环境的另外一层作用在于便捷的成果管理。特别是在进行团队协作时，如本地开发环境或工具版本不统一，团队成员间进行成果分享时则需要重新安装和运维计算环境，费时费力。在人工智能竞赛中提供统一的计算环境使得开发代码的分享、协作更加便捷，开发成果的复现效果也能够得到平等地呈现，促进赛事成果的高效产出与有效落地。

5. 数据科学平台与赛事管理系统的应用

以产品化功能赋能赛事体验

（1）提升效率的在线数据科学协同平台

人工智能竞赛作为数据科学研发的敏捷型项目，在该环境下配置数据科学协同平台和云计算资源，不仅能够保障赛事体验的各项功能需求，也是对实际数据科学研发与协作的一次真实演练。

通过高效打通数据团队管理者、算法模型研发工程师、数据分析师、业务应用人员，搭载了云计算资源的数据科学协同平台，深度参与了从任务管理、模型开发、成果管理到业务应用的数据项目全生命周期管理，可满足数据科学家、人工智能工程师、商业分析师等数据科学工作者在线完成算法建模、数据分析与可视化、结果输出等任务，并支持私有化部署和云端协同，帮助企业、高校、科研机构开展工业级数据科学应用与 AI 研发。

优质的人工智能竞赛正需要具备如下各项功能的数据科学协同平台来**确保数据科学项目的健康推进，最终的开发成果可被复现。**

- 集成体验良好、统一的云端开发环境，提供开箱即用的计算环境，无需繁琐的硬件部署和运维；
- 配置丰富的编程语言工具包和（如 Pandas、Numpy、Scipy 等）、机器学习框架（如 Scikit-Learn、Tensorflow、PyTorch 等）、以及常用的分析代码片段和自定义代码库，方便随时调用和日后复用；
- 用户可以在计算环境中自行安装额外所需的工具，并可持续复用，成为自定义的镜像环境；
- 竞赛运营人员可直接按照工具内提前规划的操作要求和指引，在赛前完成数据及资源的挂载、部署，同时便捷地进行版本管理，亦可方便参赛选手无须耗时将数据下载至本地。
- 参赛队友之间能够实现代码文件的一键分享、评论和协作编辑，降低协作门槛；
- 能够对不同版本的开发策略和代码进行敏捷的效果试验，更高效地管理工作成果；
- 为选手提供统一充分的算力，资源全面支持 AWS、阿里云、腾讯云等国内外各大主流云厂商，根据竞赛需求进行灵活弹性的跨云或混合调度；
- 成果的提交工具与竞赛系统无缝衔接，在提交处查看历史提交成绩与对应文件，把握每次提交的试验效果，为优化开发提供指导性反馈。

图 6-11 数据科学开发工具

数据来源：和鲸科技（heywhale.com）经调研统计得出。引用请注明出处。

（2）提升赛事管理便利性的竞赛系统

人工智能竞赛具有较强的技术专业性，需要专业的数据科学方法为主办方提供有保障、可预期的支撑服务，竞赛运营方也需要最大化主办方的参与度，打造透明的竞赛环境。这关系到赛事主办方对于人工智能竞赛价值的信任，也能帮助赛事主办方通过人工智能竞赛感受数据科学研发过程、积累数据科学研发的管理经验和能力。为了实现多元化的运营目标与管理目标，打造一款具有完善功能的竞赛系统则成为了优质人工智能竞赛的标配：

· 竞赛流程管理

提供一站式的竞赛信息管理功能，赛事管理员可根据赛制规划配置比赛基本信息描述、时间阶段划分、成果测评代码等。同时也可对比赛过程中产生的所有数据、报名信息和提交记录等随时进行调取查看；

赛事管理员可根据赛制要求差异化定制多轮提交、多文件提交等多项延展性功能。

· 客观评审

在竞赛系统内预置 A/B 榜、Private/Public 榜等仿作弊机制，赛事管理者可在管理后台直接调用；

- 除支持竞赛管理员使用自定义配置评审脚本和答案文件外，竞赛系统内能够预置常见的评审指标，并配有完整的使用说明；
- 竞赛管理员可直接在竞赛系统后台上传待测试的提交文件，来直接对当前的评审配置进行内部测试操作；

支持多种格式和大小的成果文件，包括表格数据、图像甚至视频文件的自动测评，因此需要具有较强的负载能力；

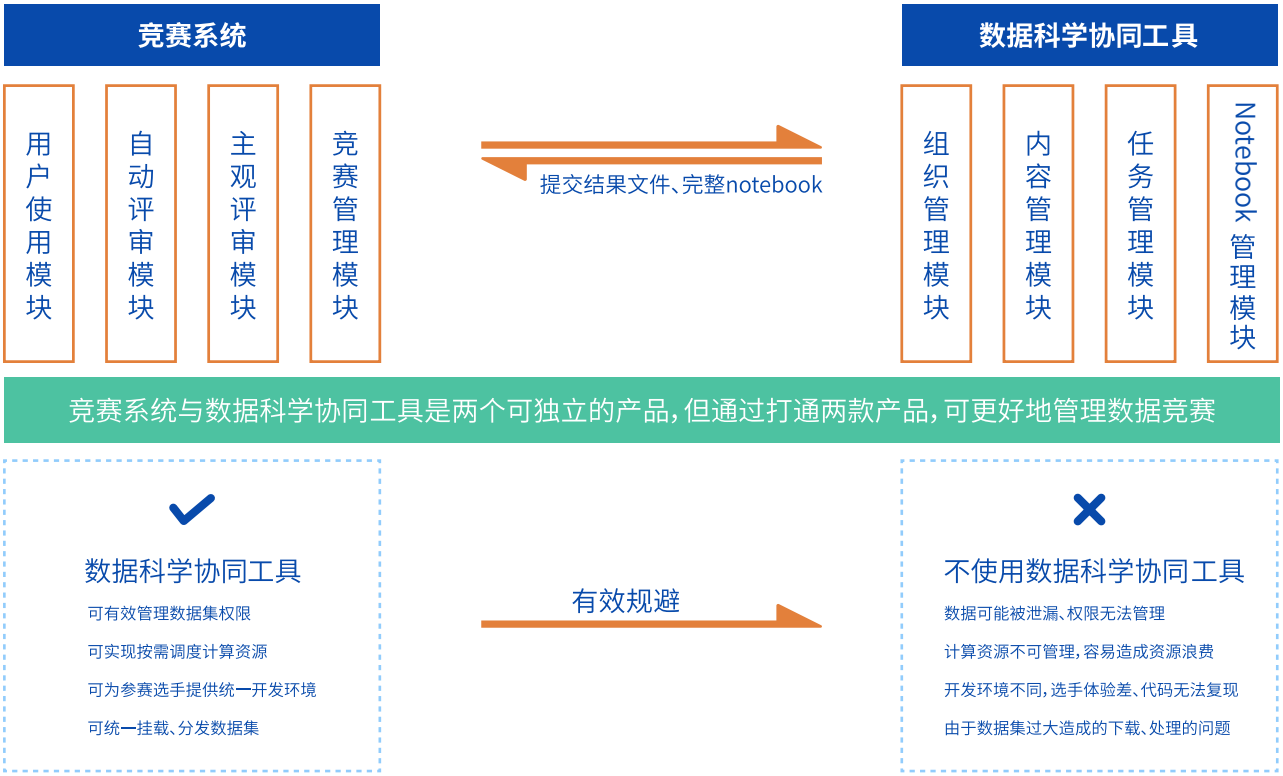
竞赛系统亦可根据评审负载弹性伸缩资源池，根据评审脚本运行时间、同时参与评审人数和提交文件的大小等维度自动计算评审复杂度并匹配相应的计算资源数。

· 主观评审

评委和竞赛管理员可直接查看选手提交的完整代码及对应的客观评审得分，并一键复现选手成果；

竞赛管理员可自行决定参赛成果的评委分配，并自主设计评分维度和规则。如可支持随机分组、集体评审的方式进行合理评审，或根据评委背景分成不同的评审组，并设置不同评审纬度和指标，最终对不同组评审成绩进行汇总。

图 6-12 数据科学协同工具与竞赛系统的配合效果



数据来源：和鲸科技（heywhale.com）经调研统计得出。引用请注明出处。

6. 即时高效的赛程运营与服务响应

即时高效的赛程运营管理与服务响应

作为整合资源的角色，赛程运营管理连接起了人工智能竞赛主办方、赛事运营平台内部团队、参赛选手、计算资源供应商、外部媒体等一系列利益相关方，来推动赛事按预期计划平稳前进。

赛程运营管理人员需要具备复杂任务的多线程处理能力和沟通协调能力，以专业的规范标准为赛事主办方和参赛选手双边提供即时、稳定、优质的服务，促进双边的共赢协作，协调内部资源高效解决赛程中存在的各项问题，并对可能存在的各项风险进行充分的预判和提出解决方案。

面向赛事主办方，赛程运营管理人员主要负责收集确认其需求，拆解、转译成分子需求，并分配给竞赛项目小组的相关成员执行具体的实施工作，并在此过程中与主办方保持紧密的沟通汇报，保证产出与其需求的一致性。内容包括但不限于：

- 在赛前协调品牌营销人员负责赛事形象的整体设计和赛事的宣传推广；
- 统筹需求分析、数据科学服务、竞赛产品各职能相关负责人员和计算资源供应商按约按时、保质保量提供服务；
- 在赛中对阶段性的进展与成果向赛事主办方定期进行同步汇报；
- 在赛后及时整理和移交竞赛成果，促进顺利结项；
- 贯穿竞赛全生命周期的成本管理、质量监控的风险防范。

面向参赛选手，赛程运营管理人员需要主动将赛程各阶段的信息进行及时的同步传达，帮助赛事主办方在选手群体中树立并传播良好的品牌形象，同时通过积极的沟通加强选手对赛题的理解和研发的积极性。另外，赛程运营管理人员也要对选手在参赛过程中遇到的疑惑和问题保持高度关注，提供及时的解答和处理，对可能发生甚至已经出现的舆论争议进行及时疏导，并适时组织有助于提升赛事质量和体验的调研活动。

图 6-13 赛事运营平台需要在赛事主办方和参赛选手间进行高效的信息和价值传递

数据来源：和鲸科技（heywhale.com）经调研统计得出。引用请注明出处。

7. 多种针对性的辅助活动保障

在实际的人工智能竞赛中，时常会出现一些意料之外的情况，需要具有灵活的应变能力。比如赛题内容所牵涉的行业场景过于细分或前沿，行业知识的缺失会导致技术能力难以施展。

以 PLAGH-MIT 医疗大数据 Datathon 为例，活动正式开始前，组委会通过提供线上讲座、视频课程、跨学科交流来帮助潜在参赛选手学习如何更好地理解和分析医疗数据，加强医务工作者和数据科学家之间的沟通协作。同时，在正式开赛前预置训练营则也能够通过邀请赛题领域内的资深人才提供竞赛基准模型（Baseline），为潜在选手提供指导借鉴和思路启发。

案例特写：中国高校计算机大赛大数据挑战赛——预选赛聚焦优质选手、分流办赛压力

赛事主办方：教育部

2018 年中国高校计算机大赛大数据挑战赛参赛人数突破 3000 人，2019 年的规模将更为盛大，运营压力和成本压力陡增。广大选手虽技术水平不可同一而论，但也都希望借由这次宝贵的机会得到锻炼。为了平衡各方诉求，2019 年的赛制在 2018 年的基础上进行了优化，在正式赛程上线前举办预选赛，利用公开数据集出具与正式赛题相关但难度更小的“文本情感分类”赛题。如此一来，新手们能够得到充分的练习和交流，预选赛设置的评分规则也能选拔出能力更突出的选手进入正式赛程环节，实现聚焦和限流的作用。

扫描上方二维码
了解更多赛事详情

案例特写：同盾科技声纹识别建模大赛——训练营克服专业壁垒，提升赛绩表现

赛事主办方：同盾科技

受制于真实应用场景下的复杂情形和算力限制等因素，语音识别目前仍面临着许多技术瓶颈。在筹办同盾科技声纹识别建模大赛时，考虑到该赛题技术难度偏大、细分领域人才偏少，可能出现竞赛参与度低、竞赛成果不达预期等潜在风险。

因此，赛事运营平台设置了训练营作为赛前先锋，在有充裕时间酝酿赛事影响力，吸引更多关注参与的同时，采取公开报名和免费参加的形式为意向参赛选手提供语音深度学习相关的知识培训，由专业的科研学者提供指导，以实践任务的方式指导和启发学员解锁部分声纹识别的技术难点。训练营上线后迅速吸引了大量数据人才的关注，并为后续赛事运营的顺利推进奠定了良好的基础。最终产出了工业级水平的竞赛成果。

扫描上方二维码
了解更多赛事详情

8. 数据科学人才社区提供人才保障

促进人才沉淀的社区运营

人工智能竞赛具有天然的社区效应，越来越多数据科学人才选择参赛以实践理论技能，并在过程中进行持续的协作、交流和讨论，形成了具有专业属性的社区雏形。

以全球最大的人工智能竞赛平台 Kaggle 为例，作为全球最大的数据科学人才社区，在与不同机构组织合作举办人工智能竞赛之外，Kaggle 还以数据集、开源代码项目分享等丰富的内容吸引人才汇集，在其数据科学的成长发展之路上不断进行助力和赋能，从而使得海量处于不同成长阶段的专业人才均对其产生了强大的粘性。

数据科学社区集聚的人才来自不同行业不同职能，能够及时响应复杂多元的竞赛需求，充分输出“数据 +”的洞察，为人工智能竞赛的成功举办提供了复合能力的全方位支持。活跃的互动交流也在社区上不断沉淀着技术先进、应用成熟的真实案例，突破固有认知的创新源源不断地发生，进而为人工智能竞赛基础上的数字化创新提供了底层支持。

图 6-14 Kaggle 数据科学人才社区的运营内容

数据来源：和鲸科技（heywhale.com）经调研统计得出。引用请注明出处。

第七章

覆盖多领域的人工智能赛事案例

赛事导航

一、政府赛事

二、科研赛事

三、产业赛事

四、高校赛事

ARTIFICIAL INTELLIGENCE

赛事导航

政府赛事

- 南京市人民政府 - 2018 全球（南京）人工智能应用大赛
- 深圳市人民政府 -2019 首届“全国人工智能大赛”(NAIC)
- 北京市医疗保障局、北京市经信局 -2020 北京数智医保创新竞赛
- 重庆市江津区人民政府 - 2020 数智重庆·全球产业赋能创新大赛
- 福州市人民政府 - 2020 数字中国创新大赛 - 数字政府赛道
- 北京市经信局 - 2020 北京数据开放创新应用大赛暨科技战疫·大数据公益挑战赛
- 四川省大数据中心 - 2021 数字四川创新大赛
- 深圳市人民政府、鹏城实验室 -2021“全国人工智能大赛”(NAIC)

科研赛事

- CCF BDCI
- NLPCC
- CCKS 竞赛
- KDDCup 2021
- KDDCup 2020
- KDDCup 2019
- NTIRE
- TREC
- IJCAI 2021
- Face Anti-spoofingChallenge

产业赛事

- 2021X-GAME 上海新能源汽车大数据竞赛 2021 中国大学生保险数字挑战赛
- 好未来 - 中国创新挑战赛智慧教育专题赛
- 中国南方电网公司 - 数字电网开发大赛
- 英特尔创新大师杯
- 中国造船工程学会 - “海上争锋”中国智能船艇挑战赛
- 易华录 - 江苏大数据开发与应用大赛
- 腾讯 - 广告算法大赛
- 极市角科技有限公司 - ECV 极市计算机视觉开发者榜单大赛
- 达观数据 - 达观杯

高校赛事

- 2018 中国高校计算机大赛——大数据挑战赛
- 2019 中国高校计算机大赛——大数据挑战赛
- 2021 中国数字人文开放创新研究大赛
- 清华大学 - 安全 AI 挑战者计划
- 上海交通大学 - 2020 第二届华为云人工智能大赛·无人车挑战杯
- 大连理工大学 - 全国水下机器人大赛
- 未来杯高校 AI 挑战赛
- DIGIX 全球校园 AI 算法精英大赛
- 海华 AI 挑战赛

一、政府赛事

南京市人民政府 - 2018 全球（南京）人工智能应用大赛

作为国内顶级规模的人工智能赛事，2018 全球（南京）人工智能应用大赛推动了产、政、学、研各界的全面联合，因其社会价值之高、调动资源之多、影响范围之广引发了各界的高度关注。南京市人民政府联合 2 家科研机构、3 所高校、15 个企业共同发布了覆盖“智能制造”、“智能驾驶”、“智能生活”、“智能医疗”、“智能城市”五大产业应用领域的 20 道赛题，设立 600 万元奖金池及 20 亿元人工智能产业风险投资基金，吸引了近 3000 名数据科学人才的参与。

和鲸社区承办由南京市政府主导的 2018 全球（南京）人工智能应用大赛，联合产、政、学、研、资各界机构组织针对五大产业应用领域方向共计发布 20 道并行赛题，所有赛题从解决产业实际困难出发，充分体现 人工智能 技术在各行各业的深度应用赋能价值，一批提高生产效率甚至改变商业模式的解决方案，从解决方案，从本次大赛中涌现。

智能城市

- 华南理工大学
  采用深度学习技术实现架空输电线路无人机缺陷检测
- 运满满
  公路干线物流区域货源供需关系预测与区域间调度
- 新狮科技
  红外视频图像中移动目标识别
- 亚信科技
  复杂业务场景下的安全身份认证算法

智能驾驶

- 智行未来汽车研究院
  无人车在复杂路口的环境感知
- 清研微视
  驾驶人倦意度及不良驾驶行为高精度辨识
- 新道格自控科技
  水面无人艇智能避障系统
- 行车宝智能科技
  基于视觉技术的车道线检测

智能生活

- 旷视科技
  多目标跨摄像头跟踪
- 南京理工大学
  基于图像的环境空气质量评价模型构建
- 端网云捷智能科技
  计算机视觉技术在无人零售（衣物）中的检测和识别应用
- 北京守敬应用技术研究院
  适应特定场景的行人跟踪再认证机制及其技术方案

智能医疗

- 筑康医疗科技
  肺部结节智能诊断
- 空天地科技
  实时监测心血管疾病智能算法
- 氪信
  人工智能社会保险反欺诈分析
- 土星信息科技
  牙根CBCT影像智能图像分割

智能制造

- 亿嘉和科技
  电力巡检机器人表计识别算法
- 天泽信息
  基于机器学习的工程机械设备故障预测系统
- 华捷艾米
  利用深度信息和VSLAM解决机器人的定位导航避障和交互
- 哈尔滨工业大学
  先进移动协作机器人技术

通过这项赛事的成功举办，南京市人民政府向外界清晰地传达：南京市人民政府对于数字化转型人工智能产业的支持力度是空前的，通过提供坚实有力的扶持政策和对接优质资本，南京有能力、有魄力、有信心把握人工智能时代的发展趋势，以更懂科技企业的方式聚集人工智能人才、提供配套设施服务、助推人工智能全面发展。

扫描上方二维码
了解更多赛事详情

深圳市人民政府 -2019 首届“全国人工智能大赛”(NAIC)

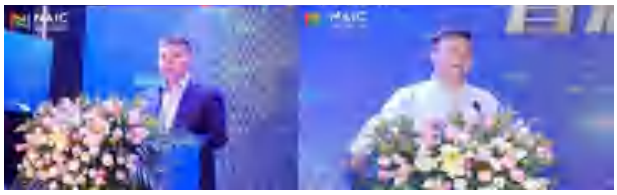
规模空前的国际化人工智能大赛推动产业创新生态闭环。人工智能应用于 4K/ 高清视频处理是一个非常有前景的领域，而 4K/ 高清视频是 5G 时代的主流应用之一。深圳市人民政府专门设立人工智能领域权威赛事——全国人工智能大赛（以下简称大赛）。大赛将立足国际视野，营造人工智能创新创造氛围，促进产业、学术、资本、人才等创新要素融合发展。大赛由深圳市人民政府主办，深圳市科创委、鹏城实验室及科技部指导成立的新一代人工智能产业技术创新战略联盟作为共同承办，多家企业与学术机构共同协办。以“AI 赋能视界”为主题，着眼于视觉处理领域的

前沿技术难题，连接海内外优秀人工智能人才，在真实的实践机会中，共同探索技术的创新突破，加速推动前沿技术在业务场景

中的落地，促进打通人工智能先进技术从科学创新到落地应用的完整闭环。

本次大赛由和鲸社区承办，吸引来自 13 个国家和地区、376 所高校、800 余家业界公司共 3000 余支团队、4000 余人报名参赛。

大赛影响力及规模位列 2019 年各类人工智能大赛之首。据了解，两大赛项共设置 536 万元奖金，其中冠军团队获得 100 万元奖金，是今年国内总奖金和一等奖金额最高的人工智能赛事。



获奖选手获得由鹏城实验室和顶尖人工智能企业开放的招聘绿色通道，以及科研经费、人才政策、研发空间等全方位支持。

扫描上方二维码
了解更多赛事详情

北京市医疗保障局、北京市经信局 -2020 北京数智医保创新竞赛

为推进信息技术在医疗保障领域的应用创新，推动医保管理与高新技术的协同发展，夯实智慧医保应用基础，帮助提升首都医疗保障精细化、智能化管理水平，结合现阶段医保信息化建设工作实际，以“创新融合、开放共享、安全平稳”为工作原则，北京市医疗保障局、北京市经济和信息化局、北京市科学技术委员会、中关村科技园区管理委员会将在北京市联合举办首届“北京数智医保创新竞赛”，主题为“创新北京、智慧医保”，英文 Slogan 为“Data enabled medical insurance”，简称“Beijing DEMI”大赛。意为此次医保竞赛通过比赛形式征集“高精尖”创新单位及前沿产品，以政商合作促进医保与高新技术融合发展。

本次大赛以党的十八大和十九大精神为指导，立足北京“四个中心”定位，坚持以提升医疗保障服务能力、改善民生保障为主线，全面发挥首都科技中心优势及医疗保障数据资源优势，深度挖掘医疗保障工作与人工智能等高新信息技术融合发展潜力，增强创新能力，提高首都医疗保障管理服务水平，推动首都医疗保障事业创新发展。

通过举办“Beijing DEMI”大赛，鼓励具备人工智能、大数据分析与挖掘等“高精尖”科技的高新技术企业、高校、研究机构等参与医疗保障管理服务，将成熟的创新科技技术转化为实践；加强政商战略合作，优选一批符合医保创新发展需要的新技术纳入医疗保障管理服务场景应用示范，不断完善首都医疗保障管理服务创新能力体系；促进“高精尖”产业发展，推进创新技术转化为实践产品并落地实施，形成线上线下融合、相互促进发展、规范安全高效的医保与高科技产业共同发展格局。

扫描上方二维码
了解更多赛事详情

重庆市江津区人民政府 - 2020 数智重庆·全球产业赋能创新大赛

重庆市江津区作为重庆工业大区，集聚了大数据智能化的相关企业 36 多家，产业规模达到 50 亿元。但由于核心技术创新能力不足，缺乏自动化、智能化等方面高技术人才，严重阻碍了工业智能化转型升级。重庆市大数据应用发展管理局、重庆市江津区人民政府联合阿里云共同启动首届“数智重庆·全球产业赋能创新大赛”，以工业智能制造为主题，大赛围绕工业制造大数据展开，聚焦生产制造、智能仓储、最优排产及订单匹配等江津企业需求场景，深入视觉计算、运筹优化等人工智能技术，召集全球开发者众智，实现工业生产降本增效，切实服务于实体经济，用技术驱动重庆智能制造升级。

算法赛以“瓶装白酒疵品质检”为课题，采集 3000 余瓶酒样的约 2 万张影像，经过数据清洗、标注等流程得到比赛数据集，总共包含 10 余种缺陷类型。要求选手研究开发基于计算机视觉的人工智能质检算法，根据相机拍摄的瓶装酒图像，判断其中是否有缺陷，并且需要准确定位缺陷的位置、类别和大小，从而达到自动化质检的目的。对于酒液缺陷，选手合理利用了酒液的动态特性，算法能够很好地将酒液中移动的缺陷目标检测出，比起未利用动态特性的方法，准确率能够提升 30% 以上，最终酒液缺陷的检测准确率能够达到 95% 以上，充分体现了人工智能算法在质检中优异的表现。

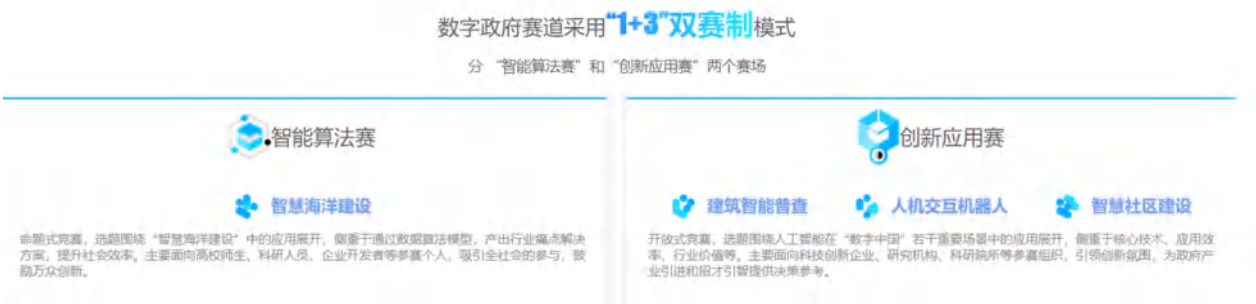
大赛共吸引来自全球 13 个国家和地区共计 3872 支队伍、4371 名选手报名，在本次大赛结束之后，由选手开发的算法将会落地产生实际价值，在实际生产中发挥重要角色，降低人力成本，创造经济效益，助力行业数字化。

大赛共吸引来自全球 13 个国家和地区共计 3872 支队伍、4371 名选手报名，在本次大赛结束之后，由选手开发的算法将会落地产生实际价值，在实际生产中发挥重要角色，降低人力成本，创造经济效益，助力行业数字化。

通过此次大赛感受到数智浪潮中澎湃涌动的力量，也在技术与产业的交流碰撞中听到了美妙的未来旋律。政府、企业、学术界，以及广大开发者，将共同构建起“新制造 新消费”的新生态，共同为产业赋能。同时，活动创造了让全国优秀人才关注江津、了解江津、来到江津的机会，通过进一步的对接交流以及企业引入，吸引高端人才来到江津、留在江津。相信更多创新力量的加入，将进一步释放重庆工业的产业活力，夯实重庆在制造设备产业的全国领先地位，使重庆江津在成为西南智能化工业中心和全国科技创新中心的道路上前进一大步。

福州市人民政府 - 2020 数字中国创新大赛 - 数字政府赛道

2020 数字中国创新大赛 - 数字政府赛道，作为数字中国创新大赛的主要赛道之一，以信息技术赋能政务服务，助力政府治理能力现代化水平提升为目标，聚焦智慧海洋、政务服务、智慧社区、城市治理四大领域，采用“1+3”双赛制模式，基于天池平台聚集全球顶级技术创新人才，发掘先进的人工智能与政务融合的创新应用成果，促进政府职能转变、再造科学决策和应急管理能力。



数字政府赛道一开始就把“方案新颖、紧扣主题、极具潜力”作为办赛的核心原则，引导参赛者从立意角度和思想高度“接轨”现代社会治理，关注福建乃至全国社会治理的热点、痛点、难点，注重技术创新与应用落地。赛事搭建了政府机构、产业资源、研究机构、投资机构的协同创新体系，投入高达 200 万的奖金及可观的落地转化资金支持。

比如智慧海洋建设赛题，针对海上商渔船碰撞事故频发、船舶违规生产作业等问题，通过创新智慧海洋应用，运用海洋大数据开发算法模型，实现对船舶安全生产行为的智能分析和预判预警，破解长期困扰主管部门的“船码不符”“一船多码”“多船一码”的难题，提升海洋船舶安全生产管理和服务水平，实现全海域护航。

建筑智能普查赛题，让人工智能技术和中国高分卫星数据深度结合，从 PB 级数据里自动化精准识别建筑物轮廓，实现高分卫星对人居建筑的每月例行监测。大赛银奖得主佳格天地的相关成果已服务于福州市“城市体检”和福建省宅基地普查工作，技术自主可控，国际领先。

无论是人才层次、专业性，还是解决方案的领先性，在相关领域表现都很出色，不少已获得相关领域国家专利、或在实践中取得优秀的成果。同时，大赛整合了全球 400+ 家高等院校及科研院所、100+ 开发者社区等资源支持和参与，共吸引了全球 16 个国家和地区的 4494 支队伍、5016 名选手报名参赛，参赛规模创下新高。

北京市经信局 - 2020 北京数据开放创新应用大赛暨科技战疫·大数据公益挑战赛

2020 北京数据开放创新应用大赛暨科技战疫·大数据公益挑战赛，由北京市经信局、CCF 大数据专家委员会联合主办，在疫情爆发的特殊背景下，快速筹备发布赛事，大赛举办过程中，产生的所有创意、模型和方案，若符合疫情期间应用需求，经自荐与组委会推荐，将随时应用于北京及全国各地的抗疫工作，特别是在湖北省等疫情严重地区应用。

本次大赛共计 439 所高校、482 家单位的 6700 余技术团队参赛，提交作品 23000 余次，最终 60 余件作品脱颖而出，其中优秀解决方案，感染路径定位、生物预警系统等部分作品实际应用，即由北京向全国推广应用，有效助力疫情防控。

四川省大数据中心 - 2021 数字四川创新大赛

为深入贯彻党中央国务院关于加快发展数字经济的战略部署，全面落实四川省委省政府“加快建设网络强省、数字四川、智慧社会，形成具有较强核心竞争力的数字经济生态体系”的发展要求，四川省大数据中心、省委网信办、省发展改革委、经济和信息化厅共同主办了 2021 届数字四川创新大赛，探索利用数据资源赋能智能时代，在提升政府治理能力和推动完善大数据生态体系等方面作一次有益尝试。

2021 届数字四川创新大赛竞赛赛题围绕政府治理体系及治理能力，推动完善大数据生态体系，利用多种开放数据资源针对性解决四川省当地需要解决的难点及痛点，助推数字经济发展。获奖者还可获得孵化办公空间资源。

深圳市人民政府、鹏城实验室 -2021“全国人工智能大赛”(NAIC)

为进一步落实科技部、工信部支持深圳建设国家新一代人工智能创新发展试验区和人工智能创新应用先导区“双区驱动”工作要求，加快推动深圳在人工智能、网络通信等科技创新领域建设步伐，由深圳市人民政府、鹏城实验室（PCNL）主办，深圳市科技创新委员会、新一代人工智能产业技术创新战略联盟（AITISA）联合承办“2021 年全国人工智能大赛”。

为配合国家“加快数字化发展，建设数字中国”战略，促进无线通信和视觉信息智能编码技术领域的原始创新积累，大赛以“AI 赋能视界”为主题，分为“AI+ 无线通信”、“AI+ 视觉特征编码”两个赛道，在 2021 年 12 月启动，面向全球开放报名参赛。

大赛立足于国际视野，营造人工智能创新创造氛围，促进产业、学术、资本、人才等创新要素融合发展，以高规格、高质量、高难度的理念加速培养人工智能领域的高精尖人才，推动人工智能技术的落地。

二、科研赛事

CCF BDCI

CCF 大数据与计算智能大赛（以下简称 CCF BDCI）由中国计算机学会 2013 年创办，是大数据与人工智能领域的算法、应用、系统、创业大型挑战赛事，面向重点行业和应用领域征集需求，以前沿技术与行业应用问题为导向，以促进行业发展及产业升级为目标，以众智、众包的方式，汇聚海内外产学研用多方智慧，为社会发现和培养了大量高质量数据人才。

2020 CCF BDCI，上半年发布疫情专题赛季（9 道赛题，算法类、解决方案类两大赛道），下半年发布正式赛季（14 道赛题，算法类、自主平台类、系统类三大赛道），以及 9 道算法类训练赛题。本届大赛共吸引了全球 25 个国家、1247 所高校、1873 家企业的 43795 人、37994 支队伍参赛，提交作品 83361 次。大赛是目前由学术组织发起的全球最大参与规模赛事，全国有超过 80 所高校将 CCF BDCI 大赛作为检验学习成果的实践推荐，每年汇编作品集供广大数据科学爱好者学习，历届优质赛题也将作为经典赛持续开放评测、训练。

赛题链接：https://www.datafountain.cn/special/BDCI2020

NLPCC

NLPCC 国际会议从 2012 年开始举办，是由中国计算机学会（CCF）主办的 CCF 中文信息技术专业委员会年度学术会议。NLPCC 专注于自然语言处理及中文计算领域的学术和应用创新，致力于推动该领域学术界和工业界研究、创新与应用的发展，成为覆盖全国、具有国际影响力的学术与创新交流平台。

LIC 比赛 (Language and Intelligence Challenge) 自 2019 年开始举办，由中国计算机学会（CCF）和中国中文信息学会（CIPS）联合主办。LIC2021 组织了机器阅读理解、对话和信息提取三项任务，并提供大型中文数据集。获胜队伍将分享 30 万元人民币的总奖金。比赛论坛及颁奖典礼将在第六届语言与情报峰会上举行。

LIC 比赛赛制非常成熟，内容为 NLP 传统内容，数据集内容完整且丰富，同时比赛对参赛选手的基础和理论功底有着较高的要求。

CCKS 竞赛

CCKS 是全国知识图谱与语义计算大会（CCKS: China Conference on Knowledge Graph and Semantic Computing）的附属竞赛，竞赛平台为 biandata。由中国中文信息学会语言与知识计算专业委员会组织和承办，根据比赛内容的不同，与包括华为、腾讯、军事科学院系统工程研究院及清华大学等多家企业、大学及研究院共同举办。

CCKS 竞赛与企业及研究所合作相当密切，其比赛领域相当广泛，比赛主题往往正切入各领域发展需要突破的关键点，同时也要求参赛人员不仅在人工智能领域有所研究，也在相关专业领域有充足的专业知识。

KDDCup 2019

全球人工智能竞赛市场已经历了 20 余年的探索，其发展起点可追溯至 1997 年首次问世的 KDD Cup，这也是数据科学领域迄今为止最重要的国际赛事之一。推出 KDD Cup 的 KDD（Conference on Knowledge Discovery and Data Mining，知识发现和数据挖掘会议）隶属于 ACM（Association for Computing Machinery，国际计算机学会），是响誉全球的顶级学术会。

KDD 2019 年，KDD Cup 首次同步开设常规机器学习、自动机器学习、强化学习三条赛道：

· 常规机器学习赛道（百度承办）：在赛题拟定上着眼于“智慧出行”，要求参考者基于城市复杂的出行情境推荐包括公共交通、出租车、自驾、骑行、步行等多种方式在内的合理的多模态出行方案。

· 自动机器学习赛道（第四范式、ChaLearn、微软和亚马逊联合承办）：要求参赛者设计时序关系数据二元分类的解决方案。

· 强化学习赛道（IBM 承办）：参赛者需要运用机器学习工具设计疟疾干预措施在模拟人群中的分配方案，以此为撒哈拉以南非洲地区的疟疾政策提供决策支撑。

连续举办 20 余年的 KDD Cup 以“创新”为核心驱动力，加之积极拥抱全球化的态度，对工业界产生了强大的吸引力。正是越来越多来自世界各地的企业以其真实业务问题与相关数据为基础提出赛题建议，从而促使 KDD Cup 通过设立更丰富的赛道来探索前沿技术在真实场景中的应用价值，并创立了总额突破 10 万美元的丰厚奖金，吸引了逾 1600 支国际参赛队伍的加盟，影响力创历年之最。

KDDCup 2020

全球人工智能竞赛市场已经历了 20 余年的探索，其发展起点可追溯至 1997 年首次问世的 KDD Cup，这也是数据科学领域迄今为止最重要的国际赛事之一。推出 KDD Cup 的 KDD（Conference on Knowledge Discovery and Data Mining，知识发现和数据挖掘会议）隶属于 ACM（Association for Computing Machinery，国际计算机学会），是响誉全球的顶级学术会。

由于各种人工智能竞赛类型的日益普及，KDD CUP 今年一共开设三个项目共四个赛道的比赛，分别是：

1、常规机器学习竞赛轨道（ML 赛道 1）“现代化电子商务平台挑战”(Challenges for Modern E-Commerce Platform)，阿里巴巴达摩院主办

2、常规机器学习竞赛轨道（ML 赛道 2）“学术图上的对抗性攻击和防御”(Adversarial Attacks and Defense on Academic Graph)，BienData 赞助

3、AutoML 机器学习竞赛轨道（AutoML 赛道） “用于图表征学习的 AutoML”(AutoML for Graph Representation Learning)，第四范式、ChaLearn、斯坦福大学和 Google 共同协办

4、强化学习竞赛赛道（RL 赛道） “在移动点播平台上学习调度和重定位”(Learning to Dispatch and Reposition on a Mobility-on-Demand Platform)，滴滴出行与 DiDi AI Labs 合作赞助

连续举办 20 余年的 KDD Cup 以“创新”为核心驱动力，加之积极拥抱全球化的态度，对工业界产生了强大的吸引力。正是越来越多来自世界各地的企业以其真实业务问题与相关数据为基础提出赛题建议，从而促使 KDD Cup 通过设立更丰富的赛道来探索前沿技术在真实场景中的应用价值，并创立了总额突破 10 万美元的丰厚奖金，吸引了逾 1600 支国际参赛队伍的加盟，影响力创历年之最。

KDDCup 2021

全球人工智能竞赛市场已经历了 20 余年的探索，其发展起点可追溯至 1997 年首次问世的 KDD Cup，这也是数据科学领域迄今为止最重要的国际赛事之一。推出 KDD Cup 的 KDD（Conference on Knowledge Discovery and Data Mining，知识发现和数据挖掘会议）隶属于 ACM（Association for Computing Machinery，国际计算机学会），是响誉全球的顶级学术会。

2021 年 KDD cup 比赛共设置三个赛题：多数据集时间序列异常检测、OGB 大规模挑战（OGB-LSC）和城市大脑挑战。数据集由主办方提供，规模较大且有代表性，例如给 OGB 大规模挑战提供的三个数据集分别在节点、链接和图级别上具有前所未有的规模和覆盖预测。

KDD cup 比赛更注重学术上的突破，聚焦于数据挖掘和神经网络的最前沿问题，参赛人员更多需要拥有扎实的基础知识，在人工智能技术本身上取得突破。

赛题链接：https://www.kdd.org/kdd2021/

NTIRE

NTIRE 全称为 New Trends in Image Restoration and Enhancement，研究图像恢复、增强新趋势，聚焦图像视频的质量改进与评估，虽然底层图像视频处理是传统方向，但 New Trends 却给这个领域带来新内涵，目前是 CV 研究社区近几年非常热门的话题，从 2017 年开始，每年与 CVPR 同期举行，包含 Workshop 与相关的挑战赛，CVPR2020 共产出 62 篇文章，已经是 CVPR 每年热门 Workshop 之一。

2021 年 NTIRE @ CVPR 2021 比赛时间在 1 月至 6 月，挑战赛分图像和视频两个赛道，由主办方提供比赛需要数据集。

NTIRE 挑战赛有众多赛道，聚集于前沿学术问题的解决，包括传统的计算机视觉问题和新出现的计算机视觉问题，历年挑战赛都为各类计算机视觉问题的解决提供了许多有价值的参考方案。

TREC

文本检索会议 TREC 由美国国家标准和技术局（NIST）等联合主办，已有 20 余年的历史，从 2017 年开始举办精准医学赛道评测，成为生物医药领域检索任务的国际权威评测。

TREC PM 2020 吸引了梅奥诊所、滑铁卢大学、康奈尔大学、苏黎世联邦理工学院、清华大学、复旦大学等 16 支世界知名团队参加。

根据组委会提供的癌症患者信息，参赛队伍需要从 3000 万篇生物医学文献中找到包含相应临床证据的文章。按照 NDCG@30 和 NDCG@5 两项搜索评价指标，组委会对搜索结果进行评分。

IJCAI 2021

IJCAI 2021 挑战赛为 IJCAI 的附属竞赛，比赛时间为 3 月至 8 月，竞赛与多个企业与机构共同举办，赛题类型独特。

大赛由提供专业数据集，如 2021 年歧义消除挑战赛由 AMiner 提供全球最大的人工标注姓名消歧数据集——WhoIsWho（https://www.aminer.cn/whoiswho），拥有超过 90 万篇论文。

IJCAI 2021 的附属挑战赛的赛题类型丰富多样，部分挑战赛专注于采用人工智能解决实际问题，部分挑战赛则专注于技术前沿，也有部分挑战赛专注于一些新奇有趣的主题，参赛者可以根据自己的能力和兴趣参与挑战赛。

Face Anti-spoofingChallenge

从 2019 年开始举办，前两届与 CVPR 合作。2021 是第三届，与 ICCV 合作由包括百度、巴黎 - 萨克雷大学、中国科学院自动化研究生在内的多家企业、大学和研究所联合主办。

2021 年挑战赛由主办方提供一个大型的 High-Fidelity Mask 数据集，即 CASIA-SURF HiFiMask（简称 HiFiMask），用于人脸反欺骗。它包括三个协议。该挑战基于协议 3，该协议评估了算法在开放集场景下的区分能力和泛化能力。换句话说，训练集和开发集只包含部分常见的掩码类型和场景，而测试集上有更通用的掩码类型和场景。协议 3 考虑了“可见”和“不可见”域以及掩码类型，它们对于实际部署更通用且更有价值。

Face Anti-spoofing Challenge 一直聚焦于人脸反欺骗这一计算机视觉热点问题，有着来自于各大企业、研究所及大学的丰富且不断更新的数据集，希望世界参赛人员能够给这一工程上的技术难点提出更优秀的解决方案。

三、产业赛事

2021X-GAME 上海新能源汽车大数据竞赛

2021X-GAME 上海新能源汽车大数据竞赛由上海市经济和信息化委员会、共青团上海市委员会、中国汽车人才研究会共同指导，上海国际汽车城、嘉定区科学技术委员会、上海青年管理干部学院共同主办。

2021 上海新能源汽车大数据竞赛以新能源汽车领域数据分析应用和算法创新为重点，围绕由上海市新能源汽车公共数据采集与监测研究中心和上海市城乡建设和交通发展研究院提供的相关数据，聚焦行业热点问题，面向全国高校征集高水平的解决方案或成熟产品。

大赛以“数联万物，车载未来”为主题，立足上海智慧城市的建设和数字化基础设施，瞄准碳达峰、碳中和工作目标，旨在实现算法创新和人才挖掘，加强政府、企业、高校和平台间的合作，推进新能源汽车大数据分析应用技术发展，打造良好的数据生态圈，促进构建绿色低碳循环发展的经济体系。

借助上海新能源汽车数据资源丰富、应用领域广泛的特点，以新能源汽车领域数据分析应用和算法创新为重点，通过融汇集成、深度挖掘，希望达到以下目的：

1. 推动新能源汽车行业大数据的应用与发展：竞赛紧扣新能源汽车使用行为、电池性能、综合能耗、交通路径、相关配套设施等行业热点问题，希望通过 AI、机器学习与大数据分析等技术的综合应用，找寻解决问题的创新思路和实践方法；
2. 拓宽整合渠道资源、推动加强渠道间合作：大赛本身可以作为打通政府、平台、企业、高校的桥梁。通过竞赛可以增强学生创新意识，发现和培养行业大数据专业技术人才。大赛的选题是根据政府委办和企业的实际诉求而确立的，大赛的成果对于政府目标和企业痛点有参考价值，从而促进政府、平台、企业、高校之间的深度合作。
3. 提升品牌效应，展现面向行业的服务能力：通过大赛与多家媒体合作，关注群体可以真切了解举办方、赞助方和其他合作方，有助于大赛各方的自身宣传。大赛不仅仅是单向服务，是可以向整个行业进行服务输送的活动。

扫描上方二维码
了解更多赛事详情

赛题链接：https://www.heywhale.com/landing/X-GAME

2021 中国大学生保险数字挑战赛

“2021 年中国大学生保险数字挑战赛”，是由中国银行保险监督管理委员会主管唯一工作日报《中国银行保险报》、中国保险学会、中国平安财产保险组成的大赛组委会主办，深圳市大数据研究院协办，知乎作为内容合作平台的校园综合赛事项目。旨在打造专业、权威及具有广泛影响力的中国大学生保险科技类赛事，为大学生提供了解保险行业数字科技战略前沿趋势，展现自我、挑战自我、增值成长的舞台。

2021 届大赛赛制全新升级，突破以往区域选拔模式，面向全球范围选拔优秀人才。大赛将聚焦产品和数字双赛道，共分初赛、复赛和全国总决赛三个赛段，并为全球参赛选手在各赛段准备多元奖励池及权威赛事证书。同时本次大赛将联合内外部行业顶级导师为参赛选手提供专业培训，优秀作品可获赛事成果研究孵化机会。

数字无限，创见未来，是本次保险数字挑战赛的主题。数字连接世界，创新带来变革。随着数据不断推动企业甚至是行业实现大跨步，企业对于优秀数据人才的需求和渴望愈发强烈，这正是现如今许多企业举办人工智能竞赛的目的——**人才的选拔与招募。**

数字挑战赛是进行数据人才选拔的敏捷途径。对于企业来说，一方面能发现更多既有数据能力又有业务感官的高新人才，另一方面也能对外树立起数据驱动增长的科技生态形象。

扫描上方二维码
了解更多赛事详情

赛题链接：https://www.heywhale.com/landing/pingancx#/pingancx

好未来 - 中国创新挑战赛智慧教育专题赛

由国家科技部指导，科技部火炬高技术产业开发中心、国家高新区管委会、北京市科学技术委员会与好未来教育集团联合承办的中国创新挑战赛智慧教育专题赛，今年就将赛题设置为了“教育手写公式识别”，旨在通过 OCR 深层次研究，突破关键技术来攻克现阶段教育科技领域内对于复杂学习场景中面临的手写公式识别这一难点，进而推动业界对真实教育场景的 OCR 研究中相关重要问题和关键技术的深入理解以及相关方法与技术的发展与突破，使得 OCR 技术能向自动批改等典型场景落地更进一步。

自比赛启动以来，共有将近 300 支队伍参加比赛，参赛队伍不仅来自清华大学、北京大学、中国科学技术大学等国内顶尖高校及科研机构，还来自阿里、腾讯、百度等一线互联网公司，参加人数众多，业内影响大。在本次的比赛中，智慧教育国家新一代人工智能创新开放平台提供了真实教学场景下收集的大量训练数据，数据量丰富、数据场景复杂，以远超各个开源学术数据集的数据体量和数据复杂性，为参赛者提供更加合理评测算法在真实场景下的鲁棒性的可能。

扫描上方二维码
了解更多赛事详情

赛题链接：https://www.kesci.com/home/competition/5f703ac023f41e002c3ed5e4

中国南方电网公司 - 数字电网开发大赛

中国南方电网公司服务广东、广西、云南、贵州、海南五省区，并与香港、澳门地区以及东南亚国家的电网相联，供电面积 100 万平方公里。供电人口 2.54 亿人，供电客户 9270 万户。2019 年南方电网公司提出把握第四次工业革命历史机遇，以数字化推进电网企业战略转型，全面启动数字电网建设。

由中国南方电网公司主办的数字电网开发大赛，本着“任务从业务中来，成果到应用中去”的原则，旨在推动新一代数字化技术与电网业务深度融合，解决电网实际业务的技术难题。题目和数据都来源于真实的电网场景，共有 7 个数字电网建设及数字化转型关键任务，主要涉及云化数据中心、智能电网和人工智能三个方向，具体包含 4 道算法比拼任务，3 道方案评比任务。

扫描上方二维码
了解更多赛事详情

赛题链接：https://www.kesci.com/landing/dgds

英特尔创新大师杯

2021 年第一届英特尔创新大师杯由阿里云和英特尔联合举办，聚焦计算机视觉 OCR 领域以及自然语言处理方向的经典技术方向，通过三个实用场景向社会各界的开发者征集解决方案。本次大赛由三个深度学习经典场景的任务组成：通用场景中文 OCR 文本识别任务，中文 NLP 地址要素解析、中文 NLP 地址相关性任务。

大赛面向全社会开放，个人、高等院校、科研单位、企业、创客团队等人员均可报名参赛。由天池平台提供相关数据集。

第一届“英特尔创新大师杯”包含多个深度学习相关的赛道，参赛选手可以自由报名参加任何赛道，发挥自己的算法技术和创造力，探索中文语义 AI 技术在各类日常生活实用场景中的应用。由天池平台提供数据集，由英特尔至强可扩展处理器为开发者提供人工智能加速服务。赛事题目包括 NLP 和 CV 两个方向，任务较为经典。

赛题链接：https://tianchi.aliyun.com/competition/entrance/531902/introduction

中国造船工程学会 - “海上争锋”中国智能船艇挑战赛

“海上争锋”中国智能船艇挑战赛由中国造船工程学会与院校联合主办，其目的是贯彻海洋强国战略，培育创新人才，助力船舶工业高质量发展。

2021 届中国智能船舰挑战赛共设置四个赛道：

S1 组：实海域智能航行。参赛团队使用具备自主智能决策能力的船艇，按规定的时间、航线和相关要求进行“智绘 2021”航行

G2 组：水下智能感知。通过组委会提供的真实环境下的全方位水下声呐图像，对不同种类水下目标进行分类标注。允许选手自行补充数据集以提高识别的泛化能力。

S2 组：虚拟仿真。设计实现遵守相关航行规则的狭窄航道自主航行算法，并集成至“智能航行仿真测试”平台中进行算法的验证测试与正式比赛；能够对大型商船进行操纵运动建模，具备自主控制模块，具备岸线检测及避碰能力，且具备抗风载荷干扰能力。

G1 组：海上智能感知。通过组委会提供的真实环境下的水面实景视频（训练集为海面船艇海试过程中录制的影像视频），检测出不同海上目标（如船只、海岸等）在画面中的位置。

中国智能船艇挑战赛聚焦于船舰在水中及水下会遇到的各类问题，对参赛者在计算机技术、船舰方面的专业知识以及工程能力都提出了要求。

赛题链接： http://cmvc.moocollege.com/

易华录 - 江苏大数据开发与应用大赛

自 2020 年启动，由江苏省工业和信息化厅、无锡市人民政府主办，江苏省大数据管理中心提供支持，无锡市大数据管理局、梁溪区人民政府、北京易华录信息技术股份有限公司共同承办。

2021 年华录杯比赛时间为 7 月至 9 月，共设置医疗卫生、智慧水利、车联网、数字运河四个赛道，大赛总奖金 100 万，每个赛道奖金 25 万。

主办方提供基于无锡市 22 个领域 62 个部门 / 地区产生的数据集 1988 类、2521 万条公共数据资源：

（1）医疗卫生赛道：预计提供 3000 张脱敏病理切片及病灶分割切片数据。为确保数据安全和算法精确性，采用数据沙箱承载，沙箱内运行计算，返回结果。

（2）智慧水利赛道：省内协同，目前已准备无锡市、扬州市两地水利部门真实样本数据 5000 张。数据采用半开放模式，选手可向组委会提交训练集，经专家评审委员会审核通过后沉淀平台。

（3）车联网赛道：提供移动视角下路面凹坑、路面裂缝样本视频及样例数据，数据来源于交通、路政等政府部门和车联网等企事业单位，为政府数据和社会数据融合的试点，预计提供 400 路视频图像。

（4）数字运河创新创意应用赛道： 选手可充分利用政府开放数据平台的数据，如无锡市公共数据开放平台。

华录杯聚焦于解决城市建设中出现的工程性问题，并提供了庞大的政府部门的公共数据资源作为比赛数据集，参赛者需要有人工智能的技术和一定工程能力，能够为在城市建设走向智能化发展的道路上提供新的思路与新的工程实践方法。

华录杯聚焦于解决城市建设中出现的工程性问题，并提供了庞大的政府部门的公共数据资源作为比赛数据集，参赛者需要有人工智能的技术和一定工程能力，能够为在城市建设走向智能化发展的道路上提供新的思路与新的工程实践方法。

赛题链接：https://www.jseedata.com/#explainSection

腾讯 - 广告算法大赛

伴随 5G 时代来临，视频广告迎来爆发式增长，在广告创意 / 投放 / 策略等广告系统的关键环节中，对视频广告内容的深度理解越发显得重要且迫切。腾讯广告算法大赛已连续举办 5 年，大赛围绕视频广告议题，从广告应用场景痛点出发，进行赛道设置。排名前三的队伍将有机会参与 ACM MULTIMEDIA 2021 GRAND CHALLENGE SESSION。

围绕视频广告议题，2021 腾讯广告算法大赛联合计算机国际学术顶会 ACM Multimedia，从广告应用场景痛点出发，开设“视频广告秒级语义解析”和“多模态视频广告标签”两大赛道，兼具算法挑战性和商业应用价值。

赛题链接：https://algo.qq.com/

极市角科技有限公司 - ECV 极市计算机视觉开发者榜单大赛

ECV 极市计算机视觉开发者榜单大赛自 2017 年启动，由青岛市人民政府指导，极市角科技有限公司、 青岛市委台港澳办、青岛市工业和信息化局、青岛西海岸新区管委和青岛市城市管理局主办，英特尔（中国）有限公司为战略合作伙伴， 中国模式识别与计算机视觉大会（PRCV 2021）提供学术支持，极市平台、OpenVINO ™工具套件提供技术支持，马达智数提供数据支持。累计吸引超过 1 万名开发者参与。

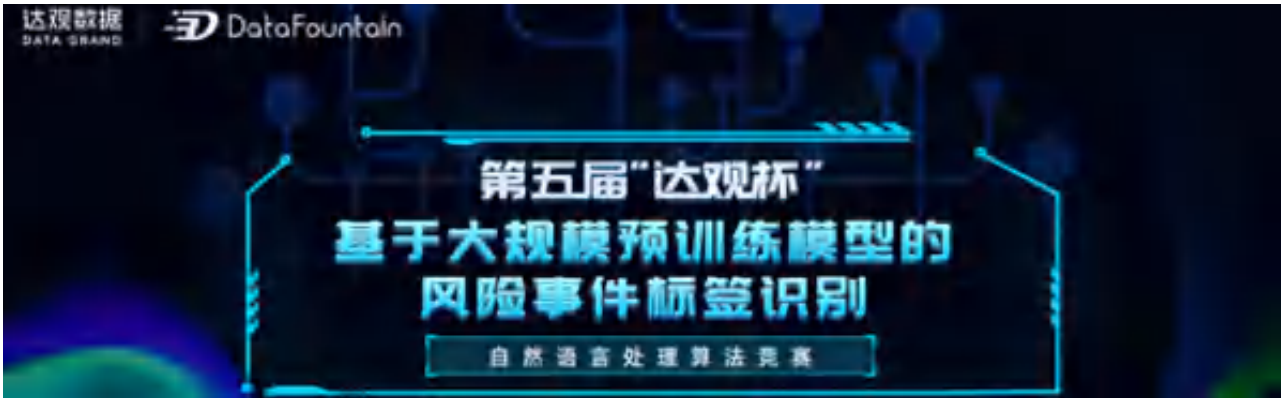
ECV- 2021 比赛时间为 6 月至 10 月，总奖金数超 100 万，围绕智慧城市、防、城管、海洋银行等实际业务场景设置比赛题，分别为：渣土车牌识别、反光衣识别、驾驶员不良驾驶识别、工程量检测、识别识别、占道经营检测、电动车进电梯检测以及人体解析分离。

赛题链接：https://www.cvmart.net/list/ECV2021

大赛为参赛者提供真实场景数据集、免费云端算力支持、便捷在线训练系统、OpenVINO ™工具套件等，帮助参赛者全程线上无障碍开发、加速模型推理，参赛者能通过在线编码训练、模型转换、模型测试等支持获得一站式竞赛体验。同时大赛也需要参赛者有从算法开发到落地应用的工程化能力，推动计算机视觉算法人才的专业工程化能力认证。

达观数据 - 达观杯

“达观杯”系列算法大赛是由达观数据发起并主办的全球性人工智能算法竞赛，由中国计算机学会自然语言处理专业委员会特别支持。每年一届，大赛参与人员为全国高校大学生、科技型企业和自然语言处理爱好者共同参与和研究实际业务场景下的人工智能技术问题。



2021 年第五届“达观杯”的主题为基于大规模预训练模型的风险事件标签识别，大赛为参赛人员提供数据集：风险事件分类的训练集规模是 10000+，包含 9 个一级标签和 35 个二级标签；大规模无标注的文本规模是亿级，可供选手选择用来进行语言模型训练。数据性质均为新闻资讯数据，并且进行了字符编码。

“达观杯“致力于通过开放真实数据，借助社会和学界的技术力量，为自然语言处理从业者搭建交流展示平台，大赛注重在传统的 NLP 背景下拓展更多的应用场景，对参数选手的基础能力和创新能力都提出了较高的要求，同时其提供的真实数据也非常有参考价值。

赛题链接：https://blog.datagrand.com/nlp-competition/

四、高校赛事

2018 中国高校计算机大赛——大数据挑战赛

2018 中国高校计算机大赛——大数据挑战赛（以下简称“大赛”）是由教育部高等学校计算机类专业教学指导委员会、教育部高等学校软件工程专业教学指导委员会、教育部高等学校大学计算机课程教学指导委员会、全国高等学校计算机教育研究会主办，由清华大学和北京快手科技有限公司联合承办，以脱敏和采样后的数据信息为基础开展的高端算法竞赛。大赛面向全球高校在校生开放，旨在通过竞技的方式提升人们对数据分析与处理的算法研究与技术应用能力，探索大数据的核心科学与技术问题，尝试创新大数据技术，推动大数据的产学研用，本次大赛鼓励高校教师参与指导。

本次大赛基于脱敏和采样后的数据信息，预测未来一段时间活跃的用户。参赛队伍需要设计相应的算法进行数据分析和处理，比赛结果按照指定的评价指标使用在线评测数据进行评测和排名，得分最优者获胜。

赛题链接：https://www.heywhale.com/home/competition/5ab8c36a8643e33f5138cba4

扫描上方二维码
了解更多赛事详情

2019 中国高校计算机大赛——大数据挑战赛

中国高校计算机大赛是由教育部高等学校计算机类专业教学指导委员会、教育部高等学校软件工程专业教学指导委员会、教育部高等学校大学计算机课程教学指导委员会和全国高等学校计算机教育研究会联合主办，面向高校学生的高水平计算机类系列竞赛，其中大数据挑战赛被列入全国普通高校学科竞赛排行榜，获得社会各界的高度关注和广泛好评。

2019 大数据挑战赛是在中国高校计算机大赛主办单位的指导下，由清华大学、南开大学与字节跳动公司联合主办，亚马逊 AWS 提供资源支持以及科赛提供竞赛平台支持，并以企业真实场景和实际数据为基础的高端算法竞赛。大赛面向全球高校在校生开放，旨在提升高校学生对数据分析与处理的算法研究与技术应用能力，探索大数据的核心科学与技术问题，尝试创新大数据技术，推动大数据的产学研用，本次大赛鼓励高校教师参与指导。

本次大赛分为报名 & 预选赛、初赛、复赛和决赛四个阶段，其中预选赛是由参赛队伍根据预选赛题在本地进行算法设计和调试并在官网提交结果进行评测；初赛和复赛均要求参赛者在和鲸社区统一提供的 ModelWhale 竞赛平台上进行数据处理、算法调试和生成结果，可使用平台提供的计算资源和工具包；决赛要求参赛者进行现场演示和答辩。

赛题链接：https://www.heywhale.com/home/competition/5cb80fd312c371002b12355f

扫描上方二维码
了解更多赛事详情

2021 中国数字人文开放创新研究大赛

数字人文回答了如何在新一轮科技革命所引发的重大知识生产变革中建设新型人文科学的重要命题，它综合了科学理性与人文诗性的研究框架，直面技术挑战的人文关怀；回到对象本身的数据研究和刻画社会场景的数字描绘，提供远读与近读的深入路径；以人文批判视野谨慎审视技术的高速发展，赋予技术以内涵；培养人文素养、科学思维和技术能力兼具的新时代人才，塑造改变社会的力量。

“时代经纬：迈向新文科的数字人文”学术研讨会由南京大学人文社会科学高级研究院与中国索引学会主办，是中国索引学会数字人文分委会 2021 年会，也是“中国数字人文”的第三届年会。本次会议旨在从多重维度就若干具体议题邀请国内外学者及专业人士展开脑力激荡与学术交流，描绘数字人文当代图景，深入探讨数字人文知识生产范式。

本次会议首次设置开放人工智能竞赛环节，面向全球征集开放数据集，集中在“和鲸数据科学协同平台”进行发布，邀请全球数字人文研究者和爱好者参加竞赛，并组织专家对竞赛结果进行评选，最后邀请获奖者参加此次数字人文大会发表论文并领取奖励。本次大赛鼓励海内外数字人文研究者及数据分析爱好者利用各种数字人文新技术对开放数据进行具备人文性的探索研究与应用。通过人文数据的创新研究竞赛，促成数字领域与人文领域研究者在竞赛合作过程中的深度协作，培育出新一代在数字与人文之间融通无间的 DH 学人，夯实中国的数字人文研究领域人才与成果基石。

赛题链接：https://www.heywhale.com/home/competition/6127152839efe300170cd454

扫描上方二维码
了解更多赛事详情

清华大学 - 安全 AI 挑战者计划

没有网络安全就没有国家安全。针对人工智能安全风险，阿里安全联合清华大学，以人工智能的对抗安全为核心，提供场景、技术、数据等支持，召集 " 挑战者 " 以攻击者或防守者的身份共同打磨 AI 模型安全；为广大安全爱好者提供数字基建安全的试炼场，在高难度的真实环境中提升技术，培养真正有安全实战能力的安全基建人才。清华大学人工智能研究院院长张钹院士表示，校企联合打造全球顶尖安全 AI 赛事，就是为了培养一流的实战型网络安全人才，推动我国人工智能领域的快速发展。

安全 AI 挑战者计划的最大特色是，以系列竞赛贯穿全年，主打高难度、高创新、高趣味、高回报的题目任务和运营玩法。每一期差异化的呈现，既满足年轻技术爱好者猎奇求新的偏好，又激发了选手剑走偏锋的奇思妙想，将思路方法持续的沉淀到技术社区，将头部企业、优秀科研院校的私有经验公有化，激发出高校学生在安全 AI 领域的创造力，为实际业务问题解决贡献力量，有助于填补和缓解 AI 安全人才缺口的问题。

赛题链接：https://s.alibaba.com/challenge

上海交通大学 - 2020 第二届华为云人工智能大赛·无人车挑战杯

人工智能作为战略新兴产业，已经开始广泛应用于多个领域，无人驾驶及机器人是其中的重要载 体。华为云联合上海交通大学举办 2020 第二届华为云人工智能大赛·无人车挑战杯，此次大赛是在华为云人工智能平台（华为云一站式 AI 开发平台 ModelArts、端云协同解决方案 HiLens）及无人驾驶小车基础上，全面锻炼和提高赛队的 AI 解决方案能力及无人驾驶编程技巧。

比赛选手拥有与华为云人工智能平台技术专家导师和上海交通大学创新中心专家导师团队进行深入沟通交流的机会，了解并动手实践华为云提供的智能硬件及人工智能平台等服务。自大赛启动以来，累计 777 人次报名，198 支团队参与。报名高校 150+，提交作品 4000 余份。人工智能竞赛作为高校推进人工智能相关学科建设的重要手段，将教学、科研、实训深度结合，以赛题的形式发布，广泛吸引校内外的人才参与竞争和交流学习，探索将理论知识转化为实践应用，帮助人才在竞赛场景中以赛促学，帮助人才认识理论学习、基础知识的意义与重要性，加强学生的理论和技能在多元场景下的应用能力。

赛题链接：https://competition.huaweicloud.com/information/1000032499/introduction

大连理工大学 - 全国水下机器人大赛

全国水下机器人大赛是一项由国家自然科学基金委员会支持，由大连理工大学和鹏城实验室等单位具体承办的水下机器人领域的重要赛事。大赛每年夏季在大连市举办，至今已成功地举办了三年，对推动我国水下机器人技术与装备的发展发挥了重要作用。

2020 年大赛组委会邀请 5 名相关领域的院士和包括 100 余名“长江学者”“杰青”“优青”在内的相关领域知名专家学者以及 150 余名产业界代表参与大赛及论坛。来自全国高校、科研机构、企业等 20 支团队参加线下现场的人机协同抓取组和自主抓取组的比赛，还有 400 余支队伍和个人报名参加了线上的目标识别组比赛。2020 届大赛较往届相比，比赛环境为真实近海水域，水下状况多变，比赛环境与技术难度更大，更加突出水下机器人的智能性，深化和拓宽水下目标检测、水声通信领域的相关研究，推进算法技术向实际产业应用进行赋能。

赛题链接：http://urpc.pcl.ac.cn/

未来杯高校 AI 挑战赛

“未来杯”于 2018 年首次面向国内外高校在读学生推出，作为国内外高校在人工智能领域认知度较高的竞赛平台之一，始终致力于实现大数据与人工智能技术及应用在高校的普及。

2018 年首届大赛，赛道包含 AI 专业组—图像算法、语音算法；未来设计组—生活设计、科幻文学，吸引了来自全国 115 所高校、3418 名选手所组成的 468 支队伍参赛。2019 年第二届大赛，组委会深挖产业需求，在保留第一届 AI 专业组赛道的基础上，还增加了以人脸识别应用为基础的智慧互娱赛道，和以房产金融数据分析为基础的智慧城市赛道，共吸引到来自海内外 221 所高校、4449 名选手所组成的 1143 支队伍参赛。2020 年第三届大赛，总报名人数 4787 人，覆盖 198 所高校，2545 支战队，虽然受到疫情蔓延的影响，但参赛规模依然在扩大。

2021 届未来杯高校 AI 挑战赛由 STEERTECH 科技平台主办，中软国际教育科技集团全程战略支持、并联合 AI TIME（清华）、IEEE 作为学术合作伙伴；ICAN 国际联盟作为创新合作伙伴、华为作为全程战略合作伙伴、华为 iMaster NAIE 作为技术合作伙伴，同时邀请清华大学、北京大学等高校人工智能领域专家对大赛进行指导。

2021 年未来杯高校 AI 挑战赛比赛时间由 5 月至 10 月，开设“AI 技术挑战赛”与“AI 学术联赛”两大赛道，总奖金分别为 100000 元和 120000 元，前者通过企业命题，真实数据集提升高校学子技术创新能力与解决行业实际问题的能力；后者通过对人工智能技术未来发展的预计，推动选手对人工智能前沿技术领域的探索及其技术在更多应用场景的实现。本次比赛的数据来自国内顶尖的科技信息情报分析引擎 AMiner。

未来杯高校 AI 挑战赛在国内外高校中有较高的影响力，提供的数据和资源都非常丰富，比赛不仅注重利用人工智能解决实际问题，同时也注重高效地利用海量的科技信息，以取得在人工智能技术自身领域的突破。

赛题链接：https://ai.futurelab.tv/contest_detail/21

海华 AI 挑战赛

海华 AI 挑战赛由中关村海华信息技术前沿研究院与清华大学交叉信息研究院联合主办，腾讯云计算协办，已举办 2020 及 2021 两届，竞赛平台为 biandata。

本届比赛一共有两条赛道，分别为中学组赛道和技术组赛道。中学组赛道只允许在校初高中学生参加，需要报名并通过审核。技术组赛道允许在校高校生（本科生及以上学历）及社会相关领域从业者参赛。

海华 AI 挑战赛致力于培养和筛选人工智能领域的相关人才，在竞赛题目上相对传统且扎实，但同时则更考验参赛人员在自然语言处理和人工智能领域的扎实基础。

赛题链接：
https://www.biendata.xyz/competition/haihua_2021/

DIGIX 全球校园 AI 算法精英大赛

DIGIX 全球校园 AI 算法精英大赛自 2019 年启动，由江苏省人工智能学会、华为终端云服务、华为南京研究所共同举办。面向全球具有正式学籍的全日制高等院校及以上在校学生。

赛题

2021 年 DIGIX 全球校园 AI 算法精英大赛在赛题设置上更加围绕实际业务场景中的热点难点，分别基于多目标多视图的用户留存周期预测、多模型迁移预训练文章质量判别、多目标优化的视频推荐、多模态多语言的搜索排序以及小样本菜单识别设置五大挑战性赛题，由主办方提供匿名化的文章质量数据和脱敏处理后的搜索数据等相关数据。

DIGIX 全球校园 AI 算法精英大赛虽然启动时间不长，但影响力在不断扩大，整体赛事规模也在快速升级，比赛内容也更加精准具体，更加注重解决实际业务问题。

赛题链接：https://www.saikr.com/vse/DIGIX/2021

第八章

人工智能竞赛生态版图与竞赛平台

和鲸社区

阿里云天池

华为云大赛平台

DataFountain

DataCastle

Kaggle（国外）

ARTIFICIAL INTELLIGENCE

人工智能竞赛是指在以真实业务问题为导向，聚合广泛的、跨学科的数据人才的参与，利用数据研发算法模型、探索解决方案的新型研发模式。

2014 年至今，全球范围内由各行业企业、顶级学术会议和第三方数据科学平台发起的各类人工智能竞赛总量已突破 1000 余场。其中，仅中国市场就已通过 12 个人工智能竞赛平台发布共计逾 **400 场**人工智能竞赛，占比近半，且年均增长率高达 **108.8%**。**236 家**企业、政务部门、高校和科研机构作为赛事主办方参与其中；吸引约 **36 万支团队**、**120 万人次**参赛；奖金规模合计高达 **2.8 亿元**人民币，赛题覆盖金融、交通出行、安防、航空天文和生物科技等 **33 个应用场景**。

利用人工智能竞赛探索数字化创新与前沿技术的落地应用方向正在成为数据科学生态必不可缺的一环。通过对这 400 场人工智能竞赛的深入分析，我们试图从更全面、更深入的角度理解中国市场的人工智能竞赛生态。

图 8-1 人工智能竞赛生态版图

数据来源：和鲸科技（heywhale.com）制作。引用请注明出处。

和鲸社区

和鲸社区（heywhale.com，原“科赛网”）是最早一批专注于AI大数据竞赛的头部平台，也是国内最大的数据科学社区之一，以“connect people with data”的使命，注册用户超20万，辐射超过100万数据科学爱好者，累计运营200余场大规模人工智能竞赛，覆盖互联网、工业大数据、智能机器人、智慧金融、智慧城市、智能电网、智慧医疗、新能源汽车等领域。

各大赛事中使用的和鲸自主研发的在线开发平台ModelWhale（原K-Lab），便于选手在统一的环境中进行参赛作品的开发、调试、提交及自动跑分与实时排名，累积用户超40万，日常支撑数千人同时在线建模，性能与体验均达国际领先水平，目前也是众多科研机构、高校和企业首选的数据科学 & 机器学习平台。

和鲸社区将高度产品化的在线开发平台和赛事运营系统开放给全行业，帮助数字化转型升级企业打造属于自己的竞赛平台，作为持续的“数字化练兵场”、“数字化人才挖掘器”和“数字化生态构建器”，用可控的成本和敏捷的方式进行一次又一次精准目标的创新升级。

官网链接：https://www.heywhale.com/home

阿里云天池

天池面向社会开放高质量脱敏数据集（阿里数据及第三方授权数据）和计算资源，吸引全球高水平人才创造优秀解决方案，有效帮助行业 / 政府解决业务痛点，并为企业招聘提供人才输送。作为中国产业AI排头兵，天池提供集品牌、生态、人才、算力为一体的数据智能解决方案，为产业创造价值。2014年至今，天池已成功运作400余场高规格数据类竞赛，覆盖全球98个国家和地区的90万数据开发者。天池平台上的竞赛课题以解决实际场景中的业务痛点为主，实战性和应用性强，场景覆盖数字政府、电商、金融、交通、物流、航空、工业、基因、电力、医疗多个领域，让AI普惠各行各业。近年，天池平台推出模型评估工具TCC（TianChi Computing的缩写），保障竞赛优胜模型质量的同时兼顾模型效率和数据安全，维护竞赛公平性并推动竞赛成果落地转化，2020年天池在TCC的基础上推出业界首创的流评测机制，推动竞赛模型评估迈向新高度。

官网链接：https://tianchi.aliyun.com/

华为云大赛平台

华为云大赛平台，是华为公司面向全球开发者的创新大赛平台，致力于为想要改变世界的开发者们提供一个实践创新的梦想平台，鼓励更多开发者基于华为平台和能力进行产品与方案的创新与应用。主要赛事有：①编程和软件开发类赛事，使用华为敏捷开发DevCloud、鲲鹏、HiQ等服务和产品，与全球编程爱好者一起探索编程与竞技的更高境界。以赛促学，以赛促建，携手为生态建

设搭建政企高校合作的桥梁。②华为云 AI 竞赛主要基于一站式 AI 开发平台 ModelArts 进行参赛作品的开发、调试、提交及自动评分，其中评分测试集不可见的方式最大化保证比赛客观公正。③网络 AI 算法类赛事，使用华为网络人工智能 (NAIE) 的端到端数据处理、特征工程、AI 模型训练平台，集成 200+ 网络电信领域 API 接口 (含自研接口如 Automl 等)，与 AI 爱好者一起探索机器学习的算法挑战。

特色：基于华为全栈 AI 能力，开放华为真实业务数据集；赛事专属 AI 训练营赋能体系；云联结政府、企业多领域办赛，提供优秀人才桥接

官网链接：https://competition.huaweicloud.com/home

DataFountain

DataFountain（简称“DF”）是国内领先的数据智能协同创新平台，为 17 万 + 用户提供了专业的大数据竞赛、人工智能数据集、开源分享社区、建模工作台、案例实训等服务。依托创新技术与国家级赛事服务能力，自 2015 年成立以来已举办 280 余场由政府、企事业单位、科研院所主办的大型赛事，是 CCF 指定竞赛平台。

DF 平台鼓励选手赛中赛后得当开源，坚持严格反作弊，关注能力提升与实践应用，为参赛选手提供了友好、公平、高成长的竞赛环境。DF 平台打造了国内多个经典赛事 IP，其中 CCF BDCI 大赛已成为国内最具生命力、参赛规模最大的大数据联赛及圈内聚首盛会。

官网链接：https://www.datafountain.cn/

DataCastle

DC 竞赛是专业的数据科学竞赛平台，基于自主研发的 DCLab 在线编程工具，主打算法建模与创新应用两种竞赛模式，聚焦大数据与人工智能，不断开拓工业互联网、物联网、区块链、网络安全、工具开发等领域。DC 竞赛利用竞赛生态，赋能政府政务数据开放、

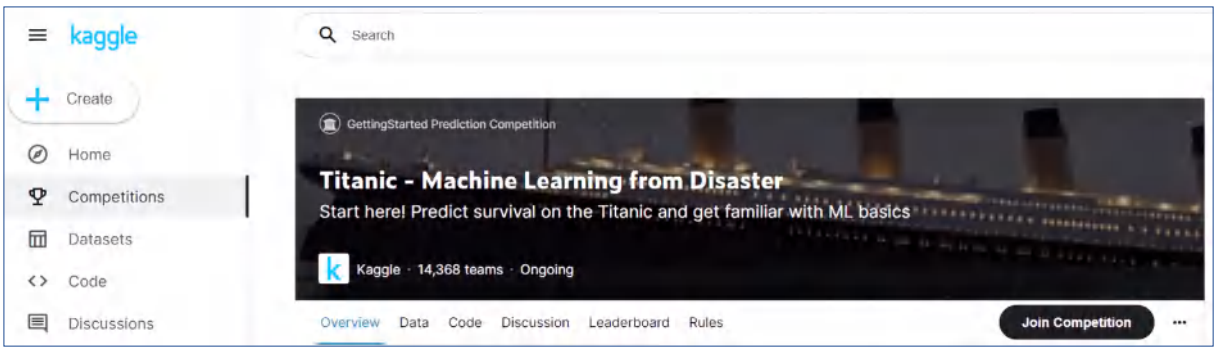
助力政府挖掘并招募优质项目；集群智为企业优化算法、精准招聘数据科学领域人才。迄今，DC 竞赛已汇集了 15 万数据科学开发者，累计上线赛题 264 道、参赛人次 28 万、作品提交 36 万次、奖金发放 7100 余万，已服务国家部委、军委、地方政府 20 余家，企业上百家。此外，DC 竞赛还联合上海市教委，面向 K12 推出了“2020AI 算法擂台赛”，不断尝试并延伸用户群体。

官网链接：https://challenge.datacastle.cn/v3/cmptlist.html

Kaggle（国外）

人工智能竞赛发展的里程碑绕不开 Kaggle，除了国际性的市场影响力和知名度外，Kaggle 更是为人工智能竞赛的平台化发展奠定了模式的基础。创立于 2010 年的 Kaggle 是目前全球最大的人工智能技术分享社区和第三方人工智能竞赛平台，面向全球用户累计举办竞赛近 400 场（含练习赛和自营赛），社区活跃用户逾百万。各类赛事主办方基于不同的目标通过 Kaggle 发布与数据相关的难题，悬赏吸引全球的人工智能人才参与竞赛。

在发展初期，Kaggle 聚集了一大批数据科学与人工智能人才，凝聚了 Your Machine Learning and Data Science Community 的社区文化，形成了讨论氛围浓厚的数据集、代码与技术分享社区，并在此基础上开始逐渐发展人工智能竞赛服务。围绕着数据科学与人工智能，Kaggle 做出了诸多努力和尝试，平台功能也日益完善，目前包含竞赛、数据集、开发工具三个子平台，配套论坛，课程与招聘服务三大模块，逐渐发展成为连接人才和应用场景双方共同致力于人工智能与数据科学创新探索的综合性社区平台。



官网链接：https://www.kaggle.com/

在中国市场，越来越多的企业、高校、科研机构和政务部门开始意识到人工智能竞赛是进行品牌宣传、人才选拔、算法创新的一种高效形式，因此投身人工智能竞赛的姿态愈发积极。在此背景下，中国市场也顺势崛起了一批人工智能竞赛平台，传统科技巨头与深耕竞赛领域的公司，凭借在资源整合、场景挖掘、工具开发、成果封装，人才转化等方面差异化的专业服务能力和经验推动着中国人工智能竞赛的专业化发展，并开始探索区别于 Kaggle 的中国特色发展道路。

第九章

人工智能竞赛的五大趋势与展望

一、奖金投入意愿攀升，参与规模稳增

二、数据红利可期，数据投入姿态积极

三、教育投入与日俱增，数据人才能力跃升

四、应用场景不断扩散，标杆行业优势初显

五、商业数字化趋势渐强，有赖技术创新落地

六、人工智能竞赛驱动的创新生态展望

ARTIFICIAL INTELLIGENCE

一、奖金投入意愿攀升，参与规模稳增

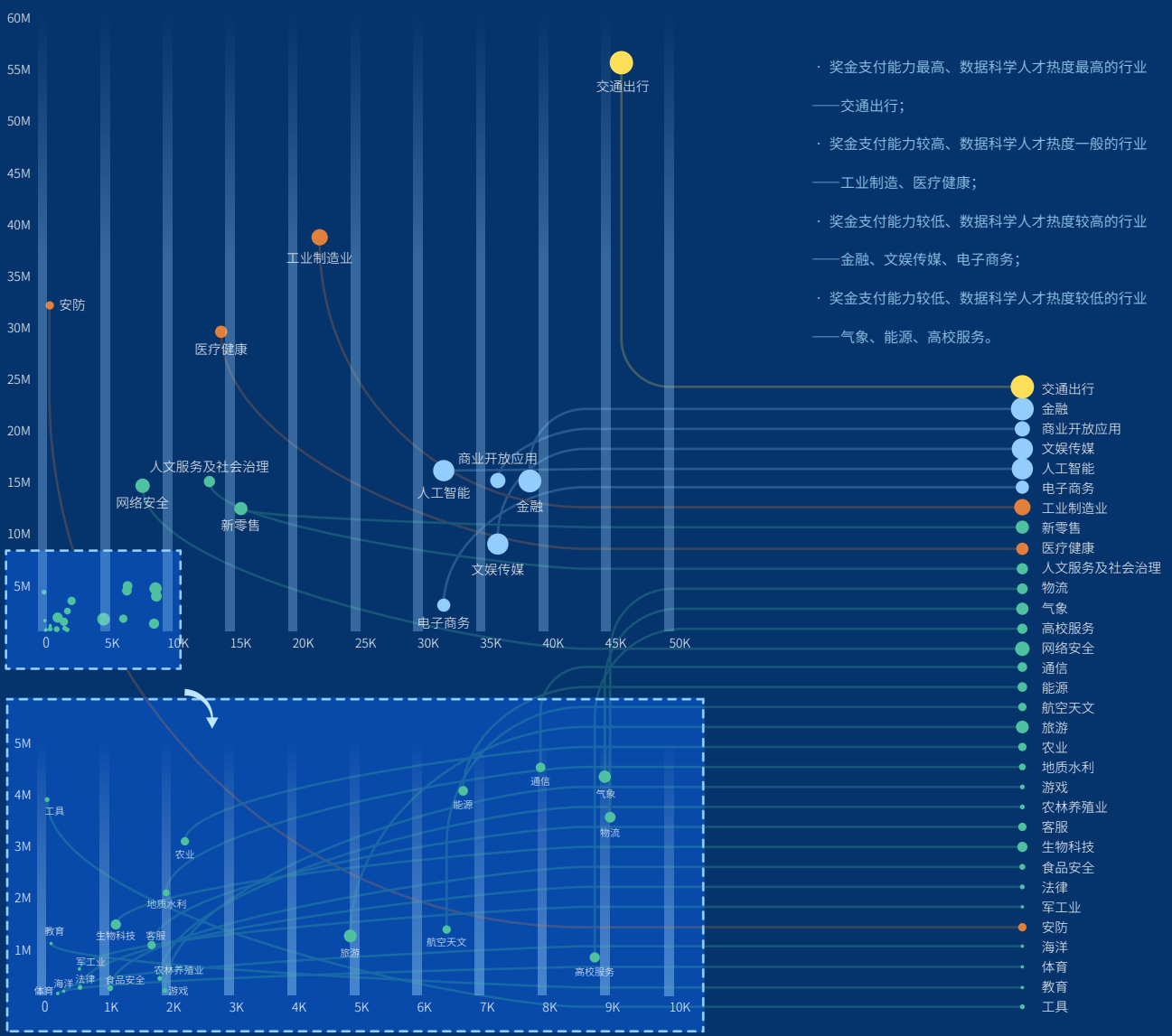
人工智能竞赛的奖金水平和参赛规模从侧面反应了各行各业的数据科学投入力度和各类技术领域内的人才结构。**自 2014 年以来，无论是参赛人数还是总奖金池规模都呈现出较为显著着上涨趋势，而人均奖金也呈同步上升态势**。由此可见各行各业在人工智能竞赛中的投入热情和力度不断高涨。

具体来说，除了具有跨行业应用价值的赛题外，赛题场景的行业分布与奖金、参赛规模主要呈现如下趋势：

- **交通出行、工业制造、医疗健康**是目前人工智能竞赛中奖金最丰厚、规模最宏大的三个领域，可见其对于数字化创新价值的重视和期待。
- **电子商务、社会公共服务、文娱传媒**等赛题场景因与大众个人生活息息相关，故而也往往具有较大的参赛规模。
- **安防、能源、气象**等赛题场景因具有一定的知识门槛，因此存在奖金丰厚但参赛者寥寥的现象。

同时，不同技术类型赛题的难易程度和发展速度也与奖金水平正向相关：**越高的奖金意味着技术难度越大，丰厚的激励有望推动尖端人才投身前沿问题的解决**。其中，计算机视觉类赛题的奖金水平一枝独秀，预计未来仍会有大量相关赛题涌现。而结构化数据挖掘类赛题则具有最广泛的受众面和最低的参与门槛，参赛规模庞大。

图 9-1 各行业赛题的奖金投入与参与规模



数据来源：和鲸科技（heywhale.com）经调研统计得出。引用请注明出处。

二、数据红利可期，数据投入姿态积极

人工智能竞赛的创新价值均以数据为基础，除了以丰厚的奖金投入来吸引优质人才的加入，越来越多的主办方也**在数据安全与法规限定的范畴内不断加大竞赛数据的投入力度，为创新提供更充足的养料**。

百度作为积极投身人工智能竞赛的先锋，已基于百度知道的真实问答文档建设了迄今为止**规模最大的中文公开领域阅读理解数据集 DuReader**，并完成了对总量类型、问答实体和观点等信息维度的标注，弥补了现有主流问答语料库对于观点类问题覆盖不足的缺陷，首批发布的阅读理解数据集包含 20 万个问题、100 万份文档及 42 万个人工撰写的优质答案，并提供开源基线系统，从而为各行各业在自然语言处理领域的创新探索奠定了坚实基础。

由中科院打造的 **CASIA-WebFace** 数据集包含了 10575 名个人的 494414 张图像，是**训练人脸识别模型的重要素材**。

此外，在由百度主导的“智能问答”和“综艺节目精彩片段预测”两场人工智能竞赛中，百度联合汽车大师和爱奇艺分别提供了汽车大师平台上的 11 万条真实问答数据、以及总长约 1200 小时的 1470 条爱奇艺电视综艺视频。这些数据对最终优质成果的诞生发挥了至关重要的作用。

携程将人工智能竞赛作为推动内外协同创新、解决实际业务问题的重要手段，其发布的赛题几乎覆盖了自身业务的核心内容，包括酒店销量概率预测、客户转化概率预测、航班延误预测、渠道销售能力预测等诸多方面。为了收获具备工业应用潜力的竞赛成果，**在经过严格脱敏处理和采取充分的数据安全保障措施后，携程为竞赛导入了海量真实业务数据**，其信息涵盖酒店基础数据、历史订单数据、价格波动数据、竞品排名数据、历史航班动态起降数据及航班延误影响因素相关数据等。

科研机构在日常研究工作中积累了大量数据，通过人工智能竞赛的开放式创新环境闭合的数据库在一定条件下进行开源，也无疑能够进一步促进这些数据的价值释放。率先将医疗大数据 Datathon 模式引进国内的解放军总医院，在 2018 年的 Datathon 中就曾发布并应用了 **2015-2017 年间在解放军总医院急诊科就诊数据库，以其作为 Datathon 竞赛的基础资源，通过跨学科、跨领域的交叉合作，真正推动了医疗大数据应用从理念到落地的探索**。

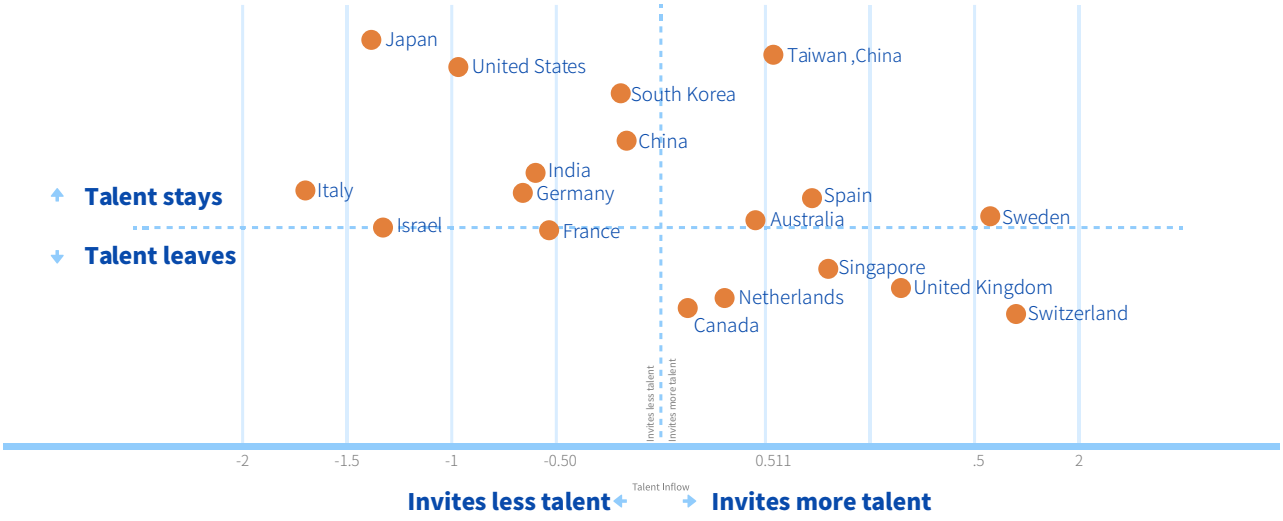
政务数据广泛覆盖自然信息、城市建设、城市管理监察、服务与民生消费等丰富的维度，是极其重要的大数据资产，开拓政务数据的创新应用场景也成为了近年来各级政务部门的重要任务。越来越多的**政务部门通过人工智能竞赛开源数据红利，探索创新机遇**。举例来说，在由广西壮族自治区人民政府发起的全球数据智能大赛（2019）中，2015-2018 年间广西 81 县十余个维度的气象和早、晚稻产量数据，及降雨、温度、光照、温差等气象数据被用以构建智能气象预测系统和精准预测水稻产量。

三、教育投入与日俱增，数据人才能力跃升

高校是学科研究和人才培育重要机构，自 2015 年以来，全国各省市地方高校积极开设人工智能专业和数据科学与大数据技术专业。截至 2021 年 12 月，全国开设人工智能专业的高校为 231 所，全国开设数据科学与大数据技术专业的高校为 558 所。

与此同时，考虑到前沿技术与商业场景之间存在的认知偏差阻碍了数据科学价值的释放，教育部近年来积极推动交叉学科的建设和复合型人才的培养，要求“创新学科组织模式，打破传统学科之间的壁垒，加强学科协同交叉融合，构建协同共生的学科体系”。而在数字化趋势下，日益完善的社会教育体系和不断催生的复合工作能力要求，继高等教育之后也催化了大批数据科学人才的诞生。人工智能竞赛日渐成为人才成长过程中进行理论实践和能力锻炼 的重要选择。另外一方面，根据剑桥大学发布的 2019 年 AI 全景报告，中国的人工智能生态已经相对较为成熟，国家数据科学人才的流失率处于相对较低的水平且人才的增长率处于相对较为稳定的水平。

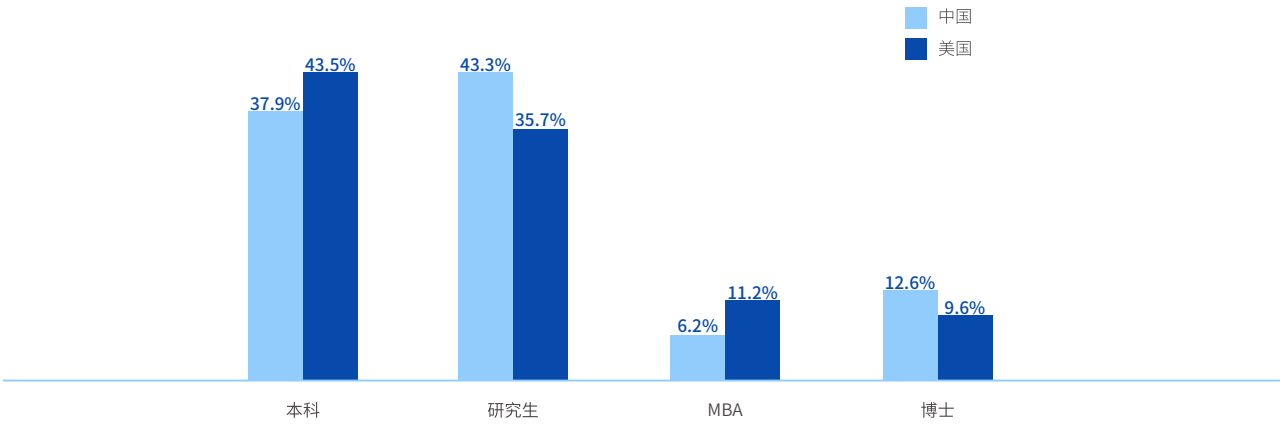
图 9-2 主要国家 AI 人才的发展流动情况



来源：剑桥大学发布的 2019 年 AI 全景报告

近年来，中国在人工智能研究领域加快发展，自 2014 年以来，中国在深度学习领域发表的论文数量开始领跑全球，这一趋势一直维持至今。同时，根据 LinkedIn 发布的《全球 AI 领域人才报告》，较之美国，中国 AI 人才拥有高学历及海外留学背景的人数占比更高，从侧面反应出人工智能竞赛的目标参赛人群整体供给质量较高，从供给层面为人工智能竞赛的价值保障奠定了良好基础。

图 9-3 中美 AI 领域专业人才最高学历分布



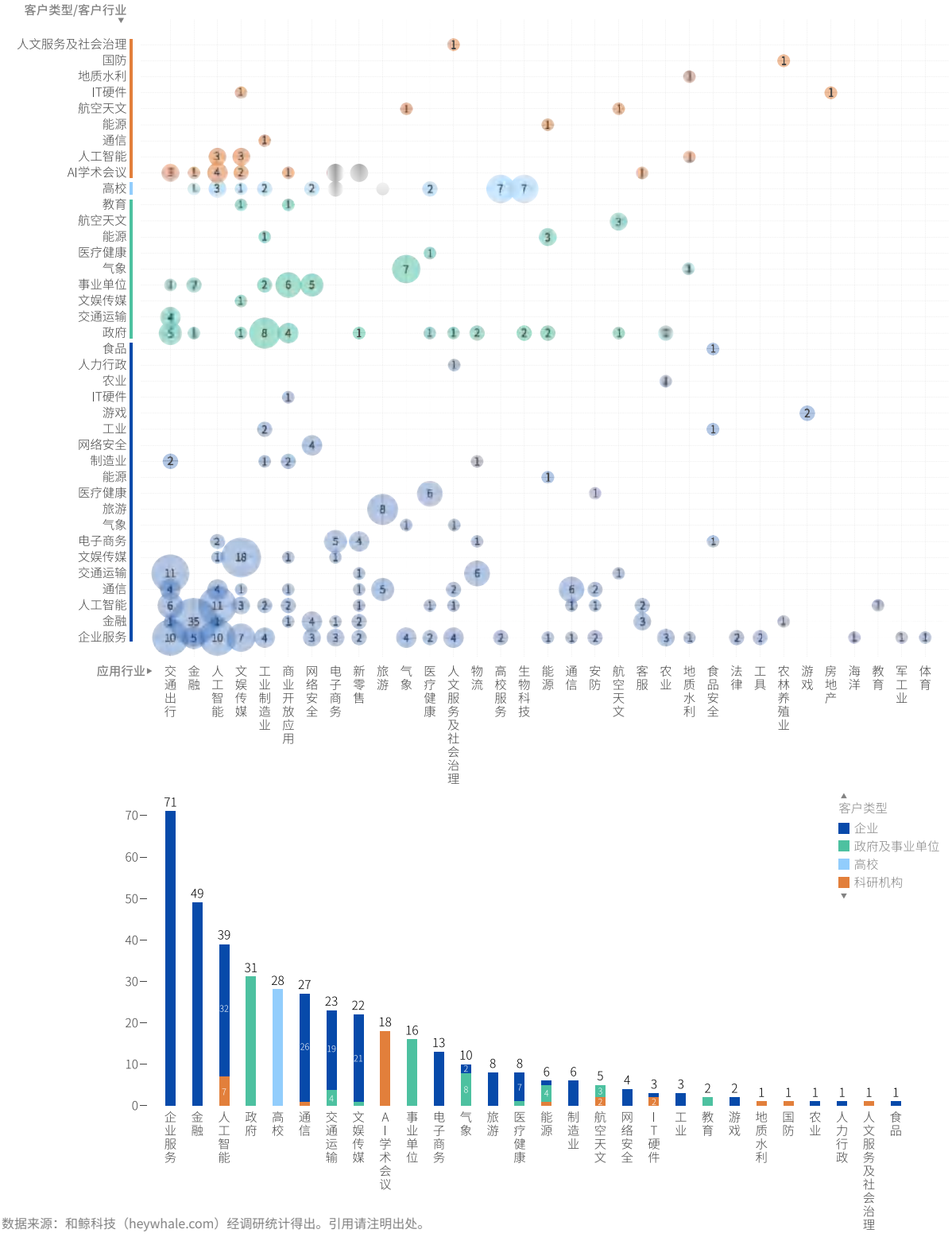
来源：LinkedIn 全球人才大数据

数据科学人才是数据驱动下技术创新和落地的抓手，人工智能竞赛因其面向不同人群的开放性和赛题内容及技术类型的丰富性，成为 了人才成长过程中重要的实践锻炼场景，提升了培养效率，而人才的不断涌现又为人工智能竞赛的规模化开展奠定了参与基础。通过和鲸调研发现，53.6% 的人工智能竞赛参与者具备硕士及以上学历。人工智能竞赛和专业人才之间互相促进、彼此成就，为数据驱动的技术创新落地和产业转型升级提供了充沛的生产力。

四、应用场景不断扩散，标杆行业优势初显

人工智能竞赛自诞生以来，始终以数据科学落地场景的探索为主要优 势特色，并对数据科学在不同场景下的应用能力和价值通过赛题 进行验证和迭代。从 2014 年至今中国市场的 600 场人工智能竞赛的赛题可以看出，除了如推理问答、人脸识别等具有跨行业场景应用 能力的赛题外，交通出行、金融、文娱传媒和工业制造是赛题场景涉及最多的四大行业。

图 9-4 赛事主办方行业类别与赛题场景行业类别的配对关系

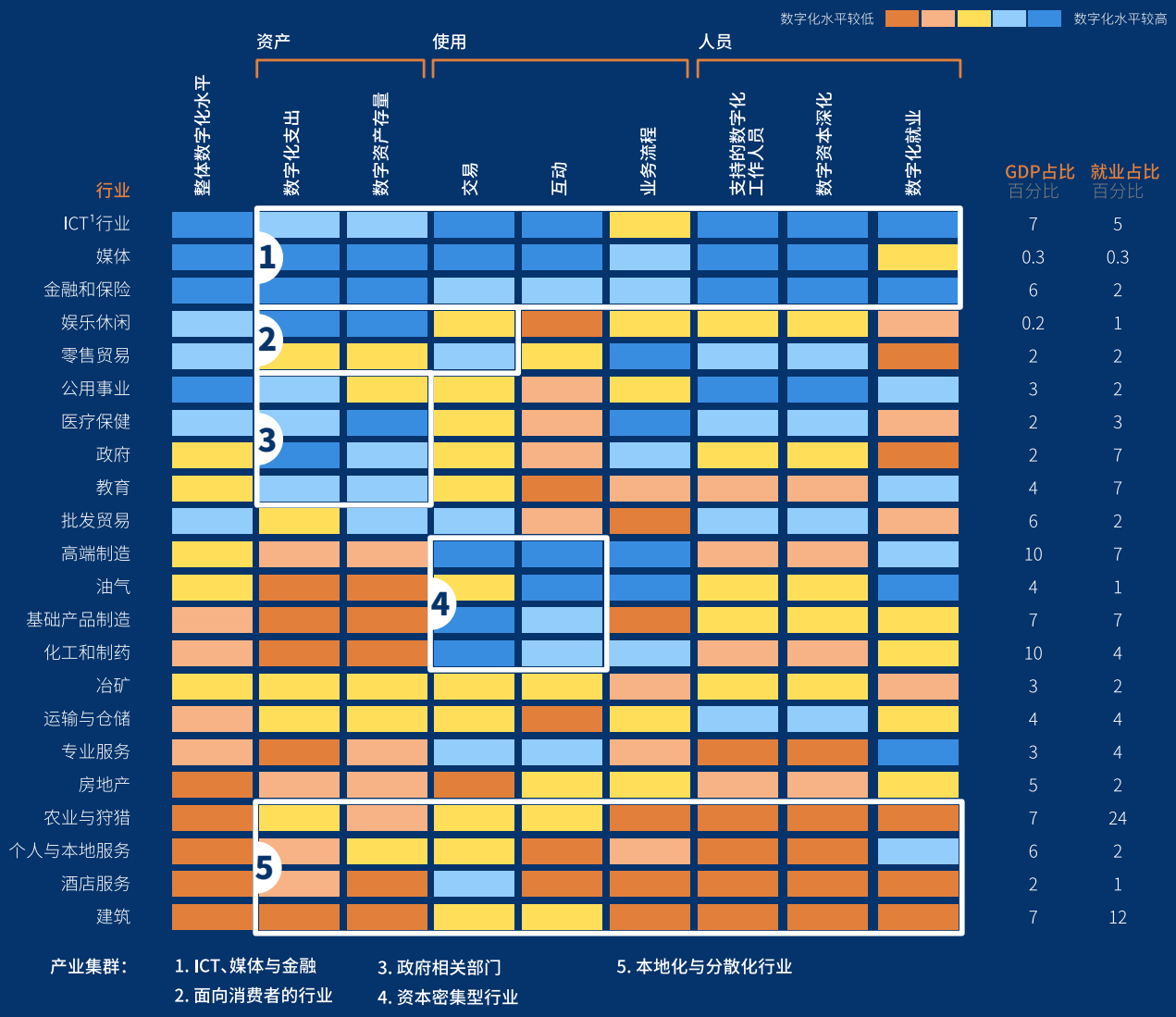


数据来源：和鲸科技（heywhale.com）经调研统计得出。引用请注明出处。

五、商业数字化趋势渐强，有赖技术创新落地

根据麦肯锡全球研究院发布的 MGI 行业数字化指数报告，在产业数字化动态发展的过程中，完备的数字资产积累、充分的资本投入与研发、可持续的人才连接与协同推进了产业数字化程度的不断加深，进而带来了较高的平均利润增长率以及劳动增长率。而人工智能竞赛在中国市场的崛起并非是对国外先进模式的亦步亦趋， 而是在技术落地应用需求增长、云计算基础不断成熟、宏观政策支持、投资机构助推、媒体宣传引导、专业人才涌现等因素共力作用的必然结果。 较为主动采用数字化战略的企业也能在后续发展中与竞争对手拉开差距。建立基于数据价值流动的生态系统，充分利用生态内不断升级的人才资源和技术资源赋能商业发展，已成为越来越多商 业领袖的共识。

图 9-5 中国 MGI 行业数字化指数



数据来源：麦肯锡全球研究院

人工智能竞赛通过快速检验人工智能技术的落地应用效果，为数字化创新的运营模式带来了新的想象空间。然而，通过竞赛来实现数据能力的锻炼和积累，需要精细化、专业化的流程管理和赛事策划，使得竞赛切实扮演好人工智能研发练兵场的角色，在吸引到大量优质人才的踊跃参与的同时，产出让人才与主办方双方都满意的成果。

六、人工智能竞赛驱动的创新生态展望

图 9-6 以人工智能竞赛为起点促进创新的螺旋上升

数据来源：和鲸科技（heywhale.com）制作。引用请注明出处。

在人工智能生态中，“数据 + 算力 + 人才 + 场景 + 算法”各环节之间需要具有紧密的连接和互动，但这个过程充满了不确定性。可以看到，人工智能竞赛可模拟、可考核，且过程及成果可见，同时风险与成本高度可控，能够串联起数据科学生态链上的各类参与者，为各方提供协作的标准界面与有效激励。 数据应用到达成熟期还有很长的一段路要走，人工智能竞赛作为人工智能探索的敏捷版和演兵场，在积累数据科学项目管理能力、和检验算法落地效果的同时，也为数字化创新的探索带来了更大的想象空间。伴随着通信技术的发展核心基础设施，“5G+ 混合云”的组合方式既能有效提升数据计算能力和传输效率，集中更大规模的数据资源，为“数据驱动”奠定基础。而人工智能竞赛引发了“产学研政资”对于数字化创新的共同关注，集合学术、人才、奖金和政策等全方位的资源支持，创造更加适宜的发展环境。

未来，相信人工智能竞赛也能够通过各界的开放式协作、开放式创新， 发现制约数字化创新效率的因素，落地改进方案。人工智能竞赛的创新价值也将得到更加广泛的认可，加速推动前沿技术的发展与产业应用的结合，在数据化、云端化、智能化的趋势下，汇聚与赋能新一代的科技人才实现更大的产业价值，成为各大产业数字化升级的强大推动力。

附录

参考文献

[1] LinkedIn，全球 AI 领域人才报告，2017 年 7 月

[2] 麦肯锡全球研究院，数字时代的中国：打造具有全球竞争力的新经济，2017 年 12 月

[3] 阿里研究院，人工智能在电子商务行业的应用和对就业影响，2018 年 8 月

[4] 清华大学，云计算和人工智能产业应用白皮书，2018 年 9 月

[5] 麻省理工学院，麻省理工科技评论，2019 年 2 月

[6] 普华永道，数字化工厂 2020：塑造制造业的新未来，2018 年 2 月

[7] WIPO，2019 人工智能技术趋势展望，2019 年 3 月

[8] Gartner，Top 10 Strategic Technology Trends for 2019: AI-Driven Development，2019 年 3 月

[9] IEC，Artificial Intelligence across industries，2019 年 3 月

[10] 亿欧智库，2019 年中国云计算行业发展研究报告，2019 年 3 月

[11] 德勤，2019 技术趋势报告：超越数字化，2019 年 3 月

[12] 腾讯云，产业互联网：构建智能 + 时代数字生态新图景，2019 年 3 月

[13] Gartner，China Summary Translation: 'Predicts 2019: Artificial Intelligence Core Technologies'，2019 年 5 月

[14] 剑桥大学，2019 年度 AI 全景报告，2019 年 6 月

[15] 云计算开源产业联盟，2019 年云计算与边缘计算协同九大应用场景，2019 年 6 月

[16] 阿里云，人工智能应用实践与趋势，2019 年 8 月

[17] 腾讯研究院，汽车行业云战略及商业模式研究报告，2019 年 9 月

[18] 埃森哲，2019 年埃森哲中国企业数字转型指数研究，2019 年 9 月

[19] 德勤，中国创新生态发展报告 2019，2019 年 9 月

[20] 信通院，2019 全球数字经济新图景，2019 年 10 月

调研参与者画像

调研对象人群画像行业分布		
年龄	20岁以下	9.4%
	20-29岁	81.5%
	30-39岁	7.1%
	40-49岁	2.0%
	50岁以上	0%
性别	男	80.5%
	女	19.5%
最高学历水平	专科及以下	2.7%
	本科	51.5%
	硕士	42.4%
	博士及以上	3.3%
TOP 5专业背景分布	计算机、通信、电子、自动化相关的理工类学科	71.7%
	统计学、数学相关的理工类学科	17.0%
	金融、经济、管理类	11.4%
	生物、化工、医学类	3.5%
	建筑、土木类	2.0%
人员身份构成	在校学生	68.3%
	在职工作人员	31.7%
在职工作人员（31.7%）中的TOP 5行业分布	互联网/通信/IT/电子商务	43.5%
	人工智能/大数据	34.5%
	金融/银行/证券	12.0%
	教育培训	5.7%
	加工制造/仪表设备	5.7%

扫描右侧二维码，获取本白皮书所采用的“人工智能竞赛数据集”

本数据集是对白皮书中所采用相关数据的抽取，包含了 2014 年至今，中国竞赛市场的 400 道赛题与 Kaggle 平台中的 172 道赛题。数据集来自于各大数据竞赛平台网站及数据竞赛官方页面，包含平台类型、比赛内容、应用行业、奖金池、客户行业等 11 个维度的字段，共覆盖 33 类应用场景、34 类主办方背景及六大类技术类型。本数据集随白皮书对外开放，我们鼓励对人工智能竞赛模式的进一步分析与分享。

Heywhale 和鲸

和鲸科技简介

和鲸科技，以 "Connect People with Data" 为使命，致力于打造以人为核心的数据科学生态。

旗下产品有和鲸社区、和鲸科赛、ModelWhale。

和鲸社区，汇聚 20 万数据科学同好一起分享、交流、协作、成长的数据科学社区。

和鲸科赛，6 年 200+ 场专业数据科学赛事，用敏捷与众创的方式助力各行业的数据人才发掘及数据应用创新。

ModelWhale，数据科学协同平台。将数据管理、建模分析、模型训练管理、算力资源管理、任务管理等功能深度整合，为科研工作者及机构解决数据安全应用、底层工程繁复、研究成果流转复现困难等问题，使数据驱动的研究更便捷高效。目前已有国家气象信息中心、清华大学、协和医学院、金风科技、中石油、默克制药、中国电信等顶尖企业 / 科研机构 30,000+ 用户通过 ModelWhale 进行数十万次科研分析，涵盖 专业科研、教学实训与科研一体化、政府及企业数据创新研究、数据开放平台等应用场景。

数据探索犹如星辰大海，和鲸科技志在与开拓者同行，相互成就、持续成长。

和鲸官网链接：

https://www.heywhale.com/

中国人工智能产业发展联盟（AIIA）简介

为进一步落实《“互联网 +”人工智能三年行动实施方案》提出的关于组建人工智能产业联盟，搭建政产学研用合作平台，构建我国人工智能产业生态的相关要求，在国家发展和改革委员会、科学技术部、工业和信息化部、中央网信办等部门的指导和支持下，中国信息通信研究院牵头会同相关单位，共同发起成立中国人工智能产业发展联盟（简称“AIIA”）。

中国人工智能产业发展联盟
Artificial Intelligence Industry Alliance

联盟着力聚集产业生态各方力量，联合开展人工智能技术、标准和产业研究，共同探索人工智能的新模式和新机制，推进技术、产业与应用研发，开展试点示范，广泛开展国际合作，形成全球化的合作平台。

中国信息通信研究院数据研究中心

CAICT 中国信息通信研究院
China Academy of Information and Communications Technology

中国信息通信研究院（以下简称“中国信通院”）始建于 1957 年，是工业和信息化部直属科研事业单位。多年来，中国信通院始终秉持“国家高端专业智库 产业创新发展平台”的发展定位和“厚德实学 兴业致远”的核心文化价值理念，在行业发展的重大战略、规划、政策、标准和测试认证等方面发挥了有力支撑作用，为我国通信业跨越式发展和信息技术产业创新壮大起到了重要推动作用。数据研究中心是中国信息通信研究院基础、重要的科研业务支撑部门，是全院知识中心和大数据中心。

北京邮电大学人工智能学院简介

北京邮电大学 人工智能学院
Beijing University of Posts and Telecommunications

北京邮电大学是国内最早从事人工智能人才培养和科学研究的单位之一，是中国人工智能学会（CAAI）的挂靠单位。学校审时度势，面向人工智能的重大国家战略和强烈人才需求，汇聚学校在人工智能领域的 100 多名优秀师资，于 2020 年 1 月 7 日成立了国内规模最大的人工智能学院。学院设置 4 个系：智能信息工程系、智能科学与技术系、智能控制系、脑认知与智能医学系；拥有 6 个本科专业：信息工程专业、智能科学与技术专业、人工智能专业、测控技术与仪器专业、自动化专业、智能医学工程专业；建设 3 个一级学科博士点：信息与通信工程、计算机科学与技术、控制科学与工程；建设 4 个一级学科硕士点：信息与通信工程、计算机科学与技术、生物医学工程、控制科学与工程；建设 2 个二级学科博士点：智能科学与技术和人工智能交叉学科。另外，学院还拥有“信息内容安全技术”国家工程实验室、“交互技术与体验系统” 文化和旅游部重点实验室等国家级、省部级教学科研及国际合作基地。

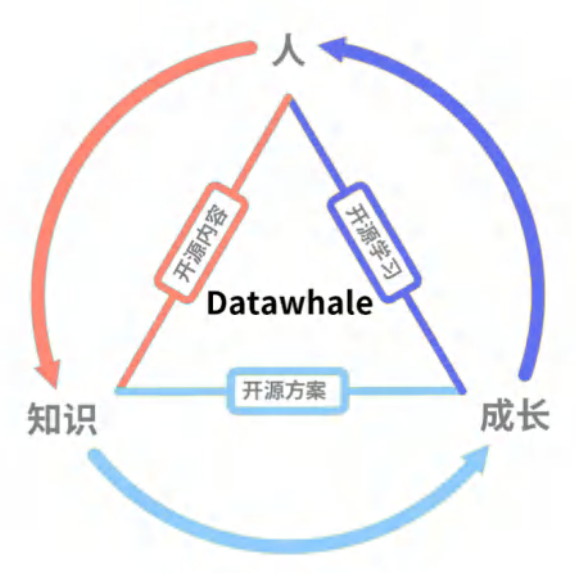
人工智能引领新一轮科技革命，已成为世界各国教育与科研的前沿热点与竞争高地。我国在“十四五”发展规划和 2035 远景目标纲要中，将人工智能作为创新发展的关键依托。

学院面向国家战略需求，立足北京邮电大学强大的“信息黄埔”实力，充分利用学院的多学科交叉特色，注重理工融合、科教融合、产教融合，守正创新，努力打造一流的人工智能人才培养和科技创新基地。

开源学习社区 Datawhale 简介

Datawhale

Datawhale 是一个专注于数据科学与 AI 领域的开源组织，汇集了众多领域院校和知名企业的优秀学习者，聚合了一群有开源精神和探索精神的团队成员。Datawhale 以“for the learner , 和学习者一起成长”为使命，鼓励真实地展现自我、开放包容、互信互助、敢于尝试和勇于担当。同时 Datawhale 用开源的理念去探索开源内容、开源学习和开源方案，赋能人才培养，助力人才成长，建立起人与人，人与知识，人与企业和人与未来的联结。



ARTIFICIAL INTELLIGENCE