人工智能

业务驱动引领大语言模型应用开发新变革

胡源，朱嘉雯

（中国电信股份有限公司上海分公司客服服务中心，上海 200120）

摘　要：重点探讨了大语言模型技术在客服中心的应用，以及业务驱动的应用开发模式带来的变革。分析了业务主导开发模式的核心意义，即推动业务需求与技术开发之间的有效沟通和协作，以提升智能应用的开发效率和质量。阐述了首次在客服领域引入“先用后训”的新型应用模式，以及自主研发平台化工具能力在高效开发中的关键作用。通过这样的实践探索，期待提升整个客服领域的工作效率和服务质量，从而推动客服领域向更加智能化、高效化的方向发展。

关键词：大语言模型；提示词工程；流程编排

中图分类号：TP393

文献标志码：A

doi: 10.11959/j.issn.1000-0801.2025060

0　引言

2022年年底，随着全球范围内ChatGPT技术的普及，基于生成式人工智能的应用逐渐开始在工作生产环境中应用。作为较早引入智能化应用的领域之一，过去几年来，上海电信客服在AI技术方面已经引入了多项应用，包括语音转文字、文字转语音、对话分类等多种基本能力，这些应用被运用于提升客服领域的服务质量监控以及降低通话时长等过程中。然而，在早期实践中发现，传统人工智能的引入通常需要大量的人工标注训练。若后续业务发生变化，几乎整体标注训练就需重新开始，这导致推广过程中存在长期运行的不确定性，影响效果。

从2023年10月开始，客服中心启动了大语言模型在客服领域的应用研究工作。在经过调研走访后，客服中心提出了全新的“业务主导模型开发”人工智能推广模式，这种模式改变了传统业务需求方和技术研发方间的工作责任分工，以全新的方式实现针对业务生产的智能应用开发。截至2024年8月，客服中心已实现了十多项大模型应用并投入生产，在应用开发效率和业务实现灵活度方面获得一线的良好反馈。

1　业务驱动大模型应用开发

本文重点从以下几方面来探讨在客服中心试点的这一种应用开发全新模式带来的改变。

1.1　业务主导大模型应用开发的本质

业务主导大模型应用开发的本质在于推动业务需求与技术开发之间的更有效沟通和协作，从而提升智能应用的开发效率和质量。在传统的生产流程中，通常业务方会向技术团队提出需求，而技术团队则负责根据需求进行分析和开发。然而，随着需求的复杂性增加，往往会出现因文字描述不够清晰或理解偏差而导致开发结果与实际需求不符的情况。因此，业务主导应用开发的目

收稿日期：2024-11-13；**修回日期：**2025-02-19

的在于让业务方更直接地参与智能应用的开发过程，以确保开发的应用能够更好地满足业务需求。

通过业务主导大模型应用开发，业务方可以在应用开发的早期阶段就参与其中，将需求细化到代码（提示词设计）层面，减少了信息传递的失真和风险。实际上，这种模式的实施也提高了业务方对技术细节的理解，并促进了业务和技术之间更深层次的合作。业务主导模型开发不仅仅是简单地将业务方需求转化为代码，更重要的是建立了一种业务人员和技术团队之间紧密合作的开发模式，共同推动智能应用的创新和发展。

在实践中还涉及建立一套灵活的开发流程和平台，使得业务人员能轻松参与到应用开发中，而无须过多依赖技术团队。这种模式的推广也能促进跨部门间的合作和知识共享，提高整体团队的工作效率和协同能力。此外，业务主导模型开发还可以带来更快的响应速度和更好的定制化服务，满足业务快速变化的需求，提高了团队的适应能力和竞争力。

1.2　先训后用还是先用后训的转变

在人工智能应用中，数据标注训练一直被视为至关重要的一环。然而，生成式预训练模型的出现，使我们开始思考先训后用与先用后训两种模式的转变。传统上，我们习惯于先进行数据标注训练，再应用模型到实际业务中，以确保模型的准确性和效果，如图1所示。然而，这种方式存在着诸多挑战和限制，如前期投入高成本、耗时等。因此，在客服领域引入大模型应用的时候，我们试点了先应用模型，再通过实际运用的反馈数据进行优化训练，即先用后训的方式。

先用后训的转变是人工应用新的思路和方法。通过先使用高参数模型生成预料数据集，再根据实际业务需求逐步优化训练，可以降低前期投入成本、提高效率，并最终实现更好的应用效果。这种方法的优势在于能够直接从实际应用中获取数据，更贴近真实场景，避免了对标注数据的依赖。同时，先用后训也能够更灵活地应对需求变化，及时优化模型，提高用户体验，如图2所示。

另外，在生成式大语言模型的应用推广过程中，我们主导推广以提示词为切入点的应用设计，以高参数体量的模型先实现面向业务的应用设计，可以最大限度降低前期设计过程中的资源

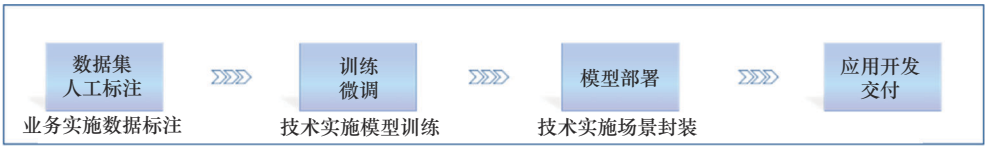

图1　传统先训后用模式

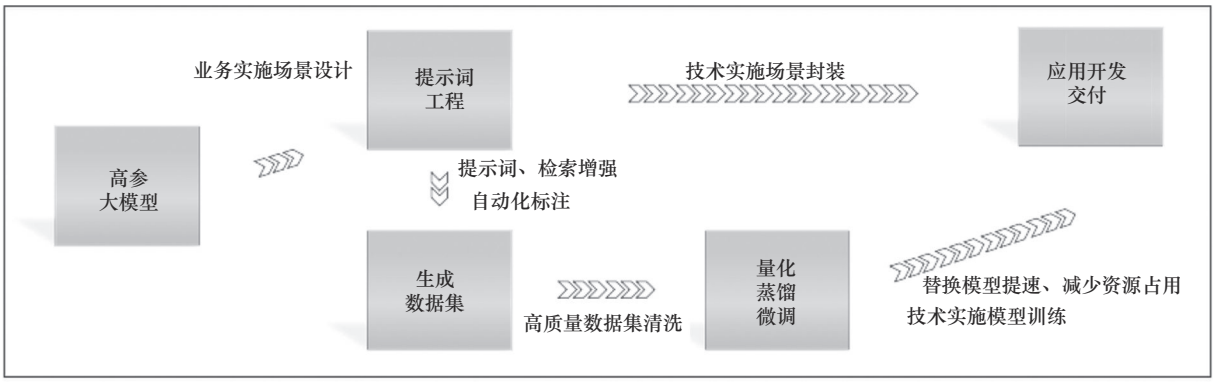

图2　引入先用后训模式

投入，鼓励一线尝试各种应用场景开发，丰富了大模型应用领域范围。

1.3 技术自研平台化能力的必要性

在传统的应用开发中，对需求工单的追踪是业绩评估的关键，这种方式鼓励用纯代码进行简单修改，然而这也导致技术与业务双方的交流仅依赖于文字工单描述，且业务方缺乏可供自我调整的工作平台。技术团队因此沉浸在眼花缭乱的细节调整中，缺乏驱动平台构建的创新性和保障稳定性的研究投入。

当客户服务中心开始扩大模型应用的推广时，新成立的团队最大的挑战无疑是研发工作量。全面采用自主研发的途径意味着每个应用开发都需要专属的研发团队予以实战。面对仅有三人的初始研发团队，追求工单量显然是雪中送炭。唯有将需求转化为对平台化能力的抽象要求，研发团队才能以静态应变的策略去应对来自不同业务的多元且不断变化的需求。

2 客服中心开发创新实践

基于以上这些在智能应用推广模式上的思考，客服中心组织了一支跨部门的大模型应用团队，从应用设计、功能研发、生产维护角度尝试在服务领域大语言模型应用试点。

2.1 全面推动提示词技术的应用与创新

本团队深入借鉴了OPENAI推荐的大模型应用部署策略，聚焦于“提示词工程”作为核心驱动力的大规模语言模型应用开发生态。利用自然语言指令引导模型执行特定业务任务，以此作为探索应用开发新路径。为了有效实施这一策略，客服中心内部精心策划并举办了十场大模型提示词应用的专项培训与交流沙龙活动，旨在使业务一线团队率先洞悉这一全新的应用开发模式，通过频繁且深入的互动讨论，我们成功培养了5个深植于业务实践的提示词应用设计团队。

这些团队成员凭借其在一线运营领域的深厚积累，仅需简短培训即可快速上手提示词的设计技巧。他们能够敏锐地从日常运营流程中发掘智能化应用的潜在价值点，进而促进工作效率的显著提升。历经超过半年的实战淬炼，团队成员不仅熟练掌握了提示词设计这一创新开发范式，还实现了个人职业生涯的重要转型，成为兼具业务洞察力与技术实现能力的复合型人才。

客服中心内逐渐培养出一个崭新的专业群体——提示词工程师。这一新兴职位不仅标志着我们对前沿技术应用的深度整合，也预示着客服领域向更加智能化、高效化方向发展的坚定步伐。

2.2 构建设计、测试、交付工具能力

要让整个“业务主导应用开发”顺利开展，“平台化”的能力是不可或缺的。传统交付中代码工作是由研发人员完成的，所以平台化的考虑明显不足。当客服中心绝大多数的代码（提示词）工作由一线设计团队进行后，自研的开发队伍更多地开始考虑整个设计配套的平台工具能力，如何提效整个设计到最终部署生产的能力配套过程。

（1）设计验证工具

研发团队开发了自助提示词设计工具（self prompt design，SPD），可以让业务人员在平台内完成应用的初始设计和验证，通过SPD的轻应用首页可以查看整个客服中心已经上线的各种提示词应用示范，提升设计效率。为了强调在应用生产上线前的测试验证，团队还开发了提示词设计助理（self prompt asistent，SPA），协助进行批量效果处理验证，实现千级的快速结果批量验证工作。

（2）低代码应用发布

为了加速上线，对于常见的生产对接模式，客服中心自研团队建立了提示词封装API的基础能力，可以快速将业务设计的提示词封装为外部可调用的API能力，实现应用上线0代码的高效。考虑应用过程可能需要的对公司IT生产接口、大

模型标签等的流程化应用，团队研发并上线了基于大模型的客服应用编排能力（魔方平台），快速将各种接口能力+大模型能力组合为最终面向生产的能力应用并实现复杂应用的快速低代码上线。

（3）自主模型微调训练

由于我们采用的是先上线应用后训练模式，在累计一定生产生成的语料数据集的同时，通过租用临时算力资源，我们已于2024年8月启动了对中国电信星辰大模型的常态化微调训练。通过场景化的专项微调训练，激发模型泛化推理能力，实现同时多场景的模型能力适配，在积极关注通用能力评估同时，实现了基于星辰大模型的推理加速及多LoRA加载等模型推理的效率提升，目前已经交付4个版本的模型权重，在维持现有高参开源模型准确率的情况下，整体资源消耗降低了50%以上。

2.3　推理提速、性能监控保障

资源能力整体架构如图3所示。客服领域最终的应用是面向生产，所以面对一线使用中的及时性、稳定性是至关重要的，但是在大模型应用中缺乏现成的有效监控和保障手段，对于技术队伍来说，一切都需要从头来。为此本文从以下几方面入手。

（1）建立异构高可用的推理能力

大模型对于算力资源消耗一直是一个大问题，前期设计中我们经常遇到忙时资源不足问题，在积极引入各种国产推理算力卡的同时，我们针对不同卡的推理极限能力、推理处理文字规模进行了测试，针对性地开发了一套异构资源下的大模型推理调度组件。依据每次生产应用请求，自动判断所需要的资源，在不同的推理节点资源中调度，获得最大的推理效率，最大限度地实现了国产算力卡、进口算力卡物尽其用。

（2）建立不同业务等级的资源使用调度

在不同算力资源物尽其用的同时，我们也发现所有请求均采用即时响应，从全天24 h角度看资源利用率，其实繁忙也就几个时段，其实资源的利用率依然很低。于是我们将业务分割为即时任务、实时任务、离线任务3类。即时任务采取请求立刻反馈方式最大限度保障资源，主要应用在填单自动生成等对延迟敏感的应用；实时任务则采用先放入队列后按一定节奏执行的方式，允许一定规模的排队以控制高峰资源占用，主要应

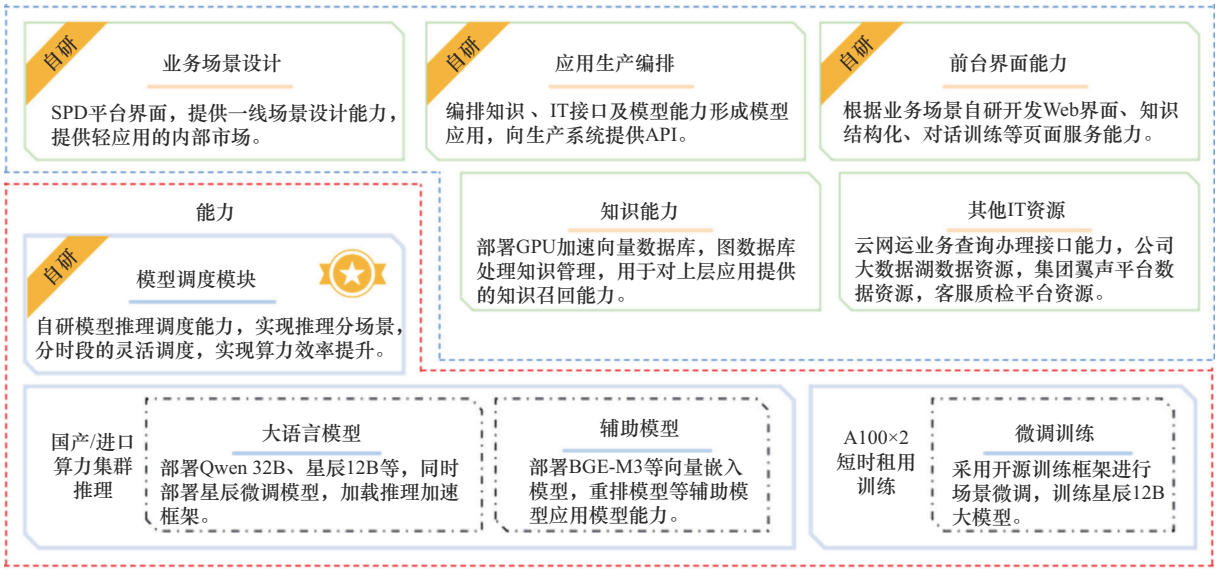

图3　资源能力整体架构

用在话务质检等允许一定时延的任务；离线任务则重点围绕每日晚上11点至次日6点之间卡资源空闲时间进行的如工单检查、个人报告生成等任务。这样分割有效地将算力资源分时统筹，实现进一步的资源使用优化。

（3）建立实时端到端质量监控

在客服领域，建立了从系统底层GPU/NPU的资源占用（如图4所示）到具体业务应用的请求规模、时延情况的全面监控能力，我们克服了对GPU/NPU算力资源缺少底层监控的直接支持手段，通过自行适配采集代码方式，成功实现了英伟达、华为、壁仞多种卡资源的算力使用率实时采集和监控功能。利用开源ELK能力并结合我们自研的推理分流API，我们实现了面向应用的全流程端到端监控（如图5所示），实时掌握某个应用的请求数量、响应时间等关键指标，真正从应用层面形成了完整的质量保障能力。

3 结束语

2024年，大模型应用的试点项目充分展示了我们全新应用设计及开发团队的创新潜力与实践成效，通过引入“业务主导开发的大模型”策略，我们成功开创了一条资源高效利用与客服领域应用设计革新并举的新路径。这一过程不仅标志着业务人员向数字化转型迈出了重要一步，也显著增强了自主研发团队的能力，验证了将人工智能深度融入业务场景的可行性与

图4 GPU/NPU的资源占用采集监控

图5 基于应用的端到端监控

优越性。

业务主导的模型开发策略不仅是技术层面的突破，更是团队协同、流程优化及业务智能化转型的综合体现。该模式通过早期整合业务专家参与应用开发，不仅确保了开发项目的精准对接市场需求，提高了开发效率与质量，还促进了跨部门间的紧密合作与沟通，加速了智能应用在实际业务中的落地生根与蓬勃发展。

接下来，客服中心将持续致力于技术创新与实践探索，不断优化大小模型等前沿技术在客服领域的应用，深化人工智能与业务流程的融合，以期为客户带来更为卓越、高效的个性化服务体验，通过不断地探索与革新，塑造更加智慧、敏捷的多模态协同客户服务模型，引领电信更高水平的服务质量和运营效率。

参考文献：

[1] SCHULHOFF S, ILIE M, BALEPUR N, et al. The prompt report: a systematic survey of prompting techniques[J]. arXiv preprint, arXiv: 2406.06608, 2024.

[2] ZHOU P, PUJARA J, REN X, et al. Self-discover: large language models self-compose reasoning structures[J]. arXiv preprint, arXiv: 2402.03620, 2024.

[3] BROWN T B, MANN B, RYDER N, et al. Language models are few-shot learners[J]. arXiv preprint, arXiv: 2005.14165, 2020.

[作者简介]

胡源（1981– ），男，中国电信股份有限公司上海分公司客服服务中心服务数字化运营中心副主任，主要研究方向为服务数字化应用开发等。

朱嘉雯（1999– ），女，中国电信股份有限公司上海分公司客服服务中心服务主办，主要研究方向为大模型应用开发上线、监控平台搭建、模型资源优化等。