大模型智能 AI 客服应用白皮书

—— AI Agent 重塑客户服务体验

#AI Agent 技术 #大模型驱动 #业务价值提升

1. 智能客服与大模型融合新纪元

1.1 大模型技术引领智能客服变革

大模型技术的崛起，正深刻地改变着智能客服领域的面貌，引领着一场前所未有的变革。传统的智能客服系统，往往依赖于关键词匹配和预设的规则库，其交互能力和问题解决能力存在明显的天花板。而大模型，凭借其强大的自然语言理解、知识推理和内容生成能力，为智能客服带来了质的飞跃。大模型能够更精准地理解用户的真实意图，即使面对口语化、模糊化甚至带有情绪的表述，也能进行有效的语义分析和情感判断 。这使得智能客服不再仅仅是机械式的应答，而是能够提供更具个性化、更富同理心的服务体验。例如，在用户表达不满时，系统不仅能识别问题，还能感知情绪，并采取相应的安抚策略，从而提升用户满意度 。此外，大模型支持更复杂的多轮对话管理，能够记忆上下文信息，进行连贯的、有逻辑的交互，从而处理更复杂的业务场景，如产品咨询、故障排查、售后支持等 。这种能力的提升，使得智能客服能够承担更多以往需要人工处理的咨询，从而显著提高服务效率，降低运营成本。

大模型的应用还推动了智能客服向更主动、更智能的服务模式转变。通过对海量用户数据的分析和学习，大模型可以帮助企业洞察用户需求，预测潜在问题，并主动提供服务和解决方案 。例如，系统可以根据用户的浏览记录和购买行为，主动推送相关的产品信息或优惠活动；或者在用户可能遇到问题时，提前进行预警和指导。这种从“被动响应”到“主动服务”的转变，极大地提升了客户体验和企业的服务价值。

同时，大模型也为智能客服的多模态交互提供了可能，支持语音、图像、视频等多种信息输入和输出方式，使得交互更加自然和高效 。例如，用户可以通过上传图片或视频来描述问题，系统则可以通过图像识别和内容分析来理解问题并提供解决方案。这种多模态交互能力，进一步拓宽了智能客服的应用场景，使其能够更好地适应不同用户群体的使用习惯和需求。总而言之，大模型技术正在重塑智能客服的核心能力，推动其向更智能、更人性化、更高效的方向发展，成为企业提升客户服务水平和竞争力的关键驱动力。

1.2 合力亿捷在智能客服大模型应用中的战略布局

合力亿捷作为深耕客户服务领域多年的技术服务商，敏锐地洞察到大模型技术带来的巨大机遇，并积极进行战略布局，致力于将大模型的先进能力融入其智能客服解决方案中，以赋能企业客户服务的智能化升级。其战略核心在于构建一个以大模型为驱动，集自然语言处理、语义理解、知识图谱、深度学习等多项智能交互技术于一体的智能客服平台 。该平台旨在解决复杂场景下的任务处理难题，提升智能客服的精准语义理解和意图识别能力，据称其意图识别准确率可高达 90% 。合力亿捷不仅关注大模型在提升应答准确性和效率方面的应用，更着眼于通过大模型实现更深层次的客户价值创造，例如通过用户画像和数据分析，提供个性化的服务和精准推荐，从而将智能客服从成本中心转变为价值中心 。公司还积极构建开放、协同的 AI 生态系统，通过与百度智能云等行业领先的 AI 技术提供商合作，结合自身在客服领域的深厚积累和行业理解，共同推动智能客服技术的创新与落地 。

在具体的战略实施层面，合力亿捷采取了多方面的举措。首先，在技术研发方面，合力亿捷积极探索和应用前沿的大模型技术，例如基于 DeepSeek 等先进模型构建其智能客服系统，以实现更低的推理成本、更强的推理能力和更快的部署速度 。同时，公司也注重自研能力的培养，例如其自研的客服 Robot，强调“独立理解能力”，能够自主完成问题理解、意图识别、逻辑应答等核心流程，并结合企业知识库进行优化，确保应答的准确性和可控性 。

其次，在行业应用方面，合力亿捷强调将大模型技术与垂直行业的特定需求深度结合，针对不同行业（如金融、政务、零售、电商等）的业务场景和痛点，提供定制化的解决方案 。例如，在金融行业，注重安全与合规强化，通过双因子验证和实时风控引擎保障交易安全；在政务民生领域，则支持多模态交互，如方言识别和材料自动填报，以提升服务的普惠性 。

此外，合力亿捷还推出了低代码 AI Agent 平台，允许业务专家通过拖拽式流程编排，快速构建和部署针对特定业务场景的智能体，从而加速大模型技术在客服领域的落地应用 。通过引入 RAG（Retrieval-Augmented Generation）技术、多模态识别能力以及实时数据反馈机制，合力亿捷的 AI Agent 能够有效降低大模型的“幻觉”风险，处理更丰富的输入信息，并在与用户的互动中持续学习和优化 。通过这一系列的战略布局和实践，合力亿捷旨在帮助客户构建“智能+情感+效率”三位一体的服务体系，实现客户服务的价值跃迁 。

2. 合力亿捷大模型智能客服的核心技术优势

2.1 行业知识增强与精准理解能力

合力亿捷在大模型智能客服领域的一个核心技术优势在于其对行业知识的深度整合与增强，以及由此带来的精准用户意图理解能力。通用大模型虽然拥有广泛的知识覆盖，但在特定行业场景下，其专业性和准确性往往难以满足企业级应用的需求。

合力亿捷通过将大模型的通用理解能力与企业私有的、结构化和非结构化的行业知识（如产品手册、政策文件、客服历史对话、用户评价等）相结合，构建了强大的行业知识图谱 。这种知识图谱不仅包含了丰富的实体和关系，还能够通过自动知识抽取工具高效构建和更新，例如，其工具可将 PDF 文档的图谱构建效率提升 70% 。

通过这种方式，合力亿捷的智能客服系统能够更深入地理解行业术语、业务流程和特定场景下的用户需求。例如，在金融行业，系统能够准确理解关于理财产品、风险评估、合规条款等复杂咨询；在电商行业，则能清晰掌握商品属性、促销规则、售后政策等信息 。在与中国联通合作的新客服项目中，合力亿捷与百度智能云共同克服了接口对接、系统融合等方面的难题，将 AI 能力融入传统客服体系，提升了对用户意图、情感和态度的精准识别能力 。

这种行业知识的增强，直接提升了智能客服的意图识别准确率。合力亿捷的解决方案通过预训练模型结合行业语料进行微调，使得意图识别准确率提升至 92% 。这意味着系统能够更准确地把握用户提问的核心，即使面对口语化、省略或者多意图混合的表达，也能进行有效的解析和归类。例如，用户可能会问“我这个手机套餐流量超了怎么办，能加个包吗？”，系统需要同时理解“流量超额咨询”和“办理流量加油包”两个意图，并结合用户账户信息给出准确的答复。

合力亿捷的认知层运用 BERT、Transformer 等深度学习模型，结合业务知识图谱，对用户输入信息进行深度语义分析和意图识别，使其在零售行业的客服方案中，意图识别准确率提升至 89% 。这种精准的理解能力是提供高质量智能服务的基础，它确保了智能客服能够给出相关、准确且有用的回答，从而有效提升用户满意度和问题解决率。此外，通过 RAG（Retrieval Augmented Generation）策略，系统能够从构建的知识库中检索相关信息，并结合大模型的生成能力，提供更具专业性和上下文相关性的答复，有效避免了通用大模型可能产生的“幻觉”或信息滞后问题 。

2.2 复杂场景交互与多轮对话管理

合力亿捷大模型智能客服在复杂场景交互与多轮对话管理方面展现出显著的技术优势，这得益于其先进的自然语言处理技术和精心设计的对话管理机制。在真实的客户服务场景中，用户的咨询往往不是单一回合的简单问答，而是涉及多个步骤、多个主题的复杂交互。

合力亿捷的智能客服系统能够支持长达 20 轮甚至 30 轮以上的上下文对话，并且在此过程中保持对对话主题和用户意图的准确跟踪，避免出现上下文丢失或理解偏差的情

况 。例如，在政务场景中，用户可能先咨询某项政策的申请条件，然后询问所需材料，接着又对材料中的某个具体条目提出疑问，系统需要在整个过程中连贯地理解用户的每一个问题，并基于之前的对话内容进行回答。合力亿捷采用“对话状态树”等技术来管理复杂的对话流程，确保在多轮交互中能够动态调整对话路径，并根据对话状态进行精准的应答和引导 。

为了实现高效的复杂场景交互，合力亿捷的智能客服系统还具备强大的上下文管理能力。这项技术能够记录和理解用户的历史对话信息，确保系统根据上下文准确理解用户意图 。例如，当用户先询问“你们有哪些平板电脑？”，接着又问“这款平板的处理器怎么样？”时，上下文管理技术使系统能够明白“这款平板”指的是上一轮提到的平板电脑，从而给出准确的回答 。

此外，系统还支持多模态交互，用户不仅可以通过文本输入，还可以通过语音、图片甚至视频等方式与客服系统进行沟通 。例如，客户可以上传故障设备的照片，系统通过视觉分析精准定位问题，并推送维修教程视频或预约工程师上门服务 。这种多模态交互能力极大地丰富了人机交互的维度，使得沟通更加直观和高效。合力亿捷的对话管理引擎采用有限状态机（FSM）与强化学习结合的混合架构，既保证了流程的可控性，又支持动态路径优化，通过实时状态跟踪，实现多轮对话的上下文连贯 。这些技术的综合运用，使得合力亿捷的智能客服能够从容应对各种复杂的业务咨询场景，提供流畅、自然且高效的交互体验。

2.3 自主学习与持续优化机制

合力亿捷大模型智能客服系统具备强大的自主学习与持续优化机制，这是其能够不断提升服务质量和适应业务变化的关键。传统的客服系统往往需要大量的人工干预来更新知识库和调整应答策略，而合力亿捷的解决方案则更多地依赖于数据驱动的自动化学习和迭代。系统通过实时分析客户意图、对话记录、用户反馈等数据，不断优化自身的语义理解模型和应答策略 。例如，基于自然语言处理（NLP）技术，系统可以实时分析客户意图并更新应答策略，从而使机器人的准确率得到持续提升，据称可提升至 92% 。这种动态优化能力确保了智能客服能够紧跟业务发展和用户需求的变化，始终保持较高的服务水平。例如，在某零售企业的实践中，通过用户反馈分析，每月更新知识条目超过 2000 条，确保了知识库的时效性和准确性 。

合力亿捷的智能客服系统还建立了“监测-分析-迭代”的闭环优化流程 。通过围绕自动解决率、客户满意度（CSAT）、平均处理时长等核心 KPI 进行持续监控和评估，企业可以洞察到智能客服的优化空间 。例如，通过质检洞察发现某些问题的回答不准确或不完善，就可以针对性地调整提示词设计、优化知识颗粒度配置，甚至补充新的知识条目到知识库中 。合力亿捷的工单管理系统结合知识库更新机制，能够快速响应客户服务需求，确保知识库的时效性和准确性 。

此外，系统还支持通过周期性的 A/B 测试来验证不同优化策略的效果，从而科学地指导优化方向，实现动态优化与价值闭环 。一些先进的模型技术，如 DeepSeek 模型所采用的增量训练技术，支持在线学习，可以将模型参数更新量减少 70%，这极大地提升了模型迭代的效率和灵活性 。通过构建从客户反馈采集到数据洞察，再到业务策略优化和价值落地的端到端数据链路，合力亿捷推动产品和服务的持续迭代升级，形成了良性的自我优化循环 。通过这种持续的自主学习和优化，合力亿捷的智能客服系统能够越用越智能，不断提升复杂问题的解决率，优化用户体验，并最终帮助企业实现服务效率和客户满意度的双重提升。

3. AI 应用与落地难点：挑战与破局

3.1 AI Agent 落地面临的四大核心卡点

尽管 AI Agent（智能体）在智能客服领域展现出巨大的潜力，能够自主理解、决策并执行任务，从而提升服务效率和用户体验，但在实际落地过程中，企业仍面临诸多挑战。这些挑战主要集中在知识库的构建与维护、复杂意图的理解与上下文关联、多轮对话的管理与流程引导，以及 AI 与人工的协同等方面。这些卡点若不能得到有效解决，将严重影响 AI Agent 的应用效果和投资回报。根据毕马威 2025 年第一季度的调研数据，虽然企业从 AI 实验阶段迈向试点阶段的比例从 37%跃升至 65%，但在生产环境中规模化部署 AI Agent 的企业比例仍停滞在 11%左右 。合力亿捷在其行业洞察中，将 AI Agent 落地难的症结归纳为四大核心卡点，这些卡点共同构成了从“实验室”走向“生产线”的鸿沟 。

下表总结了 AI Agent 落地面临的四大核心卡点及其具体表现：

核心卡点 (Core Challenge)	具体表现 (Specific Manifestations)	影响 (Impact)
卡点一：知识库构建与维护的挑战	知识获取与梳理难度大；知识更新滞后；覆盖度与深度难以平衡；冷启动与持续优化困难	AI Agent 回答不准确、过时，影响用户体验与企业声誉
卡点二：复杂意图理解与上下文关联的难题	用户表达口语化、模糊、多意图交织；NLU 技术在处理方言、术语时仍有不足；上下文关联丢失或错误	AI Agent 答非所问，对话不连贯，用户体验差

卡点三：多轮对话管理与流程引导的困境	对话流程设计复杂性与灵活性难以平衡；对话状态跟踪与维护困难；处理用户打断、澄清能力不足；场景化对话策略设计难	对话效率低下，无法完成预定任务，用户 frustration 升级
卡点四：AI 与人工协同的平滑过渡与效率平衡	人机切换时机与方式难把握；信息传递与共享效率低；人工对 AI 辅助工具接受度不高；自动化与个性化服务难以平衡	整体服务效率不升反降，用户体验受损，一线员工抵触

表 1: AI Agent 落地面临的四大核心卡点

3.1.1 卡点一：知识库构建与维护的挑战

知识库是 AI Agent 智能客服的核心大脑，其质量和完备性直接决定了 Agent 的服务能力。然而，构建和维护一个高质量、动态更新的知识库面临诸多挑战。首先，知识获取与梳理难度大。企业内部的业务知识往往分散在不同的部门、系统和文档中，格式各异，包括结构化的数据库、半结构化的 Excel 表格以及非结构化的 Word 文档、PDF 文件、邮件、聊天记录等。将这些海量、异构的知识进行有效提取、清洗、整合和结构化，需要投入大量的人力和时间成本，并且对业务专家的经验依赖性强 。

其次，知识更新滞后性问题突出。业务政策、产品信息、服务流程等处于不断变化之中，如果知识库不能及时同步更新，AI Agent 给出的答案就可能出现错误或过时，严重影响用户体验和企业声誉。例如，某电商大促期间，由于 AI 客服的知识库未能及时更新促销规则，导致错误应答率高达 17% 。再次，知识库的覆盖度和深度难以平衡。既要保证知识库能够覆盖用户可能咨询的各个方面，又要确保每个知识点的深度和准确性，这对知识库的设计和管理提出了很高的要求。过度依赖历史数据，可能导致知识库难以应对突发事件的非标准化咨询 。

最后，知识库的冷启动和持续优化也是一大难题。如何快速构建一个可用的初始知识库，并通过实际运营数据不断发现知识缺口、优化知识内容，是一个持续迭代的过程，需要有效的工具和方法论支持。企业级应用更强调“输出可控性”，即精准、高效、一致地回答问题，这需要将散乱信息转化为机器可直接调用的“标准化零件”，避免 AI 在庞杂内容中“大海捞针”或断章取义 。

3.1.2 卡点二：复杂意图理解与上下文关联的难题

准确理解用户的复杂意图并在多轮对话中保持上下文的连贯性，是衡量 AI Agent 智能性的关键指标，也是落地过程中的一大难点。用户在实际咨询时，其表达往往具有口语化、模糊性、多意图交织以及隐含情感等特点。例如，用户可能会说“我前几天买的那个手机，屏幕有点问题，能退吗？”，这句话包含了“订单查询”、“故障描述”、“退货申请”等多个意图，并且需要结合上下文（“前几天买的”）和用户信息进行理解。AI Agent 需要具备强大的自然语言理解（NLU）能力，特别是深度语义理解能力，才能准确捕捉这些复杂意图。然而，当前的技术在处理一些特殊情况时仍面临挑战，例如，当用户使用方言、行业术语或网络流行语时，语义识别的错误率可能高达 30%-45% 。

此外，上下文关联的丢失或错误也是常见问题。在多轮对话中，如果 AI Agent 不能准确记忆和利用之前的对话信息，就可能导致回答不相关、重复提问或者逻辑混乱，严重影响对话体验。例如，在长对话场景中，系统可能会遗忘用户的初始需求 。突发的话题切换也对上下文管理能力提出了考验，用户可能从咨询“套餐价格”突然转到“投诉信号问题”，AI Agent 需要能够快速适应这种变化并做出恰当响应 。解决这些问题需要更先进的 NLU 模型、更精细的上下文跟踪机制以及更强大的知识推理能力。

3.1.3 卡点三：多轮对话管理与流程引导的困境

多轮对话管理是 AI Agent 实现复杂任务处理和服务闭环的关键能力，但在实际应用中，如何有效地设计和管理多轮对话流程，并进行自然的引导，是一大挑战。首先，对话流程设计的复杂性与灵活性难以平衡。过于僵化的流程设计（如严格的菜单式导航）会限制用户的表达自由，降低交互的自然度；而过于灵活的对话管理，则可能导致对话偏离主题、效率低下，甚至无法完成预定任务。AI Agent 需要在引导用户和尊重用户意图之间找到平衡。

其次，对话状态的准确跟踪与维护至关重要。在多轮交互过程中，AI Agent 需要准确记录用户已提供的信息、当前所处的对话阶段、待完成的任务项等状态信息。如果状态跟踪出现偏差，就可能导致流程中断或错误执行。例如，在办理退货流程中，如果 Agent 未能准确记录用户已确认的退货原因，后续的流程就无法顺利进行。

再次，处理用户打断、澄清和纠错的能力有待提升。用户在对话过程中可能会随时打断、提出新的问题、要求澄清或者纠正之前提供的信息。AI Agent 需要具备良好的鲁棒性，能够优雅地处理这些情况，并回到正确的对话轨道。此外，如何在不同业务场景下设计合理的对话策略也是一个难点。例如，售前咨询场景可能需要更主动的引导和营销，而售后投诉场景则需要更谨慎的处理和安抚。AI Agent 需要根据场景动态调整对话策略，这需要大量的场景化数据和精细的策略调优。

3.1.4 卡点四：AI 与人工协同的平滑过渡与效率平衡

在智能客服体系中，AI Agent 与人工坐席的协同工作至关重要，目标是实现效率最大化和用户体验最优化。然而，实现两者之间的平滑过渡和高效协同面临诸多挑战。首先，人机切换的时机和方式难以把握。如果 AI Agent 在无法解决问题或用户表现出负面情绪时，不能及时、顺畅地将对话转接给人工坐席，会导致用户 frustration 升级，降低满意度。反之，如果过早或过多地依赖人工介入，则无法充分发挥 AI 的效率优势。如何设定合理的“智能熔断机制”，例如当对话轮次超过一定阈值、检测到用户负面情绪达到临界值或涉及敏感信息时自动转接人工，是一个需要仔细权衡的问题 。

其次，信息在不同坐席间的传递与共享效率低下。当 AI Agent 将对话转接给人工坐席时，如何将之前的对话历史、用户意图、已尝试的解决方案等信息完整、准确地传递给人工坐席，避免用户重复描述问题，是提升协同效率的关键。如果信息传递不畅，人工坐席需要重新了解情况，不仅浪费时间，也影响用户体验。

再次，人工坐席对 AI 辅助工具的接受度和使用熟练度也会影响协同效果。AI 可以为人工坐席提供实时话术建议、知识推荐、客户画像等辅助功能，但如果工具设计不符合人工坐席的工作习惯，或者缺乏有效的培训，这些辅助功能可能难以发挥应有的作用 。此外，如何平衡 AI 的自动化处理与人工的个性化服务也是一个核心问题。虽然 AI 可以处理大量重复性咨询，但在处理复杂、敏感或需要情感关怀的问题时，人工服务的价值依然不可替代。企业需要在提升自动化率的同时，保障人工服务的质量和温度，避免过度自动化导致用户体验下降 。

3.2 合力亿捷的破局之道与实践经验

面对 AI Agent 在智能客服领域落地所面临的诸多挑战，合力亿捷凭借其多年的行业经验和技术积累，提出了一系列针对性的破局之道，并在实践中取得了显著成效。这些解决方案覆盖了从知识工程优化、深度语义理解、对话策略优化到人机协同工作流等多个层面，旨在帮助企业克服 AI 应用落地的难点，充分释放大模型智能客服的潜力。其核心理念在于将复杂的业务需求与先进的 AI 技术进行深度融合与协同，通过业务流程的深度拆解和 AI 工程化的有效实施，构建稳定、可靠且可被有效监控的 AI Agent 运行体系 。

合力亿捷针对 AI Agent 落地核心卡点的破局之道：

核心卡点 (Core Challenge)	合力亿捷的破局之道 (Heli-E's Solution)	关键技术/方法 (Key Technologies/Methods)	实践效果/目标 (Practical Effects/Goals)
知识库构建与维护	知识工程优化与智	“三源融合”知识整	提升知识库质量与

的挑战	能化管理	合；增量学习与动态更新；知识图谱技术；RAG 策略	时效性，降低维护成本，减少 AI“幻觉”
复杂意图理解与上下文关联的难题	深度语义理解与上下文感知技术	BERT/GPT 等预训练模型微调；对话状态跟踪（DST）；情感分析；记忆网络	提升意图识别准确率（高达 90-98%），增强对话连贯性
多轮对话管理与流程引导的困境	强化学习驱动的对话策略优化	FSM 与强化学习混合架构；对话路径优化算法；基于 RL 的纠错模型	提升对话效率，优化任务完成率，增强系统鲁棒性
AI 与人工协同的平滑过渡与效率平衡	人机协同工作流与智能调度机制	智能熔断与无缝转接；AI 辅助人工（话术推荐、知识查询）；智能工单系统；分层响应机制	提升整体服务效率与质量，保障复杂问题处理，优化用户体验

表 2: 合力亿捷针对 AI Agent 落地核心卡点的破局之道

3.2.1 知识工程优化与智能化管理

针对知识库构建与维护的挑战，合力亿捷强调知识工程的优化和智能化管理。首先，在知识获取与构建方面，合力亿捷提供了高效的工具和方法来整合企业内外部知识资源。通过支持批量上传 FAQ、操作指南、SOP 等文档，系统能够自动完成向量索引的生成，并结合业务标签实现精细化检索 。其自研的自动知识抽取工具，据称可以将 PDF 文档的图谱构建效率提升 70% ，这大大降低了知识库冷启动的难度和成本。合力亿捷采用“三源融合”策略，即综合利用结构化数据（通过 ETL 工具清洗入库）、非结构化数据（运用 TextRank 等算法提取关键信息）和外部数据（通过 APIs 实时对接）来构建全面、动态的知识库 。

其次，在知识库的持续更新与优化方面，合力亿捷建立了动态的知识库管理机制。系统能够实时分析客户意图和用户反馈，自动发现知识缺口和过时信息，并辅助人工进行知识库的更新和维护 。例如，某零售企业通过用户反馈分析，每月更新知识条目超过 2000 条 。通过 RAG（Retrieval Augmented Generation）策略，结合业务知识

库，可以有效提升智能应答的准确率与覆盖面，保障 AI 在复杂业务场景中的专业输出能力，同时减少大模型“幻觉”的风险 。此外，合力亿捷还注重知识库的行业化与场景化适配，针对不同行业的特定需求构建领域知识图谱，例如电商行业的商品属性与促销规则图谱，金融行业的合规条款与风险提示关联图谱等 ，从而提升知识库的专业性和实用性。

3.2.2 深度语义理解与上下文感知技术

为了克服复杂意图理解与上下文关联的难题，合力亿捷在其智能客服系统中集成了先进的深度语义理解和上下文感知技术。在意图识别方面，合力亿捷采用基于 BERT、GPT 等预训练大模型，并结合行业语料进行微调，以提升对用户口语化、模糊化表达的精准理解能力，其意图识别准确率据称可达 90%以上，部分场景下甚至达到 92% ，在微信群客服场景下，AI 语义理解的准确率更是达到了 98%以上 。

这些模型经过大规模数据的训练，能够学习语言的语义、语法和语用知识，从而精准判断用户的需求。在上下文管理方面，合力亿捷的智能客服系统具备强大的多轮对话管理能力，能够记录和理解用户的历史对话信息，确保系统根据上下文准确理解用户意图，支持长达 30 轮以上的复杂对话而不丢失上下文 。例如，其采用的“对话状态树”技术，能够有效跟踪对话状态，动态调整对话路径。某电商企业采用对话状态跟踪（DST）与记忆网络结合方案，将多轮对话连贯性提升 40% 。

此外，系统还具备情感分析能力，通过分析用户文本或语音中的情感特征，识别用户的情绪状态（如满意、不满意、愤怒等），从而为后续的交互策略提供依据，例如在检测到用户不满时自动触发安抚话术或升级服务 。合力亿捷通过基于注意力机制的双向 LSTM 模型进行情感分析，情绪识别准确率突破 85% 。这种深度语义理解和上下文感知技术的结合，使得 AI Agent 能够更智能地与用户进行交互，准确捕捉用户真实需求，并提供连贯自然的对话体验。

3.2.3 强化学习驱动的对话策略优化

在复杂场景交互与多轮对话管理方面，合力亿捷积极探索和应用强化学习等先进技术来优化对话策略，提升 AI Agent 的交互智能性和任务完成能力。传统的基于规则或有限状态机的对话管理系统，往往难以应对复杂多变的对话场景和用户行为。

合力亿捷的对话管理引擎采用了有限状态机（FSM）与强化学习结合的混合架构 。这种架构既保证了核心业务流程的可控性和稳定性，又通过强化学习赋予了系统动态优化对话路径和策略的能力。通过在与用户的真实交互中不断学习和积累经验，AI Agent 能够根据不同的对话状态和用户反馈，自主调整其应答方式和引导策略，以期最大化长期奖励（如任务完成率、用户满意度等）。例如，系统可以通过强化学习来优化对话流程，减少不必要的对话轮次，将平均对话轮次从 5.3 轮降至 3.8 轮 。

此外，基于强化学习的纠错模型，可以使错误恢复成功率提高至 82%，这意味着当对话出现偏差或用户表达不明确时，AI Agent 能够更有效地引导对话回到正轨。这种数据驱动的对话策略优化机制，使得 AI Agent 能够越用越聪明，不断提升其在复杂场景下的交互效率和任务解决能力。

3.2.4 人机协同工作流与智能调度机制

为了实现 AI Agent 与人工坐席的高效协同，合力亿捷构建了智能化的人机协同工作流和调度机制。核心思想是明确人机边界，发挥各自优势，实现“AI 处理标准化、人工聚焦复杂化”的服务模式。合力亿捷提出了“智能应答—人工介入—AI 辅助”的分级响应机制 。首先，AI Agent 作为一线服务力量，处理大部分重复性、标准化的咨询，例如查询类、办理类等高频低复杂度问题，据称可承接 80%的重复性咨询 。

其次，当 AI Agent 遇到无法独立解决的复杂问题、检测到用户强烈负面情绪或涉及敏感操作时，系统会自动触发智能熔断机制，将对话无缝转接给人工坐席 。在转接过程中，系统会将完整的对话历史、用户意图分析、已尝试的解决方案等信息同步给人工坐席，避免用户重复描述，提升交接效率。再次，在人工服务过程中，AI 仍然扮演着重要的辅助角色。例如，合力亿捷的智能质检系统可以实时监测坐席的情绪波动，并自动推送话术建议，帮助人工坐席更好地与用户沟通，据称可将平均处理时长缩短 30% 。

此外，AI 还可以为人工坐席提供知识推荐、客户画像分析、多语言翻译、自动生成跟进记录等辅助功能 。合力亿捷在其金融行业解决方案中，提出了“AI 为盾、人工为矛”的协同体系，通过智能路由分配、数据闭环和能力互补来实现人机协同 。通过这种精细化的人机协作流程和智能调度，合力亿捷旨在最大限度地提升整体服务效率和质量，同时确保复杂和敏感问题能够得到妥善处理，从而在提升自动化水平的同时，保障用户体验。

4. 合力亿捷大模型智能客服实践案例深度剖析

合力亿捷将其大模型智能客服解决方案应用于多个行业，积累了丰富的实践经验。这些案例不仅展示了其技术实力，也为其他企业提供了宝贵的参考。以下将选取几个代表性案例进行深度剖析，以揭示合力亿捷如何帮助客户实现服务升级和业务价值提升。

4.1 案例一：某大型金融企业智能客服升级

在金融行业，客户服务对安全性、合规性和专业性要求极高。某大型金融企业在引入

合力亿捷的智能客服解决方案后，成功实现了服务模式的智能化升级。该方案的核心在于构建一个能够处理复杂金融咨询、保障交易安全、并符合严格监管要求的 AI Agent 体系。

首先，在意图理解与精准应答方面，合力亿捷利用大模型的深度语义理解能力，结合金融行业特有的知识图谱（包括产品信息、合规条款、风险提示等），显著提升了 AI Agent 对用户金融咨询的意图识别准确率。例如，用户咨询关于理财产品收益率、风险评估、贷款申请流程等专业问题时，AI Agent 能够给出准确、规范的解答。其次，在安全与合规强化方面，该方案特别设计了多重保障机制。例如，在涉及敏感操作如转账、密码修改时，系统会强制进行人工复核或双因子验证（如语音识别+活体检测） 。

同时，实时风控引擎会对对话内容进行持续监控，禁用“保本”、“稳赚”等违规词汇，确保服务过程的合规性 。据案例显示，该金融客户引入合力亿捷呼叫中心系统后，客服成本降低了 40%，客户满意度从 83%跃升至 95% 。这充分证明了在金融这样高度敏感的领域，通过精细化的技术设计和严格的风险控制，大模型智能客服同样能够发挥巨大价值，实现降本增效与体验提升的双重目标。

4.2 案例二：某知名电商平台智能导购与售后支持

电商平台面临着海量的用户咨询，尤其是在大促期间，咨询量更是呈爆发式增长。某知名电商平台通过引入合力亿捷的大模型智能客服解决方案，有效提升了智能导购和售后支持的效率与质量。该方案重点解决了电商场景下的高并发处理、动态知识库管理以及个性化服务推荐等难题。

在高并发处理方面，合力亿捷的弹性算力调度能力发挥了关键作用，能够根据实时咨询量自动扩容至千节点集群，确保在如“双十一”等高峰时段依然能够提供稳定、流畅的服务 。在动态知识库管理方面，针对电商促销规则复杂多变的特点（如满减、折扣、赠品等多重组合），合力亿捷提供了强大的规则引擎，能够将促销文案自动转化为可执行的逻辑，确保 AI Agent 能够准确理解和应用最新的促销信息 。

在个性化服务与推荐方面，通过分析用户的历史浏览记录、购买行为、以及实时咨询内容，AI Agent 能够构建用户画像，并提供个性化的商品推荐和购物建议，从而提升转化率和客单价 。例如，合力亿捷的美妆行业客服方案，通过用户画像技术为用户提供个性化的产品推荐和服务，提高了用户的购买意愿和忠诚度 。某电商平台的实践显示，通过将历史服务数据与实时场景结合，系统可自动生成 3-5 个最优回复方案，并依据用户画像进行个性化选择 。这些能力的综合运用，使得该电商平台能够有效应对海量咨询，提升用户购物体验，并最终促进销售额的增长。

4.3 案例三：某公共服务热线智能化转型

公共服务热线（如 12345 政府服务热线）是连接政府与市民的重要桥梁，承担着政策咨询、投诉建议、民生服务等多种职能，对话务处理的及时性、准确性和规范性要求很高。某公共服务热线引入合力亿捷的大模型智能客服解决方案后，成功实现了智能化转型，显著提升了服务效率和管理水平。该方案的核心优势在于其强大的自然语言理解能力、多轮对话管理能力以及对政策法规的精准把握。

首先，在意图识别与智能派单方面，大模型能够准确理解市民通过电话或在线渠道提出的各类诉求，即使是口语化、模糊化的表达，也能进行有效解析。例如，北京市海淀区政府利用大模型赋能接诉即办场景，实现智能化重构，逐步替代依靠人工给工单分类、打标签及识别处置分派单位等工作，提升了热线分析工作效率和智能化水平 。上海 12345 政务热线引入“星辰”政务大模型，为话务员提供智能话务总结、智能填单/派单、智能知识库问答等功能 。

其次，在知识库构建与动态更新方面，针对政策法规繁多且时常更新的特点，合力亿捷的解决方案支持快速构建和动态更新政策知识库。通过 RAG 策略，结合最新的政策文件，确保 AI Agent 给出的答复准确、权威。例如，12345 热线通过 DeepSeek 大模型动态更新政策库，问答准确率提升 10% 。此外，在多模态交互与方言支持方面，考虑到公共服务对象的广泛性，合力亿捷的解决方案支持语音、文本等多种交互方式，并针对部分地区提供了方言识别能力。例如，在某省 12345 热线部署方言模型后，语音识别准确率从 65%提升至 92% 。这些能力的提升，使得公共服务热线能够更高效地响应市民需求，提供更精准的政策解读，从而提升政府服务效能和市民满意度。

5. 未来展望：AI Agent 与智能客服新趋势

5.1 AI Agent 的演进方向与核心技术

AI Agent（智能体）作为智能客服领域未来的核心发展方向，其演进趋势正朝着更高级的自主性、更强的任务执行能力和更深度融合的人机协作模式迈进。Gartner 连续两年将“AI Agent”列为战略技术趋势之首，预示着其巨大的发展潜力 。未来的 AI Agent 将不再仅仅是能说会答的对话机器人，而是具备理解复杂指令、调用各种工具（如 API、数据库、应用程序）、自主规划并执行多步骤任务、甚至在特定场景下做出决策的能力。其演进方向可以概括为：从“理解用户”到“发起流程”，从“能回答”到“能解决”，从“人执行”到“人协同” 。这意味着 AI Agent 将更加主动地参与到业务流程中，成为企业运营不可或缺的智能伙伴。

实现这些演进方向，依赖于一系列核心技术的突破与融合。首先，更强大的基础模型能力是基石。这包括持续提升大语言模型（LLM）的理解、推理、生成和规划能力，使其能够处理更复杂、更模糊的指令，并生成更可靠、更符合逻辑的行动计划。同

时，多模态能力的深度融合也至关重要，未来的 AI Agent 需要能够理解和生成文本、语音、图像、视频等多种模态的信息，以实现更自然、更丰富的交互。

其次，工具使用与外部知识整合能力是提升 Agent 任务执行能力的关键。AI Agent 需要具备学习和调用各种外部工具（如搜索引擎、数据库查询、软件 API）的能力，以获取实时信息、执行具体操作，并将外部知识有效地融入自身的决策和行动中。RAG（Retrieval Augmented Generation）技术在这方面扮演着重要角色，通过结合外部知识库来增强模型的回答准确性和时效性 。再次，记忆与学习机制的持续进化是 Agent 实现长期自主学习和个性化服务的基础。这包括更高效的上下文记忆技术，能够支持更长、更复杂的对话；以及更强大的持续学习和自适应能力，使 Agent 能够从与用户和环境的交互中不断学习，优化自身的行为策略，并形成对用户偏好和习惯的深刻理解。最后，可解释性与可控性技术的进步，对于确保 AI Agent 行为的可靠性、安全性和符合伦理规范至关重要。我们需要能够理解 Agent 的决策过程，并在必要时进行干预和调整，防止其产生有害或非预期的行为。

5.2 大模型驱动下智能客服的未来形态

在大模型技术的持续驱动下，智能客服的未来形态将发生深刻的变革，呈现出更加智能化、个性化、主动化和情感化的特征。未来的智能客服将不再仅仅是一个被动的信息查询和问题解答工具，而是进化为一个能够深度理解用户需求、主动提供价值、并与用户建立情感连接的智能伙伴。首先，服务将更加无感和无处不在。智能客服将深度融入用户的生活和工作场景，通过多种渠道（如 App、智能音箱、车载系统、可穿戴设备等）提供无缝的服务体验。用户甚至可能意识不到正在与 AI 交互，服务会以一种自然、便捷的方式触达用户，例如，在企业微信客服中，系统自动识别客户情绪波动，无缝转接人工时同步对话记录 。

其次，个性化服务将达到新高度。基于对大数据的深度挖掘和对用户画像的精准刻画，未来的智能客服能够提供高度定制化的服务和产品推荐。例如，系统可以根据用户的消费习惯、健康状况、甚至实时情绪，动态调整服务策略，提供“千人千面”的极致体验，如结合当地血糖健康数据推荐减糖方案 。再次，主动服务和需求预判将成为常态。智能客服将具备更强的预测能力，能够基于用户行为数据和环境信息，提前预判用户可能遇到的问题或需求，并主动提供帮助或解决方案。例如，用户浏览三次价格页未下单，系统自动触发保价承诺推送 。

最后，情感交互和共情能力将显著增强。未来的智能客服将不仅仅是理性的问题解决者，更能理解和回应人类的情感需求。通过更先进的情感计算和自然语言生成技术，AI 将能够进行更具同理心的对话，提供情感支持和安慰，从而建立更深层次的客户关系。例如，系统不仅能捕捉关键词，更能通过情感分析引擎判断愤怒情绪，自动触发应急方案 。这种从“成本中心”到“价值中心”再到“情感连接中心”的转变，将是未来智能客服的核心特征。

5.3 合力亿捷在 AI Agent 领域的探索与布局

面对 AI Agent 这一未来趋势，合力亿捷积极进行探索与布局，致力于将其先进的 AI 能力与客户服务的实际需求相结合，打造下一代智能客服解决方案。其核心战略是构建一个以 AI Agent 为驱动，能够实现从“语言智能”到“任务智能”跨越的智能客服平台 。合力亿捷认识到，未来的客服系统不仅要能“理解用户”，更要能“发起流程”；不仅要“能回答”，更要“能解决”实际问题 。为此，合力亿捷在技术研发、平台构建和行业应用等多个层面进行了深入探索。

在技术平台构建方面，合力亿捷推出了自研的低代码 AI Agent 平台 。该平台允许业务专家通过可视化的流程编排和组件化设计，无需编码即可快速搭建和部署针对特定业务场景的 AI Agent。这些 Agent 不仅支持复杂的多轮对话，还具备工单流转、表单生成、系统对接等任务执行能力，支持多 Agent 协同和人机协作，旨在辅助人工坐席构建“超级客服” 。同时，该平台支持多模型接入，能够灵活适配 ChatGPT、通义千问、百度文心一言、华为盘古、智谱清言以及 DeepSeek 等主流大模型，并支持行业定制模型的快速蒸馏、训练和部署，确保企业能够持续接轨前沿 AI 能力，并根据自身需求进行定制化开发 。此外，合力亿捷还提供本地化部署方案（如 HollyONE 一体机），以满足政企、金融机构等对数据安全和网络稳定性的高要求，构建数据安全与智能服务的双重保障 。

在行业应用与赋能方面，合力亿捷将其 AI Agent 解决方案应用于零售、教育、金融、出海等多个领域，帮助客户实现全时服务进化 。例如，在连锁便利店场景，通过坐席辅助 Agent 提升人均处理能力和客服接起率；在高校运营服务中，部署基于大模型的知识库与机器人，实现多校区统一服务；在出海业务中，接入 WhatsApp Business API，支持多语言实时翻译与跨时区自动服务 。合力亿捷还强调客服人员职能的转变，从传统的“答题人”转变为 Agent 的“编排师”和“运营优化者”，通过配置业务流程、维护知识结构、优化标签体系，为系统持续注入新的业务逻辑，实现客服组织的“提效+进化” 。通过这些探索和布局，合力亿捷旨在帮助企业打破传统客服在时间、空间和人力的限制，构建具备弹性、智能和可持续增长能力的客服体系，使客服从服务的终点转变为企业效率和增长的起点 。

6. 结论与建议

6.1 大模型在智能客服领域应用的总结

大模型技术的引入，无疑为智能客服领域带来了革命性的变革。它极大地提升了智能客服的自然语言理解能力、多轮对话管理能力、知识推理能力以及个性化服务能力，使得智能客服不再是简单的问答机器人，而是能够处理更复杂场景、提供更精准服

务、甚至具备一定情感交互能力的智能伙伴。通过行业知识的增强、复杂场景交互的优化以及自主学习机制的引入，大模型驱动的智能客服正在帮助企业显著提升服务效率、降低运营成本，并改善客户体验。合力亿捷等企业在智能客服大模型应用方面的战略布局和实践案例，充分展示了这一技术路径的可行性和巨大潜力。从金融行业的合规智能应答，到电商平台的高并发导购，再到公共服务热线的智能化转型，大模型智能客服都展现出其独特的价值。未来，随着 AI Agent 技术的不断成熟，智能客服将向着更自主、更智能、更深度融合的方向发展，成为企业数字化转型和提升核心竞争力的关键支撑。

然而，大模型在智能客服领域的应用也并非一帆风顺。知识库的构建与维护、复杂意图的理解与上下文关联、多轮对话的管理与流程引导，以及 AI 与人工的协同等，依然是企业在落地过程中需要重点攻克的难题。这些挑战不仅涉及技术层面，更关乎业务流程的再造、组织架构的调整以及企业文化的适应。合力亿捷等先行者通过知识工程优化、深度语义理解技术、强化学习驱动的对话策略以及智能化的人机协同机制，为行业提供了宝贵的破局思路和实践经验。成功应用大模型智能客服，需要企业在技术选型、场景选择、数据治理、人才培养以及持续迭代等多个方面进行系统性的规划和投入。

6.2 对企业应用大模型智能客服的建议

对于计划或正在应用大模型智能客服的企业，以下几点建议可供参考：

1. **明确战略定位与业务目标**：企业应首先明确引入大模型智能客服的战略目标，是为了降本增效、提升客户体验、还是赋能业务创新。基于清晰的定位，选择合适的应用场景和技术路径，避免盲目跟风。

2. **重视高质量数据的积累与治理**：大模型的性能高度依赖于训练数据的质量和规模。企业应建立完善的数据采集、清洗、标注和管理机制，确保数据的准确性、一致性和时效性。高质量的知识库是智能客服发挥价值的基础。

3. **选择合适的技术与合作伙伴**：市场上大模型和相关技术方案众多，企业应根据自身的技术实力、业务需求和预算，选择成熟稳定、可解释性强、且易于集成和扩展的技术方案。与具有丰富行业经验和成功案例的技术伙伴合作，可以少走弯路。

4. **循序渐进，从试点到推广**：大模型智能客服的应用是一个持续迭代和优化的过程。建议企业从业务痛点明确、价值易衡量的场景入手，进行小范围试点，积累经验，验证效果，再逐步扩大应用范围和深度。

5. **关注人机协同与组织变革**：智能客服并非要完全取代人工，而是要与人工坐席形成高效协同。企业需要设计合理的人机交互流程，确保平滑过渡，并关注一线员工的培训和赋能，帮助他们适应新的工作模式，发挥 AI 的最大价值。

6. **持续监控、评估与优化**：建立完善的监控和评估体系，持续跟踪智能客服的关键

绩效指标（KPIs），如问题解决率、客户满意度、平均处理时长等。基于数据反馈，不断优化模型、知识库和对话策略，实现智能客服系统的持续进化和价值提升。

7. **关注安全、合规与伦理**：在应用大模型智能客服时，企业应高度重视数据安全、用户隐私保护以及算法偏见等问题，确保符合相关法律法规和伦理规范，建立负责任的 AI 应用体系。

通过以上策略的实施，企业可以更有效地应用大模型智能客服，提升服务能力和竞争力，最终实现可持续的业务增长。

最后编辑时间：2025 年 7 月 21 日

官网：www.hollycrm.com

扫码加微信，助您提升服务效率与客户满意度。