解决方案

第四代英特尔® 至强® 可扩展处理器
BigDL-LLM 大模型开源库
OpenVINO™ 工具套件
大语言模型 (LLM)
智慧医疗

让医疗机构高品质、低成本部署大语言模型

英特尔助惠每医疗大模型方案在至强® 平台上实现双维优化

人工智能 (Artificial Intelligence，AI) 在医疗领域的落地，正为医疗行业的信息化、数字化进程带来新一轮质变，它可以通过优化诊疗效率，改善患者体验，提升全民健康服务水平。近年来引人瞩目的大语言模型 (Large Language Model，LLM，以下简称“大模型”) 技术，以其更强的学习性能和更优的模型拟合效果，更是为 AI 在医疗领域中的应用注入了新动力。

领先的医疗人工智能解决方案提供商北京惠每云科技有限公司 (以下简称“惠每科技”) 以其临床决策支持系统 (Clinical Decision Support System，CDSS) 产品和海量医疗数据为基础，积极引入大模型技术来为医疗机构打造更高品质的医疗 AI 应用。但这并非易事，医疗机构对数据安全的顾虑，以及昂贵的 AI 专用芯片部署和应用成本是阻碍这一进程的两大“拦路虎”。

为帮助医疗机构在其广泛应用的 IT 平台上实现高质量、低成本的大模型私有化部署，惠每科技与英特尔展开技术合作，对医疗大模型在英特尔® 至强® 平台上的推理性能实施优化。双方在以第四代英特尔® 至强® 可扩展处理器为核心的硬件基础设施上，采用 BigDL-LLM 大模型开源库与 OpenVINO™ 工具套件打造了两种大模型优化方案。实际部署后的测试结果表明：在保证精准度以及不增加成本的前提下，优化后的方案可有效提升鉴别诊断、出院记录等医疗 AI 应用的处理效率，并获得了医生的认可。

“大模型等 AI 技术，正在像水、电等基础设施能力一样，在医疗机构的未来医疗服务体系中具有无可替代的价值。为帮助医疗机构应对在大模型私有化部署中面临的成本高、落地难等难题，我们与英特尔展开合作，在第四代英特尔® 至强® 可扩展处理器的基础上，以 BigDL-LLM 库和 OpenVINO™ 工具套件作为推理优化方案的左右手，双管齐下，打造高质量、低成本的医疗 AI应用并获得了预期推广成果。”

王实
CTO
北京惠每云科技有限公司

大模型技术正成为医疗 AI 发展新动力

得益于强劲算力与海量数据的加持，高速发展的大模型技术正以一系列极具震撼力的应用场景，成为 AI 领域令人瞩目的焦点。与传统 AI 技术相比，参数规模庞大 (动辄数百乃至上千亿参数量级) 的大模型不仅具备更强的学习性能和更优的模型拟合效果，还拥有高效的迁移学习能力，这能让用户在一个通用模型上完成不同类型的任务。此外，对思维链 (Chain of Thought，CoT) 的良好支持，也使大模型应用补齐了传统 AI 在逻辑推理能力上的短板。

上述优势不仅让大模型技术与应用成为了各大科技巨头争先探索的蓝海，也推动了其在社交、金融、电商以及医疗等垂直领域的迅速落地，并显现出巨大的市场潜力。有相关预测数据表明，大模型市场在未来数年都将保持 21.4% 的年复合增长率 (Compound Annual Growth Rate, CAGR)，到 2029 年或达 408 亿美元的市场规模[1]。

在医疗行业，无论是面向大众提供普惠医疗服务的智能问答与家庭医疗助手，还是有助于医护人员提升效率的 AI 导诊和临床辅助诊疗应用，或是加速医疗影像处理效能，提高大病、恶疾早期发现率的 AI 阅片等，众多医疗 AI 企业正在借助大模型来提升这些应用的性能，帮助医疗机构在诊疗服务全流程中实现更全面且优质的服务能力、更精准的结果输出以及更广泛的运用范围。而其中，深耕医疗信息化多年，具有出色医疗 AI 应用研发能力和头部优势的惠每科技，也在这一趋势中将大模型作为其技术再突破、服务再提升的重要抓手。

一直以来，惠每科技的 CDSS 产品 (如医院端核心应用 Dr.Mayson、临床科研平台 Darwin 等)，都是通过实时数据分析与事中智能提示等核心能力的打造，助力医疗机构在临床诊疗决策、病案与病历管理、诊疗风险预警以及医保费用管理等环节中提升服务质量、诊疗效率和管理效能。

而这些场景对自然语言处理 (Natural Language Processing, NLP)、计算机视觉 (Computer Vision, CV) 等 AI 能力的需求，正好让大模型有了用武之地。如图 1 所示，在惠每科技最新发布的 CDSS 3.0 架构中，新一代 AI 大数据处理平台已集成了医疗大模型。这些医疗大模型是通过海量数据在一系列大模型上重新训练而成的，不仅融合了惠每科技在医学知识库、专家系统上的雄厚知识积累，也凝集了其落地于 600 余家医疗机构所获得的丰富实战经验，已在病历生成等场景中获得了成功运用。

推动高质量、低成本的私有化部署，是医疗大模型落地的主要挑战

然而在推进医疗大模型落地的过程中，惠每科技面临着严峻挑战，其中主要的是如何帮助医疗机构实现高质量、低成本的私有化部署:

- **降低建设成本：** 传统上的大模型训练和推理工作通常需要借助专用加速芯片来完成，但这类芯片昂贵的价格往往让医疗机构望而却步，同时其普遍缺货或供货周期较长的问题也会大幅拉长方案的建设周期。

- **保障数据安全：** 行业的特殊性使医疗机构对数据安全、隐私保护极为重视，任何医疗数据都不能离开安全可控的内网环境，所以医疗大模型需要进行私有化部署。

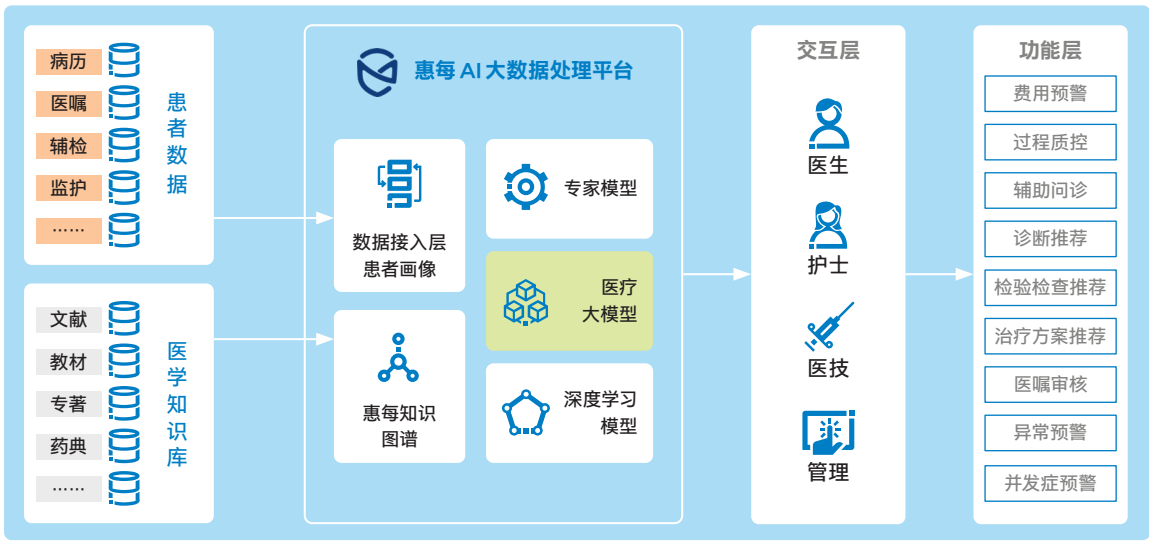

图 1 集成大模型的惠每新一代 AI 大数据处理平台架构

这些因素都阻碍着医疗大模型在医疗机构的落地与普及。

为应对这一挑战，惠每科技携手英特尔，采用 BigDL-LLM 大模型开源库和 OpenVINO™ 工具套件，在基于第四代英特尔® 至强® 可扩展处理器的基础设施上打造了大模型推理加速方案。

基于 BigDL-LLM，打造医疗大模型量化优化方案

数百上千亿计的参数规模在给医疗大模型带来更优学习效果、更精准辅助诊疗结果的同时，也对承载平台的资源，包括算力、内存等提出了更严苛的要求。这不仅会影响 AI 应用的最终运行效率，影响医护、患者、管理者以及科研人员的使用体验，更会限制更大参数规模、更优性能的大模型在医疗机构的普及。模型量化则是应对这一问题的重要优化手段之一。

对 AI 模型的量化，是指将训练好的模型的权值、激活值等从高精度数据格式（如 FP32 等）转化为低精度数据格式（如 INT4 /INT8 等），这不仅可以降低推理过程中对内存等资源的需求，从而让平台可以容纳更大参数规模的大模型，也能大幅提升推理速度，使医疗 AI 应用的运行更为迅捷。在惠每科技与英特尔的合作中，双方基于第四代英特尔® 至强® 可扩展处理器内置的指令集，借助由英特尔开发和开源的 BigDL-LLM 大模型库来实现推理加速量化方案。

BigDL-LLM 是一个为英特尔® 架构 XPU 打造的轻量级大语言模型加速库，在英特尔® 架构平台上具有广泛的模型支持，能实现更低的时延和更小的内存占用。作为英特尔开源 AI 框架 BigDL 的一部分，BigDL-LLM 不仅提供了对各种低精度数据格式的支持和优化，还可基于不同处理器内置指令集（如英特尔® AVX - 512_VNNI、英特尔® AMX 等）及相配套的软件实施推理加速，使大模型在英特尔® 架构平台上实现更高的推理效率。在本次合作中，惠每科技就使用英特尔® AVX - 512_VNNI 指令集显著加速了其医疗大模型在 INT4 低精度数据格式上的推理。

如图 2 所示，方案中 BigDL-LLM 为医疗大模型提供了两种使用方法：便捷命令（Command Line Interface, CLI）方法和编程接口（Application Programming Interface, API）方法。通过 CLI 方法，惠每科技可方便地完成模型量化并评估量化后的推理效果，由此判断该量化方案是否适用于当前这个模型。这些 CLI 命令包括使用 llm-convert 来对模型的量化精度快速转换用于预览，或者使用 llm-cli/llm-chat 来运行并快速测试量化后的模型。

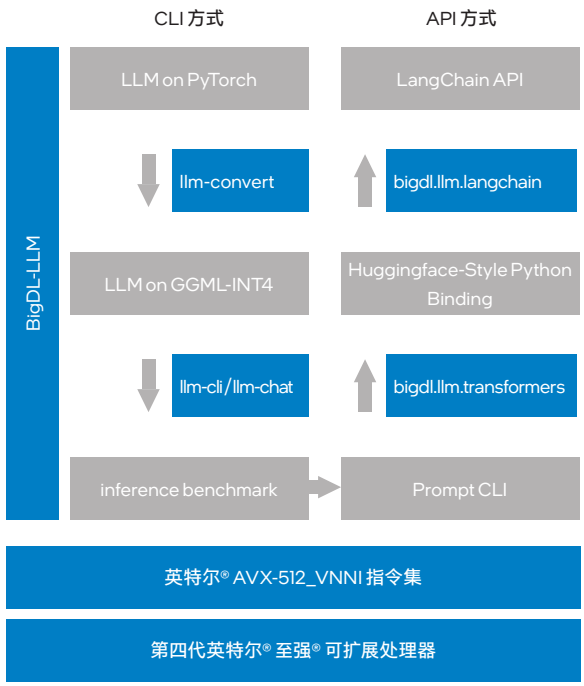

图 2 BigDL-LLM 为医疗大模型提供推理加速

另一方面，借助 BigDL-LLM 所提供的面向 HuggingFace 和 LangChain 的 API 编程接口，惠每科技能够快速地将 LLM 量化方案整合进 HuggingFace 或 LangChain 的项目代码，进而便捷地完成模型部署。作为热门的 Transformers 开源库之一，HuggingFace 上的 Transformers 模型一直是各类大模型的重要组成部分，而导入 BigDL-LLM 的优势在于能让惠每科技等用户只需修改少许代码（更改 import，并在 from_pretrained 参数中设置 load_in_4bit=True）即可快速加载模型。在使用 bigdl.llm.transformers 后，BigDL-LLM 会在模型

加载过程中对模型进行 INT4 的低精度量化，由此实现对基于 HuggingFace Transformers 的模型进行加速。

与此同时，LangChain 也是近年来大模型领域流行的开源框架之一，包括惠每科技在内的许多用户都在使用 LangChain 来开发不同的大模型应用。BigDL-LLM 同样也通过 API 编程接口 bigdl.llm.langchain 提供了便于使用的 LangChain 集成能力，让开发者能轻松借助 BigDL-LLM 来开发新模型或迁移基于 HuggingFace Transformers 优化的 INT4 模型，或是其它原生 INT4 模型。

基于 OpenVINO™ 工具套件，构建医疗大模型非量化优化方案

在量化优化方案之外，英特尔还借助 OpenVINO™ 工具套件为惠每科技打造了非量化优化方案。作为一款面向 AI 推理及部署优化的软件工具套件，OpenVINO™ 自推出以来，在帮助 AI 开发者和最终用户缩短开发、部署时间，以及充分释放丰富的英特尔® 硬件性能优势方面，始终发挥着重要作用。在最新的 OpenVINO™ 2023.11 版本中，其通过一系列新功能的加入，实现了面向大模型的功能增强。

在帮助惠每科技使用 OpenVINO™ 工具套件的 Pipeline 构建医疗大模型的高效推理服务部署之余，英特尔还借助该工具套件来助力优化模型推理流水线，通过消减模型输入和输出之间的内存副本来降低资源消耗，提升推理效率，并通过执行图的重新设计来优化模型中的组件。

以惠每科技使用的大模型 ChatGLM6b 为例，该模型的结构如图 3 所示，其流水线回路主要包含 3 个主要模块，即 Embedding、GLMBlock 层和 lm_logits。模型的流水线中有两类不同的执行图，首次推理时不需要 KV 缓存作为 GLMBlock 层的输入；从第二次迭代开始，QKV 注意力机制的上一次结果 (pastKV) 将成为当前一轮模型推理的输入。

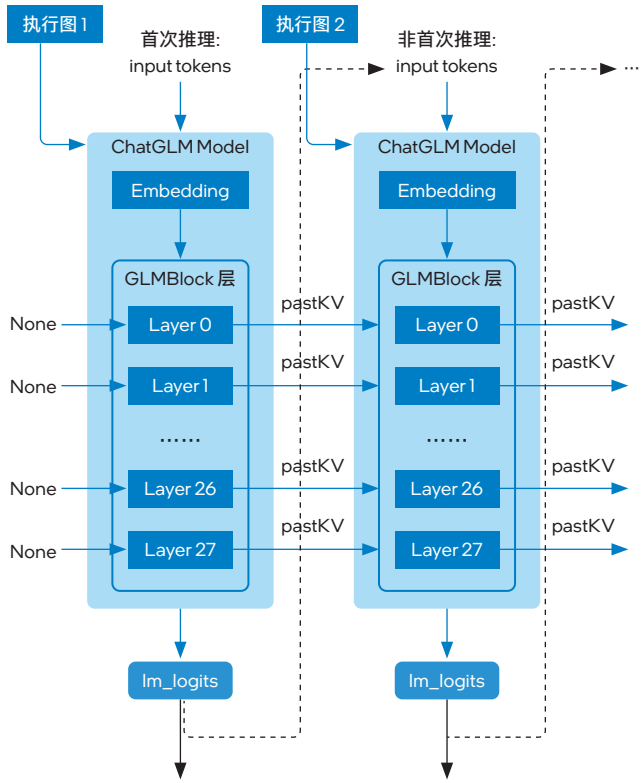

图 3 ChatGLM 的模型结构

可以看到，随着所生成 tokens 长度不断增加，在流水线推理过程中，模型输入和输出之间将存留海量的大型内存副本 (内存拷贝开销由模型的参数 hidden_size 以及迭代的次数决定)，不仅将占据大量的内存空间，庞大的内存拷贝开销也会使推理的执行效率遭遇挑战。

为应对上述挑战，基于 OpenVINO™ 工具套件的非量化优化方案执行了三个方面的优化。

优化一 利用零拷贝 (Zero-Copy) 视图来传递预分配的 KV 所需的内存副本空间。由于传统的内存拷贝需要耗费大量的处理器资源和内存带宽，因此当内存副本规模大幅增加时，会成为大模型推理效率的瓶颈。而零拷贝技术的引入，能避免数据的多次拷贝，有效实现 KV 缓存加速。

优化二 使用 OpenVINO™ opset 来重构 ChatGLM 的模型架构，从而帮助模型中的节点利用英特尔® AMX 指令集内联和

多头注意力 (Multi-Head Attention, MHA) 融合来实现推理优化。如图 4 所示，优化方案构建的 OpenVINO™ stateful 模型在 GLMBlock 层重新封装了一个类，并按图中工作流来调用 OpenVINO™ opset，然后再将图形数据序列化为中间表示 (Intermediate Representation, IR) 模型 (如 .xml、.bin)。

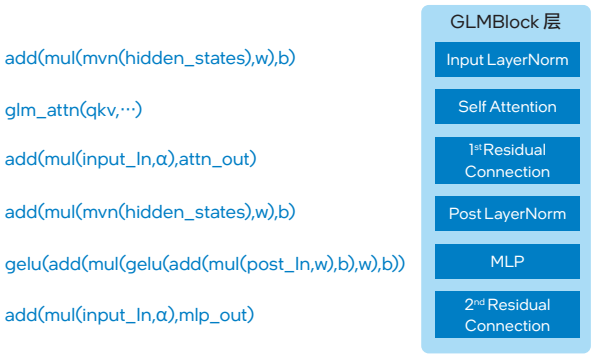

图 4 构建 OpenVINO™ stateful 模型

该优化方案一方面构建了全局的上下文结构体，用于在模型内部追加并保存每一轮迭代后的 pastKV 结果，减少相应的内存拷贝开销。另一方面，则通过采用内联优化 (Intrinsic Optimization) 的方式，实现了 Rotary Embedding 和 MHA 融合。

第四代英特尔® 至强® 可扩展处理器内置英特尔® AMX 指令集的引入，也能帮助 ChatGLM 等医疗大模型提升 BF16 或 INT8 精度数据格式下的模型推理速度。英特尔® AMX 指令集提供的内联指令能更快速地处理 BF16 或 INT8 精度数据格式的矩阵乘加运算，实现对 ChatGLM 模型中 Attention 和 Rotary Embedding 等算子的融合，从而在保证精度的同时提高运算效率、加速推理。

优化三 引入 OpenVINO™ 工具套件在 HuggingFace 上的 Optimum 接口。Optimum 是 Huggingface Transformers 库提供的一个扩展包，可用来提升模型在特定硬件基础设施上的训练和推理性能。基于 OpenVINO™ 工具套件提供的 Optimum 接口，惠每科技能在提高性能之余，将模型更便捷地扩展到更多医疗大模型推理应用中去。

这种优化方法在其它任务，包括 token-classification、question-answering、audio-classification 以及 image-classification 等中也同样适用。

效果评估

通过惠每科技与英特尔的协同优化，基于惠每科技医疗大模型构建的医疗 AI 应用无论是在应用效率还是在准确性等方面都获得了提升，并很快表现出了显著的临床应用优势与价值，包括:

- **提升医疗辅助诊疗准确性:** 通过对大量医疗数据的有效学习，医疗大模型能持续学习各种疾病特征，并借助优化方案更快、更精准地做出判断。结合惠每科技医疗知识库，为医护人员提供更加科学和准确的辅助诊疗方案和建议，优化诊疗决策;
- **提升医护与管理人员效率:** 借助基于医疗大模型构建的各类医疗 AI 应用，医护人员可以更高效地获取患者的辅助诊疗结果和病情 / 病历分析，从而能将更多时间和精力专注于患者的治疗和康复。同时医疗机构管理人员也能在诊疗风险预警、医保费用管理等环节上实现更为直观和高效的管控。

为评估优化后的医疗大模型的实用效果，惠每科技参加了由中国健康信息处理大会 (China Health Information Processing Conference, CHIP) 组织的中文临床医疗信息处理权威评测。这一评测全部使用中文真实医疗数据，覆盖诸如医疗术语识别和医疗知识问答等多个常见医疗 AI 应用场景，并采用量化的 F1 值进行排名。同时在大模型评测中，必须使用一个大模型同时完成 16 个任务的考验，非常具有挑战性。最终惠每科技从 396 支参赛队伍中脱颖而出，荣获“CHIP2023 -PromptCBLUE 医疗大模型评测”参数高效微调赛道第一名。[2]

目前，基于优化后的惠每科技医疗大模型所构建的医疗 AI 产品与应用，已在多个合作医疗机构中得到了部署与运行，并取得了不错的效果。首先，以基于大模型的鉴别诊断应用为例，这一辅助诊疗应用能体现临床医生的诊断思维链，而非简单

的记录。如图 5 所示，医生在应用中打开病程记录首页并填写患者主诉及病历特点后，后台的 3 个不同医疗大模型就会迅速执行推理，在数秒后即可生成鉴别诊断。医生可以点击查看不同大模型生成的结果，再根据自身的专业意见选择最优结果，选择【一键回填】或复制粘贴到病历相应的位置。

在此过程中，医生可对病历生成的结果进行【点赞】/【点踩】，也可在系统中反馈错误或问题，或返回病程记录页继续修改患者主诉或病历特点，之后再次通过医疗大模型进行计算和执行新的鉴别诊断推理。这些设计能有效收集医生反馈，实现大模型的增强学习。

其次，出院记录的自动生成是在合作医疗机构落地的另一项重要应用。传统上，诸如出院记录一类的流程，需要医院多个部门对多类数据进行总结并形成摘要，过程繁琐且容易出现差错。借助医疗大模型的技术优势，医生打开或保存【出院记录】时，会立即触发大模型后台计算。在数秒内得到结果后，医生即可查看包含出院诊断、入院情况、诊疗经过、出院情况和出院医嘱等内容的结果。医生可【一键回填】或复制粘贴到病历相应的位置，并对病历生成的结果进行【点赞】/【点踩】，也可点【识别错误】反馈相应问题。

上述基于医疗大模型的应用，都能与惠每科技 CDSS 系统实现无缝衔接，并可部署到既有的英特尔® 架构处理器平台。这让医疗机构无需购置专用的加速芯片或加速服务器，从而有效降低大模型部署的成本压力。来自惠每科技的数据统计表明，在某合作医院的某科室上线 1 个月后，鉴别诊断应用的使用率已达 23% 以上，出院记录自动生成应用的使用率达到 15% 以上，说明基于医疗大模型构建的 AI 应用能力已获得医生的初步认可。[3]

展望

随着 AI、大数据等新技术、新能力在医疗领域赢得越来越多的实用化落地，更多医疗机构也正通过这些前沿 IT 技术的引入来加速智慧医疗的进程，而惠每科技开展医疗大模型技术探索并基于此推出一系列医疗 AI 应用，正是这一进程的最新注脚。这些医疗 AI 应用将与惠每科技优势的 CDSS 产品一起，助力数以百计的医疗机构用户进一步提升医疗服务质量。

在此过程中，惠每科技与英特尔携手面向基于英特尔® 架构处理器（第四代英特尔® 至强® 可扩展处理器的优化）的平台展开了一系列大模型优化。无论是量化优化，还是非量化优化方案，都能在保证精度的前提下有效提升医疗大模型的推理速度，同时基于英特尔® 架构处理器的部署方案也能帮助医疗机构有效地节约成本。面向未来，惠每科技还将与英特尔一起，共同对大模型技术在医疗领域中更广泛和更深入的应用开展更多探索，例如利用大模型开展病历内涵质控等，进而推动医疗全流程的 AI 技术加持或智能化，让智慧医疗惠及更多医与患，从而普惠大众。

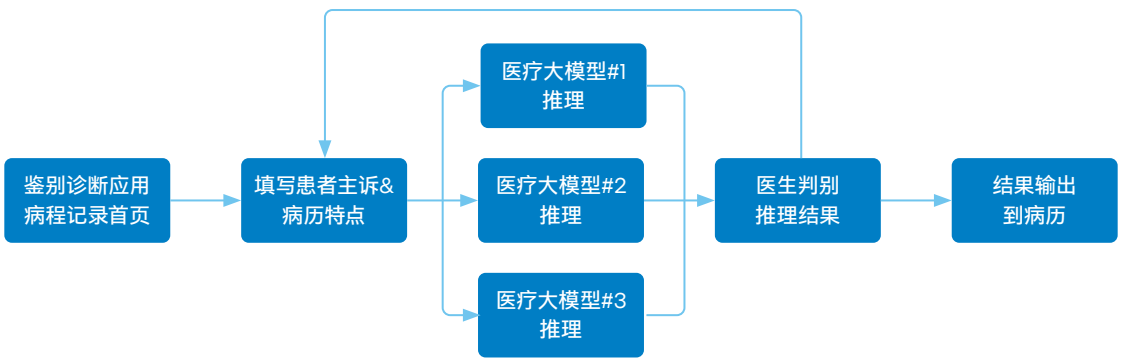

图 5 基于惠每科技医疗大模型构建的鉴别诊断应用工作流程

[1]数据援引自 Marketwatch 相关报告：https://www.marketwatch.com/press-release/large-language-model-llm-market-size-to-grow-usd-40-8-billion-by-2029-at-a-cagr-of-21-4-valuates-reports-7bbc5419

[2]数据援引自天池官网：https://tianchi.aliyun.com/competition/entrance/532132/rankingList。

[3]数字援引自惠每科技数字医学云讲坛第141期，详细信息请访问：https://www.e-chinc.com/#/ResourcesDetailVideo?id=1704682818727731202&packId=1614869950189219841