生成式人工智能(GenAI)在生物医药大健康行业应用进展报告

CMAC医学事务生成式AI联盟
理特咨询
天士力数智中药创新中心
南京柯基数据科技有限公司

2024年4月

前言

自2022年底起，OpenAI推出的ChatGPT在全球掀起了热潮。生成式人工智（GenAI）技术日新月异，各行各业正积极探索如何整合最新的GenAI技术以推动数字化转型。据统计，全球大型企业中，已有10%成功将GenAI技术应用于公司层面的平台级项目，50%正在进行小规模尝试，而另外40%仍在观望阶段。

生物医药大健康行业作为一个高度专业化和知识密集型的领域。从药物研发到临床试验，再到上市后的学术推广和患者教育等全流程应用场景，涉及到大量非结构化文本、图片和视频的处理。随着集采政策的实施和监管要求的提高，运营成本和复杂性不断上升，因此迫切需要借助人工智能来提升效率，重塑工作模式。自GenAI推出以来，国内外的药械、营养保健、医疗机构以及科研机构纷纷尝试将GenAI技术应用于不同场景，已经有一些公司和机构通过GenAI创造了全新的产品和服务，为业务增值。最近的一项调查显示，GenAI已成为大多数制药公司的首要关注点，40%的高管表示他们正计划将GenAI带来的成本节约重新投入到2024年的预算计划中。另外，60%的公司确立了使用GenAI来帮助企业降低成本或提高生产效率的目标，其中75%的公司将其视为高管层和董事会的优先事项。

2023年4月，CMAC牵头与跨国和国内生物制药企业、医药AI领先企业以及医学专家共同发布了《ChatGPT背景下的医疗健康行业数字化转型新范式研究报告》，引起了业界广泛关注。该报告结合行业实践和实际需求，从ChatGPT技术原理、技术发展、医疗健康行业国内外应用和研究进展，以及ChatGPT大模型在医药场景测试等角度，提出了ChatGPT大模型在医疗健康行业落地的挑战及可能的路径，为在ChatGPT背景下大模型如何赋能医疗健康行业数字化转型提供参考。

在过去的一年中，CMAC医学事务生成式AI联盟与数十家跨国和国内的药械企业、营养保健企业、医院、医疗科研机构等展开了深入合作。通过研讨会、咨询、概念验证（POC）、项目申报等形式，我们交流并见证了GenAI在国内生物医药大健康行业的快速发展和面临的挑战，积累了来自第一线的资料和GenAI应用落地的经验和方法论。

我们相信，2024年将是GenAI在中国生物医药大健康行业中实现规模化落地的关键一年。因此CMAC医学事务生成式AI联盟牵头，联合理特咨询、天士力数智中药创新中心、柯基数据以及生物医药大健康行业专家和GenAI技术专家，更新发布了本报告。报告着重介绍了最新GenAI技术发展和进展，生物医药大健康行业的应用场景和案例，落地挑战及方法论，以及未来展望。我们希望该报告能为GenAI在整个生物医药大健康行业的应用落地提供有益参考。

第一章：

GenAI技术进展概述

1.1 GenAI应用进展情况

当OpenAI在2022年11月30日发布ChatGPT的时候，没有人会意识到，新一代人工智能浪潮将在接下来短短数月给人类社会带来一场眩晕式的变革。自2010年代初深度学习问世以来，人工智能进入到第三次高潮。而2017年Transformer算法将深度学习推向了大模型时代。OpenAI基于Transformer的Decoder部分建立起来了GPT家族。ChatGPT一经面世便风靡全球，人们惊讶于其能够进行连贯、有深度对话的同时，也惊异地发现了它涌现了推理、思维链等体现智能的能力。

伴随AI预训练大模型持续发展，生成式人工智能（GenAI）算法不断创新以及多模态AI日益主流化，以ChatGPT为代表的GenAI技术加速成为AI领域的最新发展方向，推动AI迎来下一个大发展、大繁荣的时代，将对经济社会发展产生重大的影响。

1.1.1GenAI技术定义及背景

GenAI(Generative AI，生成式人工智能)指的是通过人工智能技术自动生成内容的生产方式。通过训练模型来生成新的、与训练数据相似的内容。与传统类型的AI主要关注识别和预测现有数据的模式不同，GenAI着重于创造新的、有创意的数据，其关键原理在于学习和理解数据的分布，进而生成具有相似特征的新数据，在文本、图像、音频、视频等多种领域都有广泛的应用。GenAI目前最引人注目的应用当属ChatGPT。ChatGPT是基于OpenAI公司的大语言模型GPT-3.5训练、调试、优化的聊天机器人应用，同一个AI模型可以处理各种各样的文字和推理任务。

ChatGPT发布仅两个月即获得1亿月活用户，超越了历史上所有互联网消费者应用软件的用户增长速度。以大语言模型、图像生成模型为代表的GenAI技术，成为新一代人工智能的平台型技术，助力不同行业实现价值跃升。GenAI大爆发的背后，普遍认为三个领域的AI技术的发展为其提供了肥沃的土壤，分别是生成算法、预训练模型和多模态技术。

第一，随着各种生成算法的不断创新突破，AI现在已经可以生成文字、代码、图像、语音、视频物体等各种类型的内容和数据。GenAI与过去最显著的区别是从分析式 AI（Analytical AI）发展为生成式AI（Generative AI）。分析式AI模型是根据已有数据进行分析、判断、预测，最典型的应用之一是内容智能推荐；生成式AI模型则是学习已有数据后进行演绎、生成创造全新内容。

第二，预训练模型，特别是以ChatGPT为代表的大模型，引发了GenAI技术能力的质变。在过去，研究人员需要针对每一个类型的任务单独训练AI模型，训练好的模型只能从事特定任务，不具有通用性。而预训练的大模型技术显著提升了GenAI模型的通用化能力

和工业化水平，让GenAI模型成为自动化内容生产的“工厂”和“流水线”。 GenAI模型，包括ChatGPT、GPT-4等大语言模型（Large Language Models，LLM）和Midjourney、Stable Diffusion等图像生成模型，又被称为基础模型（Foundation Models），其作为基于种类丰富的海量数据预训练的深度学习算法，展现出强大的、更加泛化的语言理解和内容生成能力。以大语言模型（LLM）为例，经过海量的互联网内容数据的训练，大语言模型的参数可以达到万亿甚至百万亿级别。这大大增强了语言模型的生成能力，同一个大语言模型可以高质量地完成各种各样的文字和推理任务，例如作诗、写文章、讲故事、写代码、提供专业知识等等。因此，大语言模型已经成为了各大企业竞相追逐的AI方向。

第三，多模态AI技术的发展。多模态技术让GenAI模型可以跨模态地去生成各种类型的内容，比如把文字转化为图片、视频（Sora）等等，进一步增强了GenAI模型的通用能力。

1.1.2 GenAI应用领域与案例

(1) 多模态内容生成

A. 文本生成领域

自然语言生成是一种GenAI技术，可以生成逼真的自然语言文本。生成式AI可以编写文章、故事、诗歌等，为作家和内容创作者提供新的创作方式。同时，它还可以用于智能对话系统，提高用户与AI的交流体验。ChatGPT(全名:Chat Generative Pre-trained Transformer对话生成式预训练变换模型)是由OpenAI开发的一个人工 智能聊天机器人程序，于2022年11月推出。该程序使用基于GPT-3.5架构的大语言模型并通过强化学习进行训练。

ChatGPT 目前仍以文字方式互动，可以解决包括自动文本生成、自动问答、自动摘要等在内的多种任务。Jasper已经开始为谷歌、脸书等知名公司提供文案GenAI的商业服务。

B. 图像生成领域

图像生成是GenAI技术中最为普遍的应用之一。Stability AI发布了稳定扩散（Stable Diffusion）模型，通过开源快速迭代大幅降低了AI绘画的技术使用门槛，消费者可以通过订阅旗下产品DreamStudio来输入文本提示词生成绘画作品，产品已经吸引全球 50多个国家超过100万的用户注册。

C. 音视频创作与生成

2024年2月16日， OpenAI继一年前发布ChatGPT语言大模型之后，又发布了一款基于人工智能技术的视频生成工具Sora，再次引发轰动。这是一款输入文本即可自动生成高质量视频的文生视频大模型，实现了视频生成领域革命性变革，提供了全新的视觉体验。在部分样片中，Sora还展现了对“物理规律”超强的学习能力，如能够模拟现实环境中的重力、碰撞等物理现象，可以通过直播视频功能实时传递信息，用于直播秀、在线教育、远程医疗等场合。在“现实已经不存在”的惊呼声 中，Sora确实打开了人类视频创作的新天空，它将重塑视觉内容生成的未来，同时也反映出人工智能技术远超预期的快速进步。有媒体称，Sora 不仅仅是一个工具，更是一种新的生活方式，将会对整个社会产生重要影响。

GenAI技术还可以用于语音合成，即生成逼真的语音。例如，通过学习人类的语音特征，生成式模型可以生成逼真的语音，从而用于虚拟助手、语音翻译等应用。GenAI技术可以用于生成音乐。生成式AI可以根据给定的风格和旋律创作新的音乐作品，为音乐家提供新的创作灵感。这种技术还可以帮助音乐家更有效地探索音乐风格和元素的组合。这些曲目可以用于音乐创作、广告音乐等应用。

D. 电影与游戏

GenAI可以用于生成虚拟角色、场景和动画，为电影和游戏制作带来更多的创意可能。此外，AI还可以根据用户的喜好和行为生成个性化的故事情节和游戏体验。2023年3月，腾讯AI Lab在GDC上提出了3D虚拟场景自动生成解决方案，能够帮助游戏开发者以更低成本创造风格多样、贴近现实的虚拟城市，提升3D虚拟场景的生产效率。其中重点分享了城市布局生成、建筑外观生成和室内映射生成三大能力。整个路网生成和微调过程仅需要不到30分钟，相比手动设计效率提升近100倍；而单个独特建筑的制作时间也降低至17.5分钟，大大提升了场景制作的效率。

E. 代码生成领域

经过自然语言和数十亿行代码的训练。部分GenAI模型精通十几种语言，包括Python、JavaScript、Go、Perl、PHP、Ruby等等。能够根据自然语言的指令生成相应的代码。

GitHub Copilot是一个GitHub和OpenAI合作产生的AI代码生成工具，可根据命名或者正在编辑的代码上下文为开发者提供代码建议。官方介绍其已经接受了来自GitHub 上公开可用存储库的数十亿行代码的训练，支持大多数编程语言。

(2) 翻译

GenAI可以直接应用于翻译实践之中，与传统机器翻译系统采用以句子为单位的方式训练不同，大语言模型采用以单词为单位的方式进行训练。这使得大语言模型可以理解并再现单词之间的连贯性和上下文信息，译文因而更加自然、准确。此外，传统机器翻译系统在遇到较为复杂的语言环境时，往往会出现句法和语义方面的错误，而大语言模型可以应付更为复杂的语言环境，产出更为准确、自然的译文。相比较而言，大语言模型在翻译方面展现的性能要比传统机器翻译更加突出，能够产出可与人工翻译译文相媲美的翻译作品。

(3) 内容理解与分析

腾讯会议AI小助手：只需通过简单自然的会议指令，基于对会议内容的理解，就可以完成信息提取、内容分析、会管会控等多种复杂任务。会后可以自动生成智能总结摘要，还能基于智能录制的能力，帮助用户高效回顾，提升用户开会和信息流转效率。

(4) 科研与创新(AI for Science)

GenAI可以在化学、生物学、物理学等领域探索新的理论和实验方法，帮助科学家发现新的知识。此外，GenAI还可以用于药物设计、材料科学等领域，加速技术创新和发展。

1.2 GenAI 应用关键技术

在GenAI领域中，有四种关键技术：模型训练(Model Training)、微调(Fine Tuning)、检索增强生成(RAG)和提示词工程(Prompt Engineering)。针对不同的业务目标和场景，选择适当的技术模型方法至关重要。

1. **模型训练(Model Training)：**需要大量的数据和计算资源来从头构建一个人工智能模型。它具有高度的可定制性和可扩展性，但耗时较长，成本最高。适用于全新的突破性应用，例如训练一套中医诊疗大模型。

2. **微调(Fine-Tuning)：**专注于将现有模型适应特定任务，提供了定制性和效率之间的平衡。

3. **检索增强生成(Retrieval-Augmented Generation -RAG)：**通过整合外部知识库来增强模型，非常适合需要当前或广泛信息的任务，是现阶段企业级知识库以及Chatbot建设较高性价比的主要方法。

4. **提示工程(Prompt Engineering)：**依赖于设计有效的提示来引导预训练模型，需要在提示设计方面的技能，但计算资源需求较低。这种方法不仅具有成本效益，而且非常有效，然而其潜力经常被低估。

每种方法在不同应用中都有其优势和限制，取决于数据可及性、计算资源、特定的任务、对最新信息的需求以及所需技能和成本等因素。

1.2.1模型训练

模型训练类似于AI系统开发的基础阶段(例如重新开发一个ChatGPT)。它涉及从零开始构建AI模型的过程，类似于将种子培育成长成一棵大树。这个过程非常重要，因为它奠定了AI的基本能力和智能。

主要适用的场景包括：

1. **新领域：**当涉足现有模型不适用或不足的领域时。例如，开发一种尚未被探索的新型医学诊断AI
2. **基于独特数据集应用：**在数据对特定需求具有独特性的情况下，例如公司使用客户数据来预测购买模式。
3. **创新和研究：**非常适合研究和开发，用于测试新理论或模型。模型训练是人工智能发展的基石，提供了无与伦比的定制化和创新潜力。然而，它需要大量数据和GPU计算资源和开发资源，成本很高，并带有固有的风险，因此更适用于需要定制解决方案或在人工智能应用领域开辟新天地的情况。

1.2.2 微调

微调类似于磨练技艺娴熟的艺术家，使其在特定类型中表现出色。它涉及对经过预训练的模型进行调整，即对已经从大规模数据集中学到一般模式的模型进行专门任务或数据集方面的进一步提高。这一过程对于将通用人工智能模型适应特殊需求至关重要。例如基于医学文献训练微调成一套更适合回答健康护理相关的问题。

微调主要的适用场景包括:

1. **特定任务应用：**适用于需要模型的一般理解与特定需求相匹配的任务，例如使语言模型适应医学术语。
2. **有限资源：**适用于无法负担完整模型训练所需的大量资源的情况。
3. **提升模型性能：**当您需要提高预训练模型在特定领域准确性时。

在 GenAI 中，微调是将通用模型转变为专业模型的艺术。它在效率和性能增强之间取得平衡，非常适合有针对性改进的场景。这种方法最适用于基础扎实但需要特定专业知识的情况。

1.2.3 RAG

（1）RAG介绍

RAG，即检索增强生成（Retrieval-Augmented Generation），是一种结合了信息检索（Retrieval）和文本生成（Generation）的人工智能技术。RAG是GenAI领域的重大进展，它通过整合外部知识源来增强传统的大语言模型 (LLM)。这种方法拓宽了人工智能的视野，使其能够访问和利用除初始训练数据之外的大量信息。可以将 RAG 想象为一位学者，除了拥有自己的知识外，还可以即时访问到一座全面的图书馆。

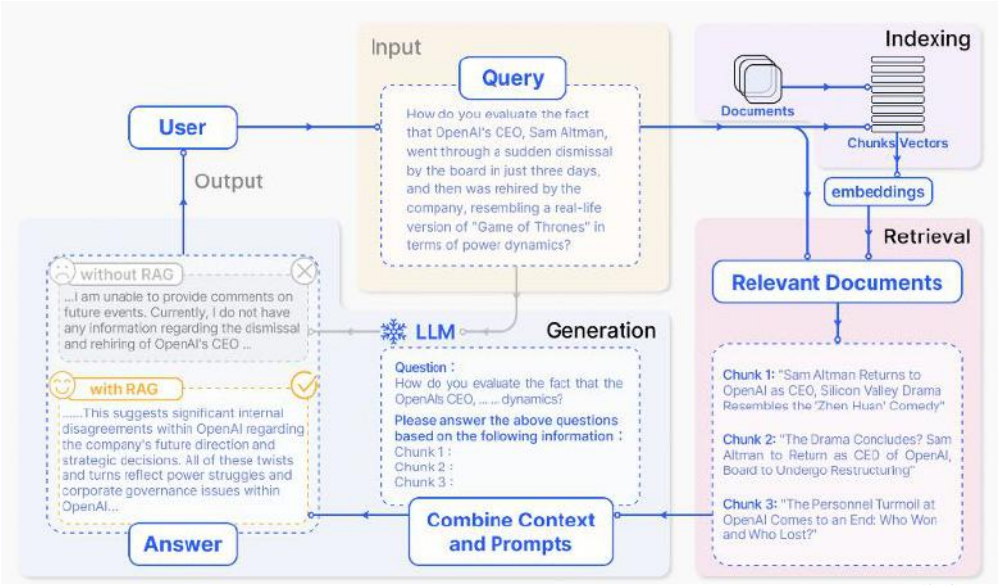


图1. RAG工作流程

上图展示了RAG架构的工作流程，整体分为五步：

1. 用户向Chatbot（LLM应用）提出问题
2. 根据问题在向量数据库(提前将知识库的文档向量化)检索匹配相关的上下文段落信息
3. 将检索结果的top_k条段落进行排序,，将提示词和组装的段落以及用户问题三者形成最终的提示词prompt
4. 将prompt提交给大模型
5. 大模型生成输出并返回给Chatbot，进而返回给用户

RAG的优势：

1. **提高答案准确性：**通过引用外部知识库中的信息，RAG可以提供更准确的回答
2. **增加用户信任：**用户可以通过引用的来源来验证答案的准确性
3. **便于知识更新和引入特定领域知识：**RAG通过结合LLM的参数化知识和外部知识库的非参数化知识，有效地解决了知识更新的问题
4. **减少幻觉问题：**RAG能够减少语言模型中的幻觉问题，使生成的响应更准确、可靠

RAG的应用场景：

1. **问答系统**：在问答系统中，RAG通过检索大量信息并生成精准、详细的答案，提高了回答的准确性和信息的丰富度
2. **内容创作**：RAG可以根据给定的主题或关键词生成丰富且有深度的文章，节省大量的时间和人力资源
3. **数据分析与挖掘**：RAG能够在大规模数据集中快速检索信息，为数据分析提供了一个强大的工具

RAG技术通过结合最新的大语言模型和外部知识库，为AI在自然语言处理领域的应用提供了新的可能性，尤其是在需要处理大量信息和提供准确回答的场景中

在RAG的技术发展中，从技术角度，呈现出以下几种范式：

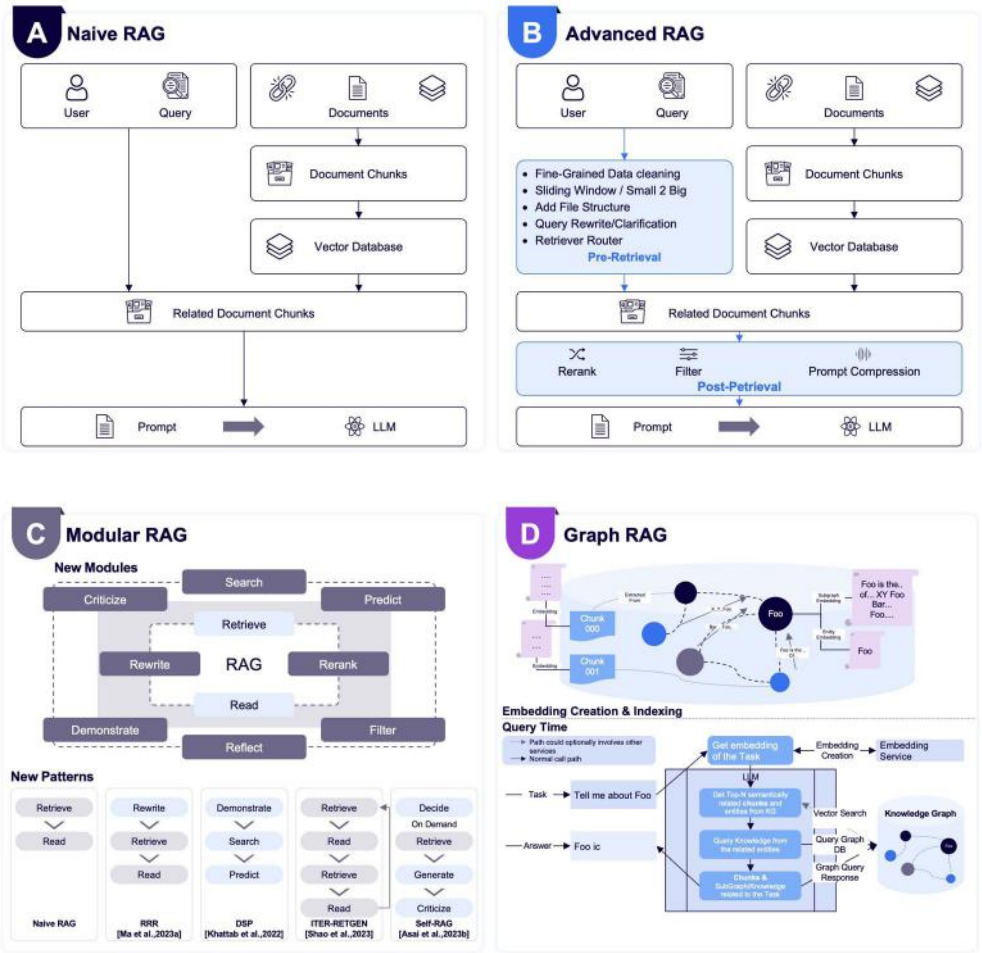


图2. RAG技术发展范式

其中Graph RAG（Graph Retrieval-Augmented Generation）是一种结合知识图谱和图数据库的检索增强技术。它通过构建图模型的知识表达，将实体和关系之间的联系用图的形式进行展示，并利用大语言模型（Large Language Model, LLM）进行检索增强。Graph RAG的核心在于将知识图谱等价于一个超大规模的词汇表，其中实体和关系对应于单词，使得在检索时能够将实体和关系作为单元进行联合建模。

Graph RAG的处理流程划分为两个主要阶段：

1. **第一阶段，**Index in time。该阶段中系统不仅将知识以图谱的形式进行存储，以便于后续的检索和引用，同时还会执行传统 RAG 流程中的 Split & Embedding操作。
2. **第二阶段，**Query Time。Split & Embedding 的操作带来的最大益处在于能够迅速锁定与查询最为相关的知识点。此外，通过利用知识图谱（KG）中知识点之间的关联关系或语义链接，系统可以快速地识别出在语义层面上相关或接近的知识。这些知识点随后被提供给大语言模型，从而使其能够生成更为贴切的答案。同时，这一过程也有助于防止语言模型产生虚假或不合逻辑的回答，提高了结果的可靠性。

Graph RAG的主要特点：

1. **知识图谱集成：**Graph RAG利用知识图谱来增强语言模型的理解能力，使得模型能够更好地理解实体间的关系和上下文信息。
2. **检索增强：**通过结合图数据库的查询能力，Graph RAG能够提供更准确、相关和多样化的信息来满足用户的需求。
3. **上下文学习：**Graph RAG支持In-Context Learning，即在向模型提出问题时，提供相关的上下文信息作为背景知识，从而生成更符合预期的响应。
4. **处理复杂查询：**Graph RAG特别适合处理复杂或多义词查询，因为它能够利用知识图谱中的结构化信息来解决歧义问题。
5. **表达和推理能力提升：**通过图技术构建的知识图谱，Graph RAG能够帮助大语言模型更好地理解实体间的关系，提升模型的表达和推理能力。
6. **适应性强：**Graph RAG技术可以适配不同的大语言模型框架，如Llama Index、LangChain等，使得开发者可以专注于LLM的编排逻辑和pipeline设计。

Graph RAG作为一种新兴的技术，正在逐渐展现出其在信息检索和处理领域的潜力，尤其是在需要处理大量结构化数据和复杂上下文信息的场景中。随着技术的进一步发展，Graph RAG有望在更多领域得到应用和推广。

1.2.4提示词工程

提示词工程（PromptEngineering，缩写为PE）是一种AI技术，它通过设计和改进AI的提示词来提高AI的表现。PE关注提示词的开发和优化，帮助用户将大模型用于各场景和研究领域。

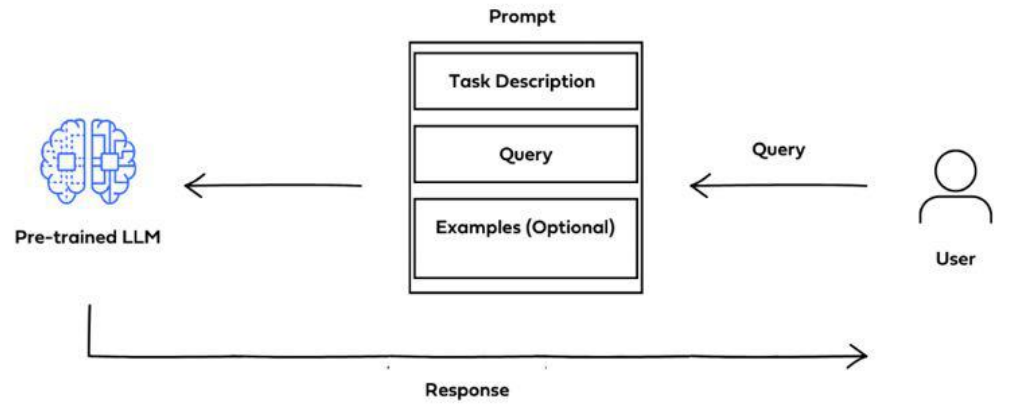


图3. 提示词工程原理

提示词（prompt）在人工智能场景下指给模型的一个初始输入或提示，用于引导模型生成特定的输出。提示词可以是一个问题、一段文字描述，甚至可以是带有一堆参数的文字描述。AI模型会基于提示词所提供的信息，生成对应的文本，亦或者图片。比如，我们在ChatGPT中输入：中国的首都是什么？ 这个问题就是提示词。掌握了提示词工程相关技能将有助于用户更好地了解大模型的能力和局限性。

主要优点

- **效率：**不需要额外的培训或计算资源，使其高效运作。
- **灵活性：**可以适应各种任务而无需改变基础模型。
- **创造力：**允许对模型的输出进行高度创造性的控制。

主要挑战

- **依赖技能：**提示工程的有效性在很大程度上取决于用户构建有效提示的能力。
- **试错：**通常涉及实验过程，可能耗时。

(1)提示词技术包含要素：

A. **指令，**想要模型执行的特定任务或指令。
B. **上下文，**包含外部信息或额外的上下文信息，引导语言模型更好地响应。
C. **输入数据，**用户输入的内容或问题。
D. **输出指示，**指定输出的类型或格式。

（2）提示词技术

1.零样本提示(Zero-Shot Prompt)

零样本提示是一种先进的自然语言处理技术，旨在让模型在没有先前见过的任务或领域中表现出色。通过零样本提示，模型能够根据用户提供的提示进行推理和生成，即使这些提示与训练数据中的内容没有直接关联。这一技术的核心思想在于通过广泛而有代表性的训练，使模型能够推广到新的输入领域，进而在没有样本支持的情况下作出准确的预测或生成。

2.少样本提示(Few-Shot Prompt)

虽然大语言模型展示了惊人的零样本能力，但在使用零样本设置时，它们在更复杂的任务上仍然表现不佳。少样本提示可以作为一种技术，以启用上下文学习。相对于零样本提示，少样本提示更专注于在有限的先验知识下进行任务推理和生成。通过少样本提示，模型可以在只有极少量相关样本的情况下，利用先前学到的知识来更好地理解和处理新的任务或领域。

3.思维链(Chain-of-Thought, CoT)

思维链提示是一种推理和生成的方法，通过将多个提示按照逻辑顺序连接在一起，引导模型实现更复杂的任务。这种方法通过逐步提供信息，促使模型在每个步骤上下文中进行思考，逐渐构建起全局的理解。链式思考提示可以用于解决需要多步骤推理的问题，例如复杂的问题回答或创造性的文本生成。这种技术提供了更深入、更结构化的信息引导，从而增强了模型的表现能力。

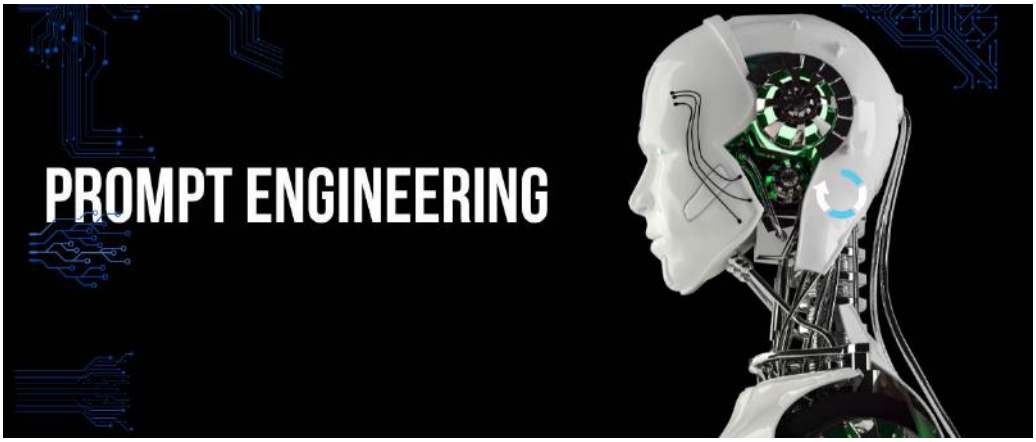


图4. 提示词工程

GenAI应用技术的选择方法可以类比于在道路建设中选择最佳路线：

1. **模型训练：**这相当于修建一条新的道路。它是一个需要大量资源、时间和数据投入的过程。虽然为创建高度定制和强大的人工智能模型铺平了道路，但这是一个庞大的任务，不总是必要或可行的。

2. **微调：**这种方法类似于修改现有的道路。在这里，您从一个预先存在的模型（道路）开始，并进行特定的调整，以更好地适应您的需求。它比修建新道路所需的资源要少，并且可以非常有效，但仍受限于原始模型的局限性。

3. **检索增强生成（RAG）：**将RAG与这个类比相结合，就好像给道路配备了动态标志，可以从各个位置获取信息。RAG结合了预训练模型的优点和获取和整合外部最新信息的能力。与模型训练和微调相比，它更具灵活性，可以适应新的信息。但是，其效率取决于外部数据源的整合和处理，这可能需要大量资源。是现阶段企业级GenAI知识库建设和Chatbot应用的性价比较高的主流方法。

4. **提示工程：**这种方法就像找到一个聪明的捷径。它涉及使用智能、有策略的提示来引导预训练的人工智能模型产生期望的结果。这种方法快速、灵活且资源高效，可以利用先进的人工智能模型的能力，而无需大量数据、计算能力或时间。这是一种创新的方式来应用人工智能的能力，往往能够以最小的投入取得令人印象深刻的成果。

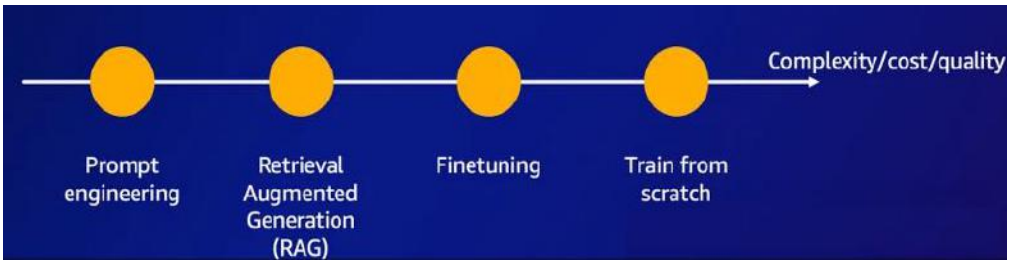


图5. 从复杂度和成本以及质量等多维度综合考虑的实际应用路径

1.2.5 LangChain

在人工智能领域，大语言模型（LLMs）如GPT-3.5和GPT-4等已经成为了自然语言处理（NLP）的强大工具。它们能够生成连贯的文本、回答问题、甚至创作诗歌和故事。然而，尽管这些模型在处理语言方面表现出色，但它们在实际应用中的潜力仍然受限。为了克服这些限制并充分发挥LLMs的能力，LangChain应运而生。

LangChain是哈里森-蔡斯（Harrison Chase）于2022年10月发起的一个基于LLM的应用开发框架开源项目，是目前大模型应用开发的最主流框架之一。它提供了一套工具和组件，使得开发者能够将LLMs与外部数据源和计算能力结合起来，从而创建更加智能和功能丰富的应用。LangChain围绕将不同组件“链接”在一起的核心概念构建，通过统一的接口简化了与 GPT-3.5、GPT-4、llama、文心一言、通义千问等 LLM 合作的过程，使得开发者可以轻松创建定制的高级用例。

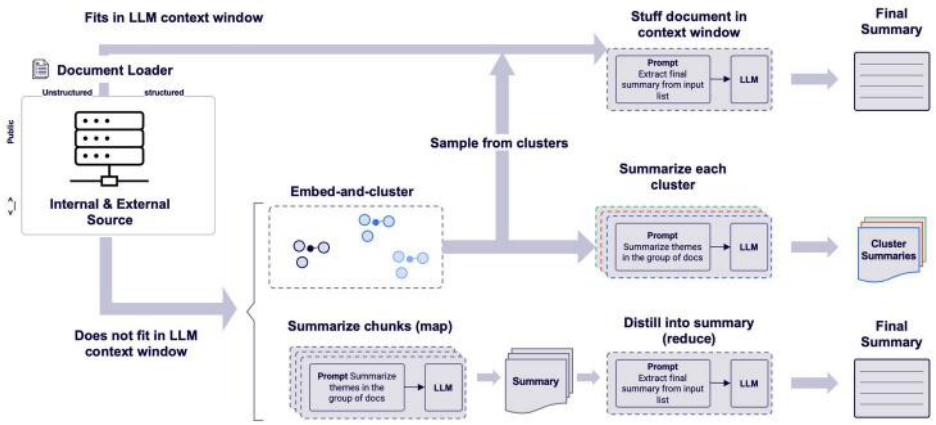


图6. LangChain工作机制

截至2024年3月，LangChain平台已实现了与包括亚马逊、谷歌及微软Azure在内 的主流云存储系统的深度整合，并且封装了涵盖新闻资讯、影视资料和气象信息的多样化API接口。此外，LangChain提供了对Google Drive文档、电子表格及演示 文稿内容的自动化总结、抽取与创建能力；同时涵盖了Google搜索与Microsoft Bing搜索引擎的网络信息检索功能。在自然语言处理领域，它成功对接了OpenAI、 Anthropic 和Hugging Face等多家知名机构的语言模型资源。

在编程与 代码管理方面，LangChain支持Python与JavaScript代码的自动生成、静态分析与调试功能，并采用Milvus与Weaviate向量数据库系统分别用于存储与检索高维向量嵌入及缓存相关对象。为加速数据访问性能，系统配备了Redis作为缓存数据存储方案，并通过Python Requests Wrapper及其他API请求手段确保了与各类服务的 无缝交互。在并发处理层面，该平台能够实时追踪并记录线程与异步子进程运行中的堆栈符号信息。截至2024年3月，LangChain已具备读取超过50种不同文档类型和数据源的强大能力，展现出广泛的应用潜力和卓越的技术适应性。

1.2.6 AI Agent

AI Agent，即人工智能代理，是一种具备环境感知、决策制定和行动执行能力的智能体，也被称为“智能业务助理”。其旨在利用大模型技术，通过自然语言交互方式高度自动化地处理专业或复杂工作任务，从而显著减轻人力负担。在本质上，AI Agent是建立在大语言模型之上的智能应用，即在大模型的基础上运行的应用程序。AI Agent不仅限于对话交流，还能整合外部工具，直接完成各种任务。

一个基于大模型的AI Agent系统可分为四个组件部分：大模型、规划、记忆和工具使用，对应需要四个能力：包含大语言模型能力、具体拆解问题的能力、具有长短期记忆控制的能力、以及具有调用外部工具的能力。AI Agent有望开启新时代，其基础架构可简单划分为Agent = LLM + 规划技能 + 记忆 + 工具使用。在这一结构中，LLM充当Agent的“大脑”，为系统提供推理、规划等关键能力。本文着重介绍了基于LLM的Agent的整体概念框架，包括大脑、感知和行动三个关键部分。

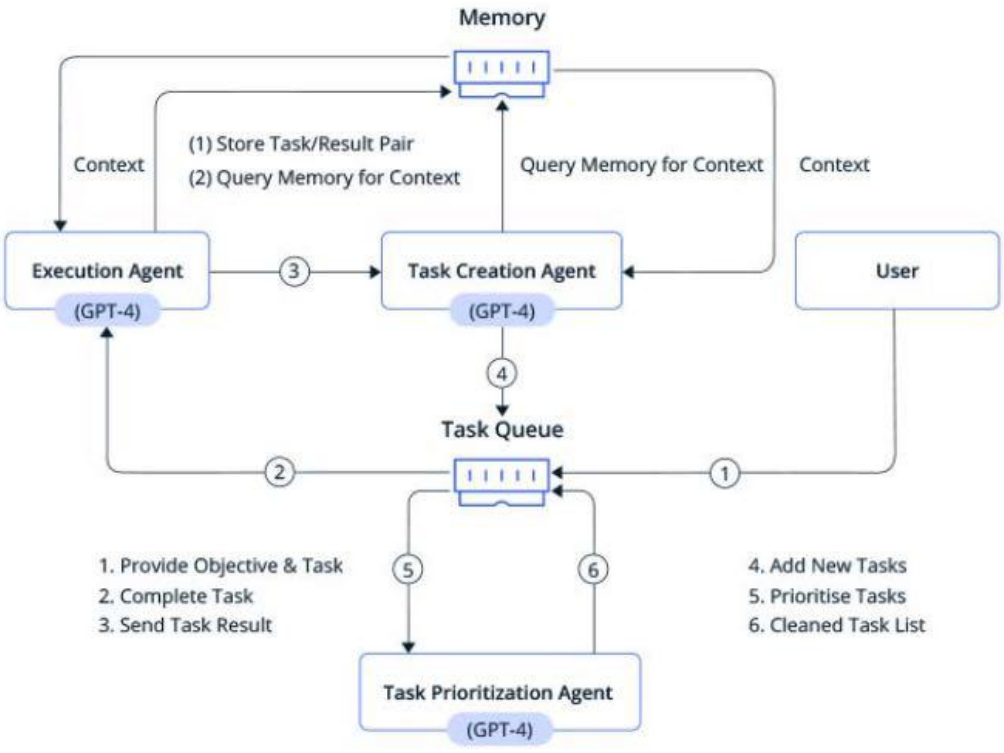


图7. Agent智能体AutoGPT原理介绍

1.3 GenAI大模型发展现状

1.3.1 国外大模型

（1）ChatGPT

ChatGPT（Chat Generative Pre-trained Transformer）是由OpenAI研发的一款聊天机器人程序，于2022年11月30日发布。它是基于GPT-3.5模型的一个专门优化用于对话生成的语言模型。ChatGPT能够根据用户输入的文本产生智能化的回答，并且具备连续对话的能力，能够捕捉用户的意图，理解上下文，并在多轮对话中提高准确率。

截至2023年12月，ChatGPT已无可争议地成为全球范围内增长速度空前的消费级软件应用典范，其用户基数在以突破1.8亿大关，并在此背景下，促使OpenAI公司的估值跃升至800亿美元的新高度。ChatGPT这一划时代产品的发布不仅引发了全球科技界的广泛关注，还强有力地激发了市场对同类竞品的研发热潮，诸如Gemini、ErnieBot、 LLaMA 以及Claude等项目应运而生。值得注意的是，ChatGPT在线服务提供了两个迭代版本，分别基于GPT-3.5和更为先进的GPT-4架构构建而成。这两个版本均隶属于OpenAI专有的生成预训练转换器（Generative Pre-trained Transformer，GPT）模型系列，该系列的设计灵感与核心技术基础源自谷歌所研发的Transformer架构。为了满足不同用户的需求， ChatGPT 允许普通用户免费体验基于GPT-3.5版本的服务；而对于追求更高级功能和持续更新内容的用户，则通过商业化品牌“Chat GPT Plus”提供基于GPT-4版本及其后续优化功能的付费订阅服务。

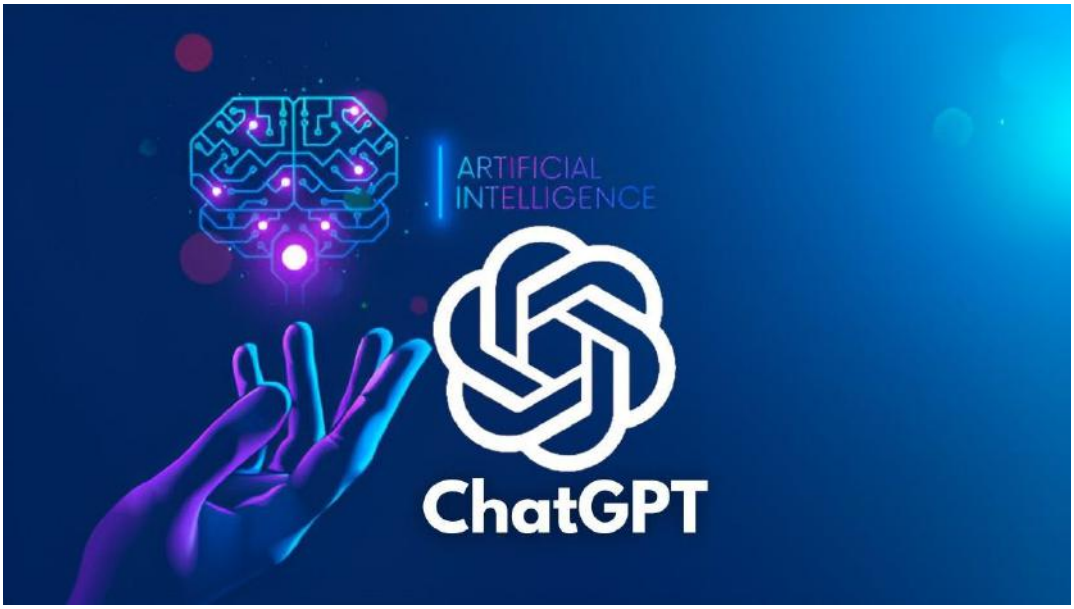


图8. ChatGPT

（2）Gemini

Gemini 是由Google DeepMind团队研发的人工智能模型，是LaMDA和PaLM2的继任者，于2023年12月6日发布。这是一个多模态模型家族，支持文本、图像、音频、视频和代码的全方位理解和生成。Gemini模型家族包含三个针对不同应用场景优化的版本：Gemini Ultra作为旗舰版，专为数据中心级别的高性能计算环境设计；Gemini Pro则定位为通用型解决方案，能在多种工作任务上展现高效性能，并且特别适用于企业级应用及云端服务；而小巧灵活的Gemini Nano，则通过高度优化实现了在资源受限的移动设备上的高效运行，为智能手机和平板电脑等移动平台带来了强大且实时的AI交互体验。Gemini的设计使其能够原生地支持多模态能力，即从一开始就在不同模态上进行预训练，并利用额外的多模态数据进行微调，以提高其有效性。

（3）Claude

Claude是由美国人工智能初创公司Anthropic开发的大语言模型，Anthropic成立于2021年，是一家专注于人工智能安全和研究的公司，旨在建立可靠、可解释、可操纵的人工智能系统。Claude模型提供了API和Slack机器人两种访问方式，其中API访问需要申请并通过后才能使用，而Slack机器人目前处于beta版本，不收费。

Claude是由Anthropic公司于2023年3月首次推出的语言模型系列的初代版本，尽管在编码、数学推理等方面尚存在一定的性能局限性，但依然在执行多样化任务上展现出了显著的能力。针对用户的不同需求，Claude推出了两种优化版本：常规版Claude和响应速度更快、成本更低廉、运行更为轻巧的Claude Instant。后者将输入上下文的处理能力提升至容纳10万token，约等同于7.5万个英文单词的内容量。同年7月11日，Anthropic发布了Claude系列的重要迭代产品——Claude2，该版本对公众开放使用，而其前身Claude1则仅限于经Anthropic审核批准的特定用户群体。Claude2实现了一次重大飞跃，将上下文窗口从原来的9,000个token扩展到了10万个token，并新增了上传PDF和其他文档的功能，使得Claude能够阅读、总结文件内容并辅助完成相关任务。11月份发布的Claude2.1版本中，聊天机器人的处理能力再度翻倍，token扩大至20万个，相当于500页的书面材料。 Anthropic在2024年3月4日正式推出了革新性的Claude3系列，Claude3分为三个型号，按功能强大程度依次排列为Haiku、Sonnet和旗舰款Opus。默认配置的Claude3 Opus token为20万个，而在特殊应用场景下，该窗口可扩展至高达100万个token,且在知识深度、数学处理和复杂任务解决方面展现出了超越GPT-4和Gemini 1.0 Ultra的实力。

（4）LLaMA

LLaMA（全称为“大语言模型Meta AI”）是Meta AI于2023年2月推出的自回归式大语言模型系列，它囊括了多种参数规模的版本，其参数量分别为7亿、130亿、330亿以及650亿。通常情况下，顶级LLM仅能通过有限或专属API途径访问，而Meta则破例在非商业许可框架下向全球研究社群开放了LLaMA模型权重的下载权限。值得关注的是，在LLaMA发布后短短一周内，其模型权重即通过BitTorrent在网络论坛4chan上被公开泄露给了公众。

2023年7月18日，Meta与微软携手推出了LLaMA系列的迭代升级产品——LLaMA-2，标志着双方在大语言模型技术领域的合作迈入了新的阶段。当日，Meta正式揭晓了三种不同规模配置的LLaMA-2模型，参数量分别达到了70亿、130亿以及700亿。尽管在架构设计上，LLaMA-2延续了前代LLaMA-1的基本框架，但值得注意的是，在构建基础模型的过程中，Meta引入了相较于LLaMA-1多出40%的数据进行训练，从而提升了模型对广泛语境和任务的理解能力。LLaMA-2产品系列不仅包括针对通用自然语言处理任务的基础模型，而且还推出了经过对话场景微调优化的变体——LLaMA-2 Chat，专为提升人机交互体验而打造。相比于其前身LLaMA-1，LLaMA-2的一大突破性举措在于所有模型权重的全面开放，并且对于广泛的商业应用场景，提供免费使用的权限，此举无疑拓宽了LLaMA-2在业界的应用范围和影响力。

LLaMA具有以下特点：

- **高效能与灵活性：**尽管参数规模相对较小，但LLaMA模型在许多NLP任务上的性能优于同等参数量级的其他模型，展现出高效的参数利用率和优秀的泛化能力。
- **开源友好：**Meta AI选择在非商业许可下向研究界开放LLaMA的模型权重，鼓励学术研究和应用开发。
- **模块化设计：**LLaMA框架体现了高度的模块化设计理念，便于开发者根据实际需求定制和集成不同的组件。
- **社区活跃：**由于模型开源，社区可以不断迭代更新模型版本，LLaMA在性能和适应性方面保持了持续进步，为人工智能领域的研究和发展注入新的活力。

（5）Mixtral

Mixtral是由MistralAI开发的一款大语言模型，它采用了专家混合（MoE）架构，这种架构通过一个网关网络将输入数据分配给被称为“专家”的特定神经网络组件。Mixtral 8x7B模型由八个各自拥有70亿参数的专家组成，这种设计提高了模型训练和运算的效率及可扩展性。Mixtral8x7B在多个领域表现出色，包括综合性任务、数据分析、问题解决和编程支持等。

2023年9月27日，Mistral AI通过官方BitTorrent磁力链接以及Hugging Face平台发布了Mistral 7B模型，该模型采用了拥有7亿个训练参数，并且严格遵循Apache2.0开源许可证，为全球开发者和研究者提供了自由访问和使用的权限。2023年12月9日，Mistral AI发布了Mixtral 8x7B，其构建在稀疏专家混合（MoE）的革新框架之上，尽管总体参数量达到了467亿之多，但得益于MoE技术的高效性，对于每个Token仅激活12.9亿相关参数进行运算。此款模型支持包括法语、西班牙语、意大利语、英语及德语在内的五大语言环境，并在多项基准测试中表现卓越，声称已成功超越了Meta公司的LLaMA 270B模型的性能水平。2024年2月26日面世的Mistral Large，则是Mistral AI的又一旗舰产品，被定位为紧随OpenAI GPT-4之后的顶级大模型。它不仅支持多种语言处理任务，还具备编程能力，在多领域应用上展示了强大的适应性和创造性。用户现可通过Microsoft Azure云端平台便捷使用这款高性能模型。Mistral Medium型号则是在广泛的多语言文本和代码数据集上进行深度训练后推出的，其综合性能评价位于Claude模型与GPT-4之间，为寻求平衡资源占用与处理效能需求的用户提供了一个理想的选择。最后，Mistral Small作为轻量化解决方案，旨在提供低延迟响应且性能不俗的小型模型。相较于Mixtral 8x7B，它在保证快速响应的同时，实现了更优的性能指标，从而在有限计算资源场景下展现出极高的实用价值。

图9. Mistral AI

（6）Stable Diffusion

Stable Diffusion 是2022年发布的深度学习文本到图像生成模型。它主要用于根据文本的描述产生详细图像，尽管它也可以应用于其他任务，如内补绘制、外补绘制，以及在提示词指导下产生图生图的转变。它是一种潜在扩散模型，由慕尼黑大学的CompVis研究团体开发的各种生成性人工神经网络之一。它是由初创公司StabilityAI、CompVis与Runway合作开发，并得到EleutherAI和LAION的支持。Stable Diffusion由3个部分组成：变分自编码器（VAE）、U-Net和一个文本编码器。

StableDiffusion算法上基于2021年12月提出的潜在扩散模型（LDM / Latent Diffusion Model）和2015年提出的扩散模型（DM/Diffusion Model，它是基于Google的Transformer模型）。2022年7月Stable Diffusion的问世则震惊了全球，相比前辈们，Stable Diffusion已经成功的解决了细节及效率问题，通过算法迭代将AI绘图的精细度提升到了艺术品级别，并将生产效率提升到了秒级，创作所需的设备门槛也被拉到了民用水准。2022年8月Stable Diffusion的开源性质，全球AI绘图产品迎来了日新月异的发展，AI绘图正在走进千家万户，舆论热潮也随之而来。2023年7月，Stability AI发布1.0版本的Stable Diffusion XL，1.0基础模型有35亿个参数，使其比以前版本大了约3.5倍。并提到在训练结束后参数稳定后会开源，并改善了需要输入非常长的提示词 (prompts)，对于人体结构的处理有瑕疵，经常出现动作和人体结构异常。2023年11月发布了Turbo版本的Stable Diffusion XL，Turbo版提取自XL 1.0而以更少扩散步骤运行。

（7）Midjourney

Midjourney是一款AI制图工具，只要关键字，就能透过AI算法生成相对应的图片，只需要不到一分钟。可以选择不同画家的艺术风格，例如安迪华荷、达芬奇、达利和毕加索等，还能识别特定镜头或摄影术语。有别于谷歌的Image和OpenAI的DALL-E，Midjourney是第一个快速生成AI制图并开放予大众申请使用的平台。Midjourney生成的作品往往带有电脑生成的痕迹，比较不会被当成假新闻素材，但对色情、血腥、暴力创作题材的审核还不够精准。

Midjourney由位于美国加州旧金山的同名研究实验室开发，于2022年3月首次亮相，于2022年7月12日进入公开测试阶段，在8月迭代至V3版本并开始引发一定的关注，而2023年更新的V5版本让Midjourney及其作品成功“出圈”。2023年4月，入选《福布斯2023年AI 50榜单：最有前途的人工智能公司》。2023年5月15日，Midjourney官方中文版已经开启内测。

（8）DALL-E

DALL-E 是一个可以通过文本描述生成图像的人工智能程序，由OpenAI发布。DALL-E通过120亿参数版本的GPT-3 Transformer模型来理解自然语言输入并生成相应的图片。它既可以生成现实的对象，也能够生成现实中不存在的对象。它的名字是2008年动画电影WALL-E（瓦力）和20世纪西班牙加泰罗尼亚画家萨尔瓦多·达利（Salvador Dalí）之混成词。自2000年代以来，已有其他许多神经网络有生成逼真图像的能力。而DALL-E的特点在于它能够通过纯文本描述生成这样逼真的图像。OpenAI尚未发布DALL-E模型的源代码，不过OpenAI在其网站上提供了DALL-E演示，可以查看部分文本描述的输出图像。

DALL-E模型最初于2021年1月5日由OpenAI发布。2022年4月，OpenAI宣布了新版本的DALL-E 2，声称它可以从文本描述中生成照片般逼真的图像，另外还有一个允许对输出进行简单修改的编辑器。根据OpenAI的公告，该程序仍处于研究阶段，访问权限仅限于小部分测试版用户。该模型有时仍会犯一些人类不会犯的严重错误。OpenAI称DALL-E 2是一个“可以从文本描述中生成原创、逼真的图像和艺术”的模型。

（9）Sora

Sora是一个能以文本描述生成视频的人工智能模型，由美国人工智能研究机构OpenAI开发。Sora这一名称源于日文“空”（そら sora），即天空之意，以示其无限的创造潜力。其背后的技术是在OpenAI的文本到图像生成模型DALL-E基础上开发而成的。模型的训练数据既包含公开可用的视频，也包括了专为训练目的而获授权的著作权视频，但OpenAI没有公开训练数据的具体数量与确切来源。 OpenAI于2024年2月15日向公众展示了由Sora生成的多个高清视频，称该模型能够生成长达一分钟的视频。同时，OpenAI也承认了该技术的一些缺点，包括在模拟复杂物理现象方面的困难。《麻省理工科技评论》的报道称演示视频令人印象深刻，但指出它们可能是经精心挑选的，并不一定能代表Sora生成视频的普遍水准。由于担心Sora可能被滥用，OpenAI表示目前没有计划向公众发布该模型，而是给予小部分研究人员有限的访问权限，以理解模型的潜在危害。Sora生成的视频带有C2PA元数据标签，以表示它们是由人工智能模型生成的。OpenAI还与一小群创意专业人士分享了Sora，以获取对其实用性的反馈。

Sora具有以下特点：

- **准确性和多样性：**Sora能够将简短的文本描述转化成长达1分钟的高清视 频，准确地解释用户提供的文本输入，并生成具有各种场景和人物的高质量视频剪辑。它涵盖了广泛的主题，如人物、动物、风景、城市场景等，可根据用户的要求提供多样化的内容。
- **强大的语言理解能力：**Sora利用Dall-E模型的re-captioning技术生成视觉训练数据的描述性字幕，提高了文本的准确性，同时也提升了视频的整体质量。此外，利用GPT技术将简短的用户提示转换为更长的详细转译，确保视频精确地按照用户提示生成。
- **以图/视频生成视频：**Sora除了可以将文本转化为视频，还能接受其他类型的输入，如已存在的图像或视频，使其能够执行广泛的图像和视频编辑任务。
- **视频扩展功能：**Sora能够沿时间线向前或向后扩展视频，允许用户根据图像创建视频或补充现有视频。
- **优异的设备适配性：**Sora具备出色的采样能力，能够应对从宽屏到竖屏的各种视频尺寸，为各种设备生成与其原始纵横比完美匹配的内容。
- **场景和物体的一致性和连续性：**Sora能够生成带有动态视角变化的视频，人物和场景元素在三维空间中的移动显得更加自然，能够很好地处理遮挡问题。

1.3.2 国内大模型

(1) 百度-文心一言

文心一言（英文名：ERNIE Bot）是百度基于文心大模型技术研发的知识增强大语言模型，被外界誉为“中国版ChatGPT”。其核心理念在于运用深度学习算法和大规模语料库，模拟人类的语言理解和生成能力，从而为用户提供智能化、个性化的服务。能够实现与人对话互动，回答问题，协助创作，高效便捷地帮助人们获取信息、知识和灵感，并且在文学创作、商业文案创作、数理逻辑推算、中文理解、多模式生成方面有很好的应用前景。

文心一言最早应该可以追溯到2010年百度成立的“自然语言处理部”，2019年3月16日，百度正式发布知识增强的文心大模型ERNIE1.0，该模型基于飞桨深度学习平台打造，通过将数据与知识融合，提升了大模型学习效率及学习效果。2019年7月31日，百度文心大模型升级到2.0。ERNIE 2.0通过持续学习框架，持续学习大规模语料中的词法、语法、语义等知识，在共计16个中英文任务上取得全球最好效果。2021年7月6日，百度发布文心大模型 3.0 (ERNIE 3.0)。ERNIE 3.0首次在千亿级预训练模型中引入大规模知识图谱，ERNIE 3.0刷新54个中文NLP任务基准，并在国际权威的复杂语言理解评测SuperGLUE上, 以超越人类水平0.8个百分点的成绩登顶全球榜首。2023年3月16日，百度新一代大语言模型文心一言正式启动邀测。2023年8月31日，文心一言率先向全社会全面开放。开放首日，文心一言共计回复网友超3342万个问题。2023年10月17日，百度世界2023大会上，李彦宏宣布文心大模型4.0正式发布，开启邀请测试。

（2）阿里-通义

阿里通义是阿里云推出的一系列人工智能产品和服务平台，旨在提供类人智慧的通用智能服务。这些产品和服务包括通义千问、通义智文等，它们支持多种API接口，使得AI应用开发变得更加简单和高效。通义千问（Qwen）是阿里云推出的一款超大规模语言模型，采用了阿里云自主研发的大规模预训练语言模型架构，通过先进的深度学习技术和海量数据训练而成。通义智文是另一个阿里云的AI产品，它可能包含了文本生成、内容理解、自动摘要、情感分析等功能，旨在帮助用户高效地处理和创造文本内容。

目前，通义千问的综合性能已经超过GPT-3.5，加速追赶GPT-4。2023年12月1日，阿里云举办发布会，正式发布并开源“业界最强开源大模型”通义千问720亿参数模型Qwen-72B。同时，通义千问开源了18亿参数模型Qwen-1.8B和音频大模型Qwen-Audio。至此，通义千问共开源18亿、70亿、140亿、720亿参数的4款大语言模型，以及视觉理解、音频理解两款多模态大模型，实现了“全尺寸、全模态”开源。自此，阿里云大模型的开源逻辑更加清晰，即通过开源的方式提供技术产品，降低门槛，推动技术普惠，为企业客户到个人开发者提供多元化、全方位的技术服务。在通义千问的基础上创建的大模型、小模型越丰富，AI生态就越繁荣。

（3）讯飞-星火认知大模型

讯飞星火认知大模型是由科大讯飞推出的新一代认知智能大模型。基于讯飞最新的认知智能大模型技术，经历了各类数据和知识的充分学习训练，可以和人类进行自然交流，解答问题，高效完成各领域认知智能需求。

讯飞星火V2.0已具备“代码生成、代码补齐、代码纠错、代码解释、单元测试生成”等能力，并且在业界参考测试集与真实应用场景均达到优异效果，逼近国外领军者。星火认知大模型V3.0的快速落地，更是推动着讯飞大模型能力迅速迫近行业前列，其在数学自动提炼规律、小样本学习、代码项目级理解能力以及多模态指令跟随与细节表达等方面进行了进一步升级，这些能力的提升将融入星火金融大模型中，为大模型在金融行业的落地应用带来全新机遇。2024年1月30日，科大讯飞发布了基于首个全国产算力平台「飞行一号」训练的全民开放大模型——讯飞星火V3.5版本。相较于上一个版本，讯飞星火V3.5版本在文本生成、语言理解、知识问答、逻辑推理、数学能力、代码能力、多模态能力等七大核心能力上均实现大幅提升，进一步逼近GPT-4 Turbo的最新水平。

图10. 讯飞星火大模型

(4) 华为-盘古大模型

华为盘古大模型是华为云推出的一系列人工智能大模型，旨在通过强大的计算能力和先进的算法，解决行业难题并释放AI的生产力。该模型涵盖了NLP大模型、CV大模型、多模态大模型、预测大模型和科学计算大模型五大类别，旨在为气象、医药、水务、机械等领域提供强大的科学计算能力。盘古大模型的研发不仅体现了华为在AI技术领域的深厚积累，也展示了华为在推进AI技术产业化应用方面的决心和能力。

在2021年4月，盘古大模型1.0就已经发布，早于今天大部分的大模型。2022年11月7日的华为全联接大会2022中国站上发布了盘古气象大模型、盘古海浪大模型、盘古矿山大模型、盘古OCR大模型等新服务。2023年7月举行的华为开发者大会上，华为云曾发布了盘古大模型3.0，是中国首个全栈自主的AI大模型，该模型已具备文生图、文生文、文生代码、文生视频等多模态能力，提供5+N+X的三层解耦架构：L0层有5个基础大模型，提供满足行业场景的多种技能；L1层是N个行业大模型，提供使用行业数据训练的行业大模型；L2层为客户提供更多细化场景模型，它更加专注于某个具体应用场景或特定业务。华为常务董事、华为云CEO张平安表示，盘古大模型聚焦产品研发、软件工程、生产供应、市场营销、客户运营等价值场景，致力于深耕行业，如政务、金融、制造、煤矿、铁路、制药、气象等。

图11. 盘古大模型

(5) 腾讯-混元大模型

腾讯混元大模型（Tencent Hunyuan）是腾讯自主研发的通用大语言模型，拥有超过千亿参数规模和超过2万亿tokens的预训练语料。该模型具备强大的中文理解与创作能力、逻辑推理能力，以及可靠的任务执行能力。

腾讯混元大模型的主要功能和技术特点包括成为腾讯云MaaS服务的基础，客户可以通过API直接调用混元，也可将其作为基础模型，为不同产业场景构建专属应用。该模型具备强大的中文创作能力、复杂语境下的逻辑推理能力和可靠的任务执行能力。其全链路自研技术是其首要特点，从零开始训练，掌握了模型算法、机器学习框架和AI基础设施。腾讯在算法层面进行了自研创新，提高了模型可靠性和成熟度，解决了大模型“胡言乱语”的问题。此外，腾讯还自研了机器学习框架Angel，提升了训练和推理速度。

腾讯混元大模型能够理解上下文含义，具有长文记忆能力，可进行专业领域的多轮对话、文学创作、文本摘要、角色扮演等内容创作。它能高效、准确地理解用户意图，解决事实性、时效性问题，提升内容生成效果。在不同场景下，如文档、会议、广告和营销，混元大模型提供了各种功能，包括文档创作、会议总结、广告素材创作等，提高工作效率并改善用户体验。

图12. 混元大模型

（6）智谱AI

北京智谱华章科技有限公司（简称“智谱AI”）专注于新一代认知智能大模型的研发，致力于在中国推动大模型领域的创新。公司与合作伙伴共同研发了中英双语千亿级超大规模预训练模型GLM-130B，并在此基础上推出了对话模型ChatGLM以及开源单卡版模型ChatGLM-6B。同时，团队还开发了GenAI模型及产品矩阵，包括AI提效助手智谱清言、高效率代码模型CodeGeeX、多模态理解模型CogVLM和文生图模型CogView等。智谱AI秉承Model as a Service（MaaS）的市场理念，推出了大模型MaaS开放平台，旨在构建高效率、通用化的“模型即服务”AI开发新范式。通过认知大模型连接亿级用户的物理世界，智谱AI凭借完整的模型生态和全流程技术支持，为各行各业带来持续创新与变革，助力加速通用人工智能时代到来。

智谱AI的产品包括ChatGLM-6B、GLM-130B、GLM系列、CodeGeeX、CogView、CogVideo等大模型。在2024年01月16日的「智谱AI技术开放日(Zhipu DevDay)」上，智谱AI推出了新一代基座大模型GLM-4。GLM-4相比上一代在整体性能上有了显著提升，十余项指标接近或达到了GPT-4水平；支持更长上下文、更强的多模态、更快的推理速度和更多并发，大幅降低了推理成本；同时，GLM-4还增强了智能体能力。

图13. 智谱AI

(7) 百川智能

百川智能公司于2023年4月10日由前搜狗公司CEO王小川创立，旨在以帮助大众轻松、普惠地获取世界知识和专业服务为使命。公司专注于通过语言人工智能技术的创新，构建中国顶尖的大模型基础设施。其核心团队由来自搜狗、百度、华为、微软、字节、腾讯等知名科技公司的AI顶尖人才组成。不到100天的时间里，百川智能公司发布了两款开源可免费商用的中文大模型Baichuan-7B和Baichuan-13B在多个权威评测榜单中名列前茅，下载量更是突破了百万。随后，公司继续发布了Baichuan2-7B和Baichuan2-13B等大开源模型。2024年1月29日，百川智能发布了参数规模超过千亿的大语言模型Baichuan3。在多个权威通用能力评测中，如CMMLU、GAOKAO和AGI-Eval，Baichuan3展现出色的能力，特别是在中文任务上超越了GPT-4。在数学和代码专项评测中，如MATH、HumanEval和MBPP，Baichuan3同样表现出色，证明了其在自然语言处理和代码生成领域的强大实力。

Baichuan3在多个医疗评测任务中表现优异，特别在对逻辑推理能力和专业性要求极高的MCMLE、MedExam、CMExam等权威医疗评测中，中文效果超过了GPT-4，成为中文医疗任务中表现最佳的大模型。此外，Baichuan3还在诗词创作、逻辑推理等方面表现出色，领先于其他大模型。

Baichuan-NPC通过强化模型基础能力，使用思维链对齐技术赋予角色模型类人的思考能力，使模型能够敏锐地捕捉上下文对话语义，生成更加符合人物性格的对话和行动，呈现出逼真的角色效果。在CharacterEval评测中，Baichuan-NPC在对话能力、角色一致性、扮演吸引力等方面显著领先，是目前中文领域最强角色模型。

第二章：
GenAI在生物医药大健康行业落地应用进展及典型案例

2.1GenAI在生物医药大健康行业主要应用场景总览

GenAI作为人工智能领域的重要分支，正逐渐引起广泛关注。GenAI以其独特的生成能力和创造性，正在改变着生物医药大健康行业的面貌，并为其带来了巨大的变革和潜在收益。GenAI已经在医疗健康开始了应用探索与落地，包括药物研发、临床研究、上市及商业化、以及用于病人诊疗等方面：

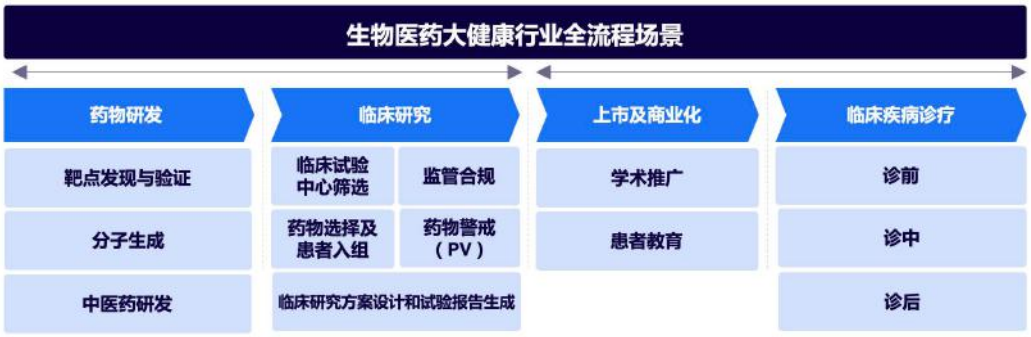


图14. 生物医药大健康行业全流程场景

在药物研发方面，GenAI可助力科研人员靶点发现及验证、药物分子生成以及为中医药研发等，加速药物发现和设计进程。在临床研究方面，GenAI可以优化临床开发的多个环节，包括筛选临床试验中心筛选、监管合规、药物选择及患者入组、药物警戒（PV）和临床研究方案设计等方面。在上市及商业化方面，GenAl主要体现于学术推广及患者教育等方面。在临床疾病诊疗方面，GenAI可以实现包括在诊前、诊中、诊后等医疗服务场景的提质和提效。

2.2 药物研发

随着人类社会对生物学、化学、物理学等的知识积累加深，我们对疾病的认识迅速提升。然而新药研发、上市的速率却没有同比例增长，其中一个重要原因是药物研发阶段耗时长、成本高、流程复杂。现如今，GenAI成为潜在的破局利器，在药物研发方面已有越来越多的应用。GenAI可以通过分析大量的基因组学、蛋白质组学、代谢组学等多组学数据，帮助研究人员更快地发现可能的靶点、药物分子、药物合成路线等，从而加速药物的发现和设计进程。

2.2.1 靶点发现与验证

药物靶点指的是药物与人体内特定分子相互作用的目标位置，也可以是参与疾病发生和发展的关键蛋白质、酶或细胞结构。靶点的发现是现代药物研发的基础，在药物研发的早期阶段，科研人员通过对疾病发生机制的深入研究，寻找与疾病相关的靶点，通过对这些靶点的深入了解，科研人员可以设计出针对性的药物，干预其功能，从而达到治疗疾病的目的。但疾病的发生非常复杂，基因冗余和多效性、代偿机制、信号反馈等，都会降低靶点被药物作用造成的影响。药物立项要经过药理学、毒理学、遗传学等多学科的检验，成功率非常低。疾病相关的靶点的早期识别和评估可以增加药物批准的机会。目前业界在致力于识别与特定疾病相关的生物学上最合理的靶点。近年来，传统的生信分析方法、多组学因子分析、知识图谱、图模型、深度学习等都较多地应用于靶点发现中，而GenAI的发展，则提供了更强有力的技术支持去寻找新的靶点、分析信号通路、以及寻求靶点与疾病相互关联的证据支撑。

2022年， AI 制药公司英矽智能（Insilico Medicine）在其靶点发现平台 PandaOmics 上增加了知识图谱的功能，可以从期刊文献中提取相关信息，将基因、疾病、化合物

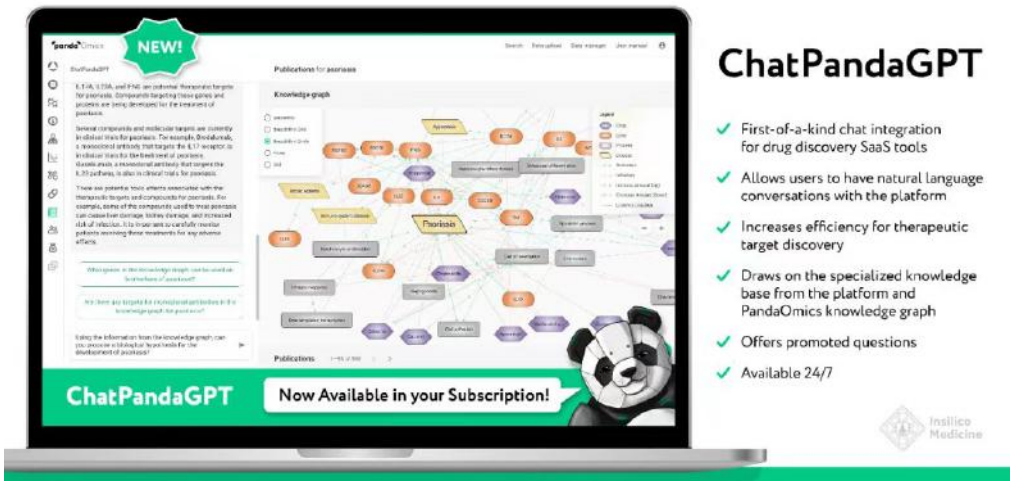


图15. PandaOmics页面

和生物过程联系起来，并将关系网络可视化形成知识图谱。2023年3月，在ChatGPT开放端口后，英矽智能将其接入靶点发现平台PandaOmics。通过将知识图谱与ChatGPT相结合，得到了具有AI问答功能的ChatPandaGPT，支持研究人员在浏览和分析大数据集的同时，高效开展基于自然语言的问答，更便捷发现潜在靶点和生物标志物。

2023年底，英矽智能发布了全球首个“由AI辅助决策的自动化实验室”，将GenAI应用于高质量自动化实验，并通过实验数据反馈推动GenAI模型迭代优化。实现在14天内完成靶点发现和验证的全自动化干湿实验闭环。目前，英矽智能宣布公司研发的抗特发性肺纤维化候选药物INS018_055已完成2期临床试验首批患者给药，这是全球首款由GenAI发现靶点并设计化合物的候选药物。

2023年4月，水木分子开源了轻量科研版BioMedGPT-10B，将文献、分子、蛋白、测序、知识图谱等数据压缩到统一的多模态大模型框架内，实现了分子性质预测、药物-靶点亲和力预测、性质预测、药物敏感性预测、分子-文本跨模态检索、分子-文本跨模态信息生成等多项任务性能优于单一专用模型。

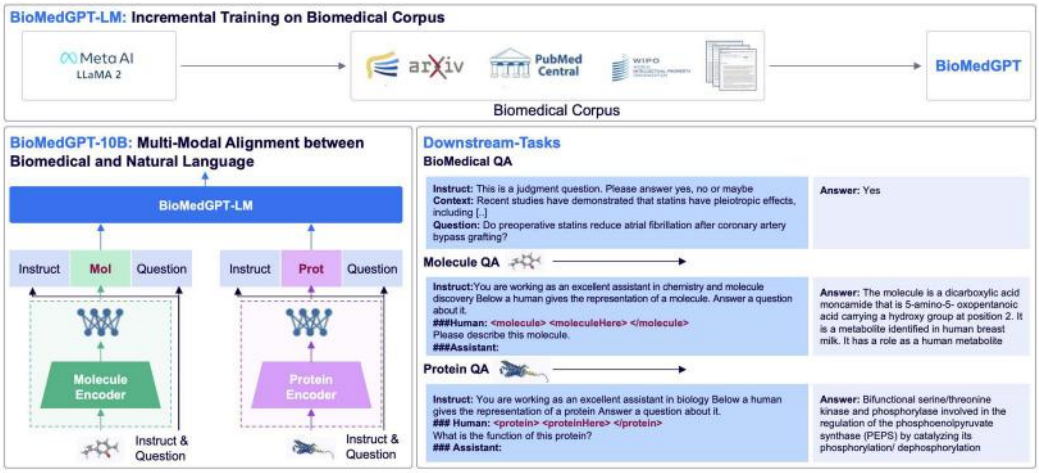


图16. BioMedGPT-10B的概述

2.2.2 分子生成

（1）大分子生成

大分子药物的作用机制主要是通过刺激机体免疫系统产生免疫物质（如抗体），从而发挥其功效，在人体内出现体液免疫、细胞免疫或细胞介导免疫。大分子药物如抗体有更好的靶向性、mRNA有望带来更好的疫苗与药物等。但这些大分子结构复杂，开发成本高，发现时间长。GenAI为此提供了新的解决方案，通过学习预测大分子（例如核酸或氨基酸）的下一个子结构，并产生有关大分子的见解，这些见解可用于新药物载体的计算机设计、以及预测其在各种药物研发检测的效果。

人类一直以来都在探索如何更高效、直接、自由地控制细胞功能、甚至生命活动，而运用 GenAI预测大分子结构使得这一切的可行性大大提升，并且其预测性能在精度、范围、耗时等方面正在不断快速提高。以蛋白质为例，利用GenAI预测结构，为解码蛋白质的三维奥秘提供了高效手段，从而能够直接按需制造蛋白质、甚至是创造未知或不存在的蛋白，扩增出近乎无限的、广阔的蛋白质序列和结构空间，对生命科学和生物医药研究范式的影响将是颠覆性的。

目前，GenAI在蛋白质解码和设计的应用主要有Transformer架构和扩散性模型两大构建思路。前者的代表是美国初创生物医药公司Profuluent在2023年1月开发的蛋白质语言模型Progen。该模型基于Transformer架构的12亿参数神经网络，提供了一种可根据所需属性生成特定蛋白质的方法，从头合成了自然界中不存在的人工酶，引起了生命科学领域的广泛关注。而后者构建思路则是采取了图像生成领域常用的扩散性模型的技术路径，更加擅长基于文本生成图像来描述蛋白质序列和结构之间的关系，并以此快速生成蛋白质的骨架结构。例如2022年10月美国斯坦福大学和微软研究院经受体内蛋白质折叠过程的启发，引入了一个折叠扩散模型，通过镜像蛋白质天然折叠过程实现蛋白质主链结构的设计，解决了直接生成结构复杂多样的蛋白质的难题。

扫描跨国大型药企与科技公司动态，2023年12月，制药巨头阿斯利康与AI 抗体发现技术初创公司Absci 签署了 2.47 亿美元协议，通过整合阿斯利康的肿瘤学研究和开发知识、以及利用 Absci 的集成药物创造平台，借助GenAI 技术来开发新的、改进的抗癌抗体疗法。2023年10月，Deepmind联合Isomorphic Labs共同发布了新一代AlphaFold模型，从上一代的预测蛋白质结构，扩展到预测蛋白质数据库（PDB）中几乎任何分子的结构，包括配体（小分子）、蛋白质、核酸（DNA 和 RNA）以及含有翻译后修饰（PTM）的生物分子。如改变氨基酸序列来改变蛋白质的性能，用于设计和开发具有特定功能的酶；预测核酸结构，加速 mRNA 疫苗等医疗创新；预测配体和蛋白质间的相互作用，帮助鉴定和设计可能成为药物的新分子等。Isomorphic Labs 正在将新一代 AlphaFold 模型应用于治疗药物设计，快速准确地表征对治疗疾病很重要的多种类型的大分子结构。

国内头部药企与AI制药企业也在GenAI生成大分子药物领域展开布局。2023年8月，深圳晶泰科技宣布与石药集团在创新药研发AI领域达成战略合作协议，结合石药集团深厚药物研发经验，利用晶泰科技开发的ProteinGPT大分子药物生成式AI模型，将“类GPT技术”应用于药物研发，覆盖抗体发现、抗体工程、抗原设计、蛋白结合剂设计等多个药物研发关键环节，一键生成符合要求的抗体或蛋白药物。荷兰-瑞士初创公司 Cradle开发的生成人工智能 (GenAI) 和合成生物学平台，旨在设计基于蛋白质的疗法和其他化合物，正在开展 12 个研发项目，关注工程酶、疫苗、肽药物和抗体，涵盖广泛所需蛋白质特性，如稳定性、表达、活性、结合亲和力和特异性。Cradle 的技术可以通过更少、更成功实验大幅加快蛋白质的设计和优化。与行业基准相比，大多数项目使用 Cradle 平台的进度要快两倍。

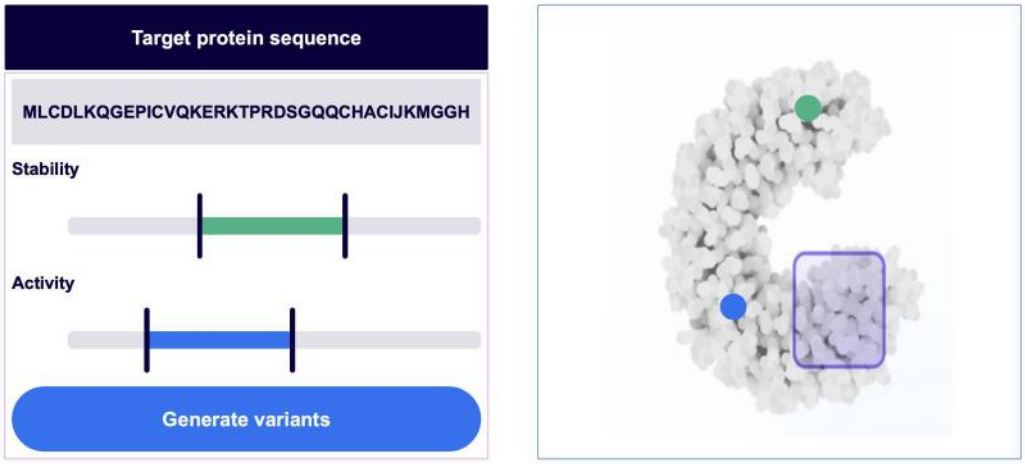


图17. Cradle公司合成生物学平台功能示意图

（2）小分子生成

小分子药物研发中的一大重要难题是如何识别并且筛选出最有可能实现所需疗效、值得进一步测试优化的化合物，传统上，药物化学家会在实验室制造化合物并进行测试，耗时长、投资大，但人工智能可以改变这个过程。GenAI通过先进的基础化学模型加速筛选过程，如同GPT-4被训练来预测句子中可能的下一个单词，这些模型可以预测小分子结构中的下一部分原子。通过多次迭代，该模型学习了小分子化学的基本原理，即使在很大程度上未探索的化学领域，这些模型也可以提供更精确的预测，医药公司可以通过这些预测来规划后续筛选。

国外大型药企与AI制药企业纷纷开展合作，各取所长。2024年1月，默沙东宣布与Variational AI公司达成合作，利用其Enki技术平台，共同合作开发小分子药物。默沙东为Enki平台提供目标产品概况（TPP），平台基于GenAI技术，可在几天时间内生成符合条件的小分子。生成物是具有多样化、选择性和可合成的先导化合物结构，从而快速进入先导化合物优化阶段。法国药物化学和新药设计AI解决方案提供商Iktos则是利用GenAI技术，降低化合物小分子筛选和生成所需的时间和成本。

其解决方案包括三个部分，一是通过Makya基于大量生物数据，来创建一种“满足所有条件”的分子，即在尽可能低的剂量下有效、安全、稳定、可申请专利且能够合成的分子；二是利用Spaya探索合成“配方”和途径；三是通过Ilaka 软件控制机器人，一次性高效合成多种化合物，不断重复上述过程，以找到更有前途的化合物。目前其拥有 50 多个已完成或正在进行的项目，合作伙伴包括强生、默克、辉瑞等跨国大型药企。回看国内，多家AI制药企业、大型药企、科技公司等，也在布局GenAI药物分子生成。英矽智能推出小分子生成AI平台Chemistry42，经过10万种公开化合物和100亿个构建块（或虚拟分子片段）的训练，生成数百个具有所需特性的化合物，被输送到管道中评估适用性，并选择满足安全性、效力、合成可用性和代谢稳定性等目标的分子。生成的分子及其后续分数将返回到生成引擎，以便模型“学习”得分高的分子类型和得分低的分子类型，重新训练生成模型以生成高分分子，已实现在一周内发现全新的先导化合物类似分子，远超人类科学家的速度。自Chemistry42推出以来，已有40多家制药公司授权该软件并将其用于自己的管道程序，以改善自己对下一个突破性疗法的探索。

2022年4月份，华为云计算技术的健康智能实验室推出了华为盘古药物分子大模型，该模型训练了17亿个小分子化合物的数据集，这一模型结合了药物分子的图形结构和SMILES字符表示法，从两个不同的角度理解分子，进而构建了一个自监督的预训练大模型。该模型适用于多个分子相关的后续任务，如预测分子属性、生成分子虚拟库以及分子的优化等。目前盘古药物模型的预训练数据集是最大的，涵盖了多个公共数据源，盘古模型采用cVAE架构，将小分子的图形表示转换成相应的化学式字符串，这样做避免了在graph2graph模型中遇到的图形生成的困难，并且相比于seq2seq模型，在训练阶段能够提供更多的信息。此外，通过设计分层的潜在空间，盘古模型在微调和化学指纹表示方面的能力得到了增强。盘古的创新网络结构不仅易于训练，还能够通过仅更新一个核心网络来支持所有药物发现任务的步骤，展现出显著的优势。

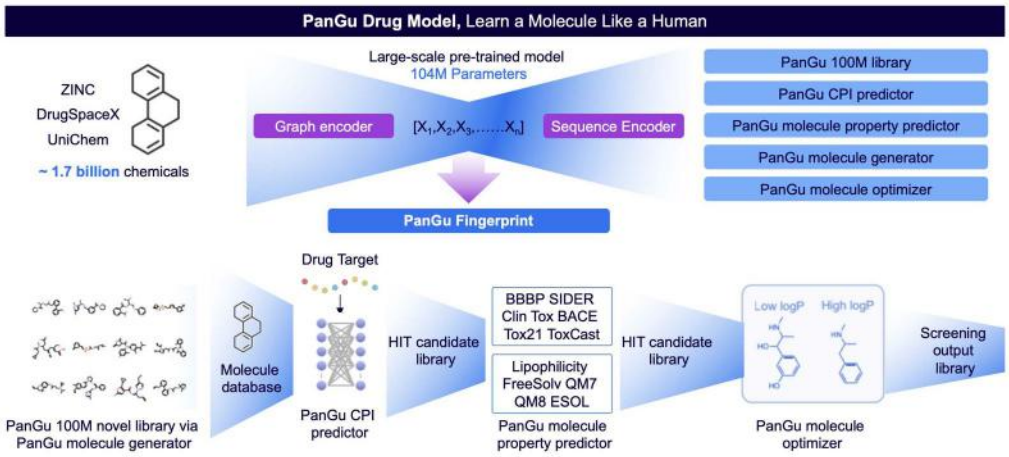


图18. 用于人工智能药物筛选和分子生成的盘古大模型

2023年底，深度势能团队联合29家单位的通力协作，发布了深度势能预训练大模型DPA-2。该模型面向丰富的下游任务，在微调DPA-2的"大模型"所需数据量整体上减少了1-2个数量级。此外，经过进一步蒸馏和压缩，深度势能团队还开发了"小模型"，该模型能够保持过去模型的精度和效率。与去年发布的DPA-1相比，DPA-2在模型架构方面有显著的更新，最大的特点是采用了多任务训练策略，可以同时学习计算设置不同、标签类型不同的各类数据集。由此产生的模型在下游任务上展现出极强的few-shot甚至zero-shot迁移能力，显著超越了过去的解决方案。目前，用于训练DPA-2模型的数据集已涵盖了半导体、钙钛矿、合金、表面催化、正极材料、固态电解质、有机分子等多个体系。

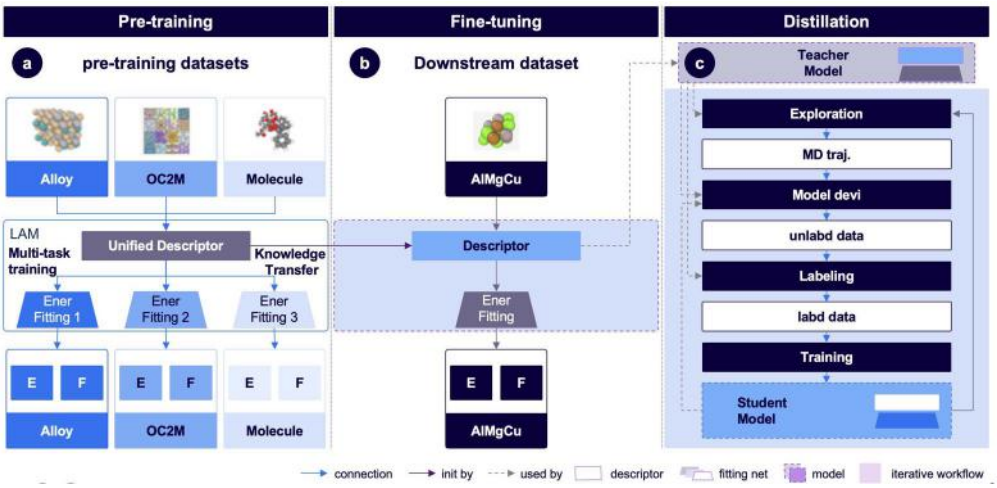


图19. DPA-2 提出的多任务预训练、微调、蒸馏全流程示意图

洛桑联邦理工学院（EPFL）和美国罗切斯特大学的研究团队，开发出了一款名为ChemCrow的语言模型代理，这款代理能够执行包括有机合成、药物发现和材料设计在内的多项化学任务。ChemCrow集成了17种由专家精心设计的工具，不仅提升了其在化学领域的表现，还赋予了它新的能力。迄今为止，ChemCrow已成功自行设计出一种驱虫剂、三种有机催化剂以及其他相关分子。通过语言模型评估和专家的评审，ChemCrow的有效性在自动执行各类化学任务方面得到了证实。

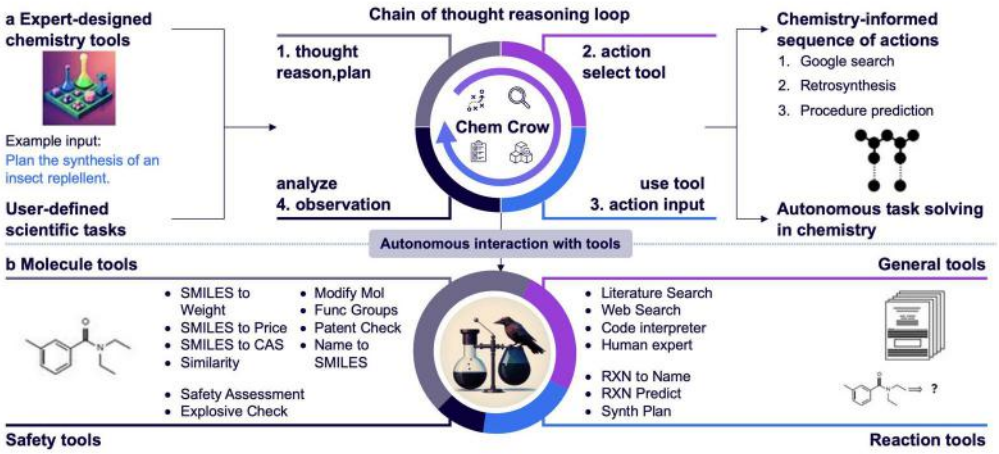


图20. ChemCrow的功能示意图

中科大MIRA Lab团队与微软研究院AI4Science团队共同开发了一种创新的分子生成模型，名为MiCaM。该模型通过构建一个含有数据驱动的高频分子片段词汇库，显著优化了药物分子的生成过程。MiCaM模型特别强调了对连接感知的高频子图（Mined Connection-aware Motifs）的利用，这些子图是通过数据驱动算法从分子库中自动提取的，它们不仅包括常见的分子片段，还细致记录了这些片段之间的连接信息。利用这一策略，MiCaM设计了一个能够同时选择分子片段并确定其连接方式的生成器，从而能够基于这些高频子图构造出全新的分子结构。在进行的两项基准测试中：一项是生成与训练集高度相似的新分子（distribution learning），另一项是创造具备特定目标属性的新分子（goal-directed）,MiCaM模型展现了其在提高分子生成效率和探索化学空间方面的显著能力。

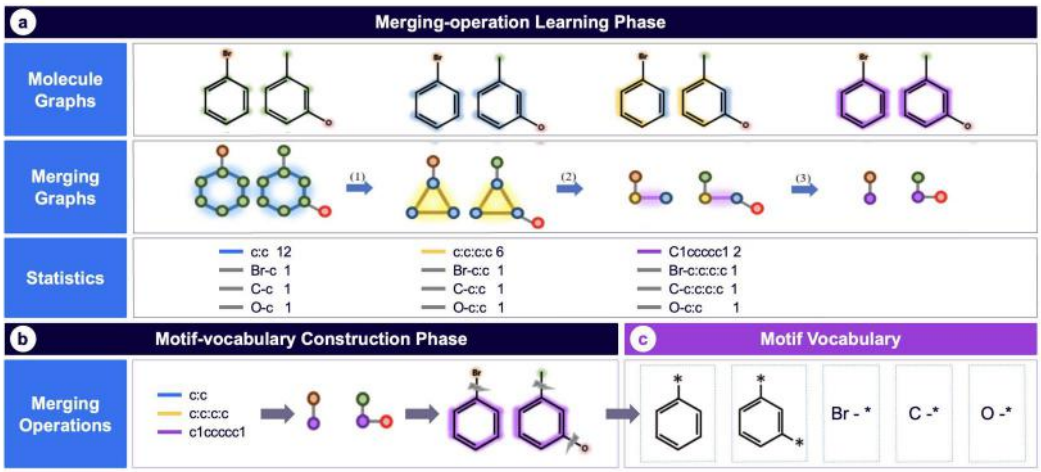


图21. 高频分子片段挖掘算法流程图

2.2.4 中医药研发

2024年全国两会期间，政府工作报告中提出“促进中医药传承创新”，有代表提议，大模型与生物医药大健康行业的结合有望革新药物研发范式，通过构建中医药AI大模型，将能够用于挖掘中药活性成分、推进中药循证工作、加速中药研发进程。中国拥有自己庞大的医学宝库，中医药的“整体观”对人体、疾病、药物的认识积累庞大的实践经验。随着人工智能技术的飞速发展，中医药领域迎来一场数字化、智能化的革命。中医药领域的大语言模型（中医药大模型）作为这场革命的代表，它通过整合和分析大量的中医药文献、药材数据库、临床研究和实践经验来训练，旨在将人工智能技术应用于传统医学知识和实践中。目前，中医药大模型主要用于中医临床辅助诊疗（病证诊断、处方推荐等）、创新研发以及中医药知识整合与普及。下面我们将盘点现有的中医大模型在创新研发以及中医药知识问答的应用。

(1) 数智岐黄 · 大模型

“数智岐黄”中医药大模型由华东师范大学、上海中医药大学、华东理工大学、海军军医大学、临港实验室与华润江中现代中药全国重点实验室联合开发，它以《黄帝内经》和《伤寒杂病论》等著名中医典籍及1000多本古籍和中医药文献为核心数据基础，以高质量中医药知识图谱为知识宝库。“宝库”中涵盖超过8万种方剂，超过2000种证候，超过9000种中药材，超过4万种中药成份，超过1.8万种靶点，超过2000种疾病。该大模型采用预训练和微调并结合检索增强生成和插件调用等技术，通过方剂推荐、中药性质解读（包括性味归经、功效与应用、药物组成、炮制方法等）和证候辅助诊断，实现中医药领域知识智能问答、健康咨询、中医药知识图谱动态交互三大核心功能，助力中医药创新研究和人才培养、临床辅助诊疗和中医养生保健，推动中医药文化传承创新发展。

图22. 岐黄问道 · 大模型

(2) 北京博奥晶方大模型开发

北京博奥晶方生物科技有限公司（以下简称“博奥晶方”）系博奥生物在中医药领域布局的产业化平台。博奥晶方通过其核心的“分子本草技术”，构建了“多弹打多靶”的中药组方筛选大模型（900多种中药、300多种食物提取物、10亿级真实基因表达谱数据、药物作用信号通路2500多万条）。博奥晶方首创基于生物芯片技术的中药组方精准筛洗大模型，用数字化技术赋能精准诊疗、中药创新药研发、药食同源健康食品开发、天然植物化妆品开发，致力于为中医药现代化和国际化开创全新的科学发展路径。

(3) 天士力数智本草大模型开发

数智本草大模型是由天士力与华为云在华为盘古大语言模型和盘古药物分子大模型基础上推出。目前数智本草大模型整合了1500+中医药典籍、4000万篇中英文文献、10TB中药以及天然产物现代化研究数据，基于数智本草大模型的数智中药问答以及报告生成平台，通过细分向量库和使用场景、优化向量库、多种检索方式结合，提升了中医药知识问答的精准性，通过药典、文献、指南、医案以及中医药现代化数据库等多种科学证据支持，深度挖掘和整合中药药理、配伍、临床应用等多维度信息，为中药研发、复方设计、药效预测提供科学、高效的信息整合，从而加速中药创新与转化。

数智本草大模型的天然产物分子大模型，是在300万天然产物及衍生物结构基础上微调而成，实现在天然产物的ADMET性质预测、分子生成、分子优化等关键任务上的性能优化，也为中药复方的深入研究和开发提供了进一步的技术支撑。同时，还可以结合天士力开发的星斗云一站式中药研发计算平台，覆盖了从疾病靶点发现、转录组学与蛋白质组学分析，到天然产物分子筛选、方剂推荐及分析等全方位中药研发流程。

同时，用于中药知识问答的中医药大模型数量也在快速增长，例如轩岐问对 · 大模型是“甘草医生”联合浙江中医药大学共同推出了中医药经方领域首个基于大语言模型的人工智能对话系统。轩岐问对是一款类chatGPT的中医垂直领域问答AI产品，其支持中医（经方）领域问题的检索与回复，包含中医基础理论、相关经典古籍、方剂配伍及临床疾病辨证选方等。中医药大语言模型项目（TCMLLM）由北京交通大学计算机与信息技术学院医学智能团队开发开发。TCMLLM拟通过大模型方式实现中医临床辅助诊疗（病证诊断、处方推荐等）中医药知识问答等任务，推动中医知识问答、临床辅助诊疗等领域的快速发展。本项目针对中医临床智能诊疗问题中的处方推荐任务，通过整合真实世界临床病历等数据得到中医处方推荐大模型。仲景中医大语言模型的灵感来自中国古代杰出医家张仲景的智慧。该模型旨在阐明中医博大精深之知识，传承古代智慧与现代技术创新，最终为医学领域提供可信赖和专业的工具。仲景中医大语言模型由复旦大学ROI Lab完成。它综合了人类记忆知识和大语言模型的语言表征能力，训练的主要内容包括患者的病因病机、诊疗方案、随访记录、处方、药物用量、治疗预期结果等。该模型采用特定的prompt模板，初步测试发现模型在妇科以外的中医临床专科领具备一定诊断和处方能力，提高模型对中医方药数据和诊断思维逻辑的推理能力。经过与文心一言、星火等大语言模型的初步对比，发现复旦同济中医大语言模型在基于300条中医方药数据构建的诊疗分解指令数据集上展现出了出色的泛化能力。

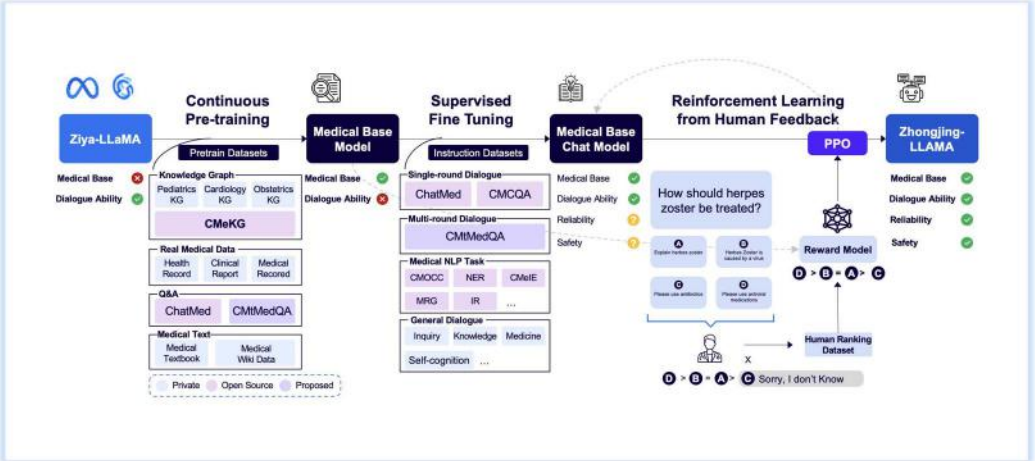


图23. 仲景大模型介绍

2.3临床研究

以GenAI为代表的基于机器学习、深度神经网络和多模态人工智能的应用有望从多个角度优化临床开发，包括筛选临床试验中心、监管合规、药物选择和患者入组、临床研究方案设计和试验报告生成、以及提高药物警戒等方面。以此，从根本上改变医疗事务部和整个生物制药或医疗技术行业管理科学发现、开发和商业化的方式。最近在GenAI的动向总体呈现两大规律，一是提升GenAI技术服务能力，通过技术优势获得大型企业合作方的青睐；二是利用GenAI赋能自身自研管线，以期转型为创新药研发中心或 药企，巩固行业领先优势。

2.3.1监管合规

在临床研究中，监管合规是一个非常重要的课题。研究人员需要遵守各种法规和规定，以确保临床研究的合法性和可靠性。然而，这些法规和规定通常非常复杂，需要花费大量的时间和精力来理解和遵守。同时，不同地区的监管要求也可能存在差异，这使得跨地域合规变得更加困难。传统上，研究人员需要手动处理监管文本，这非常耗时且容易出错。如果能够实现自动化处理监管文本，将大大提高研究人员的效率，并减少错误的发生。最后，不合规可能会导致严重的财务风险，这将对研究机构和企业造成巨大的损失。利用GenAI结合自然语言处理、机器学习、知识图谱构建等先进技术，能够从庞大的法规文本中快速提取与特定目的相关的法规，加速合规进程，并实现自动化处理监管文本，减少对第三方法律和合规支持的依赖。

2023年3月， Medidata 平台发布了包含超过 30,000 项试验与 900 万名患者的 Medidata AI，将患者层面数据直接从试验中的所有病例报告提取，确保临床试验产生合规的数据质量，对数据输入中的错误、异常值、前后不一致和错误报告中的不良事件进行排序和分类，以加快药品审批流程。此外，强生使用的GenAI项目针对新药上市合规审查的需求给予全方面的赋能，通过获取国家药监局、药物审评中心、中国食品药品检定研究所的法律法规、指导原则、相关公告等内容并定期更新，利用大模型的能力进行智能语义检索和细节内容问答，并可通过内置实体模型对法规文档进行主题分类和实体抽取。针对用户对于药品上市合规审查指导原则进行全文问答，并溯源至原文段落、针对用户对于法律法规中段落内容的提问，能通过检索问题相关的文档，可选单篇或多篇进行问答，可准确定位至相应内容，并总结回复、能帮助用户对于国内药品技术指导原则中较为专业复杂的试验设计进行分析总结。

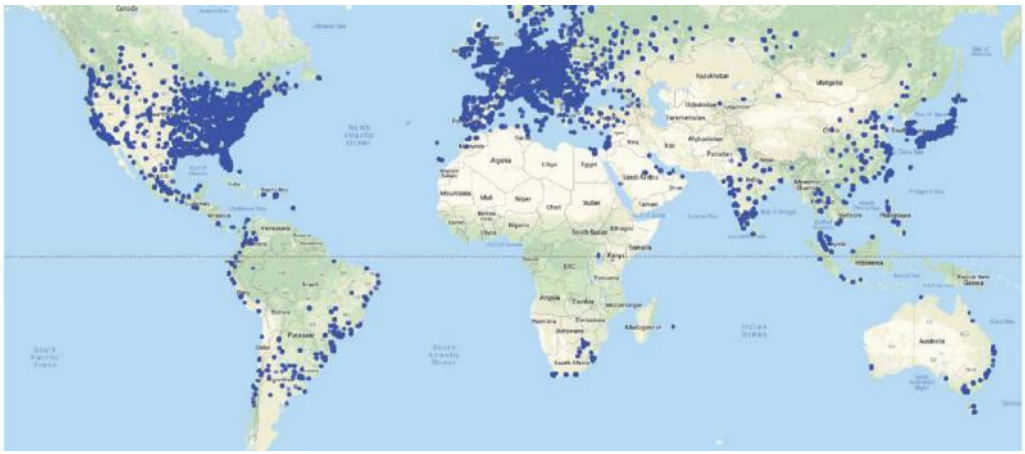


图24. Acorn AI 临床试验中心分析

2.3.2 临床试验中心筛选

临床试验中心的筛选是为了评估其在临床试验方面的资质、专业性、合作意愿和经验等方面，以确保在该中心开展的临床试验具有可靠性和有效性。这一步骤对于临床试验的成功开展至关重要。通过进行可行性研究，可以为项目组提供在该中心开展研究的依据，并提前预判可能会遇到的问题。因此，临床试验中心的筛选和评定是临床试验开展前必不可少的重要环节。生物医药大健康行业在临床试验中心筛选方面存在许多挑战和痛点，其中，信息不对称是一个显著问题，医药企业需要了解每个临床试验中心的实力、经验、设备和人员等方面的信息，但这些信息并不总是公开或易于获取。同时，时间成本高和风险控制难度大也是临床试验中心筛选的挑战，医药企业需要花费大量时间和精力去筛选临床试验中心，如若选择不合适的临床试验中心可能会导致试验失败或者延期，增加项目的风险和成本。

Acorn AI 的 Intelligent Trials 解决方案基于 20000 项临床试验的行业领先数据，提供分析平台，以提高试验的速度、成功率和质量。Intelligent Trials 解决方案助力优化试验设计，选择最优的国家/地区与研究中心，并在启动后确保试验表现良好。GenAI可以利用数据分析、智能决策支持和预测分析等技术手段，为临床试验中心的筛选提供全面支持和优化。

2.3.3 药物选择、患者入组

临床研究中的药物选择是指研究人员根据研究目的和研究设计，选择适当的药物作为研究对象，进行研究。药物选择需要考虑药物的安全性、有效性、剂量、给药途径等因素。而患者入组是指研究人员根据研究设计和入选标准，从符合条件的患者中筛选出符合研究要求的患者，并将其纳入研究中。入选标准包括患者的疾病类型、病情严重程度、年龄、性别、病史等因素，患者入组的目的是确保研究结果的可靠性和有效性。

在临床试验研究中，药物选择和患者入组是两个关键环节，但存在一定的痛点。药物选择需要确保药物的安全性、有效性和适应症范围，这需要大量的前期研究和筛选，耗时耗力。而患者入组方面，痛点主要涵盖招募合适的患者、确保患者符合入组标准、排除患者的合并症和干扰因素等方面。上述痛点可能导致临床试验进度缓慢、成本增加，甚至影响试验结果的准确性和可靠性。

在药物选择方面，GenAI可以通过结合多种数据源和模型，实现从分子到人体多层次的模拟和优化，为药物的选择和评价提供更高的精度和效率。此外，还可通过对药物分子结构的分析，预测药物在人体内的药代动力学和药效学特性，从而为药物选择提供有力支持。英国人工智能公司Benevolent.AI开发的Precision Medicine Platform系统可以通过自然语言处理，从文献、数据库、临床数据等多种来源，从中提取出有用的信息，利用人工智能和机器学习技术进行模式识别和预测，最终选择出最适合治疗特定疾病的药物。在筛选患者入组方面，GenAI能够运用多种数据来源，例如电子病历、基因组学数据和生物标志物等，对患者进行全面评估和分析。这有助于确定最适合参与临床试验的患者群体。此外，GenAI还可以运用机器学习和深度学习等先进技术，对患者的临床特征等进

行深入分析和预测，从而实现对患者的个性化匹配和推荐。在2023年9月，水木分子发布新一代对话式药物研发助手ChatDD (Drug Design) 和全球首个千亿参数多模态生物医药对话大模型ChatDD-FM 100B ChatDD-Trial可辅助临床试验研究人员找到最适合入组的患者人群。通过发现药物敏感的生物标志物，更好地理解疾病亚型，实现精准的患者分类，确保患者与试验药物更匹配，减少不必要的变量干扰，提高临床试验成功率。

图25.ChatDD辅助患者入组

2.3.4 临床研究方案设计和试验报告生成

临床研究方案设计是指在进行临床试验前，制定一份详细的计划，包括研究的目的、研究对象、研究方法、研究过程中的监测和评估等内容。而试验报告生成是指在临床试验结束后，根据试验方案设计的要求，对试验过程中的数据进行整理、分析和总结，撰写一份详细的试验报告，这份报告需要提交给相关的机构进行审批，以便将试验结果应用于后期实践。

在临床研究领域，方案设计和试验报告生成面临诸多挑战。首先，传统的临床试验设计和方案开发过程往往耗时较长，需要研究人员对大量历史数据和文献进行分析和研究，以确定合适的试验设计和终点。其次，由于临床试验设计过程中可能存在不确定性和不完善的地方，研究人员可能需要多次修改方案以达到理想的试验效果。这不仅增加了研究成本，还可能影响试验进度。此外，临床试验设计方法可能无法充分利用历史数据进行预测分析，导致试验结果的预测准确性不足。这可能会影响试验的成功率和研究成果的可靠性。

试验报告生成方面，试验报告需要整合和处理大量的结构化和非结构化数据，如试验结果、患者信息和相关文献，这对研究人员来说是一项具有挑战性的任务。其次，由于数据处理和分析的复杂性，试验报告的质量可能受到影响，如准确性、可读性和一致性等方面。

GenAI基于AI和ML的技术，通过分析过往试验数据来优化临床试验设计来构建主要终点和次要终点情境，设计端到端的临床试验；并利用AI驱动的算法缩短方案开发周期，运用预测分析预测试验结果，降低方案修改次数。此外，GenAI可实现对于历史试验的分析和解释、数据注册表和科学文献中结构化和非结构化数据库，为新的临床试验提供有价值参考。

2023年8月，英矽智能利用其自主研发的基于Transformer的人工智能临床试验预测引擎inClinico，高度准确地预测了多项临床试验II期至III期的转化结果，这项研究成果已发表在《临床药理学与治疗学》期刊上，该期刊是试验与临床医学领域权威的跨学科期刊。ConcertAI与全球性生物制药公司BMS合作，为BMS的肿瘤学临床研究提供首个完全数字化的临床试验解决方案，该解决方案将临床研究和实践整合在一起，支持更轻松的患者识别和试验同意，并有助于IRB批准和临床研究的合同谈判，消除了与数据录入重复和数据监控相关的挑战，并减轻了临床研究人员的负担。

国内本土企业也在探索尝试，上海耀乘健康科技有限公司于2022年发布Prime Create 临床研究方案生成系统以及Prime Construct 临床研究设计和建库系统，支持从临床研究初始即实现关键文档标准化、结构化、数据化。Prime Create 旨在协助医学、生物统计、临床运营等各部门专家高效撰写研究方案，便捷开展团队内及跨部门跨组织协同编辑、审阅、审批和递交工作，充分利用方案知识内容，实现知识留存的数据化、结构化，以协助临床运营相关执行文档和计划、指导文件的生成，并高度自动化对接试验建库工作，达成“方案撰写即建库”，加速从方案撰写到临床研究上线的进程。

2.3.5药物警戒（PV）

临床研究中的药物警戒是指对正在进行的临床试验中的药物进行监测和评估，以确保药物的安全性和有效性。其目的在于及早发现和解决药物的不良反应和安全问题，以保障受试者的安全和权益。

在临床试验过程中，对药物不良反应的监测和报告至关重要。然而，由于医务人员繁忙、知识储备不足或者报告流程复杂等原因，可能导致不良反应的延迟发现或者漏报。这将影响药物安全性评估，增加患者风险；其次，药物警戒涉及多个部门和专业人员，如临床研究者、药品监管部门、伦理委员会等。有效的沟通与合作对于药物安全监测至关重要。然而，在实际操作中，沟通不畅或合作不充分可能导致药物安全问题被忽视或处理不当。GenAI通过运用人工智能技术和先进的数据管理方法，分析包括药物的化学结构、药理学特性、临床试验等数据，并利用人工智能技术来模拟药物的作用机制，识别药物的潜在风险和副作用，并确保药物的安全性和有效性。同时，可将实时收集和整合各方信息，提供全面的药物警戒分析报告，帮助各部门及时识别风险、制定解决方案。目前，Labcorp推出了基于GenAI架构的临床/上市后药物警戒平台、AI舆情平台，产品通过客户的数据

图26. Labcorp 药物开发

收集整理，然后用人工智能计算机集群服务做全球的媒体、文献搜索和训练，提取相关安全信号，进行风险识别，并将安全信息推送给企业；此外，通义行业大模型通过 API 与交互式问答形式提供服务，并提供用于模型二次训练与评测的完整操控平台，与阿斯利康联合完成对药企的应用案例落地。在应对医学领域的学术文献理解方面，针对文献进行特定格式的不良反应信息的识别和总结，生成用于不良反应报告后续处理的内容，提升企业运营效率。

2.4上市及商业化

2.4.1学术推广

医药企业在营销推广方面目前存在着几大痛点。首先，“医药分离”背景下，药品进院及推广都对销售团队以及经销商人员的专业能力提出了更高的要求，医学营销推广需要处理大量的临床研究数据，这些数据往往非常复杂，需要花费大量的时间和精力进行分析和提取。其次，医学营销推广需要对不同市场的文 化、语言、习惯等进行深入了解。此外，医学营销推广还需要考虑隐私和合规性等问题，确保推广活动的合法性和合规性。

首先，在医药企业的销售端，GenAI能帮助企业内部的医药代表和MSL，优化工作效率，降低人工成本，从而在整体上赋能销售增长。柯基数据针对销售端主要面临的内容合规审核慢、以及SOP流程类问题多的痛点，为德国MNC药企市场部门打造了面向销售端的学术推广智能助手。智能助手统一构建和维护销售端知识库，覆盖临床产品、医学和SOP流程指引类知识，并与十多个销售端业务系统打通，以接口形式实时更新知识库；通过基于知识图谱和大模型GraphRAG的技术实现医学素材段落原文问答与溯源，确保学术推广的合规性。在企业微信中，以对话机器人的形式，自动回复代表90%的问题，10%无法解答的以企 业邮件的形式与各平台负责人对接并当日及时回复。通过GenAI工具，以10篇最新文章为例，升级前需要2个月的上线时间，以最新的解决方案，可实时更新发布上线。降低了人工成本80%，提升上线效率90%以上，且由于智能性大大提 升，吸引医药代表和MSL使用并提升整体销售端的使用活跃度60%以上。

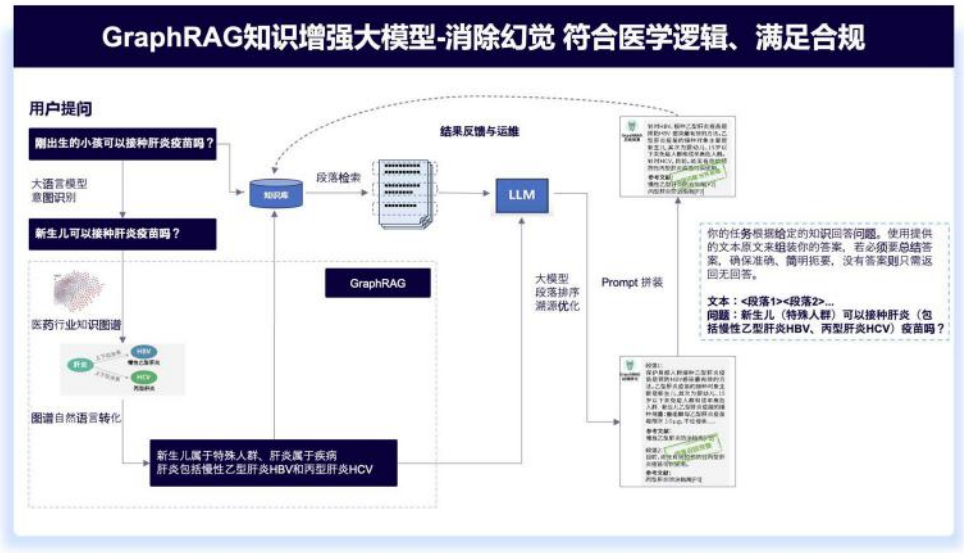


图27. GraphRAG大模型

其次，GenAI在医生端，由于医生面临着医学学术资料数量多，更新快，无法快速有效学习以提升诊疗和科研能力等相关问题。医药企业需要为医生端提供更加智 能的学术推广应用。目前，GenAI能够充分利用大量的医学学术会议内容、训练数据、知识图谱和专家经验，快速准确地回答最新的医学临床研究问题并做总结。美国医药咨询公司 ZoomRx推出了基于GenAI技术的应用程序Ferma GPT。Ferma.AI是ZoomRx开发的一个LLM，它利用了全面的生命科学数据集、精心设计的知识图谱、制药业专用的训练模型以及专业的人类智能和监督。这 种方法使Ferma.AI能够适应 制药业的具体需求，其处理和理解复杂的医学和科学语言的独特能力使其有别于传统模型。Ferma.AI可以通过提供快速、准确的 信息，使繁琐的任务自动化。FermaGPT的AACR应用程序可以梳理所有8230份提交的摘要和研究，以满足特定的请求，如识别NSCLC中的新型KRAS摘要或总结围绕前列腺癌种族差异的关键讨论。除了人工智能生成的一两段回答用户的 问题外，FermaGPT还能够列出原始材料和链接。ZoomRx在2023年4月14日至 19日举行的美国癌症研究协会（AACR）年会推出其生成性人工智能产品的公开版本，专门用于医学会议。2024年，ZoomRx计划在数据和信息发布后继续添 加。ZoomRx计划今年为大多数大型医学会议以及许多小型会议创建FermaGPT公共访问应用程序，包括AAN、ASCO、ESMO、SABCS和ASH。

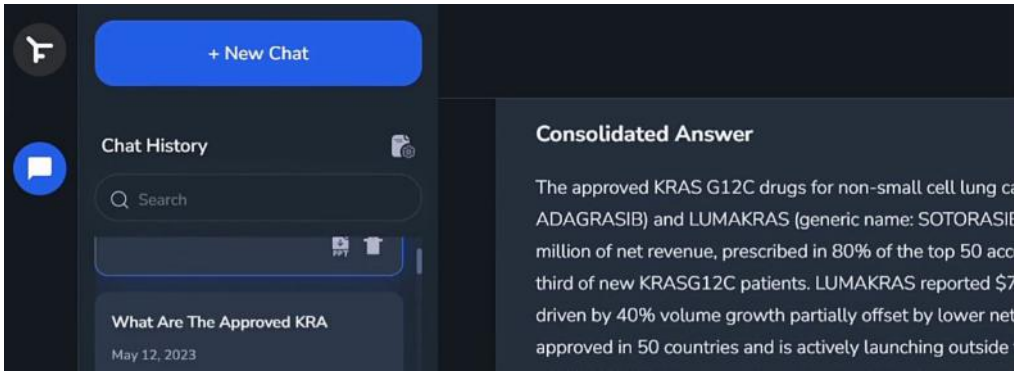


图28. FermaGPT大模型

2.4.3 患者教育

患者教育是指医疗专业人员向患者及其家属提供有关疾病、治疗和预防的信息和指导。它旨在增加患者对自身健康状况的了解，提高其自我管理疾病的能力，并促进良好的健康结果。在患者教育方面，由于医学领域涉及大量的专业术语和复杂的概念，这对患者来说可能难以理解。医生需要确保以简明扼要的方式传达信息，避免使用过于专业化的术语，以便患者能够理解和应用。

目前，GenAI可以针对不同市场的本土化特点，以及通过对目前患者的特点，针对性地生成和构建患教内容，例如图像、内容、数字广告和宣传材料。美国的人工智能工程公司Fractal Analytics提供Avalok GenAI解决方案，可以帮助营销人员创建营销内容、分析竞争情报，并通过个性化答案增强客户体验，同时确保隐私和合规性。

医蝶谷是阿里健康旗下的一款医生个人诊所云平台，专为医生打造，旨在提供便捷、高效、可信的操作平台，以便于为患者提供优质的医疗服务。医蝶谷的GenAI能力可以帮助医生制作科普视频，常规录制一个科普视频可能需要几十分钟，甚至更长。当医生拥有一个数字人模型后，医生无需多次拍摄，只需提交脚本后会自动生成视频。

另外，GenAI可以辅助医药企业搭建面向患者的健康管理用药助手，提供药品说明书相关信息、在线和线下购药渠道咨询等智能问答，及时解决患者遇到的相关问题。同时，GenAI还可以提供慢病智能护理与营养健康知识图谱智能推荐等，帮助患者更高效的获得疾病及药品、营养等相关的知识内容，提高自我健康管理的能力。

2.5 临床疾病诊疗

2.5.1 诊前

在诊前阶段，GenAI可以通过对医学诊疗数据深度学习，分析患者的基因组、生活方式和环境因素等数据，预测患者未来患病的风险，有助于实现疾病早筛，早期干预和预防性治疗，并基于患者情况进行预问诊，提高诊前与诊中链接效率。2023年4月，医联宣布推出基于Transformer架构且针对医疗应用场景调优的大语言模型MedGPT。这一模型的参数高达1000亿，训练所用医学文本数据高达20亿条，临床诊疗数据多达800万条，并由100名医生进行强化调优。在诊前阶段，利用其强大的资料检索和推理能力增强对患者疾病的预测，从而提升分诊导诊的准确性。

图29. 医联MedGPT 一致性得分

除了患病预测外，GenAI可以提高预问诊、导诊准确度和患者信息集成质量通俗易懂地给患者解释病情、提供治疗方案甚至建议生活方式、以及提供预期的结果和风险，这可以提高患者的理解力和参与度，患者能更清楚地了解自己的病情和治疗方案，从而减少不必要的误解和纠纷。DiagnaMed 推出了一款新的生成式人工智能（AI）个人医疗聊天机器人，帮助人们根据自己独特的医疗状况快速获得个性化的医疗信息，并完成从家庭到医院的预问诊和智能导诊。Dr GenAI 还能利用医疗信息（包括生命体征、实验室检测结果和身体特征）对患者进行研究，并将患者信息整合提高患者诊前信息集成能力。

2.5.2 诊中

撰写医疗文书是医疗服务人员日常工作中不可或缺的一部分，但是这项工作的主要问题存在于：首先，医疗文书的撰写需要耗费大量的时间和精力，医生需要花费很多时间来记录患者的病情、治疗方案和医疗记录等信息。其次，由于医疗文书的撰写需要遵循一定的规范和标准，医生需要具备一定的专业知识和技能，否则可能会出现错误或遗漏。此外，由于医疗文书的撰写需要手动完成，存在着一定的人为因素，可能会影响文书的准确性和完整性。

GenAI技术可以通过学习大量的医疗文书数据，自动生成符合规范和标准的医疗文书，从而大大减轻了医生的工作负担。此外，GenAI技术还可以通过语音识别技术，自动将医生的口述转化为文字，进一步提高了医疗文书的撰写效率和准确性。2023年7月，亚马逊推出“HealthScribe”，一项基于GenAI的医疗文档撰写服务工具，帮助医疗服务人员使用语音识别和GenAI技术自动创建医疗记录、成绩单和摘要。撰写医疗文书的痛点包括繁琐的记录流程、临床医生的管理负担以及电子健康记录（EHR）的准确性。采用GenAI可以简化临床记录流程，减轻临床医生的管理负担，并自动创建准确的电子健康记录。

同时，在诊中阶段，GenAI可基于患者的病历、症状和疾病历史等多模态数据，通过数据分析和智能算法为医生提供辅助诊断、指导治疗方案和预后方案。在这条路上，Glass Health发布了Glass AI发布了2.0版本，助力医生实现智能化电子病历的院外的保存和分享；并实现基于LLM+知识库的鉴别诊断DDx和治疗计划的生成。在国内云知声门诊病历生成系统以山海大模型为技术底座，应用前端声音信号处理、智能语音识别等技术，结合庞大的医疗知识图谱，可一键生成符合病历书写规范的标准病历，有效提升门诊效率和病历质量。基于对医疗场景的深刻理解和多年的技术、数据储备，云知声能够精准挖掘医疗场景落地过程中的具体痛点并给出解决方案，致力帮助医生摆脱繁重的文书撰写工作，让医生有更多时间和精力去服务患者，全面提升患者就诊体验。

图30. 云知声门诊病理撰写助手

此外，强生医疗科技公司的Monarch™支气管镜检查平台可以让医生检查传统支气管镜更难以触及的肺部区域，从而有助于早期肺癌诊断。灵活的机器人系统使用术前肺部 CT 扫描来为手术提供信息，但在这种动态环境中实时跟踪物体可能很复杂。 Monarch 研发团队使用人工智能和机器学习算法来开发和完善 Monarch 平台的导航，帮助医生在肺活检过程中引导支气管镜，使他们能够更准确地定位潜在的肿瘤。这有助于更准确的诊断和治疗。当谈到及早发现和诊断疾病时，人工智能可以成为真正的游戏规则改变者。通过将人工智能应用于心电图和超声心动图等常见诊断测试衍生或生成的数据，医疗服务提供者可以更准确地诊断疾病，防止护理延误，并有可能挽救生命

图31. 强生Monarch平台

2.5.3诊后

在诊后阶段，GenAI可以减轻医务人员负担，在线7×24小时回答患者关于病情、药物副作用、预防措施等方面的问题，并能以患者同理心的角度进行互动，实现高质量的诊后随访及慢病管理和护理，提高患者诊后体验及便捷度。2023年 12月，德国医疗科技提供商Zeiss（蔡司）推出一款基于一种概念验证（PoC）GenAI的应用程序，旨在帮助眼科医生及其临床工作人员更轻松、更全面地响应患者的询问。这个应用程序对于患者提供的健康问题、术前后疑问、健康自测情况进行自动回复，基于Zeiss所掌握的患者资料库，确保回复准确和迅速。并且，这个应用程序以个性化、关爱和抱有同理心的方式与患者互动，提供有关手术或技术的准确且预先批准的信息。在实际的应用中，反馈表示GenAI生成的回复中，有很大一部分被认为足够好，无需编辑即可直接发送给患者，展示了其在医疗环境中提供专业答案的有效性。这不仅可以帮助医生更快速、更轻松地响应患者询问，还可以让他们更多地关注患者护理，从而可能增强对视力矫正手术等未来治疗的需求，加强医疗专业人员和患者之间的沟通，最终改善患者的整体体验。

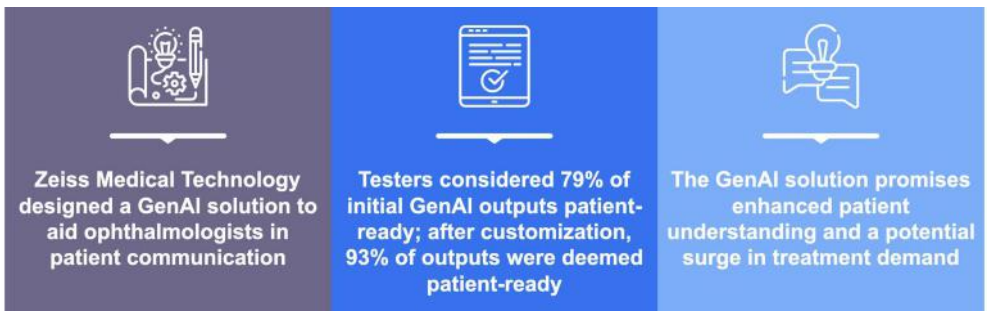


图32. Zeiss 概念验证生成人工智能应用程序

总部位于美国加利福尼亚州的Hippocratic AI成立于2022年，开发了第一个专为医疗保健设计的大语言模型（LLM）。Hippocratic AI的大语言模型优先考虑安全性和准确性，重点关注非诊断AI、面向患者的应用程序，与行业内大多数语言模型不同，公司使用基于证据的医疗保健内容进行训练，以确保生成的内容真实可靠。Hippocratic AI的目标是构建这些专家代理，确保高质量护理的获取不受人员限制和劳动力疲劳的影响，并重新定义医疗保健的护理标准。
Hippocratic AI构建高度专业化的GenAI Agent，可以为患者提供低风险、非诊断服务，通过与患者沟通并收集患者信息（药物剂量、生活习惯、人口统计数据），进行随访，执行涉及临床程序/文书工作的任务，并帮助临床导航。理想情况下，这将提高患者的依从性并降低再入院率，缩减人类护士的日常工作量；另一方面，Agent具有容量无限的优势，可以帮助减少护士倦怠并提高患者满意度，以降低医院成本，达到更好的结果。值得一提的是，这款AI语音护理助手每小时只需9美元，对于多数用户而言均可负担。

Hippocratic AI公司在建立高度专业化Agent方面投入了大量精力，构建了Polaris，一个具有多个专业医疗LLM协同工作的新型架构。该架构参数总计超过一万亿个，建立的AI Agent采用类似人类的对话方法，遵循护理协议，并跟踪其完成所需任务的进度，优势包括可准确地进行医学推理、事实检查和避免幻觉。为了进一步加强安全性，Hippocratic AI 在训练生成式人工智能医疗保健代理时，也采取了适当的人类介入措施，确保在关键时刻由人类护士进行评估和干预，增强整体系统的可靠性和安全性。
在具体操作时，以交互式对话的方式实现与患者沟通。Hippocratic AI架构包括用于语音转录的自动语音识别（ASR）、用于处理文本话语的Polaris和用于音频输出的文本转语音（TTS）。Polaris中包含一个主要LLM代理来驱动对话，以及几个专业的LLM代理，为其提供特定于任务的上下文。整体架构图如下

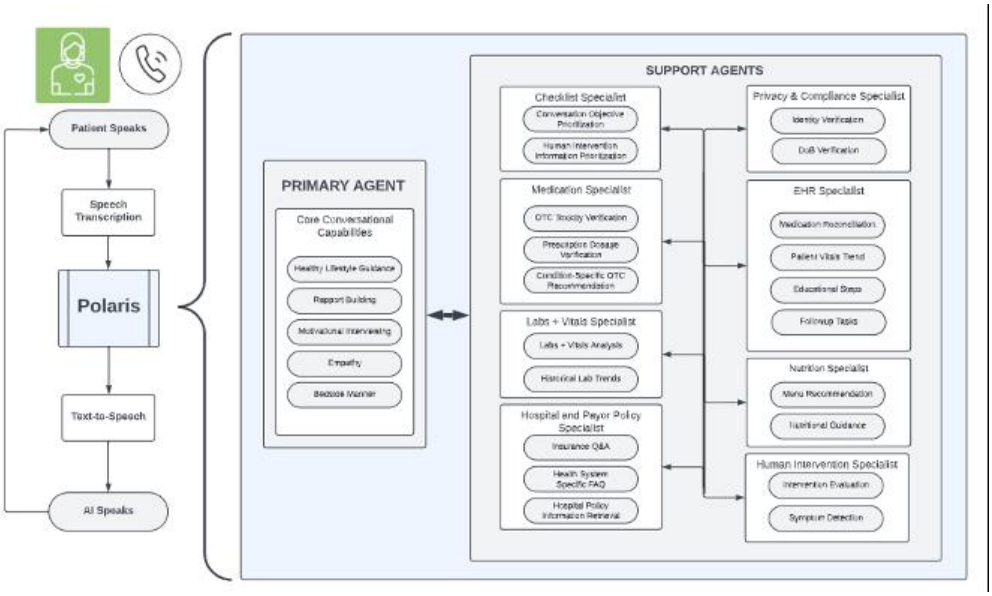


图33. Hippocratic AI应用架构图

与用户交互时，AI代理会首先自我介绍，并告知患者自己是扮演患者主治医师的护理代理。在确认患者身份后，代理将询问患者的最新情况，包括用药情况、日常饮食习惯、症状评估等内容。同时，代理还能根据患者的提问，提供一些医疗建议和解答。最后，代理会对本次对话进行总结。

2.5.4 中医诊疗

(1) 岐黄问道 · 大模型

岐黄问道大模型是由南京大经中医药信息技术有限公司（下文简称：大经中医）开发。大经中医是数智中医行业的创导者和领军者，在名老中医诊疗经验的数字化传承和中医临床智能辅助诊疗系统的开发等领域具有深厚的技术积淀。下设1家互联网医院+门诊部，已建成中医信息化、智能化“软件+硬件”的全产品布局。岐黄问道 · 大模型主要包括医疗和养生两部分，有三个子模型：知病、知症、养生，涉及的数据量包括1100多万条中医知识图谱数据，1500本中医古籍和文献数据，10万份真实中医专家医案数据，10万条脉象、舌象、经络、穴位数据和200万条真实的中医临床诊疗数据，超10万条真实临床环境下的医患对话数据集。主要的用途是根据用户提供的疾病、症状、体征信息，给出辨证（诊断）结果和治疗方案（中药处方），从而给出个性化中医健康状态辨识结果，以及食疗、茶饮、推拿、艾灸等多维度养生方案。

图34. 岐黄问道 · 大模型问答示例

（2）问止中医·大模型

问止中医·大模型是深圳问止中医健康科技有限公司旗下网站。问止科技是全球领先的人工智能中医平台型科技企业，专注于智能中医大脑研究与创新，旗下有智慧中医互联网医院、连锁智能诊所、知识付费、开放大学等业务，主要围绕中医AI打造的“问止AI联盟”旗下有400+家医疗机构、1000+名AI赋能的中医师。问止中医·大模型在中美两地历时十余年研发出的人工智能中医辅助诊疗系统——“中医大脑”，拥有海量名医智慧经验和千万有效数据案例，并不断从临床中实时学习现代最新的诊疗方法，越是在疑难症及重症领域，“中医大脑”领先人类医师的幅度越明显。问止中医· 大模型主要的训练数据主要有教科书、网上中医语料、中医名词、中医学家、基础药草、方剂学、针灸穴位、常见病症、中医试题、中医问诊对话等。问止中医训练和学习海量大数据应用于中医人工智能互联网医院的诊疗开方，创新中医药服务模式。

图35. 问止中医·大模型功能分类及问答示例

（3）华佗·大模型

华佗·大模型由香港中文大学（深圳）和深圳市大数据研究院所在的王本友教授团队训练并开源了一个新的医疗大模型。华佗·大模型使用了四种不同的数据集，包括蒸馏 ChatGPT 指令数据集、真实医生指令数据集、蒸馏 ChatGPT 对话数据集和真实医生对话数据集。通过融合 ChatGPT生成的 “蒸馏数据” 和真实世界医生回复的数据，以使语言模型具备像医生一样的诊断能力和提供有用信息的能力，同时保持对用户流畅的交互和内容的丰富性，对话更加丰富和准确。此外，为进一步提升模型生成的质量，华佗·大模型还应用了基于AI反馈的强化学习技术（RLAIF）。使用 ChatGPT 对模型生成的内容进行评分，考虑内容的用户友好程度，并结合医生的回答作为参考，将医生回复的质量纳入考量。利用 PPO 算法将模型的生成偏好调整到医生和用户之间的一致性，从而增强模型生成丰富、详尽且正确的诊断。

同时，在对患者进行个性护理方面，GenAI可以对患者持续进行跟踪护理，监控用药反应，提高消费者的依从性。GenAI可以通过微信、短信等方式，提醒消费者按时服药、监测他们的症状和副作用、提供自我护理技巧或者将用户的医疗需求与医疗资源联系起来，这可以提高消费者对于药品和服务的依从性和满意度，并帮助他们更好地管理自己的病情。

图36. 华佗·大模型问答示例

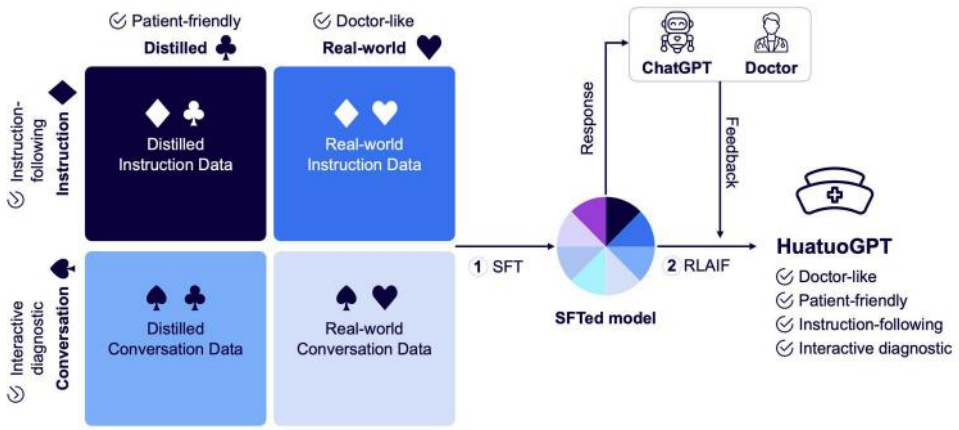


图37. 华佗·大模型介绍

2.6 现状总结

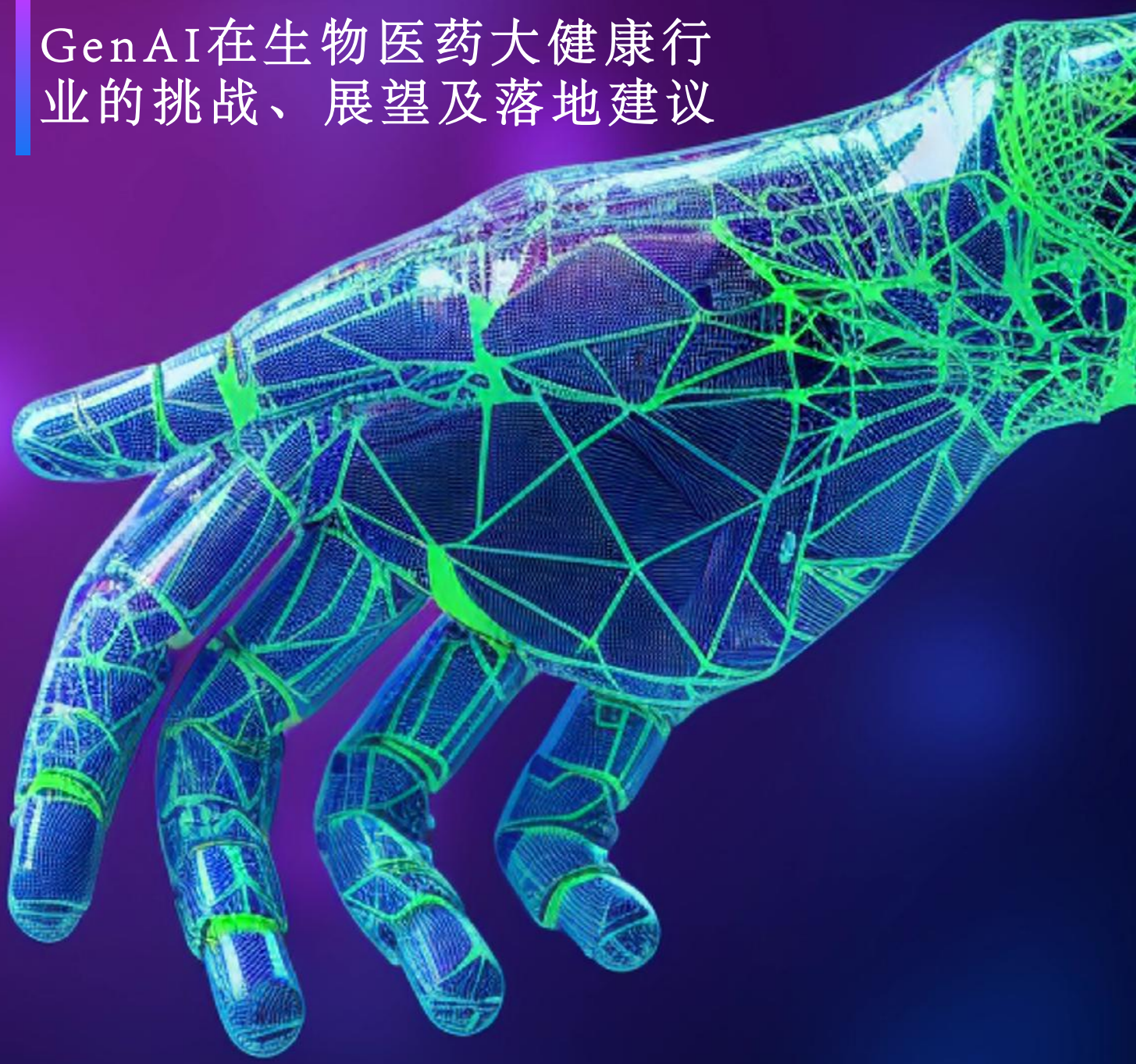
目前，GenAI在生物医药大健康行业的应用已经不局限于自然语言处理领域，包括图片、语音、视频、代码等多种模态的应用开始涌现，而大模型、生成算法与多模态等底层技术的突破成为了其质变的关键。

入局生物医药大健康领域的GenAI参与者不断增多，企业图谱加速扩张。其中包括大型互联网头部企业和科技巨头，如亚马逊和谷歌，率先利用其强大的算力和先进算法，围绕产业链进行横向拓展及纵向渗透；而GenAI技术企业则通常利用自身的GenAI技术优势进入制药场景中的一个或多个环节，通过技术优势获得生物制药企业合作方的青睐。

随着GenAI技术迭代升级及产业链完善，国内外大型生物制药企业逐渐意识到GenAI的技术应用价值，推动了生物医药大健康领域应用场景的不断丰富，目前，这些场景已覆盖了生物医药大健康价值链下全流程的应用，为整个行业带来了前所未有的机遇。

第三章：

GenAI在生物医药大健康行业的挑战、展望及落地建议



3.1 面临挑战

根据前述的分析，国内外生物医药大健康企业已经开始了针对GenAI的落地尝试，而在探索过程中还面临着如下几个比较大的挑战:

3.1.1 数据合规性、符合医学逻辑及循证溯源

技术原理角度来看，GenAI是基于深度学习技术通过概率计算生成答案，不可避免地会有幻觉(“一本正经胡说八道”)问题，而生物医药大健康行业有很强的监管以及合规性的要求，尤其是对外传递的医学内容，需要经过医学和法务的严格审核，也需要提供精准的溯源循证，因此如果直接用ChatGPT之类的GenAI产品，很多场景无法直接满足生物医药大健康行业强合规性要求。

此外生成的答案或者文章还需要符合医学逻辑的要求，需要基于通用GenAI的能力加入循证等级、指南冲突检测和因果关系等医学逻辑。

通常可以采用知识图谱结合符合RAG的方式精准定位原文段落并按照一定医学逻辑组装形成合规的答案。

3.1.2 监管合规性

国内对于GenAI有严格的监管，尤其是面向大众用户(例如患者)的GenAI应用，需要通过监管备案才能正式发布。企业在GenAI项目启动时就需要了解监管的具体要求以及GenAI应用备案的流程及周期和费用，避免由于监管问题导致系统无法上线，前期的工作功亏一篑。

3.1.3 数据安全性及私有化部署

生物医药大健康企业除了使用公开的医学指南和文献等数据，还有很多内部研发和临床试验敏感数据以及市场推广、医学循证材料等数据。ChatGPT等GenAI是一个开放的公有云平台，如果调用其API，需要将部分原文数据通过调用API传给它进行训练及交互，这会涉及到数据安全风险，以及用户交互行为的泄露风险。如果GenAI应用涉及到内部敏感数据，一般需要私有化部署的平台进行对接和权限控制，私有平台采用LLaMA或者ChatGLM之类的开源GenAI框架，同时配备相应的GPU算力资源，成本相比调用API会高不少。如果考虑成本问题，可先从基于OpenAI或者国内大模型开放API调用，处理医学指南和文献开源数据搭建GenAI应用开始进行场景技术验证，再逐步拓展到私有部署处理内部数据。

3.1.4 场景选择和成本

诚如1.2章节总结的，GenAI的应用方式主要有四种，其中像ChatGPT类通用大模型的模型训练和开发成本非常高，一般的生物医药企业是承受不起像这样巨额的初始训练成本以及系统的持续运维成本。因此需要企业需要根据自身情况选择合适的场景和高性价比的应用方式以便更好地评估ROI。业界比较主流的场景和应用方式是在知识库和Chatbot应用RAG检索增强进行升级，通常几十万的成本在几个月就能上线应用。实际落地时可选择刚上市的新产品或者重磅产品和业务部门进行合作，对技术创新的需求和接受度会更高。

3.1.5 内部利益的协同

GenAI在医药大健康行业的落地实施，需要管理层、各业务部门、IT和数字化部门、法律合规部、采购部门的共同协作。

GenAI是颠覆性的AI技术，在企业落地需要管理层从战略方向和高效分配资源提供有力支持。业务部门负责提供业务痛点和明确具体需求，IT和数字化部门则负责建立促进创新的合作伙伴关系，将GenAI技术与现有系统集成，能为业务部门提供可衡量的业务价值，同时确保数据的完整性和安全。由于GenAI的落地面临很大的合规挑战，需要尽早引入法律合规部门，确保遵守多样化的法规，保护伦理考量和患者隐私，减少落地过程中的风险。最后，采购部门需要打破传统软件的采购流程，引入更多有创新力的初创公司。如果是跨国公司，还需要了解Global总部的全球规划并争取总部资源支持。这种跨部门的协同合作，对于发挥GenAI在企业落地至关重要。

3.2 未来展望

未来，GenAI技术在生物医药大健康行业发挥关键作用，将呈现商业化进程提速、应用场景多元化与合规监管增强三大展望趋势：

商业进程稳步提速：过去，虽然GenAI在生物医药大健康行业应用可能性众多，但成功实现商业化落地的仍是核心垂直场景的应用，商业化应用侧重于病理及医学影像的医学诊断领域。随着国内生物医药大健康行业对于GenAI的市场需求逐步明确，企业底层技术的不断打磨。未来生成人工智能技术在生物医药大健康行业商业落地不断熟化，人工智能商业化落地应用将占据主导地位。企业将逐渐明确生物医药大健康行业中商业化复杂性及优先程度，持续挖掘合规且有效的商业模式，不断根据市场所需打磨产品，拓展产业协同布局，实现可持续盈利。

应用场景不断渗透：得益于硬件、算法、数据沉淀等多维度赋能，多模态GenAI技术发展，互联网和科技巨头等软硬件设备及解决方案提供商沿着GenAI医药产业链纵深的活跃度和渗透率将更高，未来GenAI在生物医药大健康行业中的应用将从点状扩展发展至面状渗透。在生物医药大健康企业端，未来将持续深化上市前研发、临床、生产及上市后商业化运营场景中的应用。在医院端，人工智能将从医学影像诊断、病理切片等辅助疾病诊断等场景渗透，逐步从影像科室应用走进多临床科室和基于GenAI面向生物制药企业的真实世界研究。在患者端，智能家庭医生、个性化健康管理和智能护理等应用将不断扩充。整个生态的应用场景也将互相融合，打通“医患药险”闭环。

合规监管愈发完善：伴随着国内生物医药大健康行业对于GenAI应用的普及率及渗透程度不断提高。政府对于GenAI在医疗大健康行业的数据安全问题重视度逐步提升，未来监管合规问题、数据安全问题变得不容忽视。生物医药大健康领域是政府监管之重，目前监管主要面向HCP及患者端的GenAI应用，未来将不断延伸至非患者端的GenAI应用当中。随着更多人工智能领域专门立法及实施细则的出台，未来将逐步形成生物医药大健康行业的体系化治理框架。

3.3 落地建议

当前GenAI在生物医药大健康行业的探索尚属早期阶段，平台层面初露头角，企业应用总体处于萌芽阶段。但不可否认的是，GenAI整合了生物学、化学、计算科学、药理学和疾病治疗等领域内容，加上生物医药大健康行业的产业链条长、参与主体多，应用空间潜力巨大。在看到机遇的同时，我们也应充分关注到，其作为新兴事物，在发展过程中不可避免会遇到一些挑战，有些甚至是前所未有的困难。因此，解锁GenAI潜力、并成功加以应用，均离不开前瞻性的考量、详细的落地路径规划、以及保障措施等全方位的考虑。

3.3.1 捕捉变化，动态调整

GenAI虽然潜力巨大，但在GenAI应用试点及推广的过程中，企业面临着一系列风险，比如知识产权、网络安全、数据隐私、数据偏差、以及错误结果等。企业应制定有效的风险管控机制，设计治理框架和相关规则，建立用于监控和管理GenAI风险的工具，并将相关政策和程序必须融入企业的文化和运营模式之中，引导GenAI处理好涉及道德、法律和技术方面的问题。随着监管环境的跟进，企业内部的指导方针也要追踪及基于监管机构最新的政策框架监测进行灵活更新，保障合规体系下的动态调整能力。

3.3.2 顶层设计，数智思维

将GenAI纳入数字化进程是工作的先决。管理层应在思想层面认识到GenAI是一个数字化策略的必选项，给予足够重视，并倾注相应的资源。GenAI可以通过自动化流程降低企业成本、通过分析大量数据来协助决策制定、通过个性化内容改善与客户的互动，并在促进创新、风险管理、质量控制和人力优化等方面有所作为，所有这些都将对企业的价值创造起到积极作用。

目前，GenAI用例已开始广泛涌现，不具备相关能力的企业将在未来丧失竞争优势。在某种程度上，各玩家起点差距不大，投入和技术门槛并非高不可攀，及时采取行动能让企业在这场革命中把握主动权，因此通过有效顶层设计，可有效保障企业在技术更迭背景下快速把握技术先机。未来企业可直接接入GenAI通用大模型能力，融合内外部数据搭建企业级的Copilot平台，可对多源异构的文档、数据库、知识图谱以及多模态图片、视频和音频等数据进行自动标签和内容生产、问答及写作、总结，未来Copilot平台将通过AI Agent智能代理进行用户意图理解并进行任务分发，例如有传统的FAQ更精准的匹配问答，从结构化的数据库和知识图谱动态生成组合答案以及分析，基于海量原始文档通过GraphRAG引擎定位原文并按照医学逻辑组装答案，基于GenAI的全库内容的总结和报告生成等，最终形成人机协同的企业级Copilot平台。

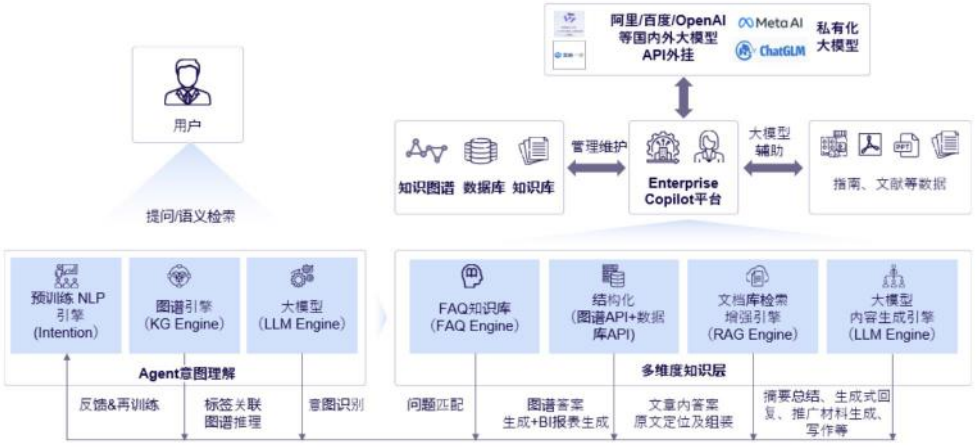


图38. GenAI顶层设计

3.3.3 目标锚定，小步快走

虽然GenAI在生物医药大健康行业落地面临很大的挑战，但只要确定目标，小步快走、实现快速迭代便可在满足合规性要求的情况下找好能带来实际价值的应用及制定好相应的技术路线，在生物医药大健康行业的应用潜力还是巨大的。
总体而言，GenAI在生物医药大健康行业潜在应用广泛，企业可选应用多，覆盖了从药物研发、临床研究、营销与临床诊疗等整条价值链。通过分析及筛选，企业可明确GenAI技术可在自身药物价值链活动的哪些环节获得最大的竞争优势，详细评估GenAI对其现有产品和服务组合的影响，分析目标用户群体特征。归纳其自身产品和服务而言最具价值潜力GenAI用例后，还需要量化相关应用的影响力和落地速度，同时兼顾模型、数据可用性、数据安全性、容错性、复杂性、可负担性和市场需求等因素。对用例的优先级进行排序，确定GenAI优先用例后，开展试点工作，收集试点工作反馈并进行多次优化，在用例影响力得到验证后，择机在整条价值链上进行推广，以便在全组织层面充分释放其全部价值。

应用场景可以首先从内部效率提升开始，可借助GenAI能力在内容建设层面探索通过GenAI来优化医学人员进行数据打标签和标注以及生产FAQ的基础工作；其次还可结合知识图谱可解释和可溯源的优势，通过通用GenAI的开放API，将生物医药大健康行业合规的开放权威数据(例如医学指南和文献、会议纪要报道、临床研究等)结合自己的业务需求进行RAG检索增强训练，能做文献智能阅读和写作以及chatbot应用并提供溯源循证的能力，满足合规性的要求情况下提升应用的智能化程度和用户体验，通过GenAI实时生成合规内容来减少生成FAQ及审核的流程。

总体而言，GenAI在生物医药大健康潜在应用广泛，企业可 锚定自身所需，通过小步快走先试先行，逐步挖掘的GenAI的应用潜力，建立人机协同的多层智能知识平台。

图39. 企业GenAI知识平台

3.3.4能力构建，组织提质

在企业能力层面，基于通用GenAI平台和开源的大模型，在生物医药大健康行业数据集上进行训练调优，构建私有化部署的企业级GenAI平台，结合知识图谱的能力，保证数据安全性的情况下处理内部数据，并可和外部数据进行融合进行合规数据的内容生成。同时基于Chatbot智能交互平台为内部员工和HCP、患者提供满足合规安全性要求的企业级智能交互方式。

在组织结构层面，为了支持全新的工作方式，可能涉及到自身的组织结构和运营模式的调整，企业需要制定实用的变革管理计划，指导组织平稳过渡转型。明确各方所应承担的责任、拥有的决策权，并重新设计职位信息，配备相应人员。企业需要更新其人才和技能培训策略，包括研发、临床研究人员、制造与质量控制人员、市场销售和医学团队、职能部门、法律合规部门员工等；确保相关员工了解如何在日常工作中使用 GenAI，如撰写提示词；掌握部署高级应用的能力，如增加GenAI专业人员招聘等。

3.3.5合作共行，优势互补

完善的基础设施和平台是GenAI发挥充分潜力的前提。目前得益于硬件、算法、数据沉淀等多维度赋能，互联网和科技巨头等软硬件设备及解决方案提供商基于其多年的科技投入和人才积累，开发了丰富的应用方案，这也是现行主流的落地模式，合作双方可以优势互补，快速部署。企业在落地云基础设施、数据平台、模型和应用时，可从使用成本、响应速度和定制化水平、潜在的收益等方面评估各类方案，确定最终合作伙伴。由于GenAI技术发展迅猛，在此背景之下，企业应着力提高自身的灵活性，研判自身数据平台和核心系统的准备情况，根据GenAI的变化不断进行调整。企业可先考虑一些风险较低的合作伙伴，这样不仅可以帮助企业更好地对自己的平台进行评估，还可以让企业在扩充自身GenAI能力的过程中不断积累经验教训。为了使GenAI应用效果最大化，企业还需要考虑与现有的AI工具相结合，例如基于现有的Chatbot工具进行GenAI的升级。