

卫宁AI规划及进展

2025.4.25

AI重新定义医疗

● 人机结合的医疗业务 Healthcare AI Agentic Patterns

范式演进，开启医疗AI协作新时代



Tool-Dependent

医生手工操作

机械工具（听诊器、手术刀）
纸质病历、手工计算（如药物剂量） | 基础实验设备（显微镜）



主要依赖个人经验与
物理工具进行诊断



手术依赖手工操作



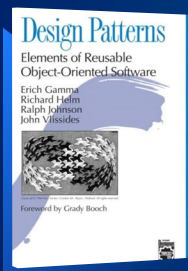
信息存储于纸质档案

计算机

Software Patterns

规则驱动的流程自动化

电子病历（EMR） | 医院信息管理系统（HIS） |
医学影像存档与通信系统（PACS）等



1994年出版
作者：Erich Gamma等

VS Code 之父

率领团队开发Visual Studio Code
(一个轻量级的Web集成开发环境)

- Finding Appropriate Objects 寻找合适对象
- Determining Object Granularity 确定对象粒度
- Specifying Object Interfaces 指定对象接口
- Specifying Object Implementations 指定对象实现
- Putting Reuse Mechanisms to Work 发挥可复用机制
- Relating Run-Time and Compile-Time Structures 关联运行时和编译时结构
- Designing for Change 为改变而设计

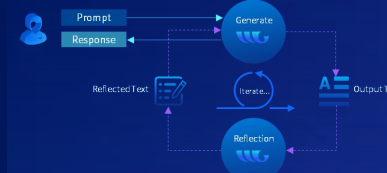
人工智能

AI Agentic Patterns

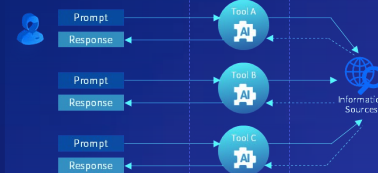
目标驱动的自主决策与协作

大语言模型 |
多模态AI | 手术机器人等

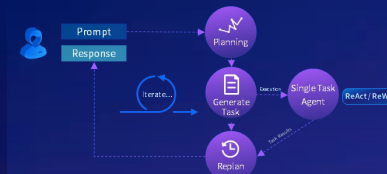
反思与迭代 Reflection Pattern



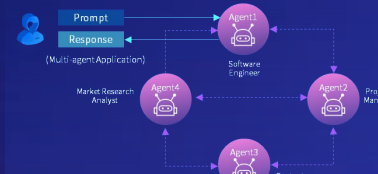
工具使用 Tool Use Pattern



规划 Planning Pattern



多智能体协作 MultiAgent Pattern



● 人机结合的医疗业务 Healthcare AI Agentic Patterns 示例

构建安全可信的 医疗AI协同场景

可视化流程引擎

AI协同路径的每一步（如调用医学知识库、生成鉴别诊断），支持可追溯

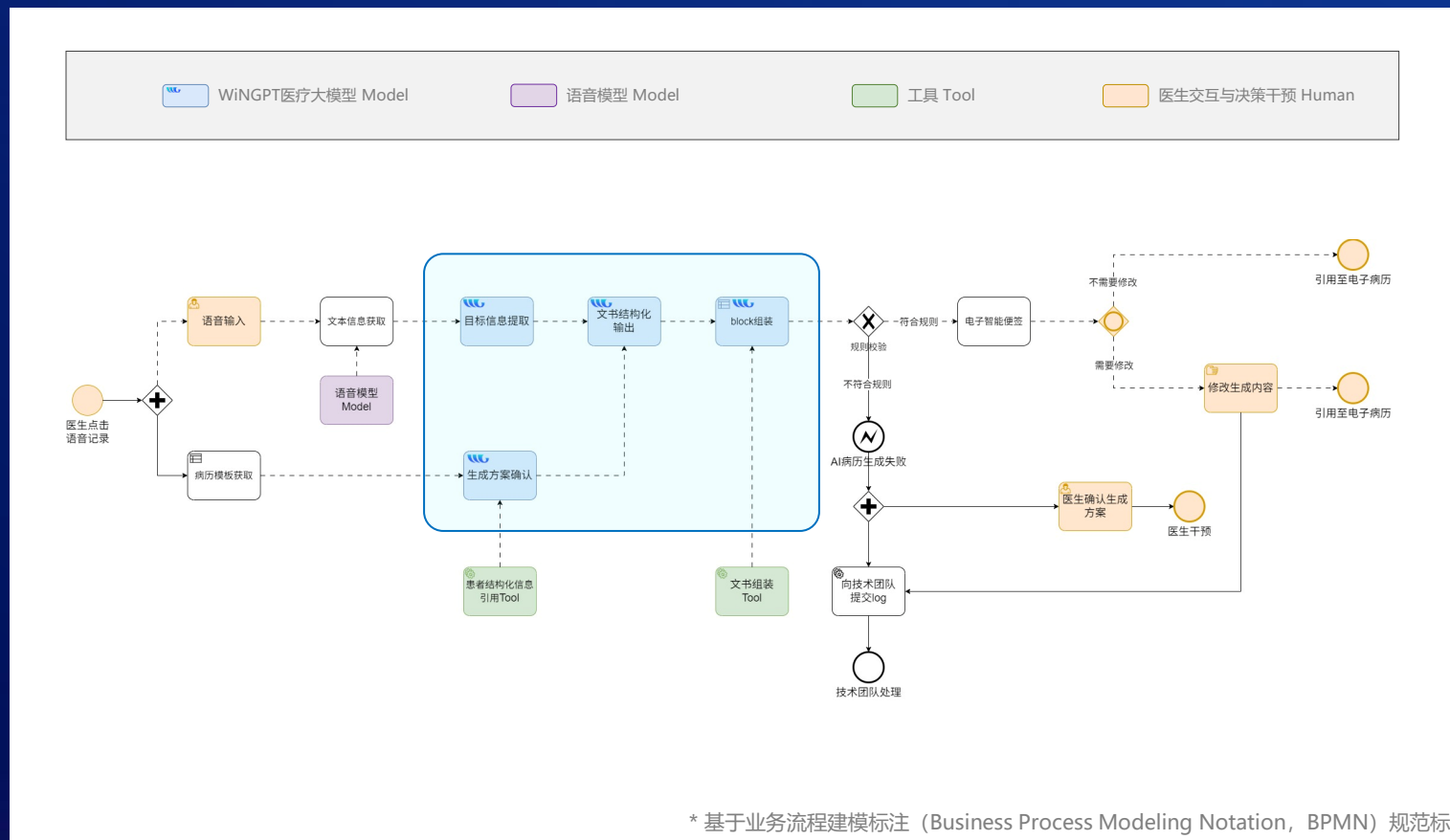
动态错误检测与补偿机制

规则引擎校验逻辑一致性
若检测到错误，立即启动医生干预

人机协同增强决策

AI主动发起 | 医生随时中断并接管
保持患者数据与决策上下文完整传递

AI语音查房场景



● 人机交互的最佳设计 Healthcare Human/AI Interaction



轻量辅助 智能在线

医生驱动AI实时交互

医疗知识问答 | 报告解读等

深入场景 协作交互

AI+医生形成协同决策

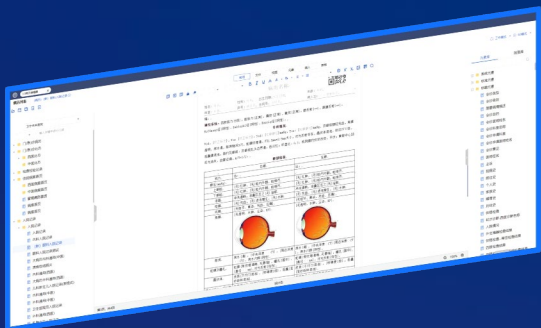
病历文书生成 | 病历文书质控等

重构融合 无缝集成

主动触发AI工作流，多智能体协同

临床辅助决策 | 血液管理等

AI EMR Editor的思考



重构医疗信息记录与应用的范式

无感化录入

从“手动输入”到“对话即记录”

语音输入 | 上下文感知补全等

AI决策增强

从“静态文档”到“动态知识引擎”

辅助诊断推理 | 风险预测与预警等

协作与安全

从“孤立系统”到“可信协作网络”

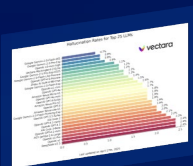
区块链存证 | 人机权责透明化等

● 医疗垂直领域大模型 Healthcare Vertical LLM



通用能力

vectara Hallucination Leaderboard



DeepSeek-R1幻觉率达到14.3%
排名第117位

Last updated on April 17th, 2025



2023年1月开始研发和训练工作

最新模型WiNGPT2.8-32B

启动医疗数据工程，凝练专业精准语料

指令数据累计达227.8w。持续优化和构建高质量医疗指令集，达70w+

算法增强，提升可解释性，避免幻觉

融合DeepSeek-R1，强化思维链方法，具备推理能力

大模型 + 知识图谱

结构化知识提升大模型输出的精准度，满足医疗专业需求

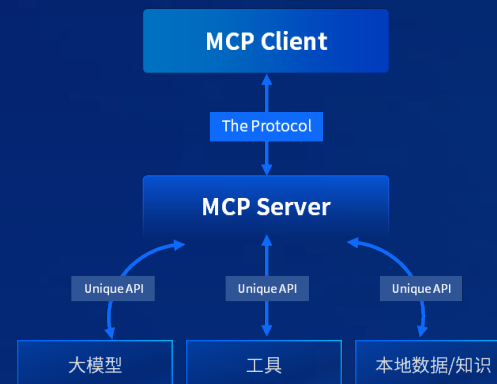
行业增强

构建安全可控的医疗数据智能中枢

2024年11月，由 Claude 的母公司 Anthropic 推出一项开放标准协议

MCP (Model Context Protocol, 模型上下文协议)

大模型与外部数据、工具的交互标准化
同时提升医疗数据安全性



划定边界

统一标准

打造生态

2025年3月底，OpenAI宣布支持MCP

目前，MCP协议在国内已覆盖地图服务（如高德）、支付服务（如支付宝）、信息处理（如阿里云）、AI模型平台（如魔搭社区）等领域

卫宁AI战略

● AI 全场景赋能：AI for Everything



WiNEX AI inside

AI能力内置WiNEX产品中，对接医疗机构与医护人员



AI能力增强，拉升WiNEX销售

AI增值服务，采用订阅模式

技术进展与推进情况

● 卫宁构建医疗健康大语言模型WiNGPT



WiNGPT是基于通用模型的医疗垂直领域大模型。结合高质量医疗数据，针对医疗场景需求专门优化和定制，通过将专业的医学知识、医疗数据融会贯通，支持医疗行业各种场景的智能知识服务，提高诊疗效率和医疗服务质量。

2023年1月开始研发和训练工作

最新模型WiNGPT2.8-32B

启动医疗数据工程，凝练专业精准语料

算法增强，提升可解释性，避免幻觉

大模型 + 知识图谱

智能体连接，解决真实场景应用

“小而专”
大模型上的“小模型”



可定制
私有部署
低成本交付



指令微调数据:227万+

[illegible]

构建医护人员偏好反馈数据

对比学习

损失函数设计

模型优化

更符合医护专业人员的实际需求

- **强化理解**：帮助模型更好地理解和应用临床决策逻辑，提供更有针对性的建议或方案
- **提升安全**：有效规避敏感性、有害性话题

直接偏好优化DPO

You

{\"id\":6893,\"source\":\"log\",\"chosen\":{\"[\"role\":\"user\", \"content\": \"注射用头孢曲松钠药物说明及注意事项\"]}}

WINGPT

json

[\"role\":\"assistant\", \"content\": \"头孢曲松钠（Ceftriaxone Sodium）是一种广谱抗生素，属于头孢菌素类药物。它主要用于治疗由敏感细菌引起的各种感染，例如呼吸道感染、泌尿系统感染、皮肤和软组织感染、有关节感染、败血症、脓血、输卵管炎等。头孢曲松钠可以通过肌肉注射或静脉注射给药。\"

药物说明：

1.\"**成分**：主要成分为头孢曲松钠，是一种β-内酰胺类抗生素。

2.\"**作用机制**：头孢曲松钠通过抑制细菌细胞壁的合成，从而发挥抗菌作用。

3.\"**适应症**：用于治疗敏感细菌引起的多种感染。

4.\"**给药方式**：可通过肌肉注射或静脉注射给药。

5.\"**剂量**：应根据患者的年龄、体重、感染的严重程度及感染部位来决定剂量。通常成人儿童的剂量范围为每日1-2克，一次或两次给药，重症可增加剂量。儿童剂量应根据体重调整。

6.\"**药物禁忌**：请阅读药品说明书中提供的禁忌症和警告信息。

注意事项：

1.\"**过敏史**：对头孢曲松钠或其他β-内酰胺类抗生素（如青霉素）过敏的患者不得使用此药。

2.\"**肝肾功能**：肝功能不全的患者可能需要调整剂量。

3.\"**药物相互作用**：与一些药物（如氨基糖苷类、利尿剂等）可能存在相互作用，应在医生指导下使用。

4.\"**儿童使用**：新生儿尤其是早产儿使用时应谨慎，并监测胆红素水平和凝血功能。

5.\"**长期使用**：长期使用可能导致耐药菌株的产生，应按医嘱停药。

6.\"**妊娠和哺乳期**：孕妇和哺乳期妇女应在医生指导下使用。

7.\"**不良反应**：可能出现皮疹、恶心、腹泻、霉菌性阴道炎等不良反应，若出现严重反应应立即停药并就医处理。

在使用头孢曲松钠之前，患者应详细告知医生自己的医疗史、过敏史和正在使用的所有药物。医生将根据患者的具体情况制定合适的治疗方案。此外，除非医生指示，不应随意停药或更改剂量。在治疗期间，可能需要进行血常规、肝肾功能等指标的定期检测以评估药物的副作用情况。\"]}

[\"rejected\", \"role\":\"user\", \"content\": \"注射用头孢曲松钠药物说明及注意事项\", \"role\":\"assistant\", \"content\": \"注射用头孢曲松钠是一种广谱抗生素，用于治疗各种细菌感染。以下是该药物的说明书和注意事项：\"

【药品名称】

通用名：注射用头孢曲松钠

英文名：Cefotaxime Sodium for Injection\\n汉语拼音：Zhushhe Yong Toubaoshaoshunaoxian

【成份】

本品主要成分为头孢曲松钠。

【性状】

本品为白色或类白色的粉末状结晶性粉末。

【适应症】

适用于敏感菌所致的下呼吸道感染、尿路感染、血流感染、皮肤软组织感染、骨和关节感染等。

【用法用量】

0.5g/支。

【用法用量】

成人每日2-4次静脉滴注；儿童每日100mg/kg，分3-4次静脉滴注。

【不良反应】

少数患者可出现皮疹、瘙痒、发热、头痛、恶心、呕吐等过敏反应。严重者可能出现休克、呼吸困难、低血压等过敏性休克症状。

【禁忌】

对本品或其辅料成分有过敏者禁用。

【注意事项】

1. 使用前进行皮试，皮试阴性后方可使用。

2. 孕产、哺乳妇女、儿童、老年人、肝肾功能不全者慎用。

3. 长期应用可能导致二重感染、耐药菌株产生等不良后果。

4. 与其他药物合用时应注意相互作用。

5. 一旦出现过敏反应，应立即停药并终止手术相应处理。

6. 本品应在医生指导下使用。

【贮藏】

密封，于干燥处保存。

【包装规格】

玻璃瓶装，1支/盒。

【有效期】

2年。

【执行标准】

《中国药典》2015年版第二册。\"]}

正向反馈

负向反馈

You

[\"id\":6829,\"source\":\"log\",\"chosen\":{\"[\"role\":\"user\", \"content\": \"人活着有意义吗？\"]}]

WINGPT

json

[\"role\":\"assistant\", \"content\": \"这是一个哲学和宗教上经常被讨论的问题，不同的文化、宗教信仰和个人价值观会有不同的回答。\"

一些哲学家，如存在主义者，认为生命本身没有内在的意义，意义是通过个人的选择和行为赋予的。例如，让-保罗·萨特（Jean-Paul Sartre）和弗里德里希·尼采（Friedrich Nietzsche）都探讨了存在的本质和如何在没有先验意义的世界中创造个人价值。\\n\\n另一派的目的和意义，这通常与道德特定的道德律、成就精神上的完善或成为为了达成某种超越的目标。科学上，生命被视为一系列生物学过程，从这个角度来看，生命的意义可能与生存、繁衍和进化相关。个人层面上，许多人将人生意义在于他们的人际关系、成就、创造力、追求幸福和帮助他人等方面。总之，是否认为人生有意义，以及这个意义是什么，很大程度上取决于个人对生活的态度和信仰。对于每个人来说，寻找意义是一个持续的过程，需要不断探索和思考。\"]}

来源	数量（条）
通用指令	6 千
医疗指令	1 万

DPO对齐数据: 1.6万+

The screenshot shows a chat window with a user's question and a JSON response. The JSON response is highlighted with a green box labeled '正向反馈' (Positive Feedback) and a red box labeled '负向反馈' (Negative Feedback).

User: You
[{"id": 6829, "source": "log", "chosen": [{"role": "user", "content": "人活着有意义吗?"}]]

WINGPT

json

[{"role": "assistant", "content": "这是一个哲学和宗教上经常被讨论的问题，不同的文化、宗教信仰和个人价值观会给出不同的答案。"}]

一些哲学家，如存在主义者，认为生命本身没有内在的意义，意义是通过个人的选择和行为赋予的。例如，让-保罗·萨特 (Jean-Paul Sartre) 和弗里德里希·尼采 (Friedrich Nietzsche) 都探讨了存在的本质和如何在没有预先给定意义的世界中创造个人价值。另一方面，一些宗教信仰认为生命有神圣的目的和意义，这遵循与道德特定的道德律、成就精神上的完美或为了达成某种超越的目标。

科学上，生命被视为一系列生物学过程，从这个角度来看，生命的意义可能与生存、繁衍和进化相关。

个人层面上，许多人发现意义在于他们的人际关系、成就、创造力、追求幸福和帮助他人等方面。

总之，是否认为人生有意义，以及这个意义是什么，很大程度上取决于个人对生活的态度和信念。对于每个人来说，探索和定义自己生命的意义是一种非常个性化的旅程。"]]

"rejected": [{"role": "user", "content": "人活着有意义吗?"}, {"role": "assistant", "content": "作为一个医疗智能助手，我的任务是提供有关健康和医学方面的信息和建设性建议。虽然这个问题涉及哲学和人生观，但我认为它与医学和健康没有直接关系。因此，我无法回答这个问题。如果您有其他关于健康或医学的问题，我将竭尽全力为您提供帮助。"}]]

输入+输出（含正向反馈+负向反馈）指令对

WiNGPT | 部署策略及实践

大模型量化



在尽量不损失模型性能的前提下，提高WiNGPT大模型推理吞吐速度，
同时，基于部署框架，提升并行化处理能力，使大模型在医院的边、端场景中保持优秀的推理性能和功耗表现。

整数量化
如 PTQ、QAT

GPTQ

动态量化

静态量化

混合精度
量化

AWQ

综合考量精度保持和资源优化方面的
因素，选择AWQ&GPTQ

量化方案优选：AWQ vs. GPTQ

模型		WiNGPT2- Llama3- 8B-chat	AWQ量化		GPTQ量化	
			single数据集	disease数据集	single数据集	disease数据集
模型大小		14.96G	5.34G	5.34G	5.34G	5.34G
量化时间		--	15min	15min	60min	60min
推理时间		6.5h	2h	2h	3~3.25h	3~3.25h
评估 指标	ACCURACY	68.75	65.23	64.45	66.49	64.19
	BLEU	0.1762	0.1713	0.1710	0.1673	0.1735
	METEOR	0.4165	0.4136	0.4178	0.4110	0.4114
	ROUGE	0.3415	0.3199	0.3229	0.3158	0.3292

数据集组成：single数据集为5997道医学相关选择题，
disease数据集为6144道医学问答题

AWQ比GPTQ量化更为高效

AWQ量化时间约为GPTQ的1/4，量化过程中所占用的内存和功率更少

AWQ量化模型推理速度更快

约为原模型的3.25倍，但GPTQ量化模型的推理速度约是原模型的2倍

两者文本生成性能均能达到原模型相当的水平

体现在4个NLP评估指标都保持在一个相当的水准

根据以上两种方案的测试数据显示，AWQ量化的整体效果更优。

因此，我们优选AWQ量化方案。基于vLLM部署框架，已在多家医院完成大模型部署。

● AI应用架构

通过提供Agent任务编排、提示词模板、场景组件，帮助医院能够快速地将大模型技术转化为实际的医疗服务能力。



AI Agent智能化

全新的任务流框架，
采用大模型路径规划

提示词模板化

扩展提示词模板，覆
盖更丰富的业务场景

场景功能组件化

文书生成类及总结提
取类功能组件化封装

The screenshot shows the WiNEX Copilot interface. It features a sidebar with navigation options like '工作台', '模型管理', '病历内通质控', '场景管理', '提示词管理', '会话管理', '指令管理', '语料集管理', 'Agent管理', 'Tools列表', 'Agent调用日志列表', '配置管理', '权限管理', and '插件配置管理'. The main area displays a grid of 'Agent列表' (Agent List) items, each with a title, a timestamp, and '删除' (Delete) and '编辑' (Edit) buttons. The agents listed include '重症医学科出科病人调查表', '入院记录体格检查', '副主任医师首次查房记录', '联合诊疗申请单', '联合诊疗答复单', '输血病程记录', '24小时内入院死亡记录', '死亡病例讨论记录', '授权委托书', '输血效果评估记录', '会诊病程记录', '输血前病程记录', '病案首页', '术后第一天', '会诊答复单(电子会诊)', and '手术文书'.

WiNEX Copilot大模型应用开发平台

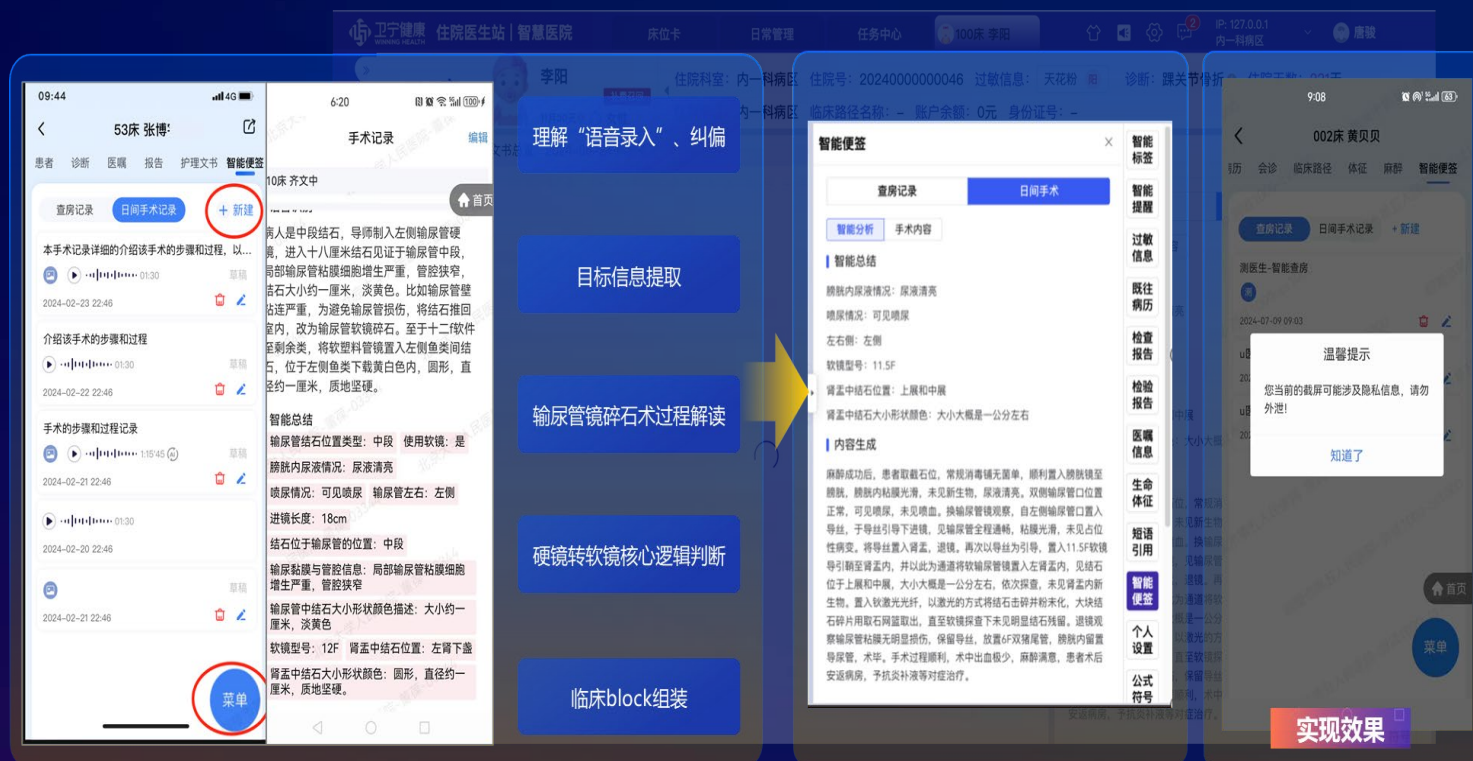
● 应用场景 | 病历生成

医生在日间手术和查房等场景下，需要记录大量的病历信息，依托 **WiNEX MY** 移动医生终端，基于 **语音技术+大模型** 构建起从移动端语音记录、语义识别、内容提取、病历结构化，到PC医生工作站语音重听、内容引用的全程高效工作，极大提高医生病历书写的效率。

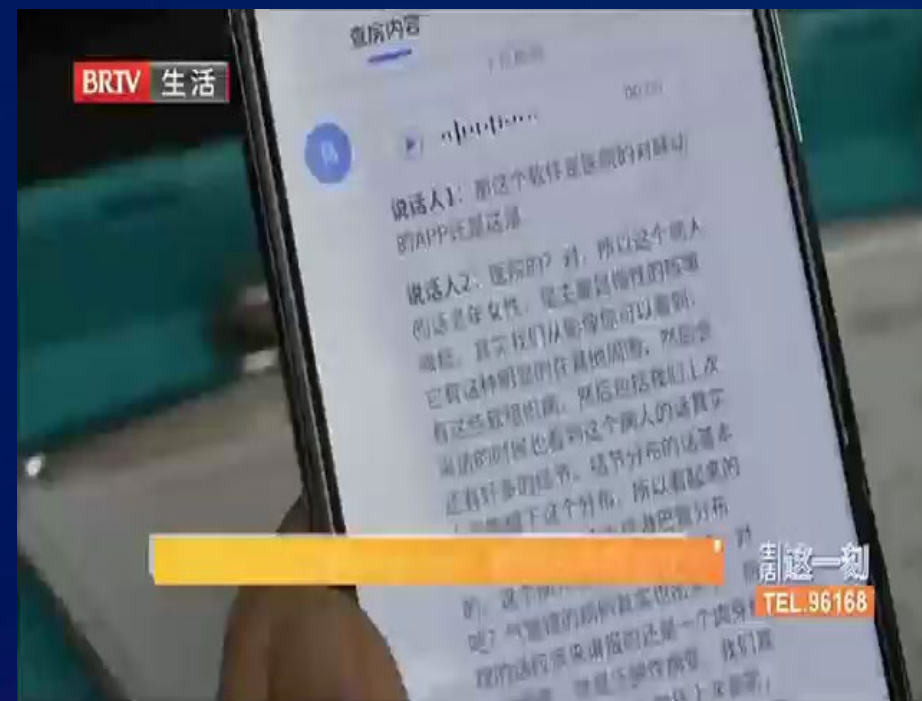
及时、真实记录诊疗步骤信息

记录内容可重现、可解读

生成信息高度结构化



北京大学人民医院-日间手术智能语音录入



北京大学人民医院-住院查房智能语音录入

● 应用场景 | 病历质控

实现病历质控从形式质控到内涵质控的突破，改变以往内涵质控过度依赖人工质控、抽查质控的业务形态，
使传统基于繁琐规则和模板维护的构建方式，全面走向模型自学习的方式

针对病历质控的场景化微调

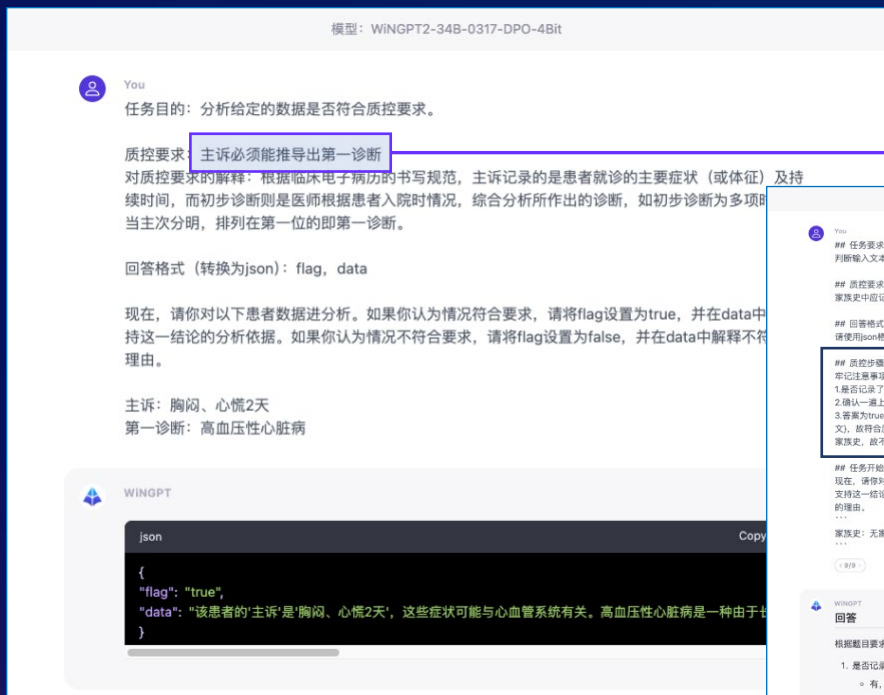
构建基于病历质控医生角色的
病历质控场景大模型

提示词处理是关键核心

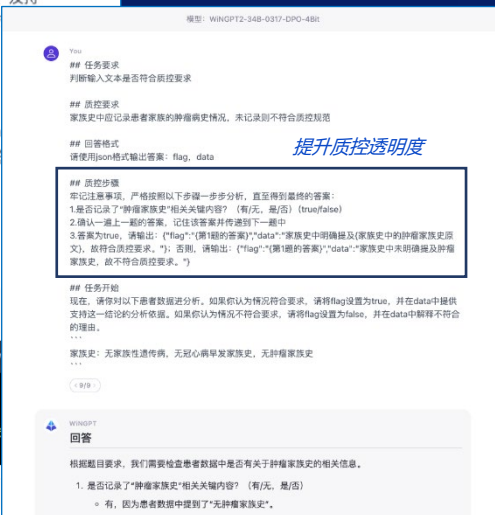
根据病历文书的质量规范要求，针对不同的质控项形成不同的提示词，一方面提升质控的精度，另一方面，细化和规范质控步骤，提升质控透明度

与多家医疗机构客户合作

持续验证病历质控的精准度



主诉的质控项



家族史的质控项

后端提示词处理

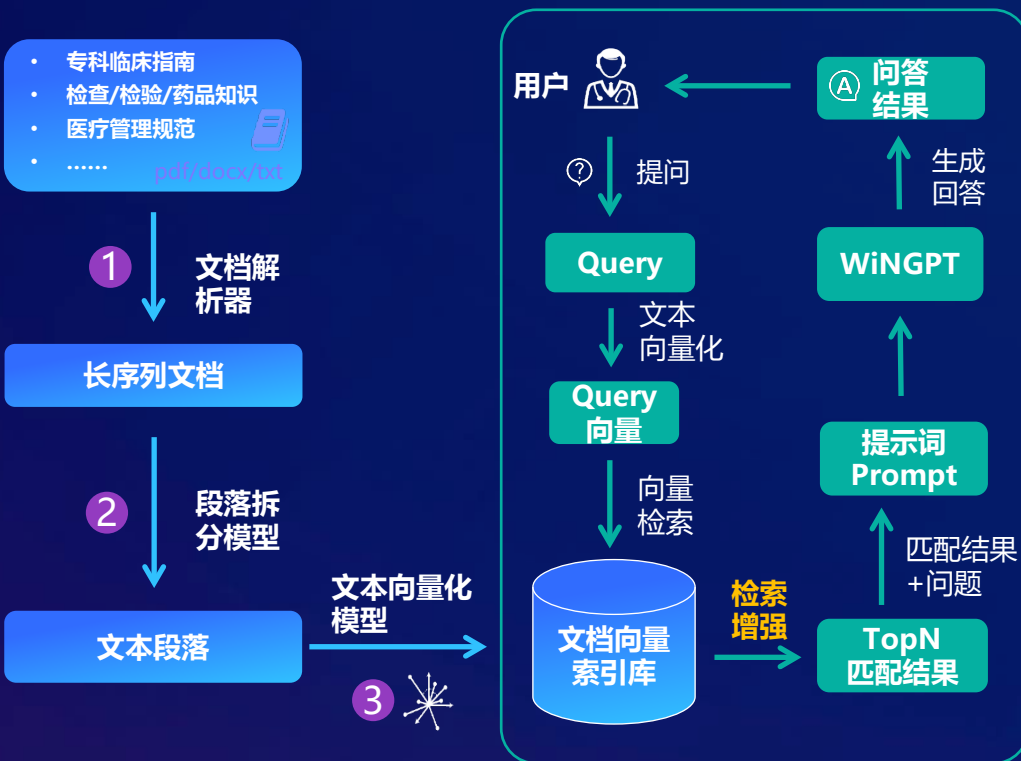


前端病历质控交互

● 应用场景 | 诊中决策支持

利用卫宁积累的临床指南、医疗管理规范、重点临床专业质控指标、检查知识、检验知识、疾病知识等以及合作的**权威知识库**，通过**RAG (Retrieval-Augmented Generation) 技术**，WiNGPT可以访问专业的医疗知识，从而减少对过时或不准确数据的依赖，**降低事实性幻觉**的风险。

同时，RAG技术的检索过程可以确保模型生成的内容与用户的查询和医疗知识图谱中的信息保持一致，从而减少忠实性幻觉的发生



基于RAG的临床辅助决策过程

知识学习更新

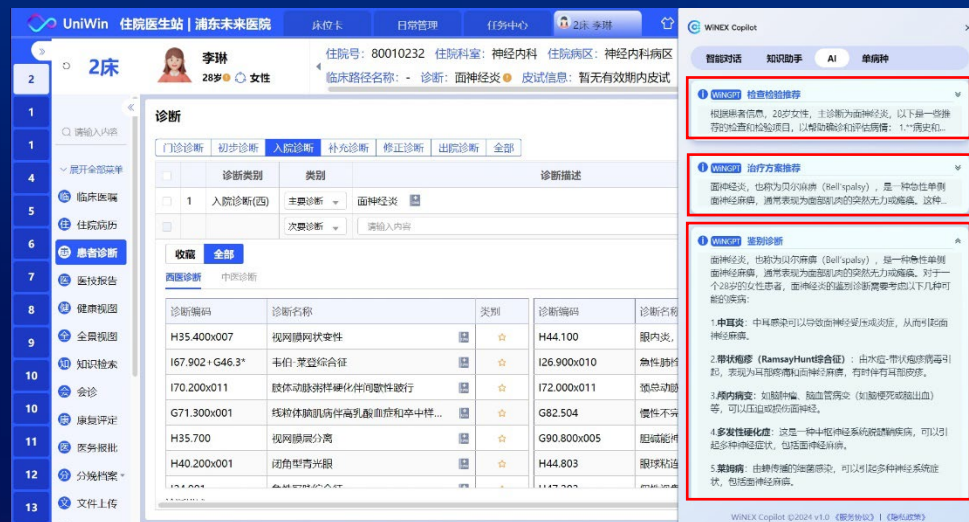
整合专业知识图谱，打造自己的知识库，持续学习，减少AI幻觉

个性化建议

多模态数据的融合和分析，提供个性化的诊疗建议和治疗方案

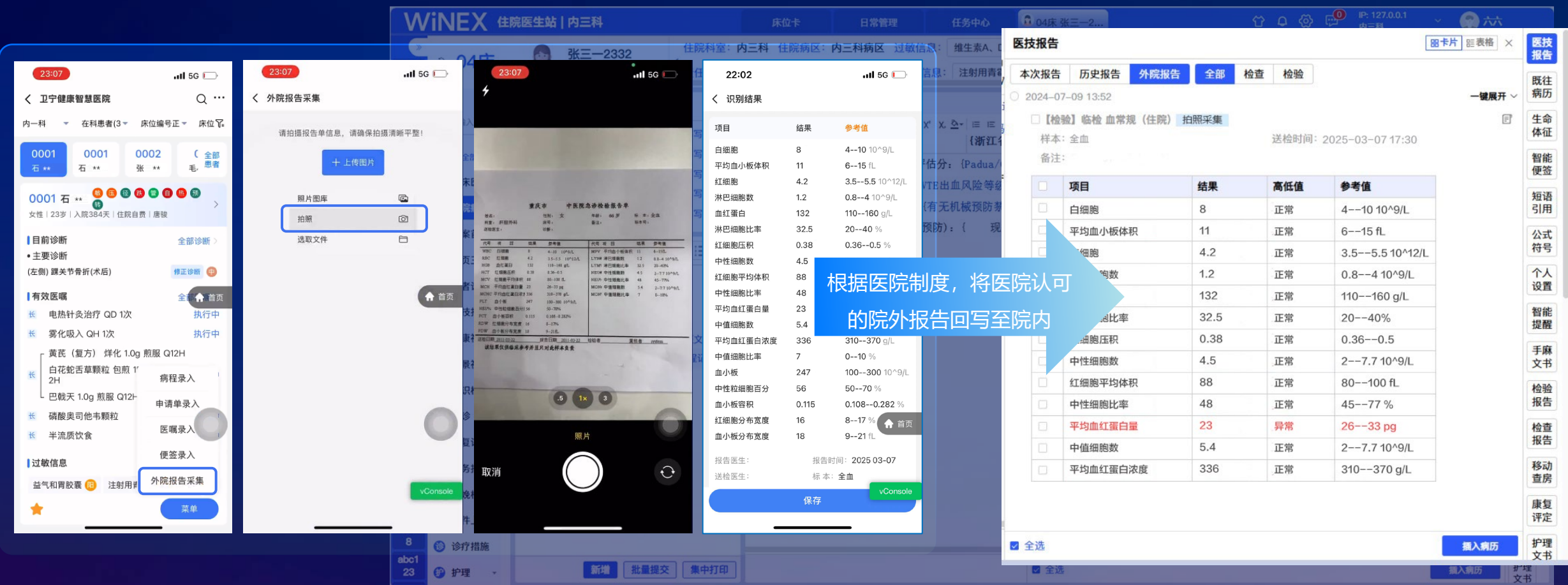
交互式工作界面

医护可以与系统进行自然语言交互，实现更人性化的智能问答和推荐功能



● 应用场景 | 报告解析

通过拍照、选择文件等方式采集院外报告，结合模型智能文字提取与识别能力，将院外报告解析为结构化报告数据并回写至院内，支持门诊/住院医生工作站查看及一键引用，同时支持与院内报告数据进行指标趋势对比



● 应用场景 | 手术用血评估与管理



聚焦患者围手术期患者术前、术中、术后核心场景，整合患者主诉、检验检查、诊断等各类信息，利用大模型，实现**各类评估与治疗建议**，以支撑患者最优的用血方案，提升预后恢复能力。

术前评估与建议

各类评估 | 治疗建议

以提升患者自身造血能力为目标，对患者术前贫血、凝血、自体采血能力等进行评估，并根据情况给出治疗建议

术中用血预测与联动

失血量预测 | 风险预测 | 关联申请单

对术中失血量、最佳血色素水平进行预测，并提供备血方案、术中用药方案，识别和管理风险，并关联业务流程，进行自动申请

术后评估与建议

最佳血色素水平预测 | 预后建议

结合术后生命体征、检验指标等信息，评估血色素水平，并给出长期预后的建议

多类血液指标评估

不同手术类型的采血评估

关键指标补充

自动关联备血申请

● 已在全国各地头部医院推进落地



北京大学人民医院

临床辅助决策 | 手术步骤智能语音录入 |
住院病历内涵质控 | 科研专病库数据自动提取与
填充 | 医学临床实验知情同意书伦理辅助审查

共同发布人医智助 (Pai Assistant)
目前已上线上述5个场景



上海市第一人民医院

影像报告助手
(涵盖引导式结论及肿瘤对比分析等)

上线使用
收到科室人员好评反馈



上海市肺科医院

病历文书助手 (常规病历 + COC专病病
历) | 病历质控助手 | 临床辅助决策 | 面
向专科专病: COC肿瘤疗效评估分析等

目前已上线病历文书&
病历质控应用



上海市皮肤病医院

病历文书助手 | 临床决策助手 |
病历质控助手

目前已上线病历文书&
病历质控应用



上海市同济医院

病历文书助手 | 超声报告生成 | 超声报告质
控 | 超声专病结构化数据抽取 | 放射报告生
成 |

内镜报告生成 | 内镜报告质控 | 内镜报告
&内镜结构化数据抽取 & 病历文书应用



中南大学湘雅二医院

输血管理辅助决策
术前: 贫血评估 | 凝血评估 | 自体采血评估
术中: 制定备血方案评估, 一键引用
术后: 患者血色素评估 | 术后凝血状况及治疗建议

预计2025年6月上线



青岛市卫生健康大数据中心

区域健康医疗大数据创新应用
共享: 检查检验指标解读
管理: 区域卫生管理决策
居民: 健康评估和指导

已推进上线, 荣获第二届全国数字健康创新
应用大赛健康医疗大数据主题赛二等奖

融合 WiNGPT / DeepSeek 等多模能力的应用场景已在全国 **100余家** 二三级医院上线

(统计截至2025年4月24日)

谢谢