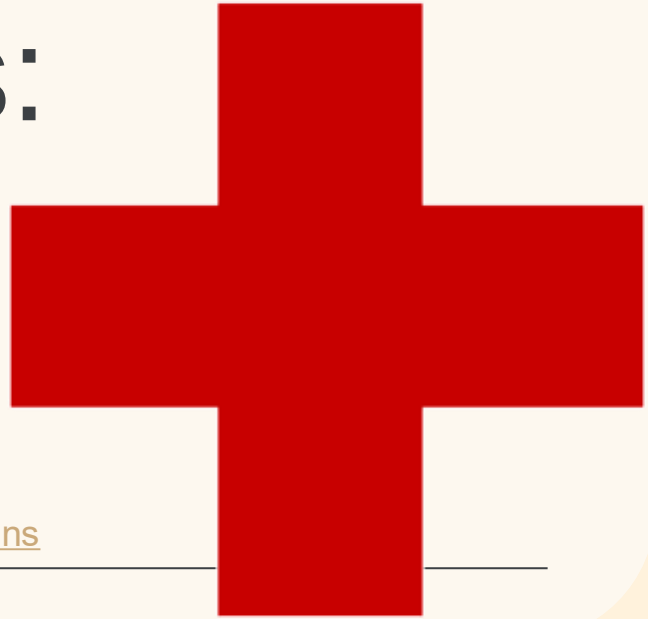
S3.2 Team Present: Agent Applications: Healthcare



Tammy Ngo (bsy6pq)
Matt Juntima (vqj9sq)
Sebastian Pop (qju9ta)

2026 Spring
LLM Agents Foundation & Applications

20260210

Agenda

1. Deep Research: A Survey of Autonomous Research Agents
(Presented by Tammy Ngo)

2. A Survey of LLM-based Agents in Medicine: How far are we from Baymax?
(Presented by Sebastian Pop)

3. Evaluating large language models and agents in healthcare: key challenges in clinical applications
(Presented by Matt Juntima)

Presented by Tammy Ngo

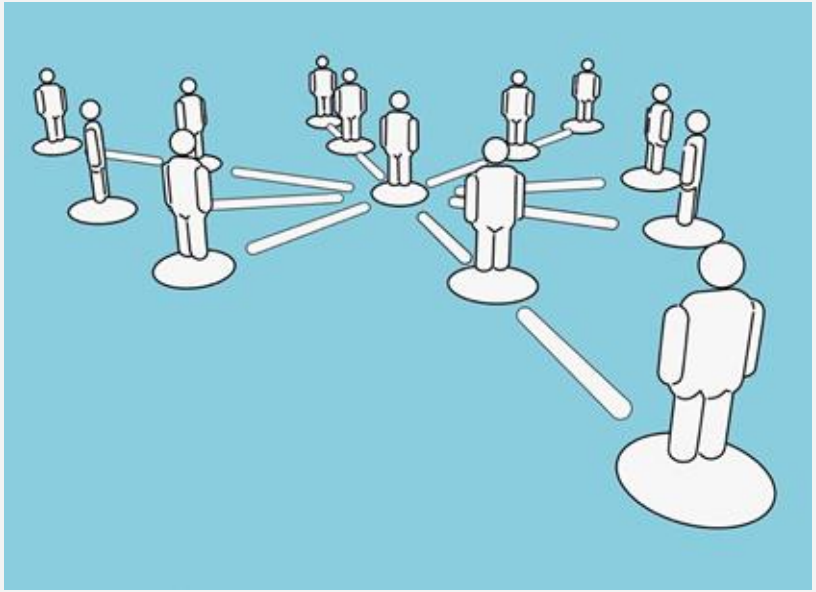
1

Deep Research: A Survey of Autonomous Research Agents

Motivation

- Large Language Models (LLMs) are powerful but limited by internal knowledge boundaries.
- Autonomous research agents extend LLMs with planning, retrieval, and synthesis abilities.

Goal: Produce **comprehensive, evidence-grounded analytical reports**.

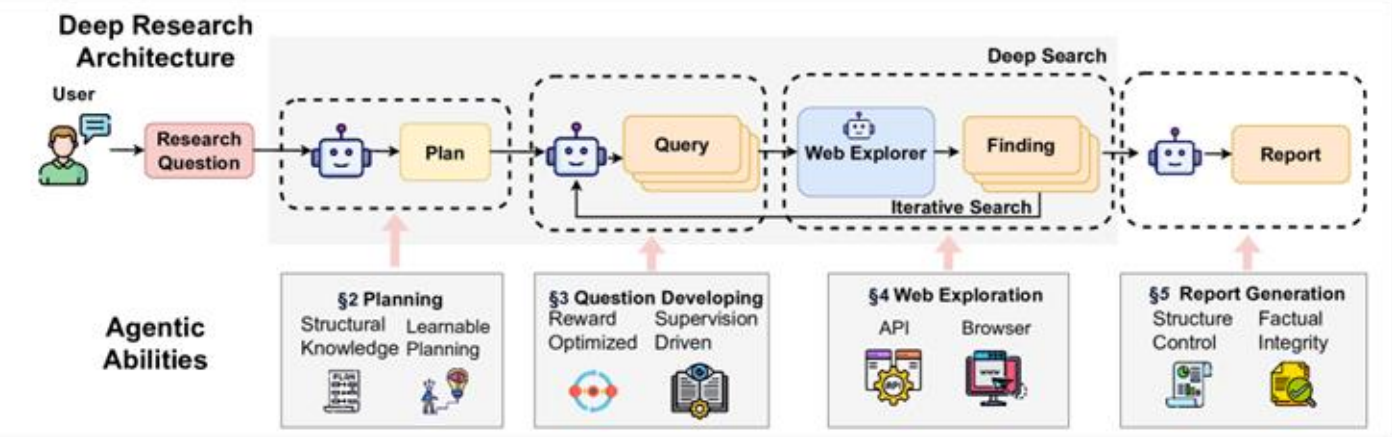


Deep Research Pipeline

Four Core Stages

1. Planning
2. Question Developing
3. Web Exploration
4. Report Generation

These stages form the standard pipeline used by modern research agents.



Research Agent Foundations

Research Agent Architecture

Core Modular Components

- Planner / Orchestrator
- Tool & Retrieval Systems
- Reasoning and Synthesis Module
- Memory (short-term + long-term)
- Evaluation and Verification

Closed-Loop Workflow

Goal → Plan → Retrieve → Analyze → Evaluate → Refine

Modular pipelines improve long-horizon task execution and reliability.

World Models in Research Agents

LLMs as Implicit World Models

- LLMs encode latent knowledge about tasks, tools, and reasoning patterns.
- Used to predict outcomes of actions during planning.

Graph-Based Structured Knowledge

- Knowledge graphs provide explicit structure for entities and relationships.
- Improve reasoning traceability and multi-step decision making.
- Combining implicit (LLM) and explicit (graph) world models improves planning reliability.

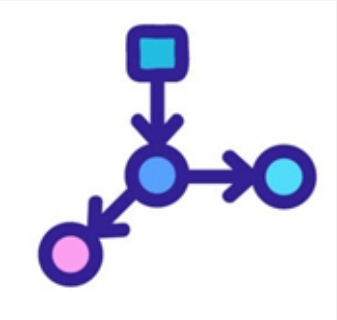


Table 1. Taxonomy of planning methods in deep research agents.

Planning	Category	Related Works
Planning with Structured World Knowledge	Planning via simulation	WebDreamer (Gu et al., 2024), Simulate Before Act (Gu et al., 2025)
	Planning via modularity	Webpilot Zhang et al. (2025b), WKM (Qiao et al., 2024),Plan-and-Act Erdogan et al. (2025)
	Planning via adaptation	Thought of Search (Katz et al., 2024), MPO (Xiong et al., 2025)
Planning as a Learnable Process	Planning via space exploration	Agentsquare (Shang et al., 2024), Agent-E Abuelsaad et al. (2024)
	Planning via self-training	Patel et al. (2024), InSTA Trabucco et al. (2025)
	Planning via preference modeling	MindSearch Chen et al. (2024), SimpleDeepSearcher (Sun et al., 2025b), Search-in-the-Chain (Xu et al., 2024b), WEPO (Liu et al., 2025), MPO (Xiong et al., 2025)

Planning

Overview & Methods

Overview

- Planning decomposes research goals into actionable sub-tasks.
- Enables coordination of tools, retrieval, and reasoning.
- Critical for long-horizon and multi-step research workflows.

Methods

Planning with Structured World Knowledge

- *Planning via Simulation*: Agents simulate candidate action sequences before execution.
- *Planning via Modularity*: Multi-module or multi-agent pipelines distribute specialized tasks.
- *Planning via Adaptation*: Plans dynamically adjust based on intermediate results and context.

Planning as a Learnable Process

- *Planning via Space Exploration*: Agents search across possible reasoning or action trajectories.
- *Planning via Self-Training*: Systems iteratively refine their own planning strategies.
- *Planning via Preference Modeling*: Plans optimized toward human-aligned or task-specific objectives.

Significant Related Works

Meta-Plan Optimization (MPO)

- Learns how to improve planning strategies across environments.
- Enables adaptive tuning of agent behavior for new tasks.
- Supports transfer of planning knowledge between domains.
- Moves planning from static heuristics to **learned, generalizable policies**.

AgentSquare

- Automatically assembles agent pipelines from modular components.
- Searches over combinations of planners, tools, memory, and evaluators.
- Identifies efficient architectures for specific research tasks.
- Reduces manual engineering of complex agent workflows.

Table 2. Taxonomy of question developing methods in deep research agents.

Optimization	Category	Related Works
Reward-Optimized Methods	Rewards for format and accuracy	DEEPRESEARCHER (Zheng et al., 2025), EVOLVESEARCH (Zhang et al., 2025d), R1-SEARCHER (Song et al., 2025), SEARCH-R1 (Jin et al., 2025), ZEROSEARCH (Sun et al., 2025a), MASKSEARCH (Wu et al., 2025b), DEEPRETRIEVAL (Jiang et al., 2025)
	Multi-dimensional reward	INFORAGE (Qian and Liu, 2025), OTC-PO (Wang et al., 2025b), IKEA (Huang et al., 2025c), AUTOREFINE (Shi et al., 2025), R-SEARCH (Zhao et al., 2025), MMSEARCH-R1 (Wu et al., 2025a), VRAG-RL (Wang et al., 2025a)
Supervision-Driven Methods	Multi-agent systems	MANUSEARCH (Huang et al., 2025b), SEARCH-O1 (Li et al., 2025a), SEARCHAGENT-X (Yang et al., 2025b)
	Supervision optimization	REASONRAG (Zhang et al., 2025a)

Question Developing

Overview & Methods

Overview

- Converts broad goals into structured research questions.
- Improves retrieval coverage and reasoning depth.
- Essential for complex analytical reporting.

Methods

Reward-Optimized Methods

- *Rewards for Format and Accuracy*:
  - Optimize for structured outputs, citations, and evidence grounding.
  - Encourages well-organized and verifiable research questions.
- *Multi-Dimensional Reward*: Joint optimization across relevance, completeness, and reasoning quality.

Supervision-Driven Models

- *Multi-Agent Systems*: Separate agents generate, critique, and refine questions.
- *Supervision Optimization*: Uses human or synthetic feedback to improve question quality.

Table 3. Taxonomy of web exploration methods in deep research agents.

Information Source	Category	Related Works
Web-based	Web scraping and crawling	**Scrapy** (Scrapy developers, 2025), **BeautifulSoup** (Richardson, 2025)
	Browser-based web agents	**WebGPT** (Nakano et al., 2021), **Selenium** (Selenium contributors, 2025)
	Multimodal web agents	**WebVoyager** (He et al., 2024), **MM-ReAct** (Yang et al., 2023), **WebArena** (Zhou et al., 2023a)
API-based	Industrial search engines	**Bing** (bin, [n. d.]), **X posts** (xpo, [n. d.]), **Google** (goo, [n. d.])
	Domain-specific search engines	**Reportify** (Reportify, nd), **YanXueZhiDe** (China National Knowledge Infrastructure, 2024), **CNKI** (CNKI, nd), **DuckSearch** (Sourty, 2024), **BraveSearch** (Brave Software, 2022), **Bocha** (Bocha AI, nd)

Web Exploration

Overview & Information Sources

Overview

- Agents gather external evidence through iterative retrieval.
- Extends knowledge beyond model training data.

<u>Information Sources</u>

Web-Based Exploration

- *Web Scraping and Crawling*: Automated extraction of documents and structured information.
- *Browser-Based Web Agents*: Navigate pages, follow links, and interact with interfaces.
- *Multimodal Web Agents*: Process text, images, tables, and mixed media sources.

API-Based Exploration

- *Industrial Search Engines*: Provide scalable, reliable web retrieval.
- *Domain-Specific Search Engines*: Access specialized databases, such as academic or technical sources.

Table 4. Taxonomy of artifact generation methods in deep research agents.

Field	Category	Representative Works
Structure Control	Planning-based Generation	Agent Laboratory (Schmidgall et al., 2025), AI Scientist v2 (Yamada et al., 2025), LongEval (Wu et al., 2025c), LongWriter (Bai et al., 2024), LongDPO (Ping et al., 2025)
	Constraint-guided Generation	WebThinker (Li et al., 2025b), Suri (Pham et al., 2024), Wan et al. (2025)
	Structural-aware Evaluation	Long2RAG (Qi et al., 2024), ExPerT (Salemi et al., 2025), Long et al. (2025), Kim et al. ([n. d.]), Huang et al. (2024)
Factual Integrity	Faithful Modeling	RAGSynth (Shen et al., 2025), BRIDGE (Dai et al., 2025), Zhou et al. (2023b), Shi et al. (2024)
	Conflict Reasoning	FaithfulRAG (Zhang et al., 2025c), DRAGged (Cattan et al., 2025), Yuan et al. (2024), Ying et al. (2023)
	Factuality Evaluation	Face4RAG (Xu et al., 2024a), SFR-RAG (Nguyen et al., 2024), Wallat et al. (2024), FaithJudge (Tamber et al., 2025), RAG-QA Arena (Han et al., 2024), MT-RAIG (Seo et al., 2025)

Report Generation

Overview & Methods

Overview

- Synthesizes retrieved evidence into structured analytical reports.
- Emphasizes traceability and citation support.

Methods

Structure Control

- *Planning-Based Generation*: Uses hierarchical outlines to guide report structure.
- *Constraint-Guided Generation*: Enforces formatting, section coverage, and citation rules.
- *Structural-Aware Evaluation*: Measures organization, coherence, and completeness.

Factual Integrity

- *Faithful Modeling*: Grounds outputs in retrieved evidence.
- *Conflict Reasoning*: Detects and reconciles contradictory sources.
- *Factuality Evaluation*: Measures evidence support and citation correctness.

Optimization

Optimization Techniques

- Prompt optimization.
- Supervised fine-tuning.
- Reinforcement learning for agent policies.
- Multi-objective optimization across accuracy, coverage, and efficiency.
- Optimization improves stability and long-horizon performance.

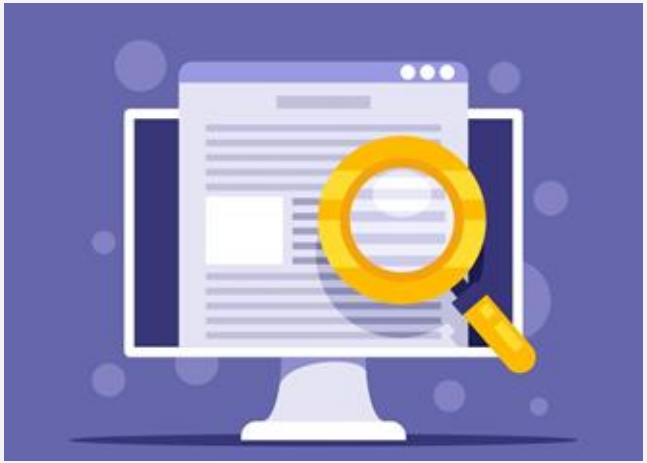


Table 5. Coverage of core Deep Research modules across major benchmarks. P: Planning, QD: Question Developing, WE: Web Exploration, RG: report generation.

Benchmark	P	QD	WE	RG	Task	Evaluation Metrics
MIND2WEB 2 (Gou et al., 2025)	✓	✓	✓	✗	Web search	Success rate, Partial Completion
BROWSECOMP (Wei et al., 2025)	✓	✓	✓	✗	Web search	Accuracy, Calibration Error
WEBARENA (Zhou et al., 2023a)	✓	✓	✓	✗	Web search	Success Rate
GAIA (Mialon et al., 2023)	✓	✓	✓	✗	Multi-step assistant tasks	EM
HUMANITY'S LAST EXAM (Phan et al., 2025)	✓	✓	✓	✗	Multidomain reasoning	Accuracy, Calibration Error
BROWSECOMP-ZH (Zhou et al., 2025)	✓	✓	✓	✗	Web search in Chinese	Accuracy, Calibration Error
MEDBROWSECOMP (Chen et al., 2025b)	✓	✓	✓	✗	Medical web search	Accuracy
GPQA (Rein et al., 2024)	✓	✓	✓	✗	Graduate-level QA	Accuracy
INFODEEPSEEK (Xi et al., 2025)	✗	✓	✓	✗	Open-domain QA	Accuracy, Information Accuracy
DEEPRESEARCH BENCH (Du et al., 2025)	✓	✓	✓	✓	Research report generation	Pairwise Agreement Rate, Overall Pearson Correlation
DEEPRESEARCHGYM (Coelho et al., 2025)	✓	✓	✓	✓	Research task sandbox	KPR, KPC, Precision, Recall, Clarity, Insight

Benchmarks and Evaluation

Overview & Benchmarks

Overview

- Research agents require evaluation beyond simple QA accuracy.
- Metrics target reasoning quality and report usefulness.

Benchmarks

Search-Oriented Benchmarks

- Evaluate retrieval effectiveness and navigation ability.
- Measure tool use and exploration strategies.

Research-Oriented Benchmarks

- *DeepResearch Bench*
  - Evaluates:
    - Report fidelity
    - Citation accuracy
    - Comprehensive topic coverage
  - Uses complex, multi-step research tasks.

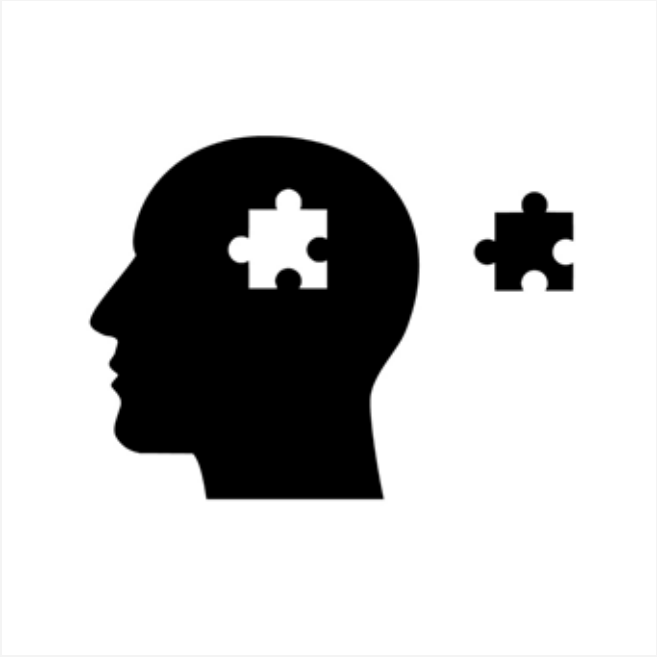
Limitations and Future Directions

Key Challenges

Plan Brittleness
- Plans may fail under uncertainty or unexpected results.

Lack of Robustness to Ambiguous Queries
- Agents struggle when goals are underspecified.

Evaluation Coarseness
- Existing metrics do not fully capture research quality.



Future Directions

Multi-Tool Integration
- Better coordination across search, code, and analysis tools.

Factuality Improvements
- Stronger verification and grounding mechanisms.

Multimodal Reasoning Capabilities
- Integrating text, images, tables, and data.

Workflow Design and Model Optimization
- More adaptive and efficient agent pipelines.

Personalization
- Tailoring research outputs to user intent and expertise.

Key Takeaways

- Autonomous research agents shift AI from **single responses** to **iterative, evidence-grounded research workflows**.
- Future progress depends on:
  - Robust planning
  - Better evaluation
  - Reliable factual reasoning

Presented by Sebastian Pop

A Survey of LLM-based Agents in Medicine: How far are we from Baymax?

Wenxuan Wang[1*] **Zizhan Ma**[1*] **Zheng Wang**[1] **Chenghan Wu**[1]
Jiaming Ji[2] **Wenting Chen**[3] **Xiang Li**[4] **Yixuan Yuan**[1]
[1]The Chinese University of Hong Kong
[2] Peking University [3] City University of Hong Kong
[4] Massachusetts General Hospital and Harvard Medical School
[1]{wenxuanwang,zizhan.ma}@link.cuhk.edu.hk [2]wentichen7-c@my.cityu.edu.hk

2

A Survey of LLM-based Agents in Medicine: How far are we from Baymax?

Motivation

Context

- Explosion of LLMs (ChatGPT, GPT-4, etc) →hype in healthcare
- Traditional LLMs ≠ enough for real medical workflows

Medical workflows require more than good answers

- Structured reasoning
- Decision making
- Tool usage

Toward LLM-based agents

- Need for LLM-based agents
- Combining language models with planning, memory, and external tools →support medical tasks

The “Baymax” vision: safe, autonomous medical assistants

Background

Scope

- Reviews 60 studies on LLM-based medical agents (2022-2024)

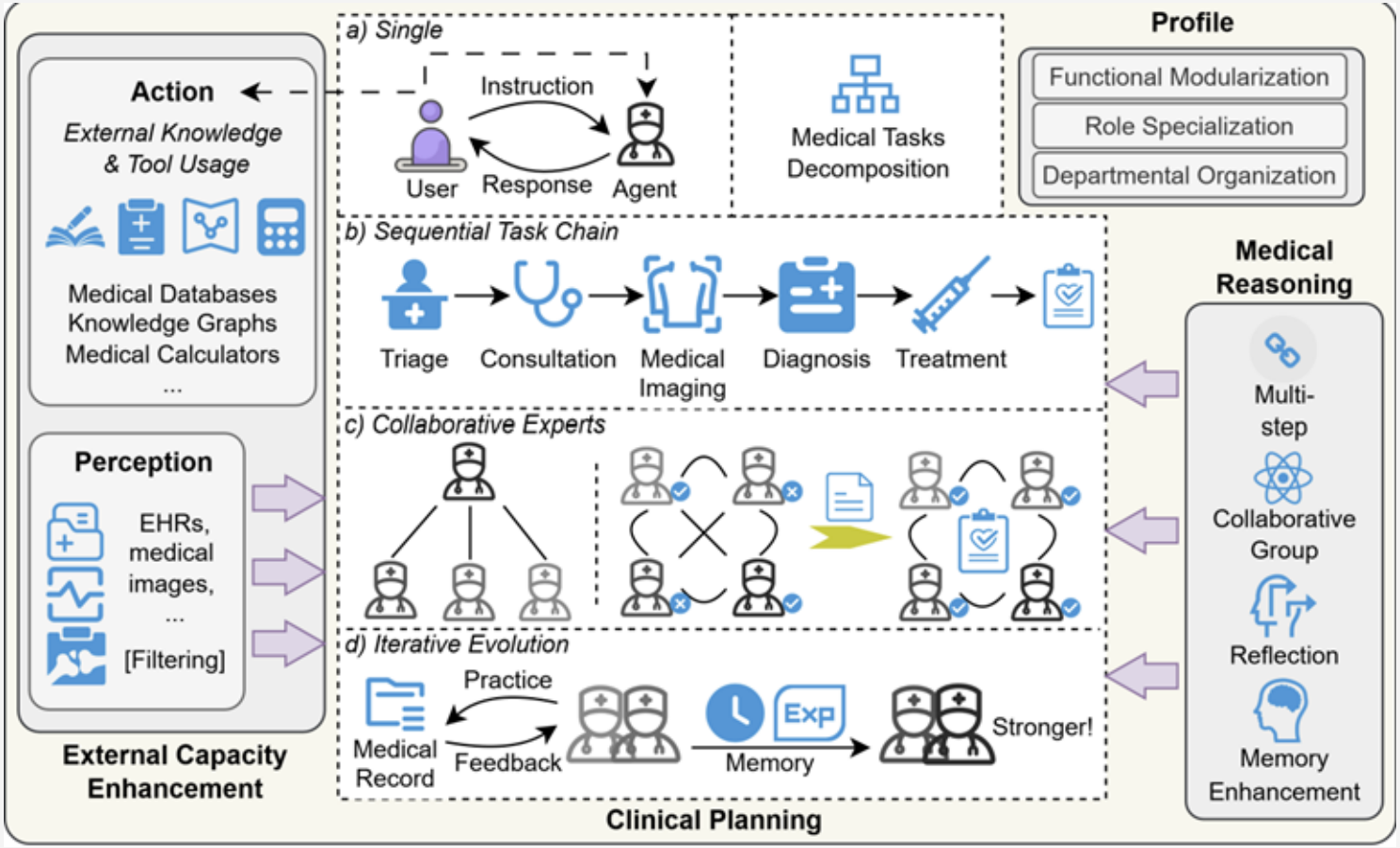
Rather than focusing on individual models

- Examines agent-based systems embedded in structured workflows

Medical settings

- These agents must operate under stricter constraints than general AI systems
    - Including handling multimodal data (clinical text, medical imaging, and lab results)

Architecture Overview



Agent Profiles

What is an agent profile?

- Define roles and responsibilities of an LLM-based medical agent

How profiles are assigned

- Rather than a single generic agent, many systems assign specific profiles
- Reflects how work is organized in real clinical settings

<u>Three common approaches</u>

- Profiles based on functional modules (e.g., data analysis)
- Profiles that mirror medical roles (e.g., specialists)
- Profiles organized by medical departments (e.g., radiology)

Why this matters

- Defining clear profiles allow agents to specialize
- Coordinate with each other
- Collaborate more effectively with clinicians in complex workflows

Clinical Planning & Agent Paradigms

What is clinical planning?

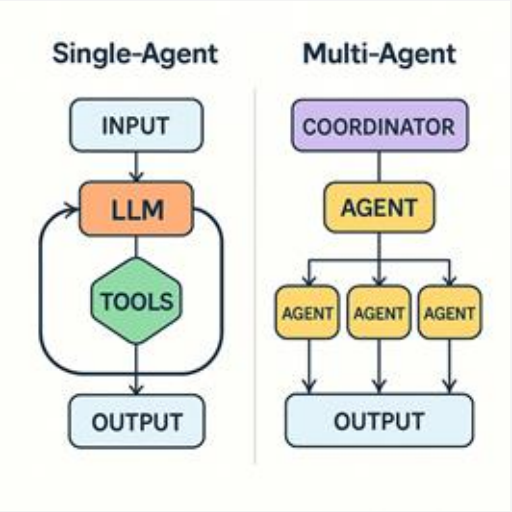
- Describes how LLM-based medical agents structure and coordinate complex medical workflows

Rather than a single response, agents break work into structured steps:

- Data collection → Diagnosis → Treatment planning → follow-up

Identifies several common planning paradigms

- In simpler settings → single agent may handle the entire task
- More complex workflows:
  - Sequential task chains (steps handled in order)
  - Collaborative expert agents (specialized agents work in parallel)
- Enables medical agents to operate
  - More reliable, transparent, and controllable way than standalone language models



Medical Reasoning

What is medical reasoning?

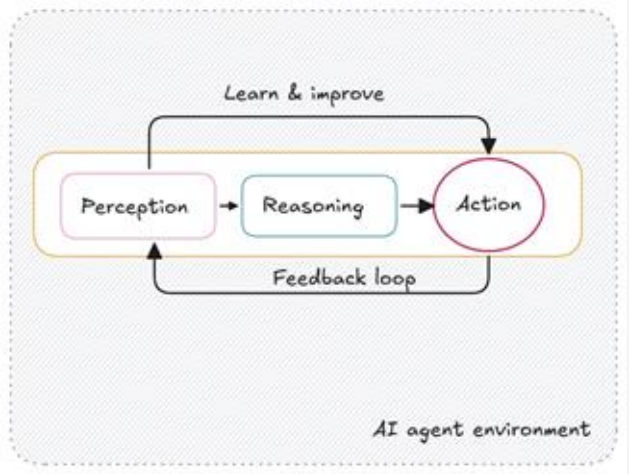
- Describes how LLM-based agents generate, evaluate, and refine clinical decisions

Unlike traditional LLMs that produce one-shot answers

- Medical agents often rely on multi-step reasoning processes
- Break down complex cases into intermediate steps

Reflection & collaboration

- Agents review their own outputs
- Receive feedback from other agents → correct errors & reduce uncertainty



External Capacity Enhancement

Focus on how LLM-based medical agents connects to real-world clinical resources

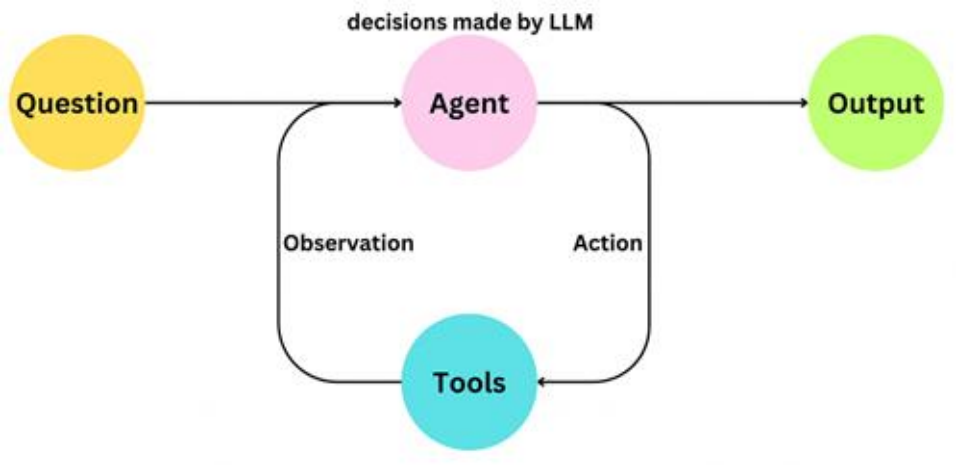
- Rather than relying only on internal knowledge

Agents access external data and tools

- Electronic health records
- Medical images
- Clinical guidelines
- Drug databases
- Use specialized tools:
  - Medical calculators
  - Knowledge graphs
- Support more accurate and context aware decisions

External grounding is important in healthcare

- Up-to-date information
- Verifiable sources are essential for safe deployment



Application Scenarios

- **Clinical decision support and diagnosis**
- **Clinical data analysis and documentation**
- **Medical training and simulation**
- **Healthcare service optimization**

Clinical Decision Support & Diagnosis

Agents mainly used to improve reasoning quality

- Rather than to replace clinicians

How agents improve decisions

- Many systems have agents with specialized roles (diagnosis, evaluation, treatment)
- Collaboration helps reduce bias and improve accuracy

Applied in areas

- Diagnosis, prescription checking, and patient interaction safety

Other Application Areas

Clinical data analysis & documentation

- *Summarizing records, predicting outcomes, generating patient-friendly reports*

Medical training

- *Agent-based simulations to evaluate and improve diagnostic and treatment skills in realistic environments*

Healthcare service optimization

- *Automating non-diagnostic tasks (patient education and administration), helping reduce clinicians workload while maintaining service quality*

Purpose	Functionality		Work	Framework Type	Tool Use
Clinical Decision Support and Diagnosis	Refine Diagnostic Reasoning		(Dutta and Hsiao, 2024)	Adaptive Planning	-
	Reduce Cognitive Bias		(Ke et al., 2024)	Collaborative Experts	-
	Task Coordination		(Wei et al., 2024b)	Sequential Task Chain	-
	Diagnosis Accuracy		(Kim et al., 2024c)	Collaborative Experts	Yes
			(Tang et al., 2024)	Collaborative Experts	-
	Domain Specific Functionalities	Clinical Trial Outcome Prediction	(Yue et al., 2024)	Sequential Task Chain	Yes
		Patient Interaction Safety	(Mukherjee et al., 2024)	Sequential Task Chain	-
		Prescription Validation	(Van et al., 2024)	Sequential Task Chain	Yes
	Diagnosis Capability		(Yan et al., 2024)	Collaborative Experts	-
	Integrated Modelling		(Fan et al., 2024)	–	-
Clinical Data Analysis and Documentation	Mortality Prediction		(Wang et al., 2024b)	Collaborative Experts	Yes
	Hospital Readmission Analysis				
	Clinical Documentation		(Lee et al., 2024)	Single Agent	-
	Patient Friendly Medical Reports		(Sudarshan et al., 2024)	Iterative Evolution	Yes
	Integrated Simulation		(Li et al., 2024b)	Iterative Evolution	-
Medical Training and Simulation	Evaluated Diagnosis and Treatment Performance		(Yan et al., 2024)	Collaborative Experts	-
	Integrated Simulation		(Fan et al., 2024)	–	-
			(Li et al., 2024b)	Iterative Evolution	-
	Medical Training	Training Environment	(Wei et al., 2024a)	Collaborative Experts	Yes
			(Wu et al., 2024)	Collaborative Experts	-
		Scenario Simulation	(Yu et al., 2024)	Collaborative Experts	Yes
Healthcare Service Optimization	Automation of Non-diagnostic Tasks		(Mukherjee et al., 2024)	Sequential Task Chain	Yes
			(Laymouna et al., 2024)	–	-
	Automation of Diagnostic Tasks		(Chadebecq et al., 2023)	Iterative Evolution	-

Evaluation & Benchmarking

Why evaluation is critical

- Medical agents operate in high-risk clinical environments, where errors can directly affect patient safety
- Careful evaluation is required to ensure reliability, safety, and clinical usefulness

Benchmark categories:

- Static Q&A benchmarks
- Workflow-based simulations
- Automated evaluation frameworks

Shows that different evaluation methods capture different aspects of medical AI agent performance

- Meaning that no single benchmark is sufficient on its own

Evaluation Attribute	Genre	Specific Names	Related Work
Benchmarks	Static Q&A Benchmarks	*MedQA*	(Jin et al., 2020)
		MedMCQA	(Pal et al., 2022)
		Pub-MedQA	(Jin et al., 2019)
		MMLU	(Hendrycks et al., 2021b) (Hendrycks et al., 2021a)
	Workflow-Based Simulation Benchmarks	*MedChain*	(Liu et al., 2024a)
		AI Hospital	(Fan et al., 2024)
		AgentClinic	(Schmidgall et al., 2024b)
		ClinicalLab	(Yan et al., 2024)
	Automated Evaluation Frameworks	*AI-SCE*	(Mehandru et al., 2024)
		RJUA-SPs	(Liu et al., 2024b)
Metrics for Task-Specific Evaluation	Exact Match Metrics	Accuracy	–
		Precision	–
		Recall	–
	Semantic Similarity Metrics	**BLEU**	(Papineni et al., 2002)
		ROUGE	(Lin, 2004)
		BertScore	(Zhang et al., 2020)
	LLM-Based Evaluation Metrics	*ChatCoach*	(Huang et al., 2024a)
		Retrival-Augmented Evaluation Framework	(Liu et al., 2024b)

Challenges & Limitation

Hallucinations

- Incorrect or fabricated outputs are dangerous in medical settings
- Errors in diagnosis or treatment can directly affect patient outcomes

Integrating multimodal data

- Medical agents must combine text, images and laboratory results
- Differences in data formats, documentation standards, and language can make it difficult

.

Evaluation methods

- Many current benchmarks rely on static tasks, and fail to capture real clinical workflows, interactive decision-making, and reasoning quality

Large-scale deployment

- Raises practical & ethical concerns (high costs, privacy and security tasks, algorithmic bias, and the lack of patient-centered design)

Future Directions

Strong focus on improving medical reasoning
- Through multi-step inference and reflection

Simulation-based training
- Key way to evaluate diagnosis and treatment decisions more realistically

Better integration with clinical systems
- Physical devices (monitoring systems and medical robotics)

Should prioritize safety before autonomy
- Current LLM-based agents should be viewed as assistive tools rather than autonomous clinicians

Key takeaways

LLM-based medical agents go beyond passive text generation

- They enable planning, multi-step reasoning, and tool use within clinical workflows

Agent architectures are essential for medical AI

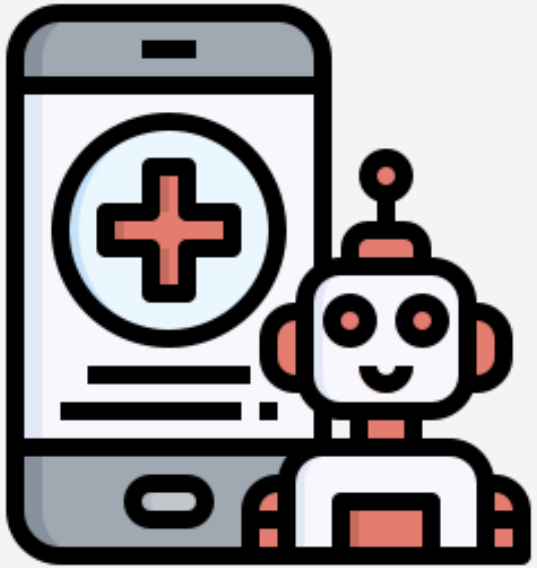
- Profiles, planning, reasoning, and external tools support safe and reliable system

Multi-agent and sequential workflows dominate practice

- Especially in clinical decisions support and diagnostic reasoning

We are not yet at Baymax

- Current systems show promise, but remain safety-focused and clinician-assisted



Presented by Matt Juntima


Review

Evaluating large language models and agents in healthcare: key challenges in clinical applications

Xiaolan Chen [1 #], Jiayang Xiang [2 #], Shanfu Lu [3 #], Yexin Liu [4], Mingguang He [1 5 6] ✉, Danli Shi [1 2] ✉

3

Evaluating large language models and agents in healthcare: key challenges in clinical applications

4

Recap

Paper Summaries

Paper 1

Modern LLM-based agents autonomously conduct complex research by…

- Integrating planning
- Adaptive query generation
- Iterative web exploration
- Structured report synthesis

to produce comprehensive, evidence-grounded analyses

Paper 2

LLM-based agents show promise in healthcare, but current systems remain assistive and require structured planning, reasoning and evaluation

Paper 3

Need for better datasets and more diverse evaluation techniques to evaluate LLMs in healthcare

Questions