

Supplementary Table 1: Search Strategy (list of queries used to search databases)

Database	Query	Study Num
Scopus	(TITLE-ABS-KEY ("agent*" OR "multi-agent*" OR "multimodal agent*" OR "multi-modal agent*") AND TITLE-ABS-KEY ("llm*" OR "large language model*" OR "generative model*" OR "foundation* model*" OR "vision language model*" OR "vision-language model*") AND TITLE-ABS-KEY ("health" OR "healthcare" OR "medicine" OR "medical" OR "clinical" OR "biomedical" OR "hospital")) AND PUBYEAR > 2021 AND PUBYEAR < 2026 AND (LIMIT-TO (DOCTYPE , "ar") OR LIMIT-TO (DOCTYPE , "cp")) AND (LIMIT-TO (LANGUAGE , "English"))	586
Web of Science	((TS=("agent*" OR "multi-agent*" OR "multimodal agent*" OR "multi-modal agent*")) AND TS=("llm*" OR "large language model*" OR "generative model*" OR "foundation* model*" OR "vision language model*" OR "vision-language model*")) AND TS=("health" OR "healthcare" OR "medicine" OR "medical" OR "clinical" OR "biomedical" OR "hospital")	240
Embase	("agent*" OR "multi-agent*" OR "multimodal agent*" OR "multi-modal agent*") AND ("llm*" OR "large language model*" OR "generative model*" OR "foundation* model*" OR "vision language model*" OR "vision-language model*") AND ("health" OR "healthcare" OR "medicine" OR "medical" OR "clinical" OR "biomedical" OR "hospital")	199
PubMed	((("agent*[Title/Abstract] OR "multi-agent*[Title/Abstract] OR "multimodal agent*[Title/Abstract] OR "multi-modal agent*[Title/Abstract]) AND ("llm*[Title/Abstract] OR "large language model*[Title/Abstract] OR "generative model*[Title/Abstract] OR "foundation* model*[Title/Abstract] OR "vision language model*[Title/Abstract] OR "vision-language model*[Title/Abstract])) AND ("health"[Title/Abstract] OR "healthcare"[Title/Abstract] OR "medicine"[Title/Abstract] OR "medical"[Title/Abstract] OR "clinical"[Title/Abstract] OR "biomedical"[Title/Abstract] OR "hospital"[Title/Abstract])	177
IEEE	("All Metadata": "agent*" OR "All Metadata": "multi-agent*" OR "All Metadata": "multimodal agent*" OR "All Metadata": "multi-modal agent*") AND ("All Metadata": "llm*" OR "All Metadata": "large language model*" OR "All Metadata": "generative model*" OR "All Metadata": "foundation* model*" OR "All Metadata": "vision language model*" OR "All Metadata": "vision-language model*") AND ("All Metadata": "health" OR "All Metadata": "healthcare" OR "All Metadata": "medicine" OR "All Metadata": "medical" OR "All Metadata": "clinical" OR "All Metadata": "biomedical" OR "All Metadata": "hospital")	173
ACM	[Title: "agent*"] OR [Abstract: "agent*"] OR [Title: "multi-agent*"] OR [Abstract: "multi-agent*"] OR [Title: "multimodal agent*"] OR [Abstract: "multimodal agent*"] OR [Title: "multi-modal agent*"] OR [Abstract: "multi-modal agent*"] AND [[Title: "llm*"] OR [Abstract: "llm*"] OR [Title: "large language model*"]	31

	OR [Abstract: "large language model*"] OR [Title: "generative model*"] OR [Abstract: "generative model*"] OR [Title: "foundation* model*"] OR [Abstract: "foundation* model*"] OR [Title: "vision-language model*"] OR [Abstract: "vision-language model*"] OR [Title: "vision language model*"] OR [Abstract: "vision language model*"] AND [(Title: "health") OR [Abstract: "health"] OR [Title: "healthcare"] OR [Abstract: "healthcare"] OR [Title: "medicine"] OR [Abstract: "medicine"] OR [Title: "medical"] OR [Abstract: "medical"] OR [Title: "clinical"] OR [Abstract: "clinical"] OR [Title: "biomedical"] OR [Abstract: "biomedical"] OR [Title: "hospital"] OR [Abstract: "hospital"] AND [E-Publication Date: (01/01/2022 TO 06/30/2025)]	
CINAHL	("agent*" OR "multi-agent*" OR "multimodal agent*" OR "multi-modal agent*") AND ("llm*" OR "large language model*" OR "generative model*" OR "foundation* model*" OR "vision language model*" OR "vision-language model*") AND ("health" OR "healthcare" OR "medicine" OR "medical" OR "clinical" OR "biomedical" OR "hospital")	12
Total Studies		1,418

Supplementary Table 2: Data Extraction Form

Concept	Definition
Basic Information	
ID	Unique ID assigned to each study.
Title	The title of the included paper.
Publication Type	Journal or conference.
Published Year	The year in which the study was published.
Clinical Task	The specific clinical application: Clinical decision support (diagnosis, QA, and treatment); Clinical monitoring and health management; Clinical documentation and report generation; Medical education and training.
Data Information	
Modality	The data modality of the study, including text, image, audio, video, signal, time series, tabular data, and genomic data.
Data Region (Setting)	The geographic region of the test dataset (e.g., Asia, North America) and the study setting (e.g., single-center vs. multi-center).
Data Type	Whether the dataset used was private or public, indicating data provenance (e.g., real-world clinical data vs. benchmark datasets).
Technical Information of Agent System	
Backbone Model	The core foundation model leveraged in the system (e.g., GPT series, Gemini, Claude, multimodal foundation models).
Multi / Single Agent	Indicates whether the system used a single-agent or multi-agent architecture.
Agent Collaboration	For multi-agent systems, specify the collaboration mode: centralized, decentralized, or hybrid.
Fusion Strategy	The data fusion strategy employed (e.g., early feature-level fusion, intermediate, late decision-level fusion, hybrid).
Tool Usage	Whether the system invoked external tools.
Agent Capability	The agent capabilities presented in the system: perception, planning, action, memory, and evolution.
Training-based or without training	Whether the system was trained/fine-tuned on data or operated without additional training (e.g., zero-shot).
Evaluation	
Evaluation Strategy (Maturity)	The evaluation strategy used in the study (e.g., automatic, manual) and the level of validation rigor (e.g., prototype, technical evaluation, expert assessment, or clinical pilot).
Specific Metrics	The specific quantitative or qualitative metrics (e.g., Accuracy, F1-score, SUS, Latency) used to assess performance.
Evaluation Dimensions	The evaluation dimensions reported in the study, covering effectiveness, efficiency, robustness, fairness, explainability, safety, and usability.

Supplementary Table 3: Basic Information of the Included Studies

ID	Title	Type	Year	Clinical Task
1	A feasibility study of automating radiotherapy planning with large language model agents	Journal	2025	Treatment and Clinical Intervention
2	Ehrgent: Code empowers large language models for few-shot complex tabular reasoning on electronic health records	Conference	2024	Question Answering
3	Colacare: Enhancing electronic health record modeling through large language model-driven multi-agent collaboration	Conference	2025	Disease-related Prediction and Diagnosis
4	Remoni: An autonomous system integrating wearables and multimodal large language models for enhanced remote health monitoring	Conference	2024	Clinical Monitoring and Health Management
5	A Conversational Agent for Early Detection of Neurotoxic Effects of Medications through Automated Intensive Observation	Conference	2024	Clinical Monitoring and Health Management
6	A Multi-agent System Approach with Generative AI for Improved Elderly Daily Living	Conference	2024	Clinical Monitoring and Health Management
7	A Proactive Agent Collaborative Framework for Zero-Shot Multimodal Medical Reasoning	Journal	2025	Clinical Monitoring and Health Management
8	Agent-guided AI-powered interpretation and reporting of nerve conduction studies and EMG (INSPIRE)	Journal	2025	Clinical Documentation and Report Generation
9	AI-Enhanced Virtual Reality Self-Talk for Psychological Counseling: Formative Qualitative Study	Journal	2025	Clinical Monitoring and Health Management
10	ChatExosome: an artificial intelligence (AI) agent based on deep learning of exosomes spectroscopy for hepatocellular carcinoma (HCC) diagnosis	Journal	2025	Disease-related Prediction and Diagnosis
11	Designing VR Simulation System for Clinical Communication Training with LLMs-Based Embodied Conversational Agents	Conference	2025	Medical Education and Training
12	Development of a GPT-4–Powered Virtual Simulated Patient and Communication Training Platform for Medical Students to Practice Discussing Abnormal Mammogram Results With Patients: Multiphase Study	Journal	2025	Medical Education and Training
13	Digital forms for all: A holistic multimodal large language model agent for health data entry	Journal	2024	Clinical Documentation and Report Generation
14	Enhancing Patient Acceptance of Robotic Ultrasound through Conversational Virtual Agent and Immersive Visualizations	Journal	2025	Clinical Monitoring and Health Management
15	Exploring Personalized Health Support through Data-Driven, Theory-Guided LLMs: A Case Study in Sleep Health	Conference	2025	Clinical Monitoring and Health Management
16	From decision to action in surgical autonomy: Multi-modal large language models for robot-assisted blood suction	Journal	2025	Treatment and Clinical Intervention

17	Integrating GPT-Based AI into Virtual Patients to Facilitate Communication Training Among Medical First Responders: Usability Study of Mixed Reality Simulation	Journal	2024	Medical Education and Training
18	Integration of multi-source medical data for medical diagnosis question answering	Journal	2024	Question Answering
19	MAGDA: Multi-agent guideline-driven diagnostic assistance	Conference	2024	Disease-related Prediction and Diagnosis
20	Mdagents: An adaptive collaboration of llms for medical decision-making	Conference	2024	Disease-related Prediction and Diagnosis
21	Medco: Medical education copilots based on a multi-agent framework	Conference	2024	Medical Education and Training
22	MMedAgent: Learning to Use Medical Tools with Multi-modal Agent	Conference	2024	Question Answering
23	Multi-Agent AI System for Adaptive Cognitive Training in Elderly Care	Conference	2025	Clinical Monitoring and Health Management
24	Timecap: Learning to contextualize, augment, and predict time series events with large language model agents	Conference	2025	Disease-related Prediction and Diagnosis
25	Towards AI-Powered Applications: The Development of a Personalised LLM for HRI and HCI	Journal	2025	Clinical Monitoring and Health Management
26	Towards Interpretable Radiology Report Generation via Concept Bottlenecks Using a Multi-agentic RAG	Conference	2025	Clinical Documentation and Report Generation
27	Survey and improvement strategies for gene prioritization with large language models	Journal	2025	Disease-related Prediction and Diagnosis
28	"Show Me How": Benefits and Challenges of Agent-Augmented Counterfactual Explanations for Non-Expert Users	Conference	2025	Disease-related Prediction and Diagnosis
29	An evaluation framework for clinical use of large language models in patient interaction tasks	Journal	2025	Disease-related Prediction and Diagnosis
30	GPT-4-based AI agents—the new expert system for detection of antimicrobial resistance mechanisms?	Journal	2024	Disease-related Prediction and Diagnosis
31	Empathic Grounding: Explorations using Multimodal Interaction and Large Language Models with Conversational Agents	Conference	2024	Clinical Monitoring and Health Management
32	Exploring the Design of LLM-based Agent in Enhancing Self-disclosure Among the Older Adults	Conference	2025	Clinical Monitoring and Health Management
33	Autohealth: Advanced llm-empowered wearable personalized medical butler for parkinson's disease management	Conference	2024	Disease-related Prediction and Diagnosis
34	Llm-based framework for administrative task automation in healthcare	Conference	2024	Disease-related Prediction and Diagnosis
35	GPTCoach: Towards LLM-Based Physical Activity Coaching	Conference	2025	Clinical Monitoring and Health Management

36	Advancements in Allergy Detection and Management: Leveraging YOLOv8 and Large Language Models for Enhanced Diagnostic Accuracy	Conference	2024	Disease-related Prediction and Diagnosis
37	Autonomous Pharmaceutical Care with Large Language Models	Conference	2024	Clinical Monitoring and Health Management

Supplementary Table 4: Study-Level Evidentiary Appraisal and Evaluation Information of the Included Studies

ID	Modality	Data Region (Setting)	Data Accessibility & Realism	Evaluation Strategy (Maturity)	Specific Metrics	Evaluation Dimensions
1	Text, Tabular, Image	Asia, North America (Multi-center)	Mixed (Real-world Clinical)	Automatic (Technical Eval)	Standard radiotherapy dosimetric indices (PTV D95 etc.); Number of optimization/planning iterations; Hallucination-detection accuracy (e.g. TPR) and self-reflection success rate	Effectiveness, Efficiency, Robustness
2	Text, Tabular	North America (Multi-center)	Public (Real-world Clinical)	Automatic (Technical Eval)	Success Rate, Completion Rate; few-shot sample efficiency, execution time, API cost; SR/CR stratified by question complexity	Effectiveness, Efficiency, Robustness
3	Text, Tabular, Time-series	North America, Europe, Asia (Multi-center)	Mixed (Real-world Clinical)	Combination (Expert Assessment)	AUPRC, AUROC, min (+P, Se); Token usage (prompt/output), API calls; sensitivity to #agents and to LLM choices; SHAP feature importance; case-study interpretability; Clinician agreement scores	Effectiveness, Efficiency, Robustness, Explainability, Usability
4	Text, Signal, Time-series, Image, Video	Europe, Asia (Single Center)	Public (Simulated)	Automatic (Technical Eval)	Accuracy, Precision, Recall, F1-score; Overall system latency	Effectiveness, Efficiency
5	Text, Audio	North America (Single Center)	Private (Simulated)	Manual (Expert Assessment)	Task completion rate, Audibility rating; Interaction duration (seconds), System response latency rating; Response sensibility	Effectiveness, Efficiency, Usability
6	Text, Time-series	NA	Public (Non-clinical)	Automatic (Technical Eval)	Rule-based validation; Simulated interaction timing vs activity patterns.	Effectiveness, Usability
7	Text, Image	North America (Single Center)	Public (Real-world Clinical)	Combination (Expert Assessment)	BLEU-1/2/3/4, METEOR, ROUGE-L, CIDEr on DVQA; VQA accuracy for expert models; blinded clinician preference rates; subgroup DVQA metrics by gender & age for bias analysis	Effectiveness, Fairness
8	Text, Tabular, Image	Asia (Single Center)	Private (Real-world Clinical)	Combination (Expert Assessment)	AIGERS Score, Accuracy, Sensitivity, Specificity, Precision, Weighted Accuracy; Semantic Similarity Score	Effectiveness, Usability

9	Text, Audio	Asia (Single Center)	Private (Non-clinical)	Manual (Expert Assessment)	Self-rated session usefulness, Problem-solving helpfulness, Willingness to return	Usability
10	Text, Signal	Asia (Single Center)	Private (Real-world Clinical)	Automatic (Technical Eval)	AUC, Accuracy, Sensitivity, Specificity; Training speed (batches per second), Accuracy vs patch size; Reproducibility of SERS substrate via R6G spectra; Retrieval-augmented answers over an exosome knowledge base (top 5 similar chunks)	Effectiveness, Efficiency, Robustness, Explainability
11	Text, Tabular, Audio	NA	NA	No Formal Evaluation (Prototype Design)	NA	NA
12	Text, Audio	North America (Single Center)	Private (Simulated)	Automatic (Prototype Design)	Qualitative review of VSP realism and SPIKES-consistent responses; Speech transcription, AI feedback generation; Heuristic stress-tests for off-topic/hallucinated replies	Effectiveness, Efficiency, Robustness
13	Text, Audio	North America (Single Center)	Private (Simulated)	Combination (Expert Assessment)	GA completion rate; NASA-TLX workload scores; Inter-rater reliability; System Usability Scale, Self-Consciousness Scale;	Effectiveness, Efficiency, Robustness, Usability
14	Text, Audio	Europe (Single Center)	Private (Simulated)	Combination (Expert Assessment)	HRI Trust Score; STT latency, LLM processing latency, NASA-TLX Score; System Usability Scale	Effectiveness, Efficiency, Usability
15	Text, Tabular, Time-series	North America (Single Center)	Private (Non-clinical)	Combination (Expert Assessment)	Oura ring sleep & activity metrics; Post-system item speed of chatbot; Ease of accessing data, Overall system usability	Effectiveness, Efficiency, Usability
16	Text, Image	North America (Single Center)	Private (Simulated)	Combination (Technical Eval)	Time to suction the actively bleeding pool, Time to suction 50% (T50) and 95% (T95) of the blood; Total tool path length (TTPL); Standard deviations of T50, T95, TTPL across 400 randomized scenes in four environments	Effectiveness, Efficiency, Robustness
17	Text, Audio	European (Single Center)	Private (Simulated)	Manual (Expert Assessment)	Subjective Assessment of Speech System Interfaces (SASSI) usability scores	Usability
18	Text, Tabular, Image	Asia (Single Center)	Public (Real-world Clinical)	Automatic (Technical Eval)	Overall accuracy; Accuracy, ROC-AUC	Effectiveness, Robustness

19	Text, Image	North America (Single Center)	Public (Real-world Clinical)	Automatic (Technical Eval)	F1-score, Precision, Recall; Ablation performance under different negation strategies and agent variants; Qualitative chain-of-thought reasoning	Effectiveness, Robustness, Explainability
20	Text, Image / Video	NA (Single Center)	Public (non-real-world Clinical)	Automatic (Technical Eval)	Accuracy, Pearson correlation; Number of API calls per setting, Average inference time per complexity level; Accuracy under different temperatures, Accuracy vs. number of agents in Group / Adaptive settings	Effectiveness, Efficiency, Robustness
21	Text, Image	Asia (Multi-center)	Public (Real-world Clinical)	Automatic (Technical Eval)	HDE (Holistic Diagnostic Evaluation), SEMA (Semantic Embedding-based Matching Assessment), CASCADE (Coarse And Specific Code Assessment for Diagnostic Evaluation)	Effectiveness
22	Text, Image	Asia, North America, Europe (Multi-center)	Public (Real-world clinical)	Automatic (Technical Eval)	GPT-4 rating score, Relative score, VQA accuracy; Tool selection accuracy; Tool selection accuracy on original tools before vs after adding the Pseudo Tool; Qualitative visual comparison / case studies	Effectiveness, Efficiency, Robustness, Explainability
23	Text, Audio	Europe (Single Center)	Private (Non-clinical)	Manual (Expert Assessment)	5-point Likert ratings, Average overall Likert score; Student's t-test on satisfaction	Effectiveness, Usability
24	Text, Time-Series	North America (Single Center)	Public (Non-clinical)	Automatic (Technical Eval)	F1 Score, AUROC / AUC; Qualitative interpretability case study	Effectiveness, Explainability
25	Text, Image, Tabular, Signal	NA	NA	NA (Prototype Design)	NA	NA
26	Text, Image	Asia (Multi-center)	Public (Real-world clinical)	Automatic (Technical Eval)	Accuracy, Correctness, Clinical Usefulness, Consistency; Model performance under different settings; LLM-as-judge interpretability dimensions	Effectiveness, Robustness, Explainability
27	Text, Genomic	North America, Europe (Multi-center)	Mixed (Real-world clinical)	Automatic (Technical Eval)	Proportion of causal genes captured at rank 1–10, Top-1 causal gene capture rate	Effectiveness,
28	Text, Tabular	North America (Multi-center)	Public (Non-clinical)	Manual (Expert Assessment)	Objective scores and subjective scores; Task completion time; Perceived Understandability of the Augmented Counterfactuals; NASA-TLX	Effectiveness, Efficiency, Explainability, Usability

29	Text, Image	North America (NA)	Public (non-real-world Clinical)	Combination (Technical Eval)	Diagnostic accuracy, Expert-rated percentage; Number of simulated conversations processed and wall-clock time, Estimated human-hours saved versus human-based evaluation; Grader-AI agreement with medical experts	Effectiveness, Efficiency, Robustness
30	Text, Image, Tabular	Europe (Single Center)	Public (Real-world clinical)	Automatic (Technical Eval)	Sensitivity, Specificity, Concordance rate; Concordance rates across three human microbiologists and three independent EUCAST-GPT-expert prompting rounds; Median number of words for reasoning and suggested confirmation steps	Effectiveness, Robustness, Explainability
31	Text, Video, Audio	North America (Single Center)	Private (Simulated)	Manual (Expert Assessment)	Active-Empathic Listening Scale, Responsiveness of Other Scale; GS-ANIM, GS-LIKE, General Agent Rating items	Effectiveness, Usability
32	Text, Image, Audio	Asia (Single Center)	Private (non real-world Clinical)	Manual (Expert Assessment)	Self-Disclosure Level Rating Scale, Length of disclosed content; 7-point Likert scale (inclination to communicate with the agent, interface preference)	Effectiveness, Usability
33	Text, Audio, Signal	North America, Europe (Multi-center)	Public (Real-world clinical)	Automatic (Technical Eval)	Classification accuracy, Confusion matrix	Effectiveness
34	Text, Tabular	North America (Single Center)	Mixed (Simulated)	Automatic (Technical Eval)	Step Success Rate (SR), Reward, Tool-use score; Average Steps	Effectiveness, Efficiency
35	Text, Time-series	North America (Single Center)	Private (non real-world Clinical)	Combination (Expert Assessment)	5-point Likert questions on user experience & advice quality; SASSI Speed factor score; SASSI overall usability score	Effectiveness, Efficiency, Usability
36	Text, Image, Tabular	Asia (Multi-center)	Public (Real-world clinical)	Automatic (Technical Eval)	Precision, Recall, F1-score	Effectiveness
37	Text, Tabular, Time-series	Asia (Single Center)	Private (Real-world clinical)	Manual (Expert Assessment)	0–5 PCNE expert score (Accuracy, Comprehensiveness; Practicality, Redundancy; Safety; Patient Compliance)	Effectiveness, Efficiency, Safety, Usability

Note: Data Region (Setting) reports the geographic region(s) and whether the study used a single-center or multi-center setting, as reported by the authors. Data Accessibility & Realism summarizes data provenance (public/private/mixed) and whether datasets were real-world clinical vs. simulated/non-clinical.

Supplementary Table 5: Technical Information of the Included Studies

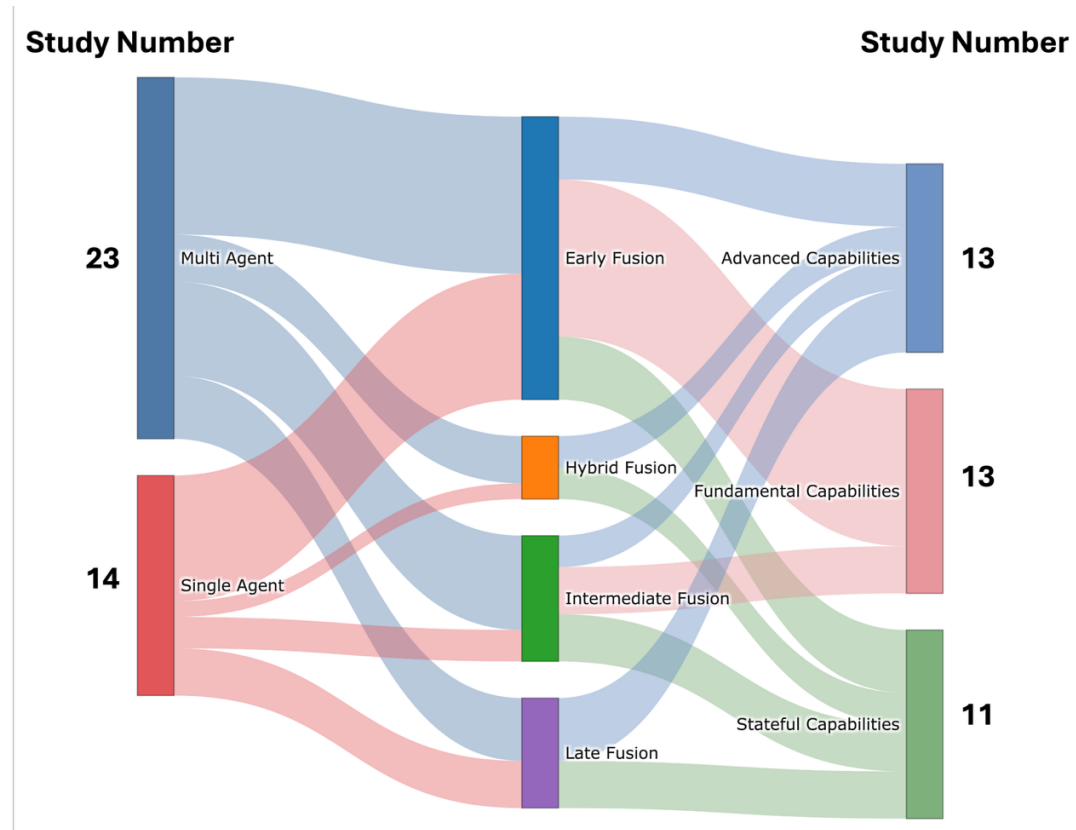
ID	Backbone	Agent	Collaboration	Fusion	Tools	Capability	Training
1	GPT-4o	Multi	Centralized	Early	Y	P/PI/A/M	No Training
2	GPT-4	Single	NA	Late	Y	P/PI/A/M/E	Few-shot Prompting
3	DeepSeek-V2.5	Multi	Centralized	Hybrid	Y	P/A/M	Auxiliary-Model Training
4	GPT-4V	Single	NA	Intermediate	Y	P/A/M	Auxiliary-Model Training
5	Blenderbot-400M-distill	Single	NA	Early	Y	P/A	No Training
6	LLaMA2, Mixtral 8×7B	Multi	Centralized	Late	Y	P/PI/A/M	No Training
7	GPT-4	Multi	Centralized	Late	Y	P/PI/A	Auxiliary-Model Training
8	Gemini 1.5-Pro	Multi	Centralized	Intermediate	Y	P/A/M	No Training
9	GPT-3.5	Single	NA	Early	Y	P/A	No Training
10	GLM-4	Single	NA	Late	Y	P/A/M	Auxiliary-Model Training
11	GPT-4o	Multi	Centralized	Early	Y	P/A/M	In-Context Learning
12	GPT-4	Multi	Centralized	Early	Y	P/A	No Training
13	GPT-4	Single	NA	Early	Y	P/A	No Training
14	Phi-3 Mini	Single	NA	Early	Y	P/A	No Training
15	GPT-4o, GPT-4o-mini	Multi	Centralized	Hybrid	Y	P/A/M/E	Auxiliary-Model Training
16	GPT-4V	Multi	Centralized	Early	Y	P/PI/A	Auxiliary-Model Training
17	GPT-3.5 Turbo	Single	NA	Early	Y	P/A	Auxiliary-Model Training
18	GPT-4V	Multi	Centralized	Intermediate	N	P/PI/A	No Training
19	Mixtral 8×7B	Multi	Centralized	Intermediate	Y	P/A	No Training
20	GPT-4V / Gemini-Pro(Vision)	Multi	Hybrid	Late	Y	P/PI/A/M	No Training
21	GPT-3.5, Claude-3.5-Sonnet	Multi	Centralized	Early	Y	P/PI/A/M/E	In-Context Learning
22	LLaVA-Med	Single	NA	Hybrid	Y	P/PI/A/M	LoRA Finetuning
23	GPT-4o	Multi	Centralized	Early	Y	P/A/M	No Training
24	GPT-4	Multi	Centralized	Hybrid	Y	P/A/M	Auxiliary-Model Training
25	LLM*	Single	NA	Early	Y	P/A	Online Adaptation
26	ChexAgent	Multi	Centralized	Intermediate	Y	P/A/M	Auxiliary-Model Training

27	GPT-4	Multi	Centralized	Early	Y	P/A	No Training
28	GPT-4 Turbo	Single	NA	Early	Y	P/A	Auxiliary-Model Training
29	GPT-4V	Multi	Centralized	Early	N	P/A	No Training
30	GPT-4	Single	NA	Early	N	P/A/M	No Training
31	GPT-3.5 Turbo	Single	NA	Intermediate	Y	P/A	No Training
32	GPT-4	Multi	Centralized	Intermediate	Y	P/A	No Training
33	LLM*	Single	NA	Late	Y	P/A/M	Auxiliary-Model Training
34	GPT-4V, GPT-3.5	Multi	Centralized	Early	N	P/A/E	No Training
35	GPT-4	Multi	Centralized	Intermediate	Y	P/PI/A	No Training
36	LLM*	Multi	Centralized	Late	Y	P/A/M	Auxiliary-Model Training
37	Qwen-7B	Multi	Centralized	Early	Y	P/PI/A/M	No Training

LLM*. The exact model version (e.g., GPT-4, LLaMA, or variants) was not explicitly stated in the original paper.

P = Perception, PI = Planning, A = Action, M = Memory, E = Evolution.

Supplementary Figure 1: Sankey Diagram of System Design Patterns



This Sankey diagram illustrates the quantitative distribution of design choices across the 37 included studies. Agent capabilities are grouped into three levels of functional complexity (categorized based on the highest level of capability):

1. **Fundamental (Perception & Action):** Baseline capability to process inputs and execute responses without retaining long-term context.
2. **Stateful (Memory):** Ability to maintain context across interactions or retrieve external knowledge (e.g., via RAG).
3. **Advanced (Planning / Evolution):** Higher-level autonomy involving explicit task decomposition or continuous self-improvement mechanisms.

Supplementary Table 6: Preferred Reporting Items for Systematic reviews and Meta-Analyses extension for Scoping Reviews (PRISMA-ScR) Checklist

SECTION	ITEM	PRISMA-ScR CHECKLIST ITEM	REPORTED ON PAGE #
TITLE			
Title	1	Identify the report as a scoping review.	P1
ABSTRACT			
Structured summary	2	Provide a structured summary that includes (as applicable): background, objectives, eligibility criteria, sources of evidence, charting methods, results, and conclusions that relate to the review questions and objectives.	P1, P2
INTRODUCTION			
Rationale	3	Describe the rationale for the review in the context of what is already known. Explain why the review questions/objectives lend themselves to a scoping review approach.	P2, P3
Objectives	4	Provide an explicit statement of the questions and objectives being addressed with reference to their key elements (e.g., population or participants, concepts, and context) or other relevant key elements used to conceptualize the review questions and/or objectives.	P2, P3
METHODS			
Protocol and registration	5	Indicate whether a review protocol exists; state if and where it can be accessed (e.g., a Web address); and if available, provide registration information, including the registration number.	P27
Eligibility criteria	6	Specify characteristics of the sources of evidence used as eligibility criteria (e.g., years considered, language, and publication status), and provide a rationale.	P27, P28
Information sources*	7	Describe all information sources in the search (e.g., databases with dates of coverage and contact with authors to identify additional sources), as well as the date the most recent search was executed.	P27
Search	8	Present the full electronic search strategy for at least 1 database, including any limits used, such that it could be repeated.	Supplementary Table 1
Selection of sources of evidence†	9	State the process for selecting sources of evidence (i.e., screening and eligibility) included in the scoping review.	P27, P28
Data charting process‡	10	Describe the methods of charting data from the included sources of evidence (e.g., calibrated forms or forms that have been tested by the team before their use, and whether data charting was done independently or in duplicate) and any processes for obtaining and confirming data from investigators.	P28, P29, Supplementary Table 2
Data items	11	List and define all variables for which data were sought and any assumptions and simplifications made.	P28, P29, Supplementary Table 2

SECTION	ITEM	PRISMA-ScR CHECKLIST ITEM	REPORTED ON PAGE #
Critical appraisal of individual sources of evidence§	12	If done, provide a rationale for conducting a critical appraisal of included sources of evidence; describe the methods used and how this information was used in any data synthesis (if appropriate).	N/A
Synthesis of results	13	Describe the methods of handling and summarizing the data that were charted.	P29, P30
RESULTS			
Selection of sources of evidence	14	Give numbers of sources of evidence screened, assessed for eligibility, and included in the review, with reasons for exclusions at each stage, ideally using a flow diagram.	P8, P27 (Fig. 7)
Characteristics of sources of evidence	15	For each source of evidence, present characteristics for which data were charted and provide the citations.	Supplementary Table 3-5
Critical appraisal within sources of evidence	16	If done, present data on critical appraisal of included sources of evidence (see item 12).	N/A
Results of individual sources of evidence	17	For each included source of evidence, present the relevant data that were charted that relate to the review questions and objectives.	Supplementary Table 3-5
Synthesis of results	18	Summarize and/or present the charting results as they relate to the review questions and objectives.	P6-20
DISCUSSION			
Summary of evidence	19	Summarize the main results (including an overview of concepts, themes, and types of evidence available), link to the review questions and objectives, and consider the relevance to key groups.	P21-24
Limitations	20	Discuss the limitations of the scoping review process.	P26
Conclusions	21	Provide a general interpretation of the results with respect to the review questions and objectives, as well as potential implications and/or next steps.	P24, P25
FUNDING			
Funding	22	Describe sources of funding for the included sources of evidence, as well as sources of funding for the scoping review. Describe the role of the funders of the scoping review.	P30

JB1 = Joanna Briggs Institute; PRISMA-ScR = Preferred Reporting Items for Systematic reviews and Meta-Analyses extension for Scoping Reviews.

* Where *sources of evidence* (see second footnote) are compiled from, such as bibliographic databases, social media platforms, and Web sites.

† A more inclusive/heterogeneous term used to account for the different types of evidence or data sources (e.g., quantitative and/or qualitative research, expert opinion, and policy documents) that may be eligible in a scoping review as opposed to only studies. This is not to be confused with *information sources* (see first footnote).

‡ The frameworks by Arksey and O'Malley (6) and Levac and colleagues (7) and the JBI guidance (4, 5) refer to the process of data extraction in a scoping review as data charting.

§ The process of systematically examining research evidence to assess its validity, results, and relevance before using it to inform a decision. This term is used for items 12 and 19 instead of "risk of bias" (which is more applicable to systematic reviews of interventions) to include and acknowledge the various sources of evidence that may be used in a scoping review (e.g., quantitative and/or qualitative research, expert opinion, and policy document).

Supplementary Note 1: Flaticon icon attributions

- Icon made by HAJICON from [www.flaticon.com](https://www.flaticon.com/free-icon/exam_18753090) (URL: https://www.flaticon.com/free-icon/exam_18753090)
- Icon made by Flat Icons from [www.flaticon.com](https://www.flaticon.com/free-icon/testing_2721288) (URL: https://www.flaticon.com/free-icon/testing_2721288)
- Icon made by Awicon from [www.flaticon.com](https://www.flaticon.com/free-icon/surgery-room_17774355) (URL: https://www.flaticon.com/free-icon/surgery-room_17774355)
- Icon made by Freepik from [www.flaticon.com](https://www.flaticon.com/free-icon/heart-monitoring_4521407) (URL: https://www.flaticon.com/free-icon/heart-monitoring_4521407)
- Icon made by iconixar from [www.flaticon.com](https://www.flaticon.com/free-icon/article_4698245) (URL: https://www.flaticon.com/free-icon/article_4698245)
- Icon made by kliwir art from [www.flaticon.com](https://www.flaticon.com/free-icon/education_15192960) (URL: https://www.flaticon.com/free-icon/education_15192960)
- Icon made by Parzival' 1997 from [www.flaticon.com](https://www.flaticon.com/free-icon/robot-assistant_10242676) (URL: https://www.flaticon.com/free-icon/robot-assistant_10242676)
- Icon made by wanicon from [www.flaticon.com](https://www.flaticon.com/free-icon/nurse_1998233) (URL: https://www.flaticon.com/free-icon/nurse_1998233)
- Icon made by wanicon from [www.flaticon.com](https://www.flaticon.com/free-icon/nurse_1998122) (URL: https://www.flaticon.com/free-icon/nurse_1998122)
- Icon made by wanicon from [www.flaticon.com](https://www.flaticon.com/free-icon/nurse_1998194) (URL: https://www.flaticon.com/free-icon/nurse_1998194)
- Icon made by HAJICON from [www.flaticon.com](https://www.flaticon.com/free-icon/retrieval_17404092) (URL: https://www.flaticon.com/free-icon/retrieval_17404092)
- Icon made by Uniconlabs from [www.flaticon.com](https://www.flaticon.com/free-icon/speech-synthesis_13063588) (URL: https://www.flaticon.com/free-icon/speech-synthesis_13063588)
- Icon made by Cap Cool from [www.flaticon.com](https://www.flaticon.com/free-icon/api_11681345) (URL: https://www.flaticon.com/free-icon/api_11681345)
- Icon made by wanicon from [www.flaticon.com](https://www.flaticon.com/free-icon/calories-calculator_2316953) (URL: https://www.flaticon.com/free-icon/calories-calculator_2316953)
- Icon made by Freepik from [www.flaticon.com](https://www.flaticon.com/free-icon/prevention_4327809) (URL: https://www.flaticon.com/free-icon/prevention_4327809)
- Icon made by Uniconlabs from [www.flaticon.com](https://www.flaticon.com/free-icon/medical-book_10620372) (URL: https://www.flaticon.com/free-icon/medical-book_10620372)
- Icon made by Freepik from [www.flaticon.com](https://www.flaticon.com/free-icon/deep-learning_8637113) (URL: https://www.flaticon.com/free-icon/deep-learning_8637113)
- Icon made by Freepik from [www.flaticon.com](https://www.flaticon.com/free-icon/note_4021693) (URL: https://www.flaticon.com/free-icon/note_4021693)
- Icon made by xnimrodx from [www.flaticon.com](https://www.flaticon.com/free-icon/excel_3382111) (URL: https://www.flaticon.com/free-icon/excel_3382111)
- Icon made by Smashicons from [www.flaticon.com](https://www.flaticon.com/free-icon/stopwatch_2738227) (URL: https://www.flaticon.com/free-icon/stopwatch_2738227)
- Icon made by Freepik from [www.flaticon.com](https://www.flaticon.com/free-icon/dna-structure_4157983) (URL: https://www.flaticon.com/free-icon/dna-structure_4157983)
- Icon made by Shuvo.Das from [www.flaticon.com](https://www.flaticon.com/free-icon/audio_10434673) (URL: https://www.flaticon.com/free-icon/audio_10434673)
- Icon made by Iconjam from [www.flaticon.com](https://www.flaticon.com/free-icon/biomedical_14784300) (URL: https://www.flaticon.com/free-icon/biomedical_14784300)
- Icon made by Victoruler from [www.flaticon.com](https://www.flaticon.com/free-icon/x-ray_2118468) (URL: https://www.flaticon.com/free-icon/x-ray_2118468)
- Icon made by Flowicon from [www.flaticon.com](https://www.flaticon.com/free-icon/video-lesson_6174998) (URL: https://www.flaticon.com/free-icon/video-lesson_6174998)
- Icon made by Freepik from [www.flaticon.com](https://www.flaticon.com/free-icon/medical-report_10154472) (URL: https://www.flaticon.com/free-icon/medical-report_10154472)
- Icon made by Prosymbols Premium from [www.flaticon.com](https://www.flaticon.com/free-icon/ecg-monitor_18265449) (URL: https://www.flaticon.com/free-icon/ecg-monitor_18265449)
- Icon made by Sir. Vector from [www.flaticon.com](https://www.flaticon.com/free-icon/medical_11916882) (URL: https://www.flaticon.com/free-icon/medical_11916882)
- Icon made by Parzival' 1997 from [www.flaticon.com](https://www.flaticon.com/free-icon/health_10242634) (URL: https://www.flaticon.com/free-icon/health_10242634)
- Icon made by juicy_fish from [www.flaticon.com](https://www.flaticon.com/free-icon/xray_10476427) (URL: https://www.flaticon.com/free-icon/xray_10476427)
- Icon made by phatplus from [www.flaticon.com](https://www.flaticon.com/free-icon/microscope_3035512) (URL: https://www.flaticon.com/free-icon/microscope_3035512)

- Icon made by Freepik from www.flaticon.com (URL: https://www.flaticon.com/free-icon/eeg_9851782)
- Icon made by Iconjam from www.flaticon.com (URL: https://www.flaticon.com/free-icon/functional-genomics_17791485)
- Icon made by Uniconlabs from www.flaticon.com (URL: https://www.flaticon.com/free-icon/data-gathering_8439019)
- Icon made by Eucalyp from www.flaticon.com (URL: https://www.flaticon.com/free-icon/technology_4365271)
- Icon made by Eucalyp from www.flaticon.com (URL: https://www.flaticon.com/free-icon/statistics_1459073)
- Icon made by Freepik from www.flaticon.com (URL: https://www.flaticon.com/free-icon/stamp_5454576)
- Icon made by Rooman12 from www.flaticon.com (URL: https://www.flaticon.com/free-icon/technology_12070801)
- Icon made by Elzicon from www.flaticon.com (URL: https://www.flaticon.com/free-icon/read-email_12478650)