Chao Ding

Email: dingchaochao58@gmail.com | Google Scholar |Homepage| Location: Shanghai, China

PUBLICATIONS

** Equal contribution / co-first authorship.*

Medical AI / Clinical AI

1. **Ding, C.*,** et al.
   Conversational Fluency Does Not Ensure Clinical Safety in Patient-Facing Medical AI. Manuscript under review at ***NEJM AI***.
2. **Ding, C.***, Bian, M.*, Yuan, M.*, et al.
   Advancing medical AI through benchmarking and competition for specialty triage. PDF DOI
   npj Digital Medicine, IF: 15.1, 2026.
3. **Ding, C.***, et al.
   SafeMed-R1: Clinician-Audited Safety and Ethics Alignment for Medical Large Language Models.
   arXiv:2605.28338 arXiv preprint, 2026.
4. Jiang, Y.*, **Ding, C.***, et al.
   Med-GRADE: Medical Grading and Rubric-based Assessment of Doctor-Patient Encounters.
   Manuscript in submission to EMNLP 2026

AI Agents / Multimodal Reasoning

5. Ding, J.*, **Ding, C.***, Jiang, Y.*, et al.
   Beyond Knowledge to Agency: Evaluating Expertise, Autonomy, and Integrity in Finance with CNFinBench. Project Code
   KDD 2026, accepted.
6. Jiang, Y.*, Li, Q.*, Xu, B.*, Sun, H., **Ding, C**., et al.
   IBISAgent: Reinforcing Pixel-Level Visual Reasoning in MLLMs for Universal Biomedical Object Referring and Segmentation.
   arXiv:2601.03054 CVPR 2026, accepted.

Clinical Epidemiology / Predictive Modeling

7. **Ding, C.*,** Lu, R.*, Kong, Z., Huang, R.
   TyG index, depression, and cognitive dysfunction: NHANES with machine learning support. PDF DOI
   Journal of Affective Disorders, IF: 4.9, 2025.
8. **Ding, C.***, Yuan, M.*, Cheng, J., Wen, J.
   Smoking types and stroke risk: development of a predictive model for identifying stroke risk. PDF DOI
   Frontiers in Physiology, IF: 3.4, 2025.
9. **Ding, C.,** Kong, Z., Cheng, J., Huang, R.
   U-shaped relationship between TyG index and depression using machine learning. PDF DOI
   Heliyon, IF: 3.6, 2024.

CLINICAL TRAINING

Putuo District Central Hospital, Shanghai University of Traditional Chinese Medicine | Standardized Residency Trainee *Jul. 2022 – Jul. 2025*

- Completed multi-year standardized residency training across emergency medicine, cardiology, respiratory medicine, pediatrics, obstetrics, neurology, and related departments.
- Participated in outpatient, inpatient, and emergency-care settings, gaining first-hand experience in history taking, physical examination, preliminary assessment, differential diagnosis, patient communication, and referral decision-making.
- Managed and observed high-risk clinical presentations including chest pain, acute dyspnea, obstetric bleeding, pediatric fever, trauma, infection, and neurologic emergencies, strengthening clinical judgment in red-flag recognition and time-sensitive escalation.
- Developed a clinical understanding of early-stage triage bottlenecks, including incomplete patient narratives, limited consultation time, uncertainty in referral thresholds, and communication gaps between patient symptoms and clinical risk.

RESEARCH EXPERIENCE

Shanghai Artificial Intelligence Laboratory | Research Assistant *Dec. 2024 – Present*

Project 1: Patient-Facing Multi-agent | *arXiv / Manuscript under review at* ***NEJM AI*** | *Dec. 2025– May. 2026*

- Built a simulated patient–doctor evaluation environment integrating patient simulators, frontier LLM doctor backbones, an automated clinical judge, and a SafeMed-R1 in-loop safety controller.
- Formalized delayed escalation as a temporal safety failure and evaluated 5 LLMs across 150 multi-turn consultations, using first-escalation turn and missed-escalation rate against a 40-case clinician reference benchmark.

• Implemented a five-action controller—PASS, REWRITE, ASK-MORE, ESCALATE, REFUSE—reducing mean escalation turn from 6.28 to 3.78 rounds and missed escalation from 62.5% to 12.5%, while revealing GPT-5.1 over-intervention behavior.

Project 2: SafeMed-R1: Medical LLMs | arXiv:2605.28338 Code *Dec. 2024 – Dec. 2025*

- Developed SafeMed-R1, a safety- and ethics-aligned medical LLM based on Qwen3-32B, using SFT and RL/GRPO with clinician-audited reasoning traces, safety/ethics supervision, and red-team jailbreak stress testing.
- Built the Clinical Trust Signals pipeline for model supervision, including expert-reviewed QA-CoT construction, rubric scoring, adversarial re-answering checks, and benchmark evaluation; achieved 79.6% macro-averaged clinical accuracy and reduced unsafe outputs under adversarial testing.
- Deployed SafeMed-R1 as both a standalone aligned medical model and a modular turn-level governance model for patient-facing dialogue, connecting training-time safety alignment with real-time escalation control.

Project 3: MedBench / MedTriage: Benchmarking and Specialty Triage | PDF DOI accepted by ***npj Digital Medicine* /** *Dec. 2024 – April.2025*

- Built MedTriage from hospital intake records, online guidance dialogues, and outpatient clinical notes; designed leakage-prevention cleaning and strict multi-label exact-match department recommendation.
- Developed MedGPT-Guide, a retrieval-augmented triage model using BGE-m3 retrieval, 10 relevant + 10 random demonstrations, CoT prompting, candidate-order perturbation, self-consistency voting, and ensemble aggregation; improved accuracy from 0.5319 to 0.7830.
- Contributed to MedBench infrastructure for LLMs Homepage, a cloud-based evaluation infrastructure spanning 700k+ expert-curated tasks, 24 primary and 91 secondary specialties, and dedicated tracks for LLMs, multimodal models, and clinical agents..

Project 4: multi-agent / agentic Benchmark| Project Code *KDD 2026, accepted* *Dec. 2025 – April.2025*

Expertise, Autonomy, and Integrity Evaluation for High-Stakes Financial LLM Agents

- Co-developed CNFinBench, evaluating high-stakes financial agents across expertise, autonomy, and integrity with 29 subtasks, 11,947 single-turn QA instances, and 321 multi-turn adversarial dialogues.
- Designed agentic workflow evaluation covering requirement parsing, path planning, API / database operations, tool invocation, multi-agent collaboration, and result verification; contributed to Chain-of-Attack evaluation with 17 attacker personas and 7 attack strategies.
- Helped develop HICS to quantify behavioral compliance drift across 22 open- and closed-source models, revealing execution-chain degradation and rapid multi-turn attack collapse.

EDUCATION

Shanghai University of Traditional Chinese Medicine | Master of Medicine, *expected Dec 2026*

Relevant coursework: Medical Statistics; Research Methodology and Scientific Writing; Immunological Techniques; Medical English; Clinical Practice

Jiangxi University of Traditional Chinese Medicine | Bachelor of Medicine, *2015-2020*

Training in anatomy, physiology, pathology, pharmacology, medical imaging, immunology, genetics, biochemistry, microbiology, and cell biology.

HONORS & AWARDS

Provincial	• *3rd Prize*, 18th Challenge Cup Shanghai Collegiate Extracurricular Academic Science & Technology Works Competition	*2023/09*
University-Wide	• Yifang Innovation Award*, 3rd Prize*	*2026/05*
	• Academic Scholarship, Shanghai University of Traditional Chinese Medicine	*2023/09*