

Reimagining Pain Management with AI: Opportunities and Risks of Large Language Models

Prof. Anand Rao

Distinguished Service Professor of Applied Analytics and AI
Heinz College of Information Systems and Public Policy
Carnegie Mellon University

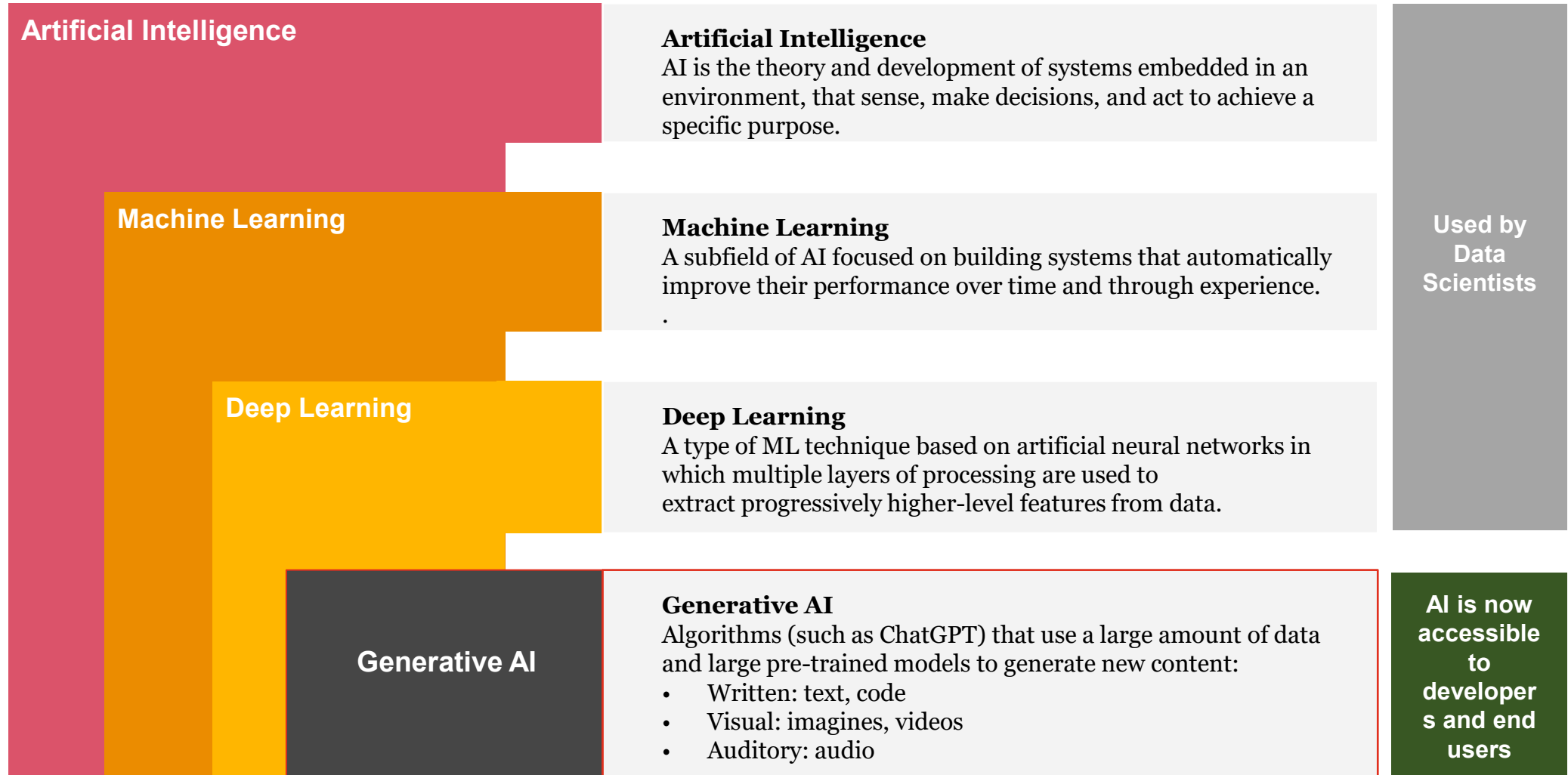
Email: anandr2@andrew.cmu.edu

LinkedIn: <https://www.linkedin.com/in/anandsrao/>

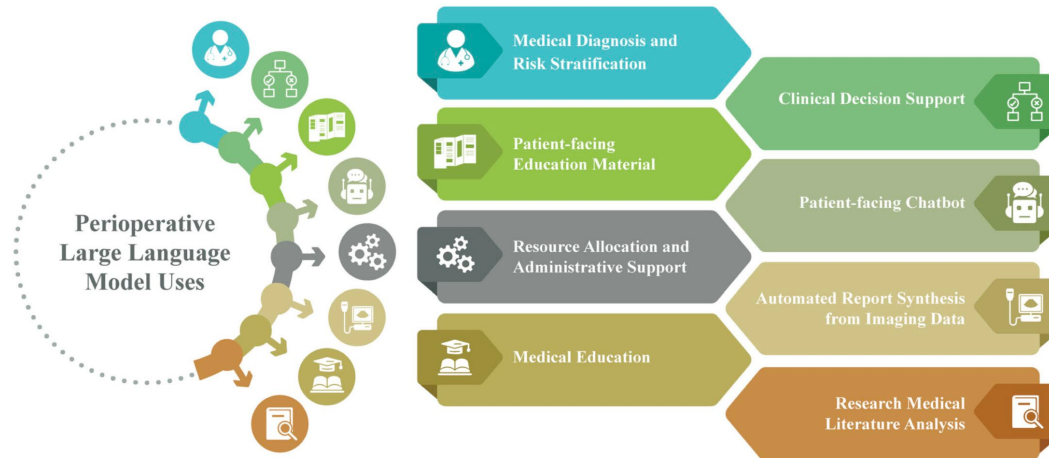
Agenda

- From AI to Generative AI to Agents
- Why now & where LLMs help
- Evidence from pain & peritop
- Safety playbook: evaluate + guardrails
- What's next & how to start

From AI to Generative AI



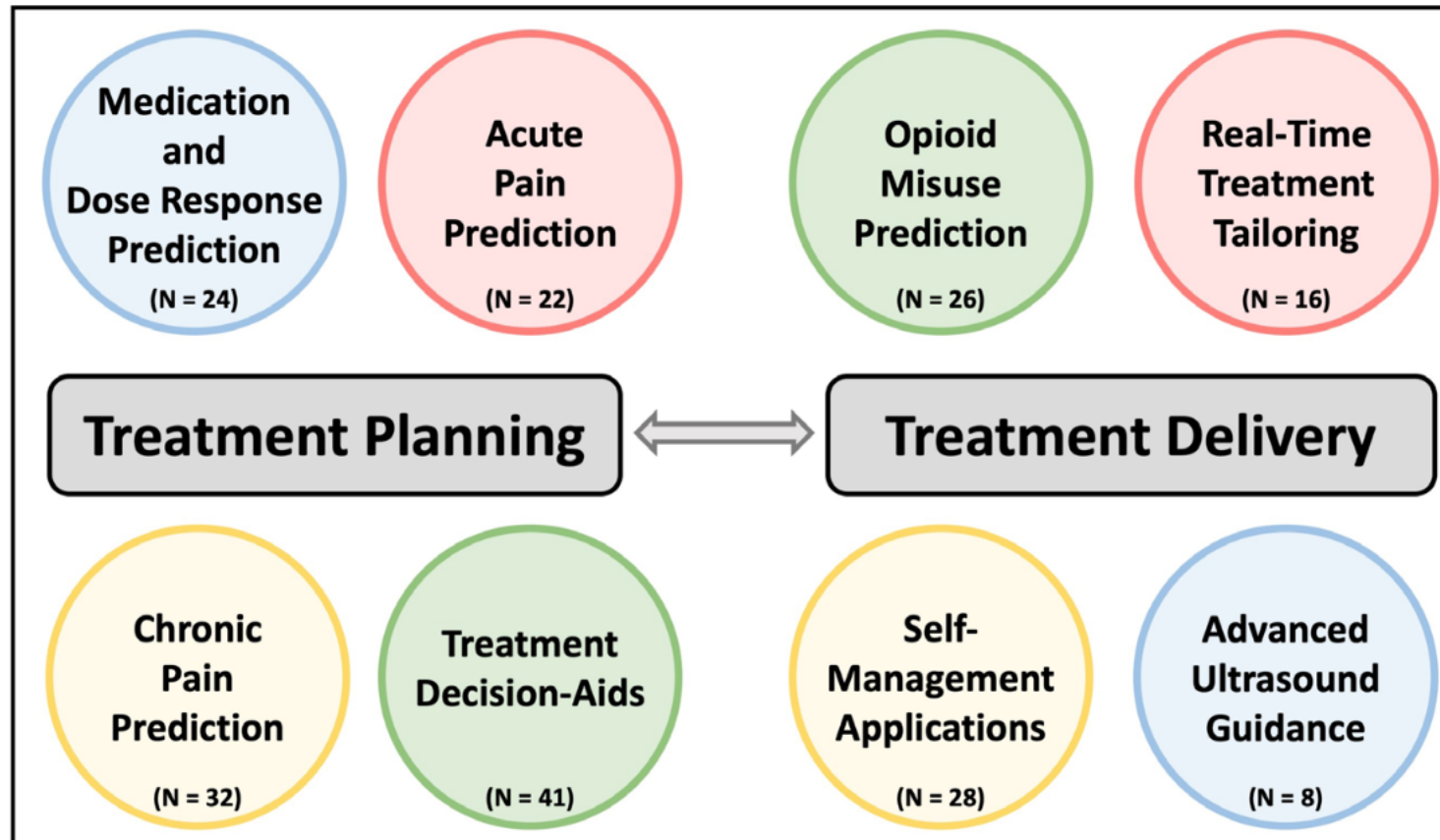
Why Now: Pain, Workflows, Readiness



- **Documentation burden; narrative signal unlocked** — Pain histories are long and fragmented; LLM summaries reclaim time and surface key descriptors while preserving clinician judgment, empathy, and procedural focus.
- **Intake to structured, searchable summary** — Convert free-text narratives into timelines, medication lists, coded problems, and red-flag citations, enabling faster retrieval and consistent documentation across rotating teams.
- **Surface safety and workflow flags early** — Automatically highlight anticoagulation, chronic steroid use, infection risk, sleep-apnea concerns, and prior complications to streamline pre-procedure planning and consent discussions.
- **Augmentation with protocols and oversight** — Position LLMs as co-pilots grounded in local guidelines, with human review and audit trails—never autonomous orders or unsupervised patient messaging.

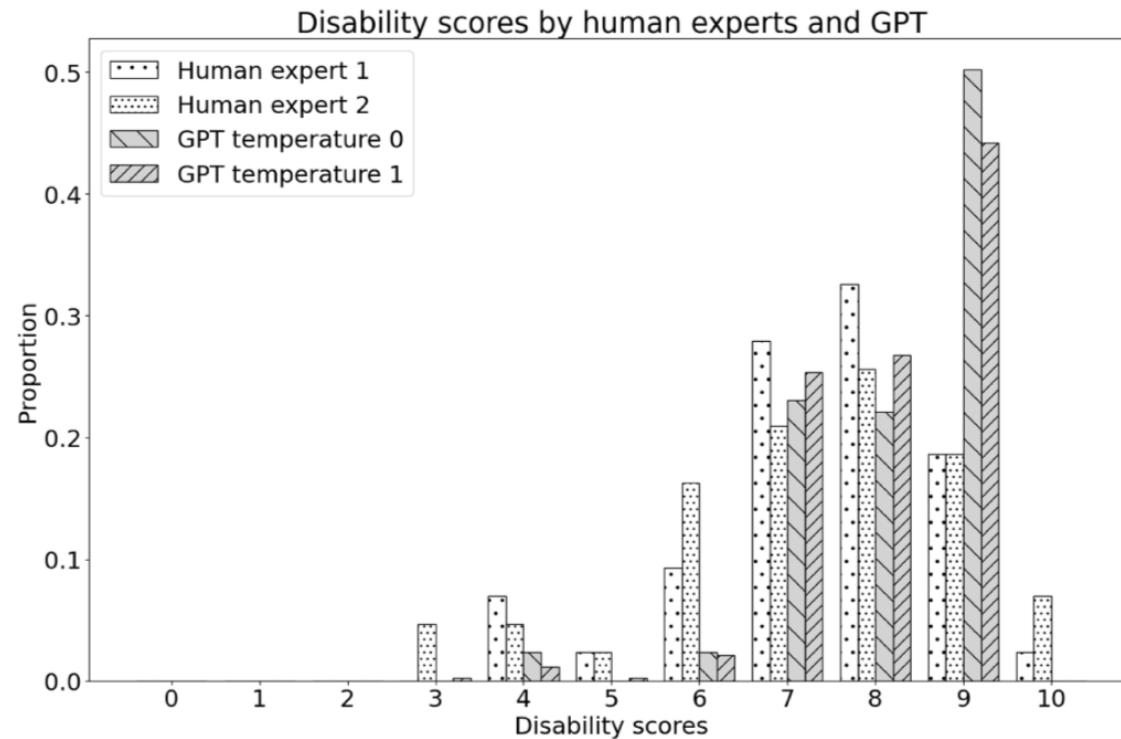
Source: Hondjeu, A. R. M., et al. "[Large Language Models in Perioperative Medicine: Opportunities and Challenges](#)." *Can J Anesth*, 2025. Ke, Y. H., et al. "[Real-world Deployment and Evaluation of PEACH](#)." *arXiv preprint*, 2024

AI in Pain Medicine: Planning to Precision Delivery



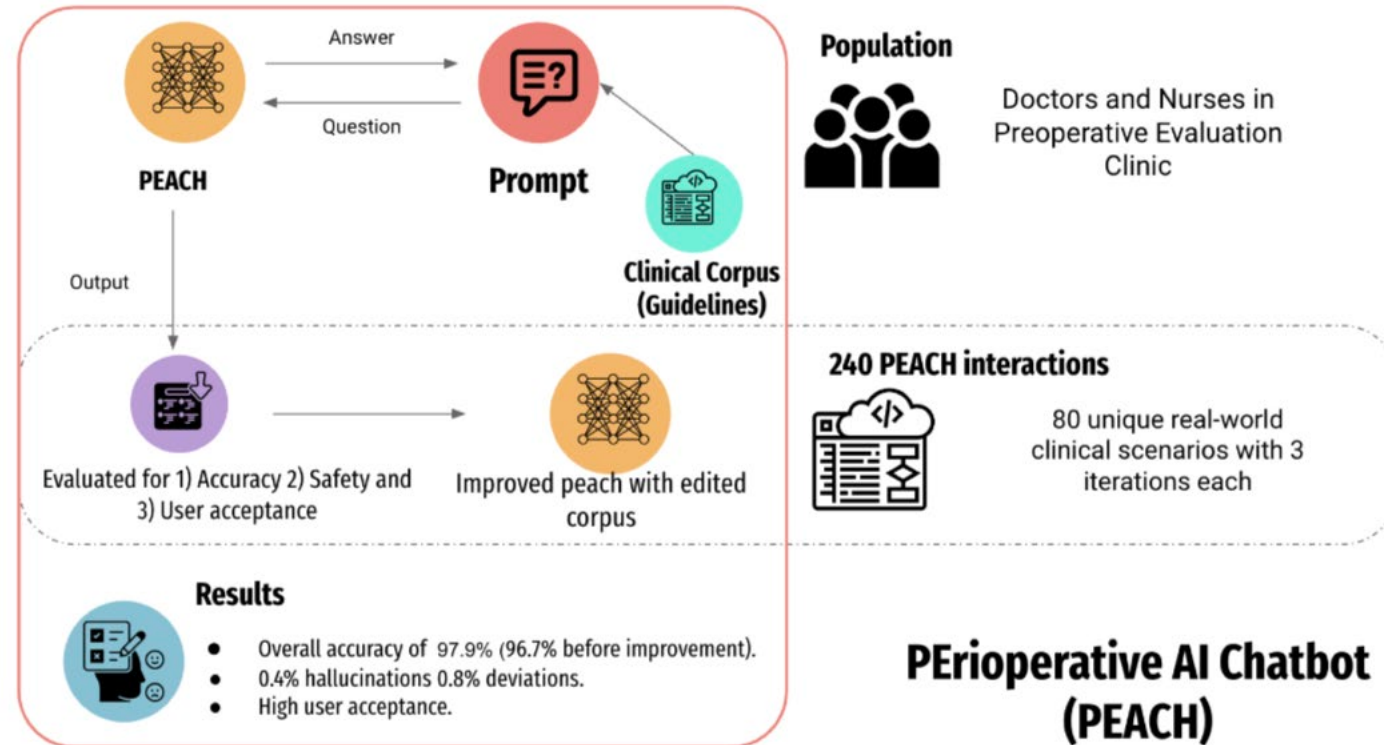
- AI supports both **upstream planning** (e.g. predicting pain or misuse) and **downstream delivery** (e.g. real-time tailoring)
- Most common use: **AI-based decision aids** (n = 41) and **chronic pain prediction** (n = 32)
- Emphasis on **personalization** and reducing opioid reliance

Do LLMs Match Experts? (Pain Narratives)



- **Feasibility with small sample** — 43 fibromyalgia narratives (Spanish); GPT-4 scored severity/disability against two experts using predefined prompts and outputs.
- **High weighted agreement values** — Weighted agreement ≈ 0.94 – 0.95 ; Gwet AC2 ≈ 0.79 – 0.84 ; Krippendorff's α 0.45 – 0.57 vs experts (below inter-expert α).
- **Errors small and significant** — RMSE ≈ 1.20 (severity) and 1.44 (disability); significantly lower than naïve baseline ($P < .001/.01$), indicating non-trivial predictive value.
- **Slight, consistent overestimation** — Mean overestimation 0.66 – 0.83 (severity) and 0.25 – 0.37 (disability); treat as **method signal**, not practice-changing evidence.

Perioperative Co-Pilot: Workflow + Evidence

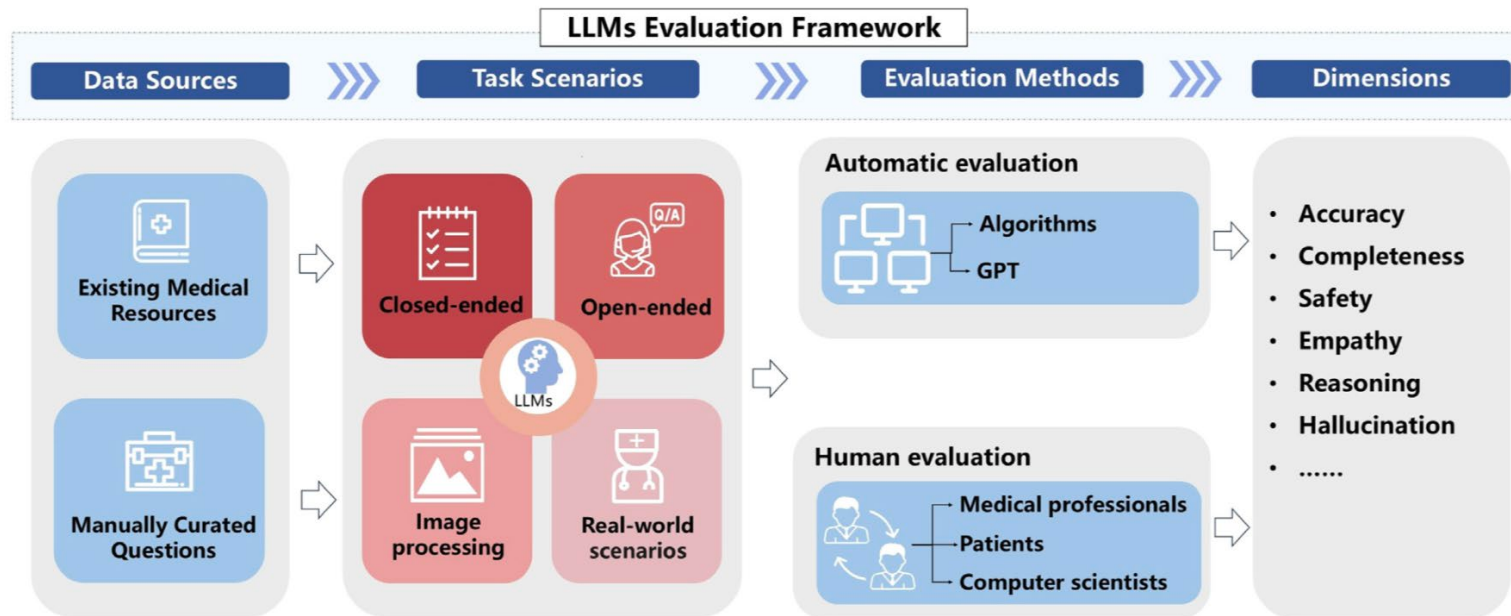


Accuracy 97.9% | Hallucination 0.4% | Deviation 0.8% | 10–15 s |
95% CI 0.952–0.991 | $p=0.018$.

- **Protocol-grounded, human-in-the-loop** — Clinicians finalize decisions first; chatbot answers are constrained to local guidelines with inline sources inside a secure enclave.
- **Strong accuracy and safety profile** — After a minor protocol fix, observed accuracy 97.9% with 0.4% hallucinations, 0.8% deviations, and typical 10–15-second response time.
- **Robust human comparison** — Planned $n \geq 73$; executed 80 questions/240 iterations; agreement with attendings ~91%; accuracy tested above a 95% null ($p=0.018$).
- **What to copy Monday** — Ground in your protocols, log everything, require sign-off, and grade outputs weekly for accuracy, deviations, and hallucinations.

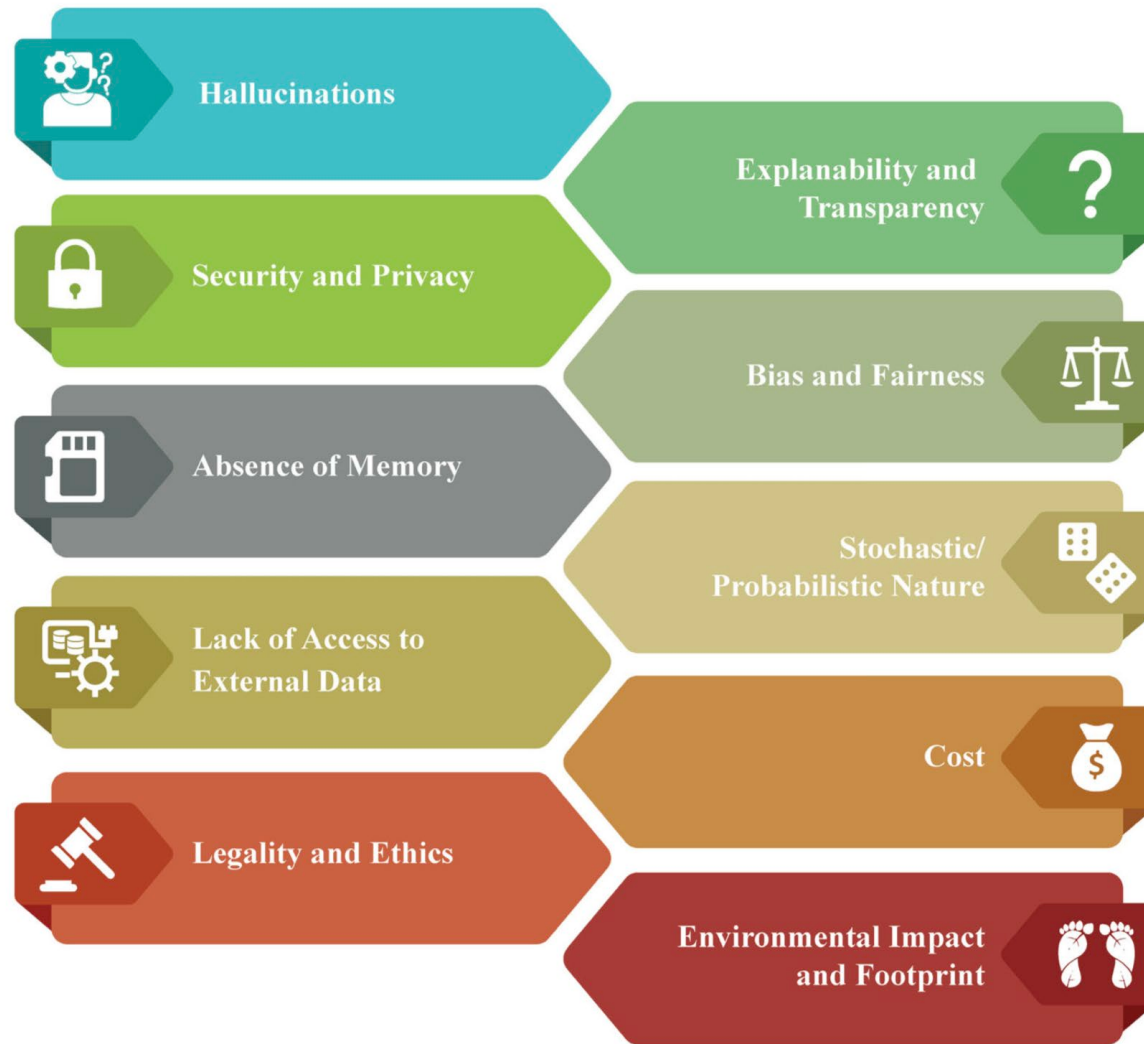
Source: Ke, Yu He, et al. "[Real-world Deployment and Evaluation of PErioperative AI CHatbot \(PEACH\)](#)." *arXiv preprint*, 2024

How we evaluate LLMs in Clinic



- **Beyond raw accuracy metrics:** Include completeness of key facts, and readability of outputs at patient-appropriate grade levels for safe communication.
- **Safety and harmful content:** Track hallucinations, dangerous recommendations, and inconsistent reasoning with standardized grading rubrics and reviewer consensus.
- **Stability and empathy measures:** Assess test-retest variability under identical inputs and clinician-rated empathy of patient-facing drafts.

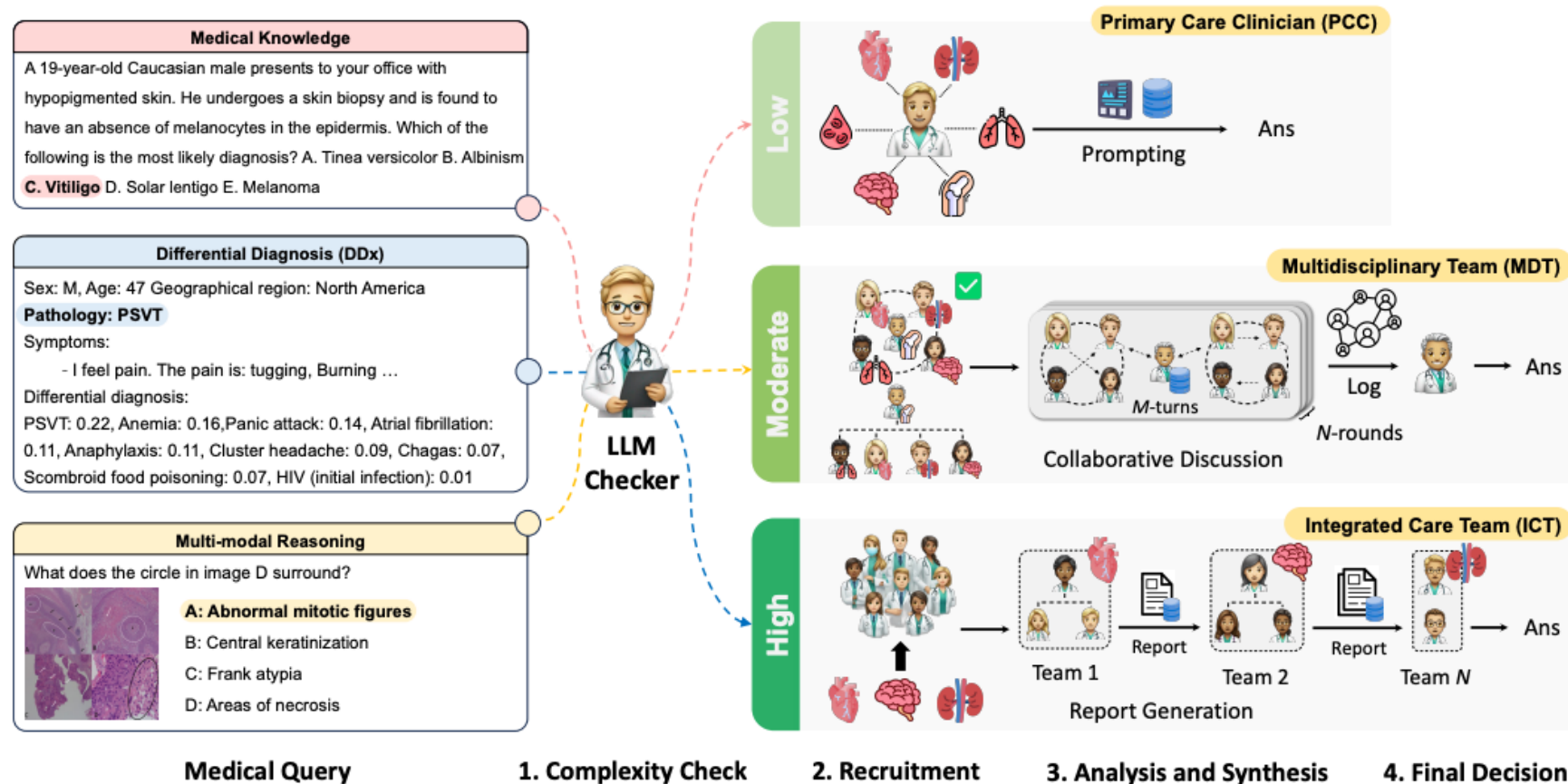
Risks & Guardrails that work



- **Grounding beats hallucinations** — Constrain answers to local guidelines with inline citations; allow “no-answer”; require clinician sign-off; grade accuracy/safety weekly.
- **Protect privacy and legality** — Keep PHI in a secure enclave; role-based access and audit logs; retention limits; IRB/consent for any patient-facing output.
- **Measure and mitigate bias** — Track subgroup errors and chance-corrected agreement; calibrate thresholds; redact sensitive attributes; run periodic blinded chart reviews for drift.
- **Stability, efficiency, sustainability** — Temperature = 0 and version-locked prompts/models; retrieval as the source of truth; prefer smaller models; cache responses to cut cost/footprint.
- **Phase-in path:** synthetic testbeds → silent deployment → measured roll-out.

Source: Hondjeu, A. R. M., et al. “[Large Language Models in Perioperative Medicine: Opportunities and Challenges](#).” *Can J Anesth*, 2025.

When a Single Co-Pilot Isn't Enough: Agent Teams

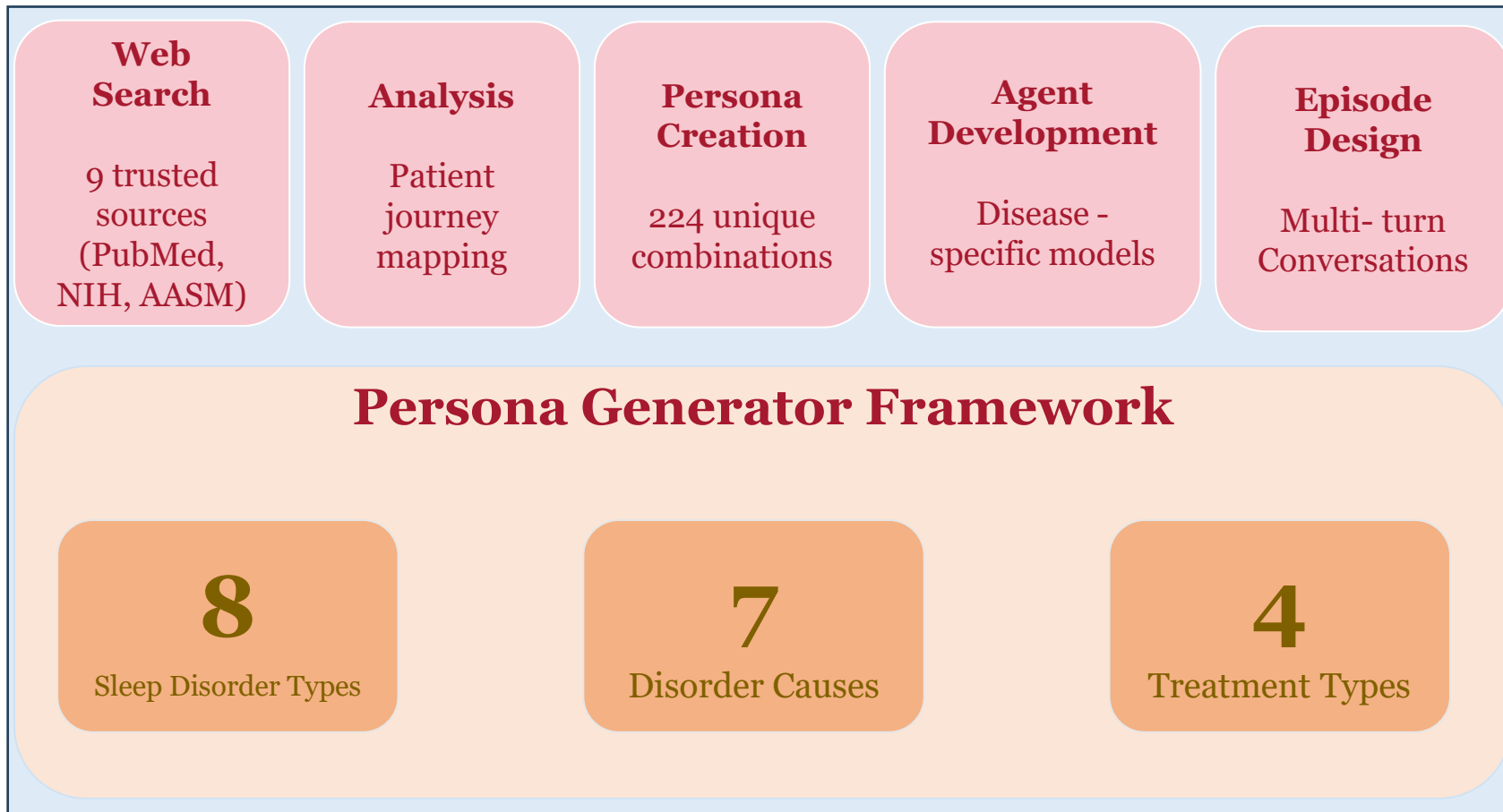


- Triage by complexity → right team
- Protocol-first, tools-second
- Standardizes complex care while keeping a human final decision.

Source: Kim, Yubin, et al. "[MDAgents: An adaptive collaboration of LLMs for medical decision-making](#)." *Advances in Neural Information Processing Systems* 37 (2024): 79410-79452.

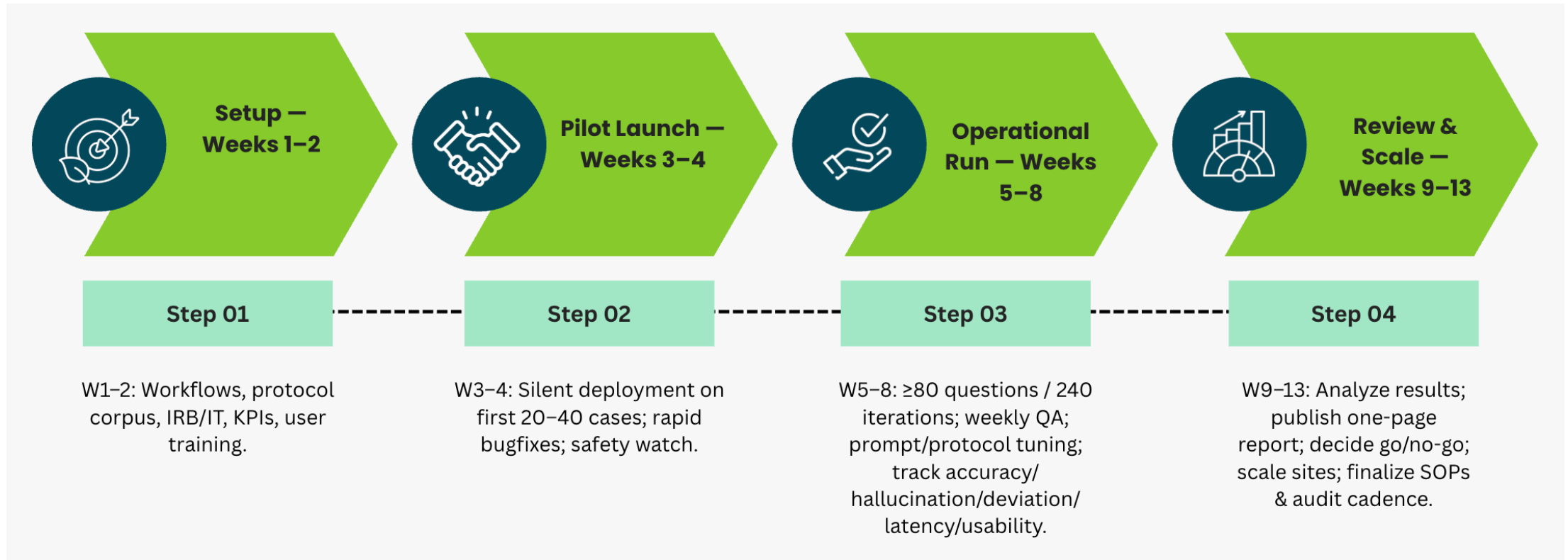
Safer, Faster Testing: Synthetic Patients & Episodes

SleepAgents: Multi-Agent Simulation of Sleep Disorders



- **Train and test with no PHI** — Build realistic synthetic pain cases using public evidence and expert input; simulate longitudinal patient journeys safely.
- **Catch risks before going live** — Run agents through test cases first to measure hallucinations, deviations, and guideline compliance.
- **From sandbox to clinic** — Use simulation results to tune prompts and agents, then safely graduate to silent deployment—like what was done in PEACH.

90-Day LLM Co-Pilot Roadmap



Key Takeaways

- **Ground in your protocols; clinicians sign off.** (No autonomous actions.)
- **Measure before and during:** accuracy + completeness; aim hallucination & deviation <1%; readable patient drafts.
- **Start small, then scale:** 2 workflows • 13 weeks • share results.



Ask me anything!