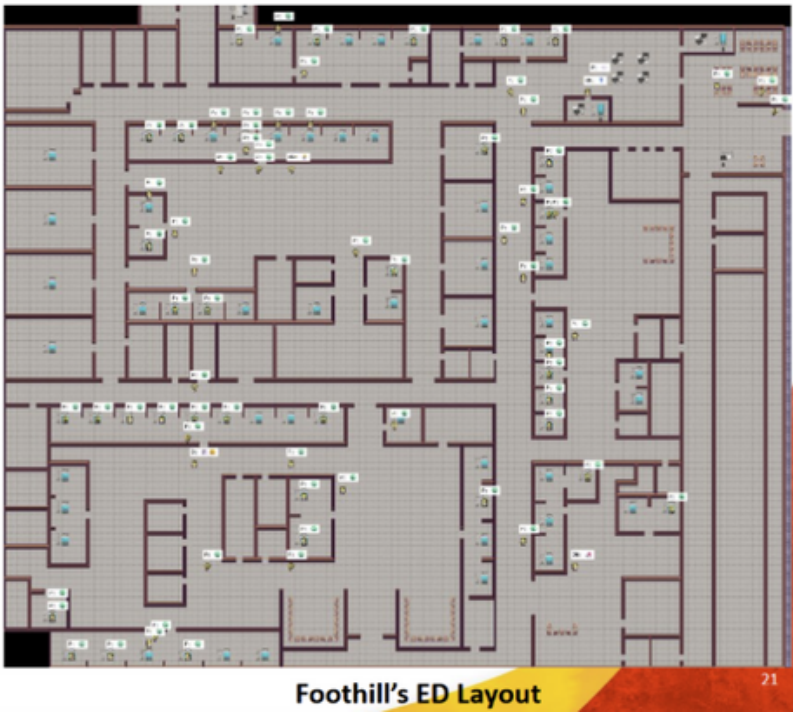
SCHULICH School of Engineering
UNIVERSITY OF CALGARY CUMMING SCHOOL OF MEDICINE
#1 | Q4 2025

LLM4ED Quarterly Newsletter

ED Simulation Project



Expanded Foothill's ED layout with added beds and functional zones, refined patient-flow paths, and support for high-load simulation and metric reporting.

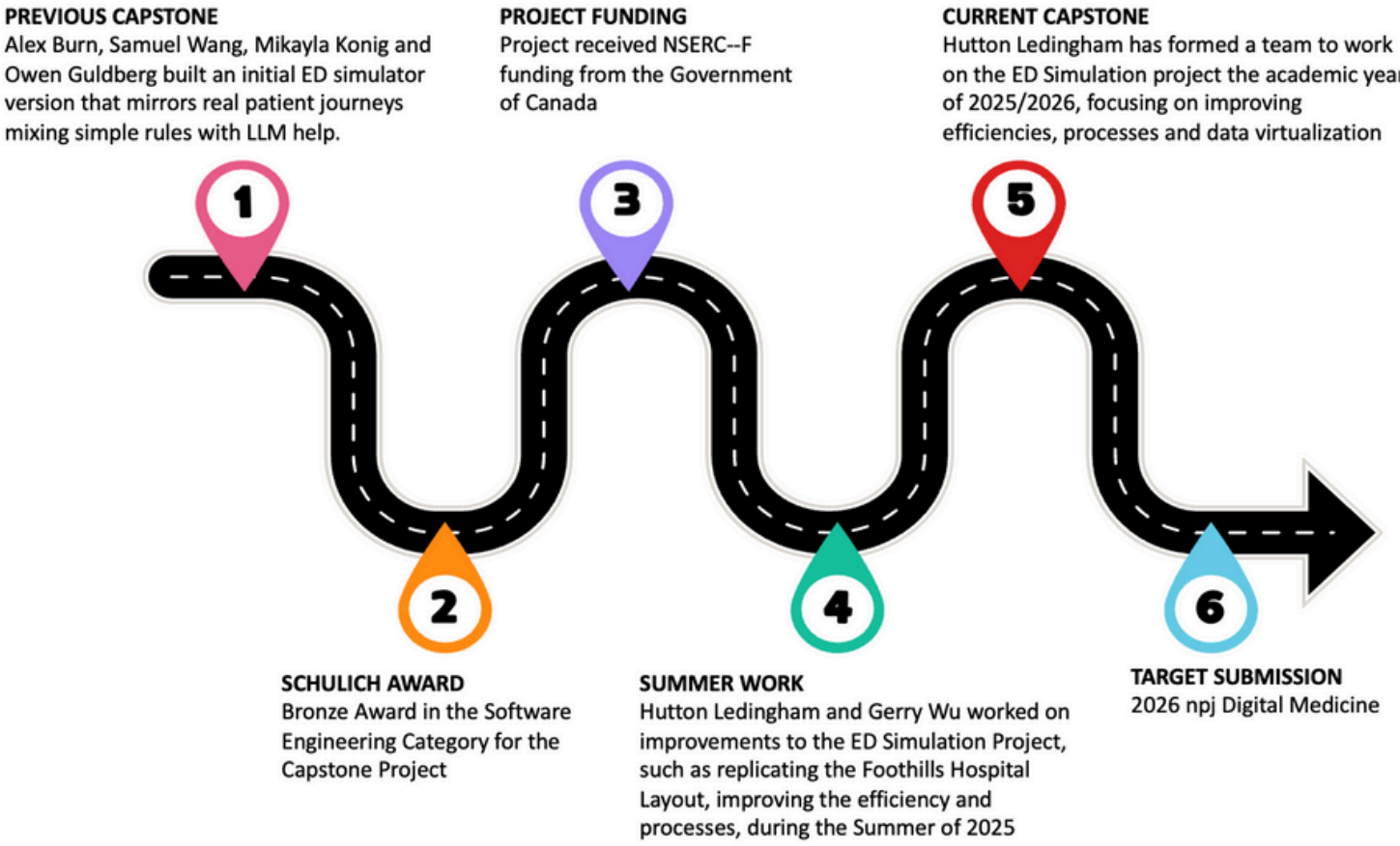
RECAP

Over the summer, the ED concept was transformed into a working, easy-to-tune simulation by **Jiajun Gerry Wu** and **Hutton Ledingham** that mimics Foothill's floor plan and day-to-day workflow. Patient arrivals follow real statistics. Triage uses an LLM to assign CTAS and route people to the correct injury zones. Each role (currently includes triage nurse, bedside nurse, doctor, and patient) can speak and behave consistently through simple prompt templates. We can also change staffing and beds on the fly, swap map layouts in seconds, and customize visuals. This helps better illustrate what's happening in the ED Sim. The patient's care journey is modelled end-to-end: bedside nurses transport patients for tests, doctors review and consult, CTAS-1 pages are immediately sent, and patients may decide to leave if wait times are too long, reflecting real behaviour. Behind the scenes, the system logs the duration of every patient's stay in each area and state, grouped by CTAS, and exports clean CSV files. Altogether, the sim is moving towards a more realistic, data-driven testbed that allows us to iterate quickly and explore how staffing, layout, and policies impact throughput and wait times.

CURRENT WORK & WHAT'S NEXT

The latest LLM4ED update streamlines the emergency department simulation by using large language models only when they improve clinical realism, while routine steps run on lightweight logic and templated flows, and the LLM intervenes only for nuanced decisions. The end-to-end ED pipeline, including Registration, Triage, Waiting, and Patient Care, now records key time points such as Arrival, PIA, Disposition, and Leaving, and mirrors real case mix using CTAS distributions. This structure supports decision-relevant metrics that operations leaders are familiar with, such as Wait time from Arrival to PIA, Treatment time from PIA to Disposition, and Total ED time from PIA to Leaving. Performance is first tested on a simplified map before scaling to a full ED layout, and companion Python visuals recreate common hospital dashboards to highlight bottlenecks and the impact of different policies.

Next, we scale to the large ED layout, finalize the data pipeline, and release the first CTAS-stratified visuals for wait time, treatment time, and total ED time. We will profile runtime across LLM and non-LLM components and tighten thresholds to keep latency low without sacrificing conversational realism. Concurrently, the system is maturing into a usable tool with save/loadable scenarios (staffing, beds, policy switches), real-time and post-run dashboards, exportable reports, and plain-language documentation. Future work includes structured content-validity checks and walkthroughs with ED staff, incremental role expansion beyond clinicians (e.g., porters) and tighter interfaces with external services (e.g., diagnostic imaging, inpatient units, and specialist consults), each step moving the simulator closer to a true-to-life ED while preserving decision-readiness. These pieces feed the paper (target: npj Digital Medicine), which will highlight stable large-map runs and figures from the new visualization stack aimed at operational insight, not just a slick demo.



The Roadmap of ED Simluation

LLM Summarizer Project

RECAP

"Dual-stage and Lightweight Patient Chart Summarization for Emergency Physicians" has been accepted to the IEEE Annual Congress on Artificial Intelligence of Things (IEEE AIoT), led by **Jiajun (Gerry) Wu** and **Swaleh Zaidi.** We built a two-stage, on-device system for ED chart summarization, demonstrating that small language models on low-power IoT hardware can perform well without cloud assistance. The design pairs retrieval-augmented generation with a dual-device setup and task-tuned prompts. The retrieval stage trims the search space and eases the generator's load, while the summary itself has two parts: a tight "must-know" section and a brief, complaint-specific narrative that ensures key details aren't missed, even when concise. We also introduce an evidence-based evaluation that scores summaries without requiring gold references, allowing us to judge factual accuracy, completeness, and clarity directly from the underlying notes.

CURRENT WORK & WHAT'S NEXT

The LLM EHR Summarization effort is now running on two coordinated tracks because great summaries require two key elements: a credible quality measurement method and a model aligned to that standard. Splitting the work lets us move faster and with clearer accountability. One team builds a benchmark for clinical usefulness, the other tunes the model to meet it. This led to stronger physician trust, lower cognitive load, and summaries that improve real decision-making rather than just scoring well on abstract benchmarks.

In addition, we would like to introduce two new ongoing projects. The first project, the Multi-agent evaluation system, captures what different roles value (physician, nurse, patient), quantifies reliability, and turns "sounds good" into evidence that holds up. The second project is to improve Preference Optimization, which trains the model to respect those preferences while accounting for real constraints.

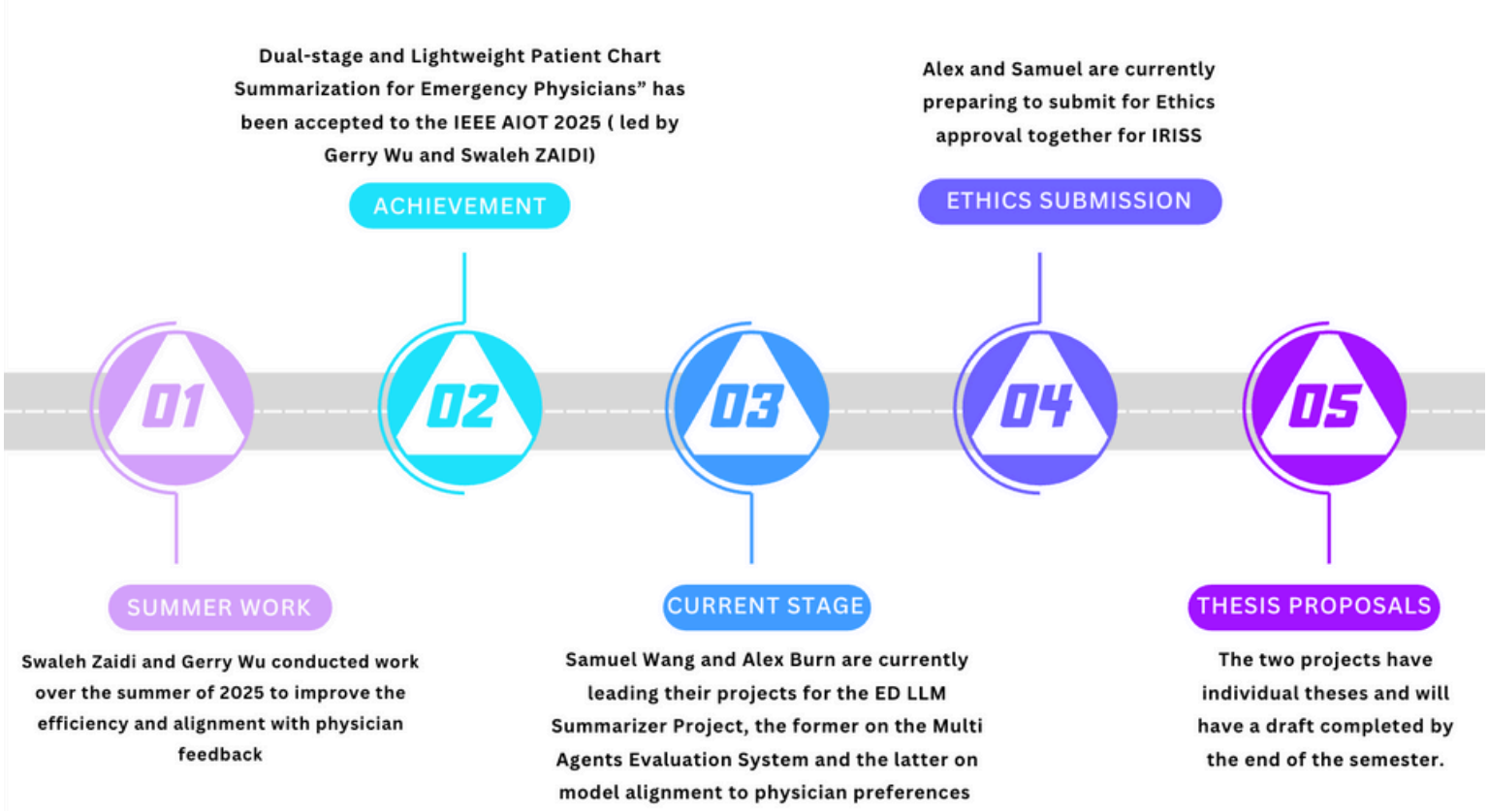
- Alex and Samuel are leading the **ethics application process**.
- The goal is to **collect feedback** from doctors, nurses, and patients through short interviews and surveys.
- **They are conducting a literature review covering four key areas:** Clinical summarization, Small language models (SLMs) on edge devices, Retrieval-augmented generation (RAG) for electronic health records (EHRs), Reference-free scoring methods

MULTI-AGENT EVALUATION SYSTEM (LED BY SAMUEL WANG)

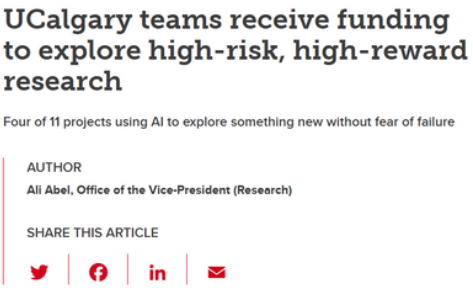
- **CEESQI:** Build the Canadian ED EHR Summary Quality Index as a role-aware rubric.
- **Role views:** Capture how physicians, nurses, and patients each define a "good" summary in a multi-agent evaluation setup.
- **Independent scoring:** Each role scores summaries separately to preserve their distinct priorities.
- **Gold-Standard Metric (GSM):** Aggregate those signals into a single adaptive GSM with provisional weights.
- **Refinement:** Refine weights using reliability statistics and preference learning to improve fairness and stability.

PREFERENCE OPTIMIZATION (LED BY ALEX BURN)

- **Alignment goal:** Ensure the model matches the target GSM standard for clinically useful summaries.
- **Method survey:** Evaluate and compare DPO, RLHF, IPO, KTO, ORPO, and related variants to identify or refine methods that best capture real clinician preferences.
- **Training focus:** Produce concise, decision-ready ED summaries that both score well under the GSM and read naturally to physicians.
- **Outputs:** Deliver a repeatable alignment pipeline, a comparative analysis of optimization strategies, and guidelines for scaling physician feedback collection.



The Roadmap of LLM Summarizer Project



UCalgary teams receive funding to explore high-risk, high-reward research

Four of 11 projects using AI to explore something new without fear of failure

AUTHOR

Ali Abel, Office of the Vice-President (Research)

SHARE THIS ARTICLE

UCalgary News Highlighted Our Work

OPTIMIZING OPERATIONS OF EMERGENCY DEPARTMENTS USING MULTI-AGENT LARGE LANGUAGE MODELS

Improving patient care and reducing clinician burnout are top of mind for Drs. Drew and Holodinsky are using LLMs to find ways to optimize how emergency departments operate.

CHECK OUT NEWS

Meet Our Team

Jiajun (Gerry) Wu

THIS IS GERRY, PHD STUDENT IN DENOS LAB!

Jiajun Wu received a Bachelor of Information Technology (Honours) in Information Resource Management, with a minor in Computer Science from Carleton University in 2022. He is currently a PhD student in the Department of Electrical and Software Engineering at the University of Calgary, Calgary, AB, Canada. His research interests include federated learning, edge computing, and large language models. His paper has received the Best Paper Award at the 2025 IEEE International Conference on Autonomous and Trusted Computing.