

MedEvoEval

Evaluating Continual Evolution of Doctor Agents through Simulated Clinical Episodes

Hui Zhang

Beijing Institute of Technology

Sydney · Poster Session 1



**Does experience
help this agent?**

Doctor agents need two levels of evaluation

WITHIN AN EPISODE

How does the agent reach a decision?



Ask questions

Request tests

Finalize

Clinical quality + process costs

ACROSS EPISODES

What changes with experience?

- 01 Improvement**
Repeated clinical episodes
- 02 Transfer**
Unseen held-out cases
- 03 Retention**
Earlier cases after updates

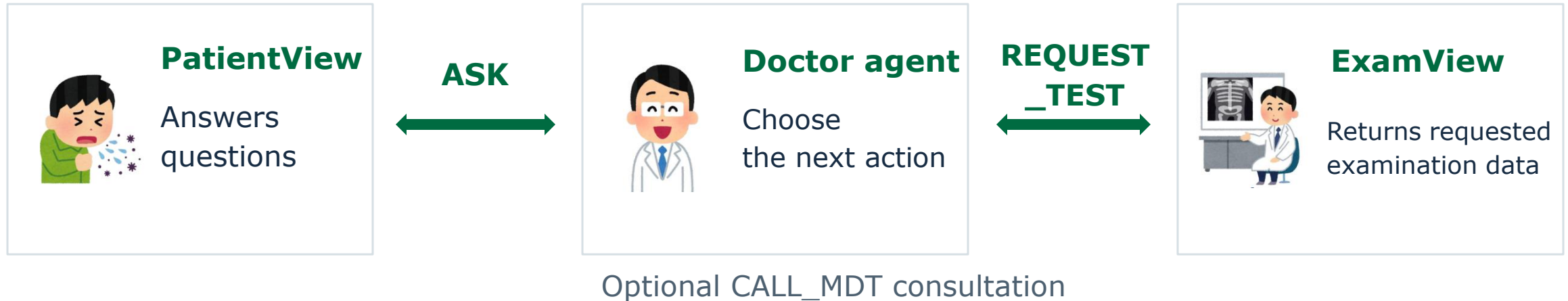
700

processed outpatient episodes

500 longitudinal + 200 held-out

Evidence arrives through actions

The doctor starts with a brief complaint.



ManagerView evaluates the final output

$$S = 0.5 \text{ Diagnosis} + 0.3 \text{ Evidence} + 0.2 \text{ Management}$$

$$U = 0.6 S - 0.2 (T/10) - 0.1 (E/3) - 0.1 R$$

T: turns E: examinations R: fraction of examination requests with no nonempty result

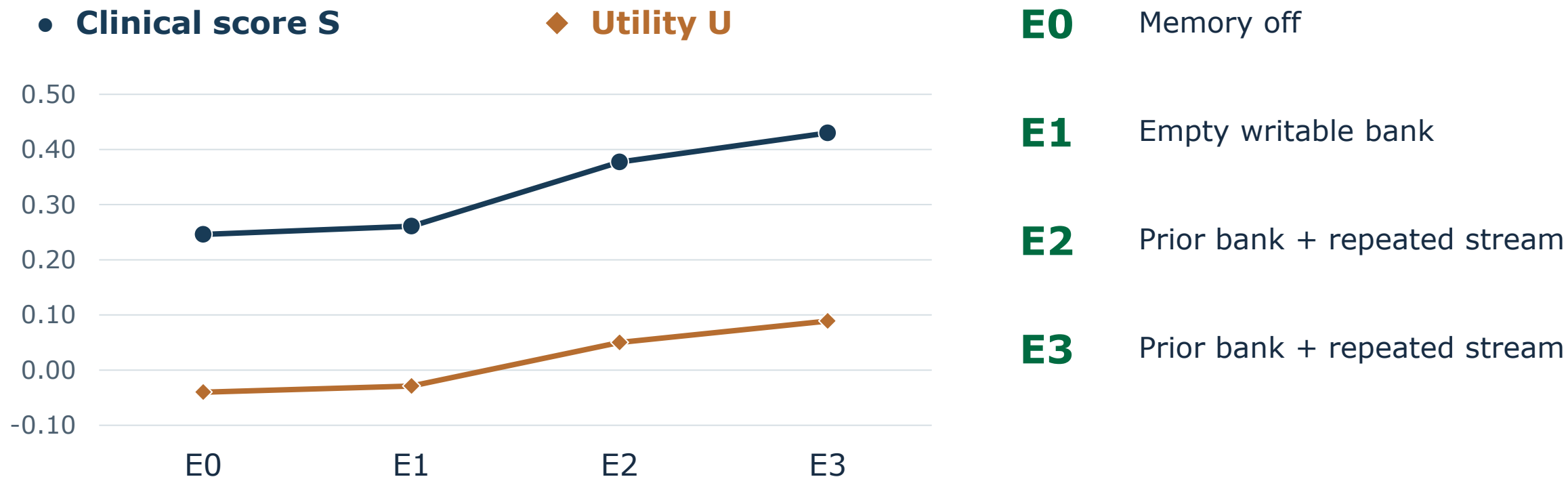
A writable experience bank

Success, correction and rule cards

TF-IDF top-3 retrieval

Repeated-stream gains need a held-out test

500 cases per condition



Gains combine repeated exposure and memory accumulation.

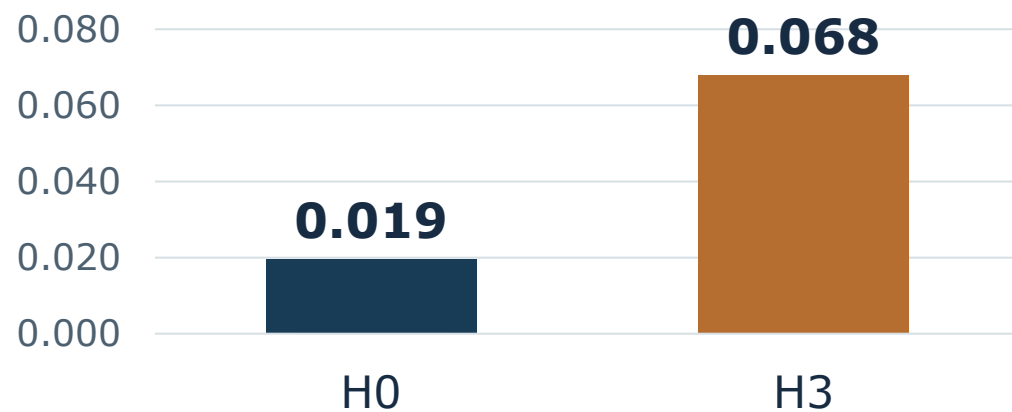
Transfer test: 200 held-out cases with no content overlap.

Held-out gains center on utility and efficiency

200 paired cases, Qwen-Turbo. H0: memory off. H3: mature memory + sequential write-back.

+0.0483

UTILITY

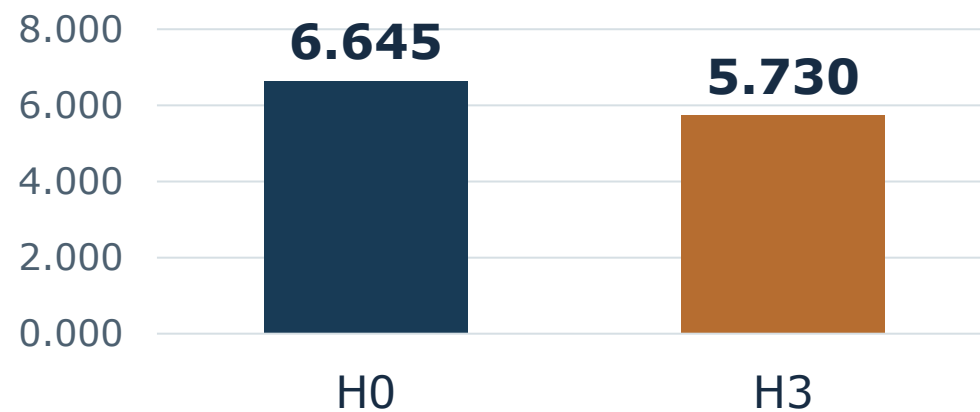


95% CI [0.0185, 0.0784]

Holm-adjusted p = 0.017

-0.915

INTERACTION TURNS



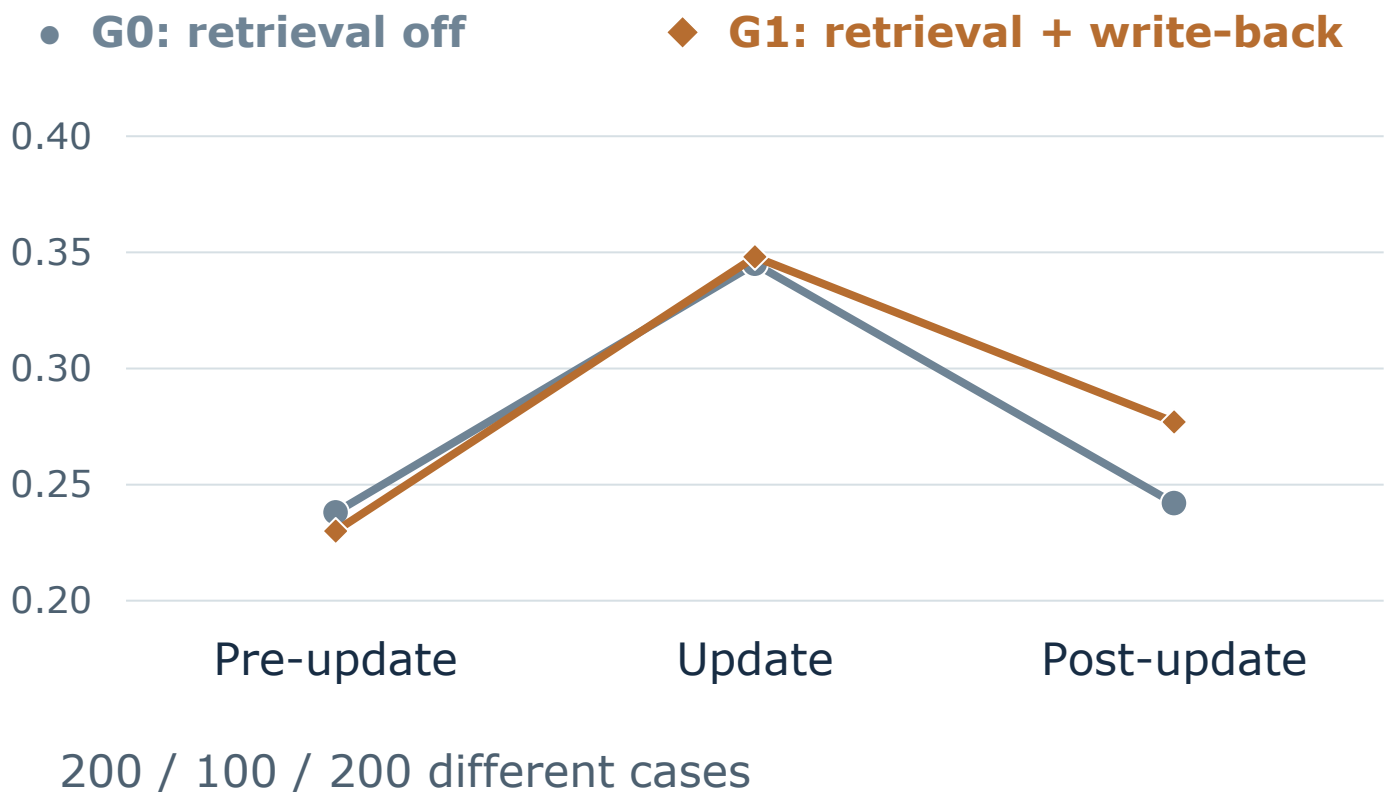
95% CI [-1.290, -0.540]

Holm-adjusted p < 0.001

Only utility and turns pass Holm correction over 7 endpoints. Individual paired-bootstrap CIs.

Updates and retention need separate tests

Clinical score across update phases



Backward transfer

450 paired earlier-case comparisons

$$\Delta S = -0.0087$$

95% CI [-0.034, 0.017]

$$\Delta U = -0.0099$$

95% CI [-0.027, 0.008]

Neither passes Holm correction.

Small negative retention estimates do not establish an absence of forgetting.

Evaluating agents that learn from experience

**Connect clinical decisions, resource use,
and behavior over time.**

Clearest held-out gains: utility and interaction efficiency.



Paper



Code + Data



NeurIPS Page