THE EXPERT EVAL LOOP: A DESIGN FRAMEWORK FOR EXPERT EVALUATION IN HEALTHCARE AI

***Hrozian Y.**, Founding Designer, Lumos AI, San Francisco, California, USA*

Abstract. *This paper argues that the evaluation of healthcare and life-science AI has become a design discipline, not only an engineering one. It introduces the Expert Eval Loop, a four-stage framework (Calibration, Elicitation, Adjudication, Drift Monitoring) for designing the interfaces through which domain experts contribute the irreplaceable judgment AI evaluation requires. Production patterns illustrate how interface design directly shapes evaluation quality.*

Keywords: *Expert Eval Loop, expert evaluation, healthcare and life-science AI, human-AI interaction, interface design, design discipline.*

AI [1] is increasingly part of how patients and clinicians make healthcare decisions, from LLMs interpreting lab results to ambient scribes drafting visit notes. The path from a wrong AI answer to a harmed patient is short. As capabilities scale, the limiting factor on safe deployment shifts from model performance to evaluation quality. Most published work treats evaluation as engineering [2]. Where ground truth is expensive and contested, the interface through which an oncologist, geneticist, or pathologist provides judgment becomes the load-bearing surface of the entire system [3]. The Expert Eval Loop names the four interface-design stages of expert evaluation (Table 1).

Table 1 – The Expert Eval Loop

Stage	What design must do	Risk if mis-designed
Calibration	Align experts on task definition and edge cases before live work begins.	Low inter-rater agreement; useless aggregate scores.
Elicitation	Capture expert judgment with low cognitive overhead and structured affordances.	Rushed or shallow ratings; survivorship bias toward easy cases.
Adjudication	Surface disagreement between experts and resolve through structured deliberation.	Silent averaging hides contested cases that matter most.
Drift Monitoring	Detect when evaluation results diverge from deployment reality over time.	Eval scores stay stable while real-world performance degrades.

Calibration aligns experts on task definition and edge cases before live work. Production calibration decomposes evaluation into single-task steps, delivers instructions one idea per screen rather than as flat documents, and specifies subtle criteria for mechanical checking, not interpretation. An oncology rubric criterion like *"The treatment plan includes starting with Letrozole 2.5mg daily"* is directly verifiable; rhetorical phrasing like *"the treatment plan is appropriate"* is not. Experts work through known-answer cases as *calibration practice* before contributing to live work; per-step quizzes verify their understanding along the way.

Elicitation captures expert judgment under structured affordances. The rating scale itself is a design choice: a 1-to-4 scale eliminates the neutral midpoint and forces a choice; the 1-to-5 default permits hedging. Friction is placed only where signal is needed. Passing scores advance automatically. Non-passing scores require justification, because a score without explanation is unactionable.

Adjudication is the most under-designed stage. Most teams default to silent averaging: disagreeing ratings collapse into a single aggregate score, hiding exactly the contested cases that matter most. Better adjudication interfaces operate as a tiered review chain. Production platforms use three tiers: a *primary expert* produces the case judgment; a *promoted reviewer* (a primary expert demonstrated to be reliable, now authorized to review and edit others' work) examines and may correct the rating; and a *third-tier reviewer* handles cases that remain contested. Each tier needs different interface support: fast capture, diff-style visibility, or disagreement summaries layered over case detail. This operationalizes Amershi et al.'s [4] error-and-recovery guidelines (G7-G11) inside eval workflows.

Drift Monitoring closes the loop. Expert calibration decays over weeks and months of use, silently, in ways output metrics alone won't catch. *Gold case injection* (seeding known-answer cases into the live work queue at low frequency) lets the platform compare expert performance against ground truth [5] without disrupting workflow. When drift is detected, recalibration drills refresh expert alignment. Choices like which cases to inject, how often, and who gets notified determine whether the platform catches drift in time to act.

For UX teams designing evaluation infrastructure for healthcare and life-science AI, the takeaway is that eval quality is a design output, not a downstream metric. The framework was developed in this domain but applies wherever expert judgment is the load-bearing surface of evaluation. Naming the stages and assigning interface responsibility to each is how the design discipline of eval becomes methodical rather than ad hoc. Further research can examine how this discipline holds up over longer timescales and how AI tools might assist experts at each stage of the loop without replacing their judgment [3].

References

1. Kaluhin, N., Vovk, O., & Chebotarova, I. (2024). The impact of artificial intelligence on future of humanity. Jóvenes en la ciencia, (26). https://www.jovenesenlaciencia.ugto.mx/index.php/jovenesenlaciencia/article/view/4235/3716.

2. Anthropic. (2026). Demystifying evals for AI agents. anthropic.com/engineering/demystifying-evals-for-ai-agents.

3. Szymanski, A., Anuyah, O., Li, T.J.J., & Metoyer, R.A. (2026). Key Considerations for Domain Expert Involvement in LLM Design and Evaluation: An Ethnographic Study. In Proceedings of IUI '26. ACM. doi.org/10.1145/3742413.3789105.

4. Amershi, S., Weld, D., Vorvoreanu, M., et al. (2019). Guidelines for Human-AI Interaction. 2019 CHI Conference on Human Factors in Computing Systems. (p. 1-13). ACM. doi.org/10.1145/3290605.3300233.

5. Chen, S., Chen, Y., Li, Z., et al. (2025). Benchmarking Large Language Models Under Data Contamination: A Survey from Static to Dynamic Evaluation. In Proceedings of EMNLP 2025. Association for Computational Linguistics. aclanthology.org/2025.emnlp-main.511/.