Discussion Leader Cover Letter

Sumin Hong, 04.23.2026

Paper: Shaikh, O., Sapkota, S., Rizvi, S., Horvitz, E., Park, J.S., Yang, D., & Bernstein, M.S. (2025). Creating General User Models from Computer Use. UIST '25. ACM. https://doi.org/10.1145/3746059.3747722 (Best Paper)

(a) What the paper states

General topic. Technology that comes to "understand us—from our preferences and habits, to the timing and purpose of our everyday actions," given that "current user models remain fragmented, narrowly tailored to specific applications, and incapable of the flexible, cross-context reasoning required to fulfill these visions."

Specific behavior/activity. Everyday computer use—screenshots, emails, clicks, and other unstructured interaction traces—as the observable signal from which a General User Model (GUM) infers confidence-weighted natural-language propositions about the user's knowledge, beliefs, and preferences.

Research questions. The paper states no numbered RQ list; evaluation targets three equivalent questions:

- Are propositions generated by a GUM accurate and well-calibrated across unstructured inputs?
- Do the architectural components (retrieve, revise, audit) each contribute to that calibration?
- When deployed longitudinally, does a GUM-based assistant (Gumbo) execute suggestions of value that users "wouldn't think to request explicitly"?

(b) Problem • Claims • Evidence • Statistical Analysis

Problem. Existing user models are siloed per app, depend on explicit dialogue or clarification, and cannot reason across contexts; AI systems therefore ground poorly in implicit situations and produce one-size-fits-all responses.

Claims.

- An observe → audit → propose → retrieve → revise architecture can turn unstructured interaction data into a coherent user model.
- The resulting propositions are calibrated—confident when correct, unconfident when wrong.
- A Contextual-Integrity-based audit suppresses the large majority of privacy-sensitive inferences.
- A proactive assistant built on a GUM (Gumbo) completes useful actions in the wild.

Evidence.

- Email accuracy (N = 18). Emails streamed into the GUM; participants rate each proposition. Mean accuracy 76.15%; propositions at max confidence (10) rated 100% accurate. Ablations confirm retrieval, revision, and audit are each necessary.
- Privacy audit. Only 7 / 180 propositions from the full pipeline were flagged as contextual-integrity violations.
- Field deployment (N = 5, 5 days). After 24-h burn-in, Gumbo observed participants' screens for 4 active days. Accuracy 79%; 25% of suggestions rated 6–7 on a 7-point scale; 2/5 participants asked to keep the system.

Statistical analysis.

- Calibration via Brier score (full pipeline 0.17 vs. higher for ablations; 0.28 in the field).
- Accuracy reported per confidence bin, not only as aggregate mean—exposing over- vs. under-confidence.
- Retrieval and revision effects estimated via pairwise win-rate on participant ratings.
- Field utility combines 7-point Likert ratings with qualitative coding of semi-structured interviews (hybrid design).