SYMPOSIUM ON COMMUNICATION & INTELLIGENCE · UNIVERSITY OF CHICAGO · MAY 2026

Distinguishing Human from AI

A Process-Level Exploration

Nykko Vitali & Jason Mitchell · Department of Psychology, Harvard University · nvitali@fas.harvard.edu

SCAN FOR WORKING PAPER

Large language models (LLMs) can produce text that mirrors the complexity and nuance of human authors (Jakesch et al., 2024). An empirical question that arose after their widespread adoption is how often can a given human discern text from them versus another human. Most Turing Test style experiments in Human/AI detection work often collapses turn-by-turn judgments into a single accuracy number. We instead model detection as a sequential-inference problem. In a pilot study, 247 detectives exchanged messages with a human or LLM partner across five conversational motifs (warm, playful, guarded, contrarian, bland), logging a binary Human/AI label and 0–100 confidence after every turn. After 7.5 minutes the respondent makes one final judgment about the detective. The respondent in the LLM condition was Gemini 3 Flash with a two-step generation pipeline (tactic selection then response). Detection, in our framing, is not a single end-of-conversation choice but a sequence of evidence-weighted updates that we can model directly.

We ask whether detectives discriminate accurately turn by turn (*sensitivity*), whether their judgment shifts as the conversation unfolds (*bias*), what rule best describes their turn-by-turn updating (*belief updating*), and how much of the available text-based signal they actually use (*signal use*).

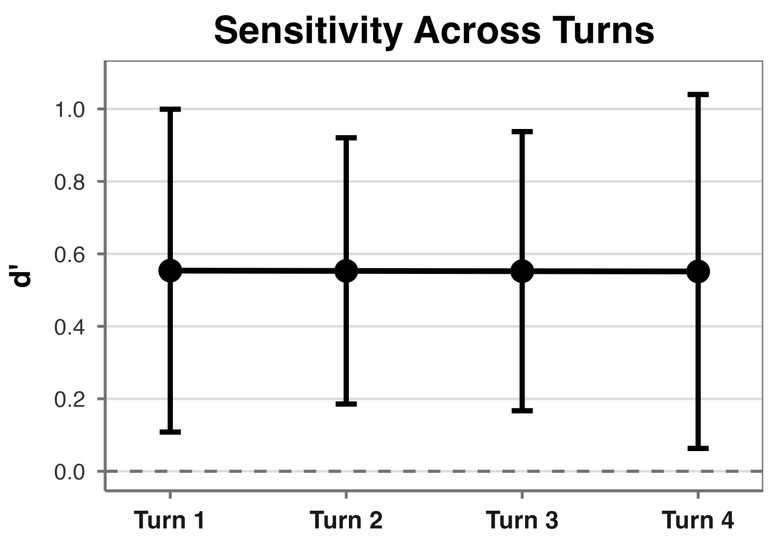

Fig 1. Detectives discriminate above chance, but sensitivity stays flat. Probit mixed-effects SDT fit to 915 turn judgments across 247 sessions. The truth main effect (b = 0.553, p = .003) confirms above-chance discrimination, but the truth × turn interaction is null (b = −0.001, p = .994). Detectives discriminate AI from human better than chance, but they do not become better doctors as the conversation continues.

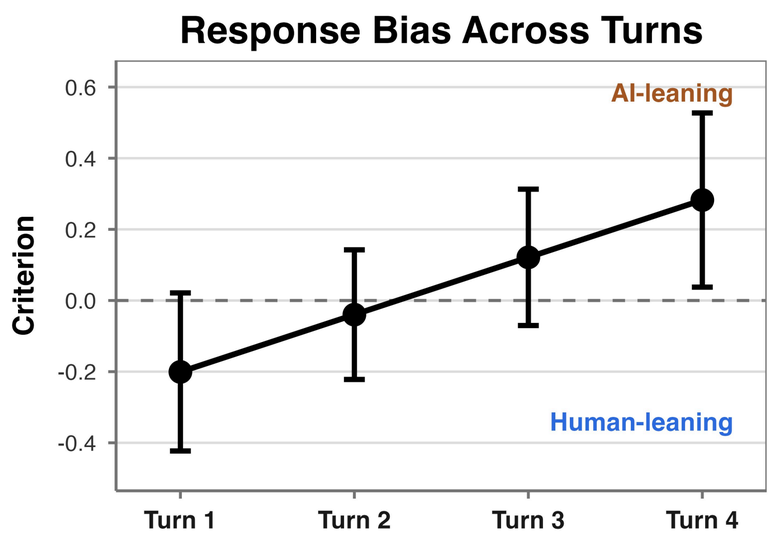

Fig 2. Suspicion increases even though sensitivity does not. Same SDT model, decision criterion across turns. Criterion moves from c = −0.20 (T1) to +0.28 (T4); the turn main effect is significant (b = −0.18, p = .001), motif main effect is not (p = .162). Detectives became more willing to call their partner AI, even though their discrimination ability stayed flat.

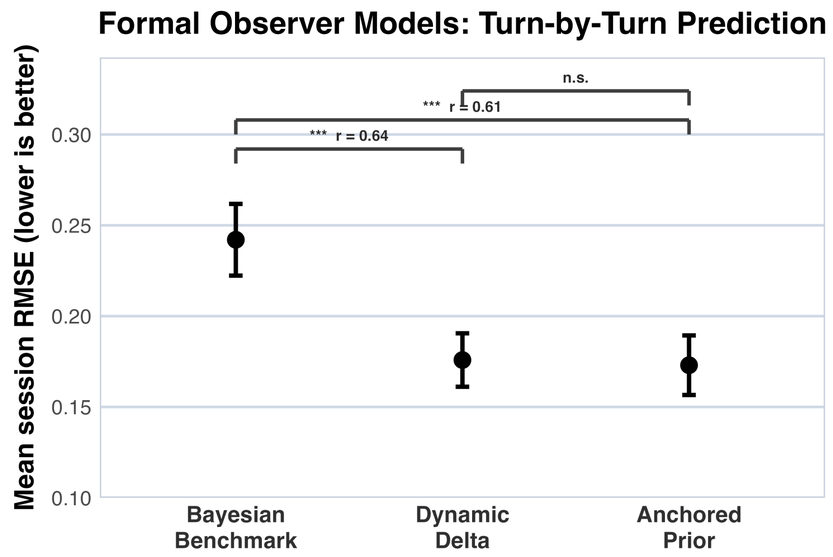

Fig 3. Detectives' beliefs are sticky. Three observer models predict per-turn belief from the same decoder evidence. The Anchored Prior, which carries the previous turn's belief forward unchanged, fits nearly as well as the Drift-Adjusted Decision (RMSE .207 vs .202), and a normative Bayesian observer fits worst (RMSE .272). Wilcoxon: DAD vs Bayesian p < .001, r = .64. Pattern holds within both AI and human conditions (r = 0.65 and r = 0.62 for DAD vs Bayesian). Detectives updated less like normative evidence integrators and more like anchored, bounded reasoners.

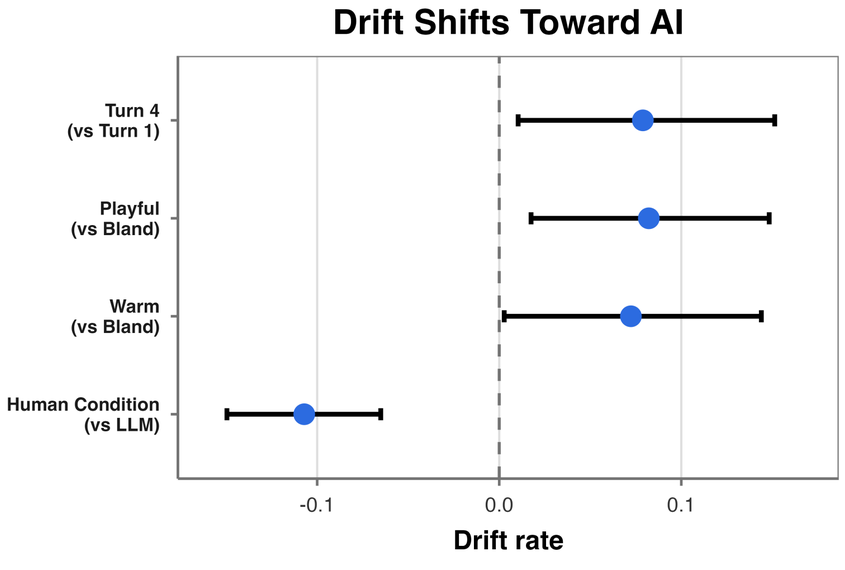

Fig 4. Condition and motif effects appear in evidence accumulation. Wiener Drift-Diffusion Model (brms) with drift varying by condition, motif, and turn (877 judgments, 244 sessions; reference: AI partner, Bland motif, Turn 1). Drift shifts toward AI for Playful (b = +0.084), Warm (b = +0.074), and Turn 4 (b = +0.080), and toward human for the Human partner condition (b = −0.106). All four 95% CrIs exclude zero.

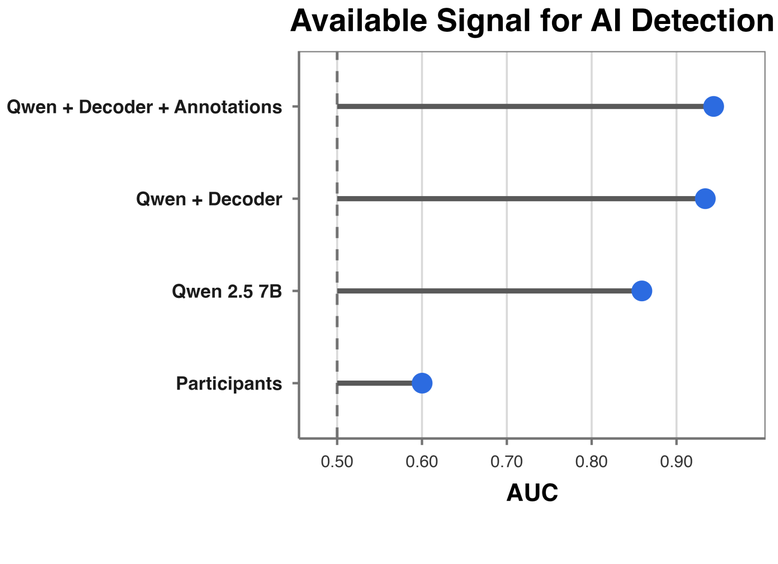

Fig 5. The text contains more signal than detectives use. Logistic ridge decoders trained on linguistic features and sentence-embedding PCs (5-fold session-wise CV). The decoder reaches AUC = .898; adding LLM-rated cues raises this to AUC = .934 (Brier −8.2%, p = .002). Detectives reach AUC ≈ .61 (62.3% accuracy). Removing word count from the Qwen feature set drops AUC from .859 to .813, so message length matters, but the bulk of the diagnostic signal is in the text itself, and detectives don't use it.

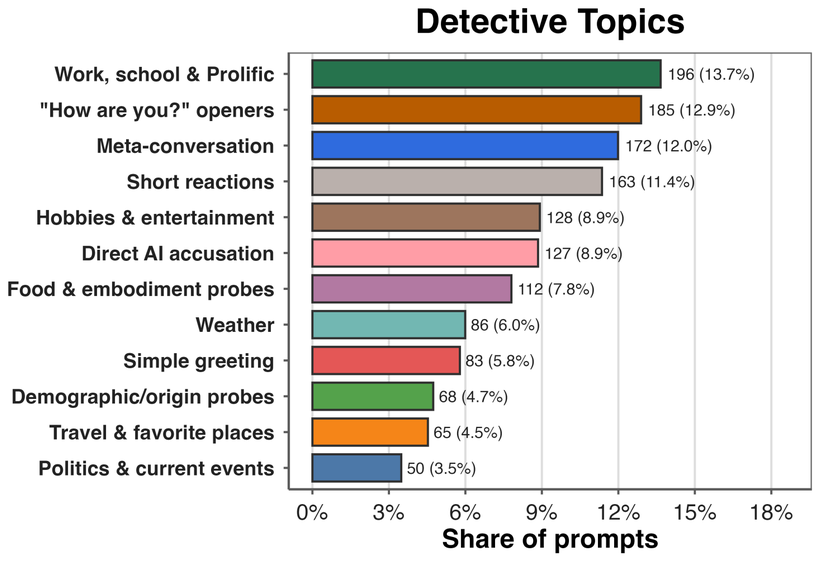

Fig 6. Detectives mostly asked ordinary conversational questions. BERTopic clustering of 1,435 detective prompts. The top clusters are work/school/Prolific (13.7%), "how are you?" openers (12.9%), and meta-conversation (12.0%); 8.9% of prompts directly accuse the partner of being AI. Most available signal emerged through ordinary chat rather than explicit AI tests.

DISCUSSION

Prior work in AI detection has largely been focused on outcome instead of process (Jones et al., 2025). We instead model conversations as a dynamic sequential-inference process. Across the panels above, our work shows convergence with previous work, but with notable divergences given the novel structure of our experimental design. Consistent with previous work (Jannai et al., 2023), detectives can tell humans from LLMs (*Fig 1*), but their underlying discrimination sensitivity does not actually improve with more evidence. What changes instead is how skeptical they grow over time (*Fig 2*). While static evaluations suggest people default to a "seeing-is-believing" bias (Köbis et al., 2021), our dynamic setup gives evidence that prolonged interaction breeds suspicion. Our results highlight the danger of relying on raw accuracy scores, which systematically conflate true detection ability with baseline skepticism (Batailler et al., 2022; Zloteanu & Vuorre, 2024).

Stepping into a completely new direction, we map this psychological process with normative models. Converging with classic cognitive models of social inference (Epley & Gilovich, 2006; Tamir & Mitchell, 2010), their turn-by-turn beliefs are best described not by a normative learner but by an Anchored prior (*Fig 3*). Furthermore, extending drift-diffusion models to human-AI interaction (Bavard et al., 2024; Hutcherson et al., 2015), we find that the bias toward suspicion appears in the evidence accumulation process, and not only in final choices (*Fig 4*). Discriminative models recovered substantially more Human/AI signal from the messages than detectives used in their judgments (*Fig 5*), aligning with prior text-generation research (Ippolito et al., 2020). This suggests that detection failures result from more than an absence of signal. Instead, we find people appear to underweight available cues and rely heavily on anchored priors and conversational fit. Topic structure shows the cues are scattered across everyday content rather than concentrated in explicit AI-tests (*Fig 6*). Our results provide evidence that human-AI detection can be modeled as a dynamic process of sequential inference in which people show modest sensitivity, shifting suspicion, and bounded use of the evidence available to them.

KEY TAKEAWAYS

1. **Detectives discriminate, but weakly.**
   They detected Human/AI differences above chance, with d′ ≈ 0.55.
2. **More conversation did not improve sensitivity.**
   Across the first four turns, discrimination stayed essentially flat.
3. **What changed was suspicion.**
   The decision criterion shifted steadily toward AI judgments over time.
4. **Beliefs were anchored.**
   Bounded and sticky updating models fit better than a Bayesian benchmark.
5. **The signal was there, but underused.**
   Text decoders recovered far more Human/AI information than detectives used.

REFERENCES

Jones et al. (2025) · Jakesch et al. (2024) · Jannai et al. (2023) · Köbis et al. (2021) · Batailler et al. (2022) · Zloteanu & Vuorre (2024) · Epley & Gilovich (2006) · Tamir & Mitchell (2010) · Bavard et al. (2024) · Hutcherson et al. (2015) · Ippolito et al. (2020)