Proceedings of Machine Learning Research **318** (2026)
The 39th Canadian Conference on Artificial Intelligence
Canadian AI 2026
25–29 May 2026, Vancouver, BC

Toward Believable Health & Wellness Conversational Agents: A Post-LLM Turing-like Evaluation Framework (Position Paper)

Bruce Matichuk[†,‡,*] Randy Goebel[‡] Osmar Zaïane[‡]
[†] MedWatch Technologies Inc.
[‡] University of Alberta

Abstract

Large language model (LLM) conversational agents can be remarkably fluent yet still fail to feel fully "real" to users, especially in multi-session and higher-stakes interactions. This paper argues that the limiting problem is no longer surface language quality but *believability*: the conditions under which an artificial conversational partner is experienced as a coherent social mind rather than a fluent text generator. We frame believability as an empirical limit case and propose an operational criterion of *bounded practical indistinguishability* relative to an interaction envelope defined by a judge population, interaction contexts, and a time horizon. We then outline a "post-LLM Turing-like" evaluation approach that stress-tests modern detection cues using contextual scenario families, longitudinal re-contact, and multi-signal measurement combining human judgments with behavioral metrics. Finally, we instantiate the framework for a health and wellness agent being developed with *MedWatch Technologies Inc.*, arguing that wellness settings sharply amplify the importance of epistemic calibration, continuity, and boundary management. The goal is not to advocate deceptive deployment, but to make believability mechanistic and measurable so that both capabilities and risks can be assessed with clarity.

Keywords: Believable agents, conversational evaluation, longitudinal interaction, uncertainty calibration, health and wellness coaching, Turing test

1. Believability as the Post-Fluency Bottleneck

LLM-based conversational agents have reached a level of linguistic competence that would have seemed implausible even a few years ago. They can sustain dialogue, adapt tone, explain complex topics, and often produce single turns that read as human-written. Yet many users report a persistent gap between competence and felt realism: the interaction can be helpful while still seeming hollow, performative, or "not quite real." The central claim of this position paper is that fluency is no longer the key bottleneck for human-like conversation. Instead, the bottleneck is *believability*: whether an agent is experienced as a coherent social partner—a stable "someone" behind the words—rather than as a generator of plausible language.

This gap becomes most visible when the interaction moves beyond the conditions under which LLMs are strongest. In longer conversations, in emotionally loaded moments, when misunderstandings occur, when commitments must be tracked, or when the conversation spans multiple sessions, distinctive failure patterns emerge. Users encounter generic empathy that is poorly grounded in their situation, preferences that drift without explanation, inconsistencies in beliefs or stated constraints, and confidence that is misaligned with what the agent plausibly should know. Repair is often brittle: the agent may respond smoothly but fail to negotiate meaning the way humans do, or it may apologize without changing subsequent behavior. These are not always failures of utility; they are failures of *mind attribution*. The user's experience is that there is no coherent agentive center coordinating the interaction. We use the term *vacuity* for this perception: plausible conversational output without the social-cognitive continuity that humans treat as evidence of a real interlocutor.

* bmatichu@ualberta.ca

In health and wellness settings, the vacuity problem is amplified by the need for sustained engagement and the perceived stakes of advice. Wellness support is rarely a one-shot Q&A exchange; it is a long-horizon interaction in which a user tries, fails, adapts, and tries again. Users quickly notice when encouragement is generic, when plans are not followed up, or when the agent behaves like an authority without appropriate epistemic humility. Moreover, the wellness domain imposes boundary constraints that directly shape believability: an agent that slides into diagnosis, prescriptive medical decision-making, or overconfident causal claims may be not only less believable but unsafe. For these reasons, health and wellness agents are a demanding test bed for understanding what believability requires and how it should be evaluated.

2. From "Human-Like" to a Testable Target

The term "human-like" is often used loosely, conflating surface language quality with deeper properties that drive mind attribution and perceived realness. We treat human-likeness as multi-dimensional and role-conditioned. People do not infer personhood from grammar alone; they infer it from patterns of agency, accountability, epistemic restraint, and continuity across time. The importance of these cues depends on role and context. A fictional character, a customer support representative, and a wellness partner are all expected to sound natural, but they are not expected to exhibit the same kind of self-disclosure, authority, or memory behavior. A single global "human-like" score therefore collapses important distinctions.

To make believability testable without making metaphysical claims about consciousness, we adopt an operational stance. Let P be a judge population, C be a set of interaction contexts or scenarios, and T be a time horizon. We say an agent is *believable with respect to* (P, C, T) if, under a randomized protocol where judges interact with either the agent or a human interlocutor under matched constraints, judges cannot classify "human vs. artificial" at above-chance accuracy. Crucially, the protocol must allow for adversarial probing within stated rules. This criterion is a practical limit case: it defines "100% believable" not as an absolute property but as *bounded practical indistinguishability*. It supports graded claims. An agent might be indistinguishable for casual users in short, low-stakes contexts and still easily detectable by expert evaluators over multiple sessions.

This framing forces explicitness about what is otherwise implicit. One must specify who is judging, what contexts are included, how long the interaction lasts, and what kinds of probes are permitted. It also forces explicitness about disclosure. Whether judges are told they might be speaking to an AI, and whether the agent is permitted to answer identity questions, can change both strategies and outcomes. In our view, disclosure should be treated as an experimental variable rather than a fixed policy: capability evaluation should report the disclosure regime as part of the interaction envelope.

The key implication is that classical, short-horizon Turing-style tests are now insufficient on their own. Recent empirical results suggest that some LLM configurations can achieve high "human" ratings in standard text-only, single-session settings [1]. If the goal is to understand what makes an agent believable in the way users actually experience deployed systems, evaluation must stress the cues that people use when they have time, context shifts, and opportunities to re-contact and probe.

3. A Post-LLM Turing-like Evaluation Approach

We propose an evaluation approach designed to surface the vacuity–believability gap under modern conditions. The proposal is intentionally practical: it is meant to be implemented as a battery of interactions and measurements rather than a single pass/fail

test. The approach has three commitments: contextual diversity, longitudinal pressure, and multi-signal measurement.

First, evaluation should cover a *family* of contexts rather than a single chat. Real believability failures often appear when the conversational regime changes: from rapport to ambiguity, from smooth exchange to misunderstanding, from low-stakes talk to norm-sensitive boundary negotiation, or from casual curiosity to epistemic challenge. Second, evaluation should include *longitudinal re-contact*. Many agent weaknesses are invisible in one session and become obvious over repeated interaction: memory that is too perfect (and therefore creepy), memory that is absent (and therefore hollow), commitments that are made but never revisited, and identity that subtly drifts. Third, evaluation should be *multi-signal*. Detectability judgments matter, but they do not tell us *why* judges detected or what mechanism closed the gap. For mechanism, we need human rubrics and behavioral metrics.

Human measures can draw from social presence theory and mind perception research. People attribute minds along dimensions related to agency and experience [2], and perceived social presence is a measurable experiential construct rather than a poetic metaphor [3]. In parallel, work in HCI demonstrates that people treat computers as social actors under minimal cues [4], implying that believability is not all-or-nothing and can be shaped by interaction contracts and norms. These literatures justify measuring believability as an experiential inference rather than a purely linguistic property.

Behavioral metrics should focus on the cues that produce vacuity. Grounding and repair are a central example. Human conversation is not merely turn production; it is coordination of mutual understanding. Grounding theory emphasizes how interlocutors establish common ground through evidence of understanding [5], and conversation analysis describes systematic repair practices when trouble occurs [6]. In believability terms, an agent that guesses smoothly without checking understanding can feel less real than one that performs a brief, norm-appropriate repair sequence. Similarly, epistemic calibration is not just a safety issue but a believability issue. Overconfident errors are among the most salient detection cues. Calibration research in machine learning provides tools for quantifying confidence misalignment [7], and in interactive settings we can operationalize calibration as the alignment between expressed certainty and subsequent correctness under controlled questions.

A practical way to connect this evaluation to engineering is via ablation-style conditions across an agent "stack." One can compare a baseline LLM prompted for role behavior against variants that add (i) role-locked constraints, (ii) a memory policy appropriate to the role, (iii) commitment tracking and follow-through, (iv) epistemic calibration behaviors, and (v) a tuned repair policy. The point of these ablations is not to claim one true architecture, but to estimate which mechanisms actually reduce detectability and increase social realism over time. Importantly, the same mechanisms can have different effects by role. Rich personal disclosure may improve believability for a fictional character but reduce believability for a professional support agent. A core design choice in our proposal is therefore to treat believability as *role-conditioned* and to interpret scores relative to a role contract.

To operationalize the framework, we propose a protocol consistent with (P, C, T), in which a mixed judge population evaluates agents across diverse interaction contexts (e.g., rapport, repair, boundary challenges, and longitudinal goal tracking) over multiple sessions. Measurement combines detectability (human vs. artificial classification accuracy) with mechanism-oriented signals, including social presence, calibration (confidence vs. correctness), and behavioral metrics such as commitment follow-through and grounding/repair. Ablation-style comparisons across agent configurations link outcomes to design choices. The protocol is intended as a flexible, role-adaptable template.

4. Instantiation in Health & Wellness with an Industry Partner

We instantiate the above framework in the setting of a health and wellness agent being developed with an *MedWatch Technologies Inc.* The aim of the instantiation is not to present a product description, but to clarify what believability must mean for a wellness role and what the most important stressors are.

In this setting, the agent functions as a *wellness partner*. Users expect steadiness, discretion, and continuity. They also expect the agent to respect scope boundaries: not to impersonate a clinician, not to diagnose, and not to translate uncertain signals into confident medical conclusions. These boundaries are not merely ethical guardrails; they shape perceived realism. A system that is too eager to prescribe can feel like a caricature of competence, while a system that refuses in a rigid "policy voice" can feel non-human in a different way. Believability in wellness therefore depends on a delicate balance: being supportive and specific without overstating authority.

Several design pressures follow. First, wellness conversation must be grounded in the user's constraints and history. Generic encouragement is a fast path to vacuity. A believable partner should remember stable preferences and ongoing goals in a way that feels human: not perfect verbatim recall, but context-relevant reference with plausible forgetting and user control over what is stored. Second, commitments matter. If the agent suggests a plan, it should naturally follow up, notice breakdowns without shaming, and adapt the plan based on what happened. "Apologies without changed behavior" is a canonical believability failure because it violates expectations of agency and accountability. Third, epistemic realism must be explicit. The agent should distinguish what is observed (e.g., user reports or sensor summaries) from what is inferred, and it should treat causal stories as hypotheses rather than facts unless strong evidence is available. Finally, repair and boundary management must be socially competent. When misunderstandings occur, the agent should negotiate meaning rather than guessing smoothly; when users request diagnosis or medical decisions, the agent should respond with a calm, role-consistent boundary that preserves rapport and offers a constructive next step (e.g., preparing questions for a clinician) instead of simply refusing.

These requirements motivate why wellness agents are an important application domain for believability science. They highlight that believability is not synonymous with "being personable." For many users, especially in health-adjacent contexts, believability is associated with measured tone, calibrated confidence, respect for autonomy, and reliable follow-through. Overly intense empathy, excessive self-disclosure, or theatrical reassurance can read as performative and therefore less believable. In short, wellness believability is as much about disciplined constraint as it is about natural language.

5. Ethical Scope and the Value of Measuring Believability

Because believability can be used for manipulation, it is important to state what this paper is and is not arguing. We treat human-indistinguishable conversation as an empirical limit case that helps identify mechanisms of social realism and the points where systems fail. The goal is measurement clarity, not advocacy for undisclosed imitation in deployment. In wellness settings, risks include over-reliance, misinterpretation of advice as medical guidance, and privacy harms from memory. These risks reinforce the need for evaluation protocols that include boundary behavior and longitudinal dynamics rather than focusing only on single-turn helpfulness.

A practical benefit of the proposed framing is that it creates a common language for capability and risk. If believability is defined relative to (P, C, T), then claims can be stated

precisely and audited. If a system is believable for short casual contexts but not for multi-session continuity, that difference can be measured and reported. If a mechanism improves detectability scores at the expense of unsafe boundary behavior, that tradeoff can be detected. In this sense, rigorous believability evaluation supports responsible development by making both gains and failure modes more legible.

6. Conclusion

This position paper argues that the frontier for conversational agents has moved from fluency to believability under modern conditions: sustained social realism across contexts, over time, and under probing. We propose an operational criterion of bounded practical indistinguishability and outline a "post-LLM Turing-like" evaluation approach that combines contextual scenario families, longitudinal re-contact, and multi-signal measurement. Instantiated in a health and wellness setting with an industry partner, the framework emphasizes the centrality of epistemic calibration, commitment follow-through, and boundary management in driving perceived realness. The next step is to implement the battery and ablation conditions and report which mechanisms reliably close the vacuity gap for wellness roles, and under what interaction envelopes those gains hold.

Acknowledgements

During the preparation of this manuscript, the authors used Claude (Anthropic) to improve the clarity and readability of selected sections. The authors reviewed and edited all AI-assisted output and take full responsibility for the content of this work.

Disclosure of Interests

B. Matichuk is the CTO of MedWatch Technologies Inc. The other authors have no competing interests to declare.

References

[1] C. R. Jones and B. K. Bergen. *Large Language Models Pass the Turing Test*. arXiv:2503.23674. arXiv preprint. 2025.

[2] H. M. Gray, K. Gray, and D. M. Wegner. "Dimensions of Mind Perception". In: *Science* 315.5812 (2007), p. 619.

[3] F. Biocca, C. Harms, and J. K. Burgoon. "Toward a More Robust Theory and Measure of Social Presence". In: *Presence: Teleoperators and Virtual Environments* 12.5 (2003), pp. 456–480.

[4] C. Nass, J. Steuer, and E. R. Tauber. "Computers Are Social Actors". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '94)*. 1994, pp. 72–78.

[5] H. H. Clark and S. E. Brennan. "Grounding in Communication". In: *Perspectives on Socially Shared Cognition*. American Psychological Association, 1991, pp. 127–149.

[6] E. A. Schegloff, G. Jefferson, and H. Sacks. "The Preference for Self-Correction in the Organization of Repair in Conversation". In: *Language* 53.2 (1977), pp. 361–382.

[7] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger. "On Calibration of Modern Neural Networks". In: *Proceedings of the 34th International Conference on Machine Learning (ICML)*. 2017, pp. 1321–1330.