PsychEvolve: Subtle Cues Must Keep Earning Their Influence if Full-Duplex Agents Are to Reason Psychologically and Evolve

Junjie Chen[1*], Wanru Li[1*], Yanbin Hao[1✉], Yi Tan[2✉],
Tianle Niu[3], Yulang Chen[4], Pengyu Liu[1], Linfeng Zhang[3], Meng Wang[1]
[1]HFUT, [2]NTU, [3]SJTU, [4]HKUST

Project Page: https://junjie-chen.info/psychevolve-page

[*]Equal contribution. [✉]Corresponding authors.

Abstract

Full-duplex conversational agents can now see and hear users while speaking with them. Through video avatars, 3D avatars, or robots, they increasingly respond with nonverbal behavior as well as speech. This makes them promising partners for psychological support, where what matters is often what a person shows but does not say. Human listeners miss fleeting cues and forget much of what they notice. An agent could in principle attend to every frame, relate each cue to its own behavior, and remember what it learns about a person, and so reason psychologically and evolve. Current systems are still far from this, and we expect that better perception alone will not close the gap. Subtle cues are ambiguous, the agent changes what it observes, and memory turns passing readings into traits. *We argue that if full-duplex agents are to reason psychologically and evolve,* ***subtle cues must keep earning their influence.*** Following Brunswik's lens model, an agent should rely on a cue only as much as the cue has proved to reveal this person's appraisals, should not count its own influence as confirmation, and should trust the cue less when later evidence disagrees. We support this ***position*** by tracing how conversational agents have evolved toward full duplex, together with a brief bibliometric analysis of the field. To make the position concrete, we sketch ***PsychEvolve***, a blueprint spanning data, architecture, and psychological grounding that raises open questions rather than claiming solutions, and we illustrate it with the design of an ***expert-in-the-loop*** avatar platform for psychological support.

1 Introduction

> *"He that has eyes to see and ears to hear may convince himself that no mortal can keep a secret. If his lips are silent, he chatters with his finger-tips; betrayal oozes out of him at every pore."*
>
> — Sigmund Freud (1905)

Freud's observer needed only eyes and ears. Full-duplex conversational agents now have both, and like human listeners they use them while the other person is still talking. Rather than waiting for a turn to end, they listen and respond in overlapping streams (Défossez et al., 2024; Roy et al., 2026; Xie et al., 2026), increasingly with live video as input (Wang et al., 2026b; Cui et al., 2026). This shared timeline is what makes humanlike embodiment a promising path rather than a cosmetic layer. A turn-based agent's face has nothing to respond to while the user speaks, whereas a full-duplex agent can nod, look, and react in the moment, whether as a streaming video avatar (Ki et al., 2026), a 3D or Gaussian-splatting avatar driven by generated motion (Chen et al., 2025, 2026; Nagano et al., 2026; Moon et al., 2024), or a humanoid robot (He et al., 2024; Cheng et al., 2024; Hu et al., 2026a). Such agents are promising partners for emotional support (Sharma et al., 2023; Li et al., 2023), well-being coaching (Spitale et al., 2025; Jeong et al., 2023), and expert-supervised therapy (Garety et al., 2024; Heinz et al., 2025), which we refer to collectively as psychological support. In these settings, much of what matters is never put into words. A client may slow down, glance away, or rub their fingers as a sensitive topic approaches, without remarking on any of these small signs at all.

Human listeners often miss such cues. Micro-expressions last a fraction of a second and often appear in only the upper or lower face (Porter and ten Brinke, 2008), and people recall only about a tenth of a conversation immediately after it ends (Stafford et al., 1987). An agent need not share these limits. ***In principle***, it could attend to every frame without fatigue (Xie et al., 2022; Li et al., 2025; Hao et al., 2026), relate each cue to the words it accompanied and to its own concurrent behavior, and remember what it learns about one person across sessions (Zhang et al., 2026). Such an agent could then ***reason psychologically***, asking what a slowed answer suggests about this person's goals and concerns (Scherer, 2009), and ***evolve***,

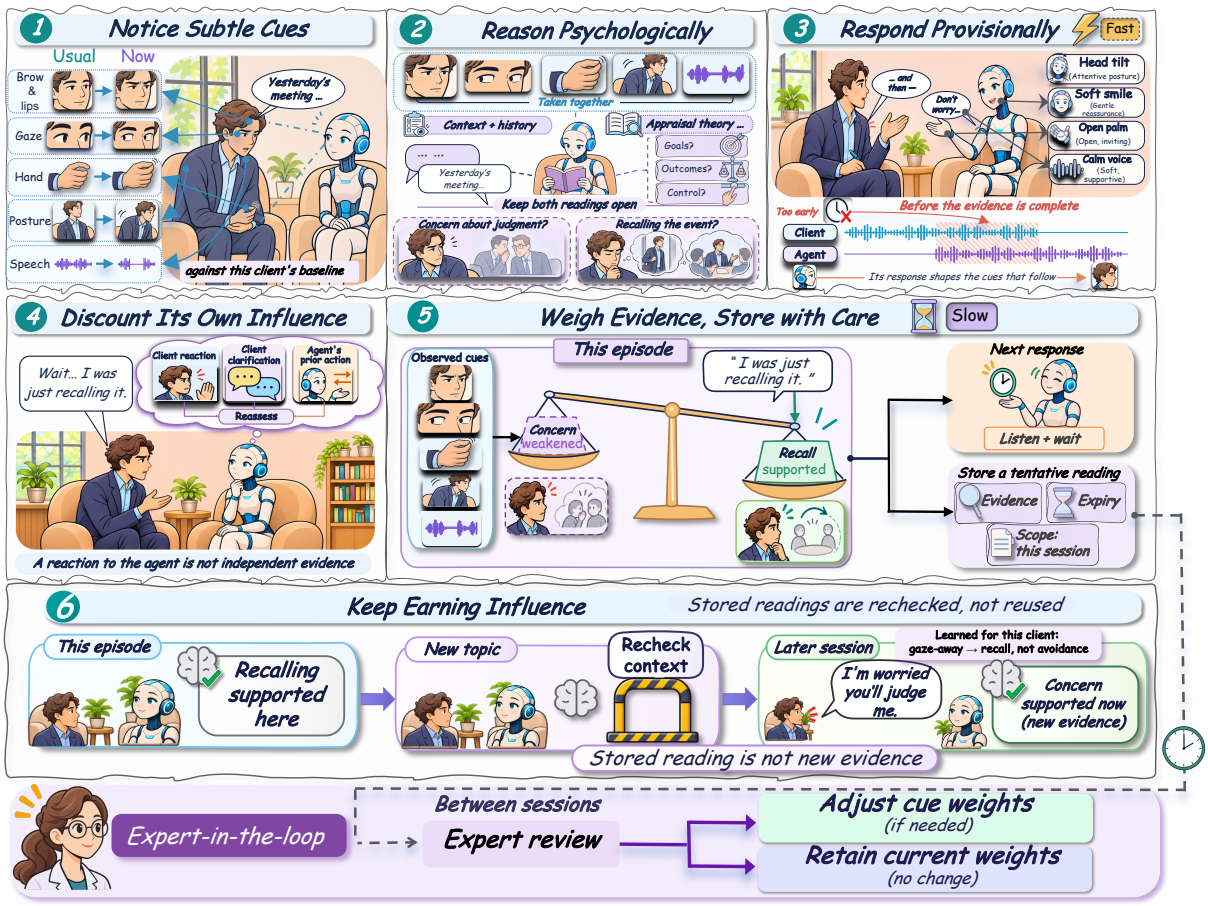


Figure 1: How a subtle cue keeps earning its influence for one client. The fast pathway observes subtle cues, compares possible readings, and responds provisionally (1–3). The agent reassesses the client's reaction together with its own prior action (4). The slow pathway revises the reading with an expert in the loop and decides what to store (5). Stored readings must then be re-earned by new evidence in later sessions rather than reused as facts (6).

learning which cues are informative for this person and in which situations (Mischel and Shoda, 1995).

Current systems are still far from this vision. Speech language models struggle to use expressive vocal cues (Zhou et al., 2026) and often infer a speaker's state from the words alone (Yang et al., 2026), multimodal models can report cues that are absent from the input (Zeng et al., 2026), and human raters can barely tell generated listener motion apart from motion generated for a different partner (Kucherenko et al., 2023; Nagy et al., 2026). We expect that better perception alone will not close this gap, for three reasons. ❶ ***Subtle cues are ambiguous.*** Neither facial configurations nor micro-gestures map reliably onto particular emotions in isolation (Barrett et al., 2019; Liu et al., 2021b): a perfectly detected finger rub may still reflect anxiety, concentration, or simply cold hands. ❷ ***The agent changes what it observes.*** It acts while the user is still speaking, and listeners' responses affect how well speakers tell a story (Bavelas et al., 2000), even when the listener is a robot (Park et al., 2017). An agent that reads the user's next move as evidence about the user alone may mistake its own influence for the user's state. ❸ ***Memory turns passing readings into traits.*** Memory-augmented agents over-apply stored memories (Hu et al., 2026b), miss changes in a user's state (Jiang et al., 2025), and can mistake a transient state for a stable one (Fei et al., 2026), much as people infer lasting traits from merely situational behavior (Gilbert and Malone, 1995).

Freud was confident that a watchful observer could read every secret from a person's fingertips. Modern evidence disagrees. People judging lies from truths are right only 54% of the time (Bond and DePaulo, 2006), and a lens-model meta-analysis traces this limit to the weakness of the cues rather than to observers using the wrong ones (Hartwig and Bond, 2011). Watching more closely does not make a weak cue strong. For an agent that never stops watching, the risk is the re-

verse of Freud's promise: *the more it sees, the more confidently it may read secrets that are not there.*

We argue that if full-duplex agents are to reason psychologically and evolve, subtle cues must keep earning their influence. By *influence* we mean the weight a cue carries in what the agent infers, how it responds, and what it remembers. Brunswik's lens model gives *earning* a precise sense (Brunswik, 1956). An inference is accurate only when the observer relies on each cue as much as that cue is actually valid. For a full-duplex agent, validity is personal and situated: a glance away may signal avoidance in one person and recall in another (Mischel and Shoda, 1995). A cue therefore earns influence by proving, over interaction, that it reveals this person's appraisals in this kind of situation. Behavior that the agent itself elicited does not count as independent confirmation, and influence should be lost as readily as it is gained when later evidence or the person's own account disagrees. Acting on a reading, repeating it, or storing it changes only how the reading is used, never how well the available evidence supports it.

We develop this ***position*** in three steps. ❶ To support our position, we trace how conversational agents have evolved from text dialogue to full-duplex audiovisual interaction and which psychological theories each stage has drawn on (Section 2). Every stage has read emotion largely through coarse cues, and a reproducible bibliometric analysis suggests that research on subtle cues, empathetic dialogue, and user memory has each grown rapidly, while work that combines them for psychological support remains rare. Both observations point to the gap that our position addresses. ❷ To make the position concrete, we sketch ***PsychEvolve***, a blueprint spanning data, architecture, and psychological grounding, and use it to raise open questions rather than to claim solutions (Section 3). It would run on three clocks. Within a turn, a fast pathway sustains full-duplex dialogue, coordinates its voice and nonverbal behavior, and responds provisionally before the evidence is complete. Across turns, a slower team of agents weighs cues against this client's baseline and passes back hypotheses that carry their evidence and expiry. Between sessions, the slowest loop keeps professionals ***expert-in-the-loop***: they may review stored readings and correct how much each cue is worth. Figure 1 follows one client through this loop. ❸ We illustrate how the blueprint could be put into practice with the design of an avatar platform for psychological support (Section 4). It builds on avatar-based interventions (Leff et al., 2013; Garety et al., 2024; Morency et al., 2015). Clients can confirm, correct, or withdraw what the agent has understood about them, experts can correct that understanding with stated reasons, and changes take effect only in a later session with the client's consent. The platform is designed to support professionals rather than to make any diagnoses about clients.

We believe our position bears on a broader question that deserves the community's attention:

Open Question

How can we build psychologically grounded, fine-grained, multimodal, efficient, continually evolving, self-correcting, and full-duplex conversational agents in which subtle cues keep earning their influence, to better understand and support human emotions and psychological well-being?

We suggest one test for any answer: does learning from subtle cues make an agent more accurate about the person, or simply more confident?

2 What Have Conversational Agents Borrowed from Psychology, and What Have They Missed?

A trajectory shaped by psychology. Artificial intelligence has drawn on psychology from its beginnings. Simon and Newell, two of the field's founders, published early accounts of bounded rationality and human problem solving in *Psychological Review* (Simon, 1956; Newell et al., 1958). The perceptron, an ancestor of today's neural networks, appeared in the same journal as a model of how organisms perceive and remember (Rosenblatt, 1958), and the backpropagation work that revived neural networks came from a program on the microstructure of cognition (Rumelhart et al., 1986). Conversational agents follow the same pattern. Figure 2 traces how they have widened what they can perceive and express, from scripted text to agents that listen while speaking, see the user, render a face or body, and remember across sessions. Appendix A reviews each of these stages and its psychological roots in more detail. ***Psychology has inspired AI, and AI may in turn help psychology.*** Our position aims at this second direction: agents that *reason psychologically* to better understand and support the people who are seeking help.

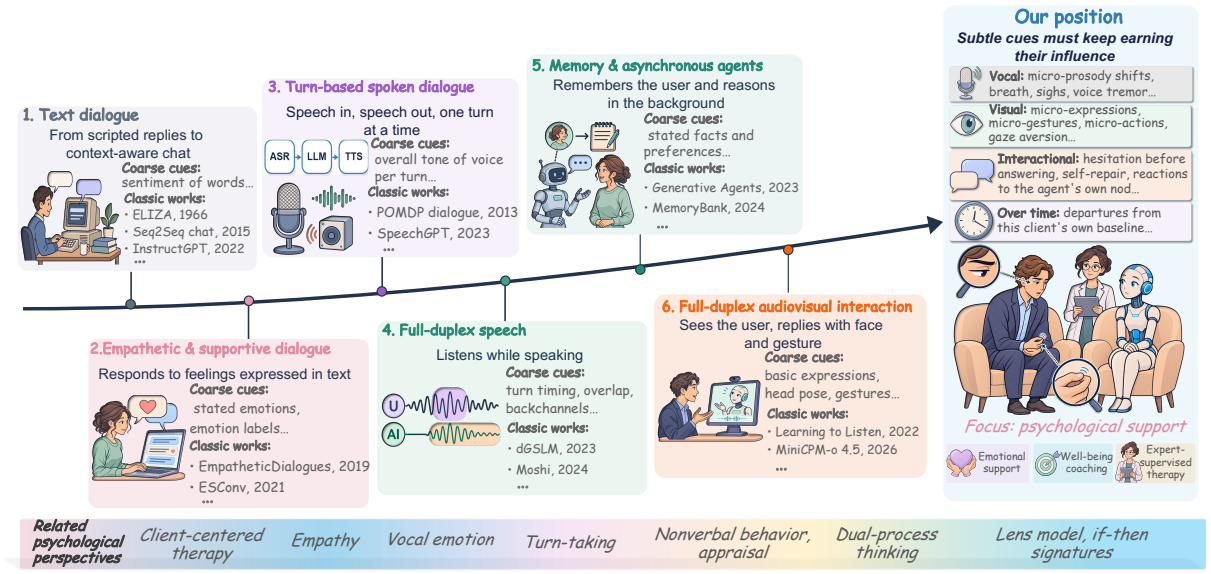


Figure 2: From coarse to subtle cues. Each stage of conversational agents can read emotion, but only through coarse cues in its own modality. Our position concerns subtle vocal, visual, interactional, and longitudinal cues in psychological support. Years mark approximate starts.

What did each stage borrow? ELIZA's reflective replies echo client-centered therapy (Weizenbaum, 1966; Rogers, 1957), and neural models later learned to hold context and follow instructions (Vinyals and Le, 2015; Ouyang et al., 2022). Empathetic dialogue taught agents to recognize and acknowledge feelings (Rashkin et al., 2019; Davis, 1983), and emotional support systems were grounded in Helping Skills Theory (Liu et al., 2021a). Spoken dialogue systems modeled uncertainty about the user explicitly (Young et al., 2013) and later moved to speech language models (Zhang et al., 2023a), giving access to the voice, a long-studied channel of emotion (Scherer, 2003). Full-duplex models brought in the overlaps, pauses, and backchannels described by conversation analysis (Sacks et al., 1974; Nguyen et al., 2023; Défossez et al., 2024). Memory-augmented agents that keep reflecting quietly in the background echo dual-process accounts of human thinking (Park et al., 2023; Zhong et al., 2024; Evans and Stanovich, 2013). Full-duplex audiovisual interaction now adds the face and body, as agents see the user and respond with their own face and gesture while still listening (Ekman and Friesen, 1969; Zhu et al., 2025; Ki et al., 2026; Cui et al., 2026), much as attentive listeners nod and smile at a speaker.

What has been missed? Each stage widened the channels an agent can read, yet the cues it reads have remained largely coarse: words, emotion labels, the tone of a whole turn, or a basic facial expression, each of which can be read on its own. Psychology has mostly been borrowed to decide how an agent should behave, and rarely to decide how much a cue should be trusted. To our knowledge, few systems at any stage ask whether a cue is valid for this person, whether the agent itself produced it, or when its influence should lapse. These are the three conditions that our position asks agents to consider. But what makes a cue subtle rather than coarse, and is the field already moving toward such cues? We first look at what the literature counts show, and then we return to answer the first of these two questions in turn.

Is the field moving toward subtle cues? Figure 3 gives a coarse, keyword-based estimate (Appendix B). We split the years into three periods at the field's turning points: neural dialogue before large language models (2016–2019), the arrival of large pretrained models (2020–2022) (Brown et al., 2020), and the period after instruction-tuned chat models reached the public (2023–2026) (Ouyang et al., 2022; OpenAI, 2023). The last period shows a burst. As abilities emerged at scale (Wei et al., 2022), conversational agents became usable in practice, and research followed. Empathetic dialogue, psychological support agents, and full-duplex dialogue grew several-fold. Their combination has not kept pace. Work on subtle cues in dialogue grew only modestly, and each added requirement, psy-

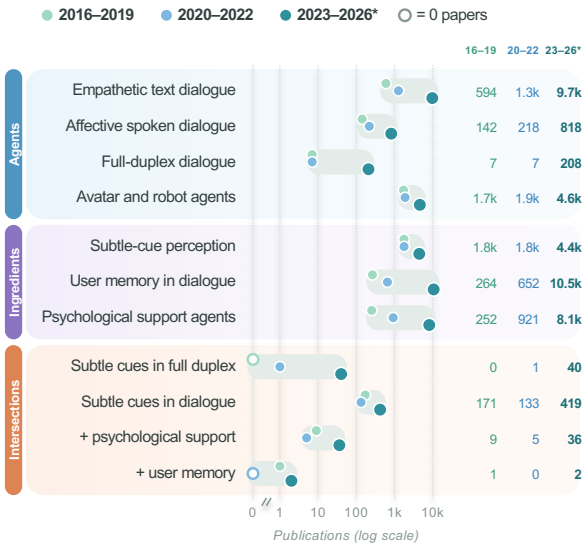


Figure 3: Publications per period by topic, estimated from keyword searches of titles and abstracts in a single source, OpenAlex (Appendix B). Counts are coarse estimates and show broad trends only. A row starting with "+" adds one condition to the row above. Hollow dots mark zero counts. Exact counts are on the right. *The data for 2026 cover only part of the year.

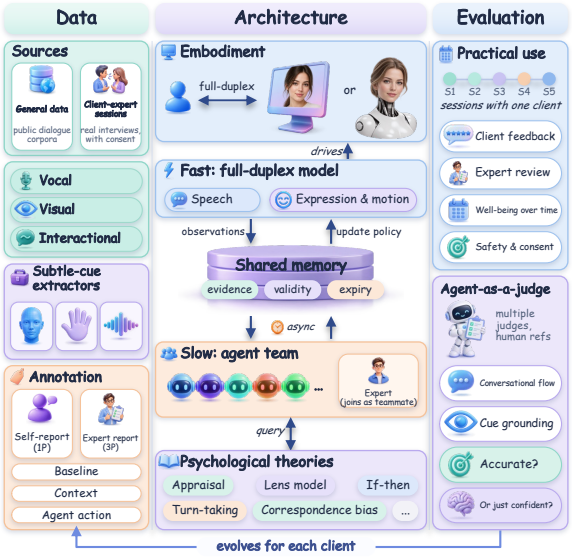


Figure 4: The proposed PsychEvolve framework. Psychology-oriented data (left) feed a fast full-duplex model that drives an avatar or humanoid robot, a shared memory of hypotheses, and a slower agent team that consults psychological theories and that experts may join (center). Evaluation (right) rests mainly on practical use, with agents serving only as auxiliary judges.

chological support and then user memory, shrinks the count until almost nothing remains. We read every paper in the smallest intersection and found none that clearly combines all four of them.

What makes a cue subtle? Following the lens model, a *cue* is an observable behavior that is probabilistically related to an unobservable state (Brunswik, 1956; Hartwig and Bond, 2011). We call a cue *subtle* when it is brief or low in intensity, carried by how something is said or done rather than by what is said, and interpretable only relative to the ongoing interaction and the person's own baseline. Psychology already offers fine-grained coding systems for such behavior, such as FACS for facial action units (Ekman and Friesen, 1978) and BAP for body actions and postures (Dael et al., 2012). Subtle cues span vocal, visual, interactional, and longitudinal channels, from fleeting micro-expressions (Porter and ten Brinke, 2008) and unintentional micro-gestures or micro-actions (Liu et al., 2021b; Guo et al., 2024) to a change in how a client responds to the agent itself. No list can be complete, because the same behavior may be telling for one person and ordinary for another. Coarse cues can often be read in isolation, but subtle cues cannot, which is why we argue that they must keep earning their influence over time.

3 A Blueprint, Not Yet a System: What Would PsychEvolve Need?

From position to blueprint. Return to the client from our introduction, who slows down and rubs their fingers as a sensitive topic approaches. What would an agent need in order to notice this, respond in the moment, and learn over many sessions how much such a cue means for this person? We sketch one possible answer, PsychEvolve (Figure 4). It is a blueprint rather than a finished system, and we present it through the questions it raises, organized by data, architecture, and evaluation. Three requirements from Section 1 frame each question. **R1**: a cue's validity is personal and situated. **R2**: behavior the agent itself elicited is not independent evidence. **R3**: influence must be lost as readily as it is gained when later evidence disagrees.

Whose cue is it? Learning what a cue means for one person starts with data that know whose cue it is. The largest face-to-face corpus to date, Seamless Interaction, records more than 4,000 hours of dyadic conversation and pairs participants' self-reports with annotations by trained observers (Agrawal et al., 2025). Yet such corpora target general social behavior, their dense psychological labels cover only a small fraction of the footage,

and they rarely record how each person usually behaves or what the other party did just before. How, then, should data for psychological support be built? We envision combining general dialogue data with consented sessions between clients and professionals. Subtle-cue extractors could turn raw streams into candidate evidence, drawing on facial action units (Ekman and Friesen, 1978; Baltrušaitis et al., 2018), body actions and postures (Dael et al., 2012), micro-gestures (Wang et al., 2026c), vocal change (Scherer, 2003), and the timing of the exchange (Sacks et al., 1974). But which cues deserve extraction, and when does a self-report (1P) outweigh an expert's report (3P), or the reverse? Annotations that also mark the client's baseline, the context, and the agent's preceding action would let these questions be studied, as R1 and R2 require.

How should the agent act before it is sure? Full-duplex models now listen while they speak (Défossez et al., 2024) and can drive a face or body in real time (Ki et al., 2026; Chen et al., 2026). In our blueprint, a fast full-duplex model holds the conversation and drives either a photorealistic avatar or a humanlike robot, and its replies are provisional by design. This raises a question that current systems rarely face. How should a live model be steered by a memory that keeps changing beneath it? Before each response, the fast model could read the latest entries in a shared memory and adjust both what it says and how it drives voice, expression, and motion. How often should it read memory, how should a stale or uncertain entry soften its behavior, and how should it record its own actions so that later reasoning can properly discount them?

What should the agent remember? Existing agent memories store what a user said or prefers (Zhong et al., 2024; Park et al., 2023), and they tend to over-apply stored memories and miss changes in a user's state (Hu et al., 2026b; Jiang et al., 2025). A memory for subtle cues may need to hold hypotheses instead of facts, each with its evidence, its estimated validity for this client, and the conditions under which it expires. A reading that is reused but never supported again should decay rather than harden into a trait, as R3 asks. What exactly should such a hypothesis contain, and how quickly should it fade if it is never supported again?

How might a team of agents read one person? Dual-process accounts of cognition (Evans and Stanovich, 2013) have inspired agents that pair a fast responder with a slower reasoner (Yu et al., 2026; Wang et al., 2026d). These systems coordinate tasks and tool use. We ask whether the slow side could instead coordinate hypotheses about a person. Imagine a team of collaborating agents (Wu et al., 2023) working asynchronously on the shared memory, some specializing in one channel of cues and others consulting psychological theories, such as appraisal (Scherer, 2009), the lens model (Brunswik, 1956), if-then signatures (Mischel and Shoda, 1995), turn-taking (Sacks et al., 1974), or the correspondence bias (Gilbert and Malone, 1995). Related systems already ground affect in appraisal (Quan et al., 2026), weight modalities by their reliability for each utterance (Wang et al., 2026a), or let a counselor evolve its skills at test time (Yan et al., 2026). Fewer ask how much each cue is worth for one client. How should these agents divide the work, and how should they resolve disagreement without letting the single most confident voice win?

Where do experts fit? A professional might join the team as its slowest member. Between sessions, an expert could review stored hypotheses and correct how much a cue is worth for a client, in line with proposals for supervised, collaborative use of language models in psychotherapy (Stade et al., 2024; Sharma et al., 2023). The expert would not steer the live conversation, and each correction would update validity in memory. Yet how much expert time can such a loop demand, and how should an agent weigh an expert's reading against the client's own account of themselves?

How would we know it works? Psychological support is complex and subjective, and its effects unfold over weeks. Static benchmarks can test pieces of the problem. VideoFDB, for example, scores conversational flow and visual grounding in full-duplex audiovisual dialogue (Mazumdar et al., 2026). They cannot show whether an agent's understanding of one person improves. We therefore argue that evaluation should rest mainly on practical use, through repeated sessions with the same client, client feedback, expert review, change in well-being, and adherence to safety and consent, following the logic of clinical trials of conversational agents (Heinz et al., 2025). Model judges can assist (Zheng et al., 2023; Zhuge et al., 2024), for instance by scoring flow and cue grounding with several judges against human references, as VideoFDB does. Because such judges may share the

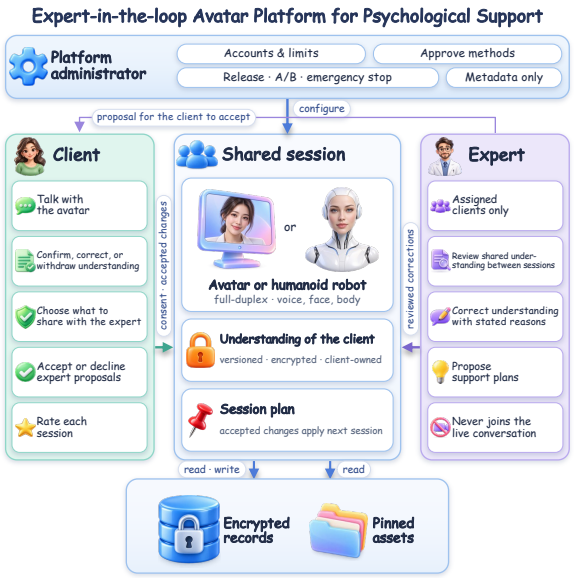


Figure 5: Roles in the envisioned platform. The client talks with the avatar and controls what is shared and accepted. The expert reviews and corrects the agent's understanding between sessions and never joins the live conversation. An administrator configures the platform and can see only the metadata of each session.

biases of the models they assess, we treat them as a complement to human judgment, not a replacement. Which outcomes should count, over what horizon, and how can such studies stay safe? Across all of these questions, the test remains the one posed in our introduction: does learning from subtle cues make the agent more accurate about the person, or simply more confident in its own readings?

4 Who Decides What the Agent Believes? An Expert-in-the-Loop Design

From blueprint to a design. To make our position concrete, we ask what it would demand of a system that people actually use. Conversational agents are increasingly studied as a way to extend such support beyond the therapy room (Li et al., 2023; Heinz et al., 2025), and avatars have already entered clinical practice. In AVATAR therapy, a clinician voices a digital avatar of a distressing voice and gradually hands control to the client, an approach tested in randomized trials (Leff et al., 2013; Garety et al., 2024). Inspired by this work, we designed an expert-in-the-loop avatar platform for psychological support (Figure 5). It departs from AVATAR therapy in one respect that matters for our position. The agent, not a clinician, forms an understanding of the client, and that understanding remains open to correction by both the client and an expert. The design illustrates PsychEvolve rather than validating it, and it never makes clinical diagnoses on its own behalf or without an expert.

What can the client do? The client talks with an avatar or humanoid robot and also governs what the agent comes to believe. Every reading the agent stores is shown to the client, who can confirm, correct, or withdraw it, and a withdrawn reading loses its influence at once, as R3 requires. The client decides what to share with an assigned expert, accepts or declines each proposed change, and rates each session. This control matters for disclosure. People disclose more to a virtual human when they believe no one is watching (Lucas et al., 2014), so expert involvement should never be hidden and should depend on what the client chooses to share.

What can the expert do? The expert acts on the slowest clock of PsychEvolve. Between sessions, an expert reviews the understanding that an assigned client has chosen to share, corrects it with stated reasons, and proposes support plans, in line with proposals for supervised use of language models in psychotherapy (Stade et al., 2024). The expert never joins the live conversation, which keeps the fast pathway responsive and keeps clinical judgment out of split-second decisions. A correction does not change the agent directly. It becomes a proposal that the client may accept or decline.

How do changes reach the next session? Each session starts from a pinned plan, so nothing changes silently while the client is talking. Accepted corrections and plans take effect only in the next session, and every version of the agent's understanding is kept, encrypted, and owned by the client. An administrator configures accounts, approves which methods may run, and can roll back a release or stop it in an emergency, but sees only metadata and never the content of a session. These rules turn the abstract requirements of Section 3 into routines that clinicians and clients can audit.

What remains open? The design leaves several questions for empirical work. How often should a client be asked to confirm a reading before confirmation itself becomes a burden? And how can expert review scale to many clients without turning into a rubber stamp? Appendix C shows the platform's interface and its main functions.

5 Alternative Views

Objections worth taking seriously. A position is only as strong as the objections it survives. We consider five and say where each is right.

Why not simply ask the person? Self-report is the most direct evidence, and R3 lets it overturn a reading. Yet people cannot always say what they feel and talk less freely when asked too often. Cues and questions are partners. A cue that has earned some influence can tell the agent when a gentle question is worth asking, and the answer then shows how much that cue deserved.

Can a cue's validity be learned for one person? Subtle cues are weak (Hartwig and Bond, 2011), and one client offers few sessions, so a skeptic may doubt that person-specific validity can ever be estimated. Bayesian accounts of affect and dialogue suggest one route, starting from population priors and updating them with each person's evidence (Hoey et al., 2013; Young et al., 2013). Slow learning is itself a safeguard, since a cue should keep little influence while evidence is scarce.

Should agents read unspoken cues at all? Inferring emotion from faces is scientifically contested (Barrett et al., 2019), an agent that records every glance could profile a person more closely than any friend, and people disclose less when others may review what they share (Lucas et al., 2014). These concerns argue against confident inference from cues, which is exactly what our position rejects. We therefore treat consent, client ownership of stored readings, the right to withdraw them, and a non-diagnostic scope as requirements of the design rather than as optional extra features.

Should AI offer psychological support at all? Some argue that support for vulnerable people should come only from humans, and we share much of this concern. Trials of chatbots report benefits but leave safety and long-term effects open (Li et al., 2023; Heinz et al., 2025). Our position does not claim that agents should replace professionals. It claims that if agents are used in this setting, as they increasingly are, their beliefs about a person must be earned, visible, and correctable.

Won't scale make this unnecessary? Larger end-to-end models may learn to handle cues well without explicit memory, and our requirements do not depend on modules. Yet more data of the same kind may not help. Recorded conversations do not separate the user's state from the listener's influence, so a model trained to imitate them has little reason to learn R2. We therefore see R1 to R3 as tests that any model architecture should pass.

6 Discussion

Seeing more is not knowing more. As the Freud quote that opens Section 1 puts it, a watchful observer can read every secret. Full-duplex agents may soon watch more closely than any human listener, a capacity that we have argued cuts both ways. If such agents are to reason psychologically and evolve, subtle cues must keep earning their influence, one person and one situation at a time.

Why won't this emerge on its own? Today's full-duplex models generate chunk by chunk from an interleaved stream of user and agent tokens (Défossez et al., 2024; Nagano et al., 2026), and what they have emitted is not revisited later. This is why PsychEvolve lets the live model act at once while keeping revisable hypotheses about each person outside it, where both the person and experts can review and correct each of them between sessions.

What should the community change? We call for data that follow a person over time and record the agent's own actions alongside self-reports and expert reports. We encourage models that ask how much a cue should be trusted, not only how accurately it is detected. And we recommend evaluations that ask, beyond how a single response scores, whether an agent's understanding of one person improves across sessions and stays open to correction by the person concerned and by experts.

Why psychological support first? We argue that psychological support should come first because subtle cues matter most there and misreading them costs most. Its tradition of supervision and consent also makes it a suitable place to develop such agents before they spread to everyday settings with far weaker safeguards for the people who rely on them.

What remains open? PsychEvolve and our platform are early and incomplete sketches that we hope others will test, question, and improve. One test may help judge any answer: does the agent become more accurate about the person, or simply more confident in its own readings?

Limitations

This is a position paper, and its claims rest on argument and prior evidence rather than on new experiments. We have not shown that current systems violate R1 to R3, nor that PsychEvolve would satisfy them. These are hypotheses for future work. Our bibliometric analysis relies on keyword searches of a single source, OpenAlex, whose precision varies across topics and whose coverage has grown sharply since 2025. It supports only coarse comparisons across topics. The validity of a cue may also differ across cultures and populations. The platform is a design that has not been evaluated with clients or clinicians, and the questions we raise about data, memory, collaboration among agents, and expert review remain open. Finally, our scope is psychological support rather than diagnosis or treatment, and our arguments may not transfer to clinical settings without further careful validation with both clients and clinicians in practice.

Ethics Statement

Agents that read subtle cues in psychological settings carry clear risks. Continuous capture of faces, voices, and bodies could enable profiling or surveillance, and readings stored over time could harden into labels that follow a person. A visible human reviewer may also reduce what clients choose to disclose (Lucas et al., 2014). Our position and design respond to these risks by treating consent, client ownership of stored readings, the right to withdraw them, and a non-diagnostic scope as requirements. Any collection of client and expert sessions should require ethical approval and informed consent, data minimization, and secure storage. Any deployment must also comply with the laws of each country and region where it operates, including data protection, health, and AI regulations, and with the usage policies of the models and data it relies on. Experts should review only what clients choose to share, and no change should reach a client without their agreement. Such agents should complement professional care and never replace it, especially for people in crisis. The faces of avatars and robots in our figures are generated and do not depict real people. AI assistants were used to help edit the text and draw icons, and the authors take full responsibility for the content and claims of this paper.

References

Vasu Agrawal, Akinniyi Akinyemi, Kathryn Alvero, Morteza Behrooz, Julia Buffalini, Fabio Maria Carlucci, Joy Chen, Junming Chen, Zhang Chen, Shiyang Cheng, Praveen Chowdary, Joe Chuang, Antony D'Avirro, Jon Daly, Ning Dong, Mark Duppenthaler, Cynthia Gao, Jeff Girard, Martin Gleize, and 65 others. 2025. Seamless interaction: Dyadic audiovisual motion modeling and large-scale dataset. *Preprint*, arXiv:2506.22554.

Nalini Ambady and Robert Rosenthal. 1992. Thin slices of expressive behavior as predictors of interpersonal consequences: A meta-analysis. *Psychological Bulletin*, 111(2):256–274.

John W. Ayers, Adam Poliak, Mark H. Dredze, Eric Craig Leas, Zechariah Zhu, Jessica B. Kelley, Dennis J. Faix, Aaron Michael Goodman, Chris Longhurst, Michael A. Hogarth, and Davey M. Smith. 2023. Comparing Physician and Artificial Intelligence Chatbot Responses to Patient Questions Posted to a Public Social Media Forum. *JAMA Internal Medicine*, 183(6):589.

Alan Baddeley. 2000. The episodic buffer: a new component of working memory? *Trends in Cognitive Sciences*, 4(11):417–423.

Tadas Baltrušaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. 2018. OpenFace 2.0: Facial behavior analysis toolkit. In *IEEE International Conference on Automatic Face & Gesture Recognition (FG)*, pages 59–66.

Rainer Banse and Klaus R. Scherer. 1996. Acoustic profiles in vocal emotion expression. *Journal of Personality and Social Psychology*, 70(3):614–636.

Lisa Feldman Barrett, Ralph Adolphs, Stacy Marsella, Aleix M. Martinez, and Seth D. Pollak. 2019. Emotional expressions reconsidered: Challenges to inferring emotion from human facial movements. *Psychological Science in the Public Interest*, 20(1):1–68.

Janet Beavin Bavelas, Linda Coates, and Trudy Johnson. 2000. Listeners as co-narrators. *Journal of Personality and Social Psychology*, 79(6):941–952.

Timothy W. Bickmore and Rosalind W. Picard. 2005. Establishing and maintaining long-term human-computer relationships. *ACM Transactions on Computer-Human Interaction*, 12(2):293–327.

Charles F. Bond and Bella M. DePaulo. 2006. Accuracy of deception judgments. *Personality and Social Psychology Review*, 10(3):214–234.

Cynthia Breazeal. 2003. Emotion and sociable humanoid robots. *International Journal of Human-Computer Studies*, 59(1-2):119–155.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda

Askell, et al. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901.

Egon Brunswik. 1956. *Perception and the Representative Design of Psychological Experiments*. University of California Press, Berkeley.

Tanya L. Chartrand and John A. Bargh. 1999. The chameleon effect: The perception-behavior link and social interaction. *Journal of Personality and Social Psychology*, 76(6):893–910.

Junjie Chen, Fei Wang, Zhihao Huang, Qing Zhou, Kun Li, Dan Guo, Linfeng Zhang, and Xun Yang. 2025. Towards seamless interaction: Causal turn-level modeling of interactive 3d conversational head dynamics. *arXiv preprint arXiv:2512.15340*.

Junjie Chen, Fei Wang, Kun Li, Yiqi Nie, Xun Yang, Yanbin Hao, Linfeng Zhang, and Meng Wang. 2026. Evolvingavatar: Interactive 3d head generation that adapts as conversations unfold. *arXiv preprint arXiv:2609.35616*.

Xuxin Cheng, Yandong Ji, Junming Chen, Ruihan Yang, Ge Yang, and Xiaolong Wang. 2024. Expressive whole-body control for humanoid robots. *Preprint*, arXiv:2402.16796.

Herbert H. Clark and Edward F. Schaefer. 1989. Contributing to Discourse. *Cognitive Science*, 13(2):259–294.

Junbo Cui, Bokai Xu, Chongyi Wang, Tianyu Yu, Weiyue Sun, Yingjing Xu, Tianran Wang, Zhihui He, Wenshuo Ma, Tianchi Cai, et al. 2026. Minicpm-o 4.5: Towards real-time full-duplex omni-modal interaction. *arXiv preprint arXiv:2604.27393*.

Nele Dael, Marcello Mortillaro, and Klaus R. Scherer. 2012. The body action and posture coding system (BAP): Development and reliability. *Journal of Nonverbal Behavior*, 36(2):97–121.

Mark H. Davis. 1983. Measuring individual differences in empathy: Evidence for a multidimensional approach. *Journal of Personality and Social Psychology*, 44(1):113–126.

Alexandre Défossez, Laurent Mazaré, Manu Orsini, Amélie Royer, Patrick Pérez, Hervé Jégou, Edouard Grave, and Neil Zeghidour. 2024. Moshi: a speech-text foundation model for real-time dialogue. *arXiv preprint arXiv:2410.00037*.

Paul Ekman and Wallace V. Friesen. 1969. The repertoire of nonverbal behavior: Categories, origins, usage, and coding. *Semiotica*, 1(1):49–98.

Paul Ekman and Wallace V. Friesen. 1978. Facial action coding system. Consulting Psychologists Press.

Erik Ekstedt and Gabriel Skantze. 2022. Voice Activity Projection: Self-supervised Learning of Turn-taking Events. In *Interspeech (USB)*, pages 5190–5194.

Robert Elliott, Arthur C. Bohart, Jeanne Cherry Watson, and Leslie S. Greenberg. 2011. Empathy. *Psychotherapy*, 48(1):43–49.

Jonathan St. B. T. Evans and Keith E. Stanovich. 2013. Dual-process theories of higher cognition: Advancing the debate. *Perspectives on Psychological Science*, 8(3):223–241.

Qingkai Fang, Yan Zhou, Shoutao Guo, Shaolei Zhang, and Yang Feng. 2025. LLaMA-omni 2: LLM-based real-time spoken chatbot with autoregressive streaming speech synthesis. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 18617–18629, Vienna, Austria. Association for Computational Linguistics.

Tianxiang Fei, Mingyang Song, Mao Zheng, and Xiang Yu. 2026. Memory beyond recall: A dual-process cognitive memory system for self-evolving LLM agents. *Preprint*, arXiv:2606.09483.

Kathleen Kara Fitzpatrick, Alison M. Darcy, and Molly Vierhile. 2017. Delivering Cognitive Behavior Therapy to Young Adults With Symptoms of Depression and Anxiety Using a Fully Automated Conversational Agent (Woebot): A Randomized Controlled Trial. *JMIR Mental Health*, 4(2):e19.

Sigmund Freud. 1905. *Fragment of an Analysis of a Case of Hysteria (1905 [1901])*, volume 7.

Philippa A. Garety, Clementine Edwards, Hassan Jafari, Richard Emsley, Mark A. Huckvale, Mar Rus-Calafell, Miriam Fornells-Ambrojo, Andrew I. Gumley, Gillian Haddock, Sandra Bucci, Hamish J. McLeod, Jeffrey McDonnell, Moya Clancy, Michael Fitzsimmons, Hannah Ball, Alice E. Montague, Nikos Xanidis, Amy Hardy, Tom Craig, and Thomas Ward. 2024. Digital AVATAR therapy for distressing voices in psychosis: the phase 2/3 AVATAR2 trial. *Nature Medicine*, 30(12):3658–3668.

Premanand Pralhad Ghadekar, Ajinkya Matre, Piyush Pise, Harshada Sawai, Manoj Mule, and Sanika Shedbale. 2025. Realistic AI based mental health detection and analysis model using AR agentic AI & Conv-BiLSTM architecture. In *IEEE International Symposium on Sustainable Energy, Signal Processing and Cyber Security (iSSSC)*, pages 1–6.

Daniel T. Gilbert and Patrick S. Malone. 1995. The correspondence bias. *Psychological Bulletin*, 117(1):21–38.

Dan Guo, Kun Li, Bin Hu, Yan Zhang, and Meng Wang. 2024. Benchmarking micro-action recognition: Dataset, methods, and applications. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(7):6238–6252.

Ying Guo, Xi Liu, Cheng Zhen, Pengfei Yan, and Xiaoming Wei. 2025. Arig: Autoregressive interactive head generation for real-time conversations. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 12956–12965.

Yanbin Hao, Pengyu Liu, Xing Wei, Xun Yang, Dan Guo, and Meng Wang. 2026. A new multi-domain benchmark for micro-action recognition and detection. *arXiv preprint arXiv:2606.14096*.

Maria Hartwig and Charles F. Bond. 2011. Why do lie-catchers fail? a lens model meta-analysis of human lie judgments. *Psychological Bulletin*, 137(4):643–659.

Tairan He, Zhengyi Luo, Wenli Xiao, Chong Zhang, Kris Kitani, Changliu Liu, and Guanya Shi. 2024. Learning human-to-humanoid real-time whole-body teleoperation. *Preprint*, arXiv:2403.04436.

Michael V. Heinz, Daniel M. Mackin, Brianna M. Trudeau, Sukanya Bhattacharya, Yinzhou Wang, Haley A. Banta, Abi D. Jewett, Abigail J. Salzhauer, Tess Z. Griffin, and Nicholas C. Jacobson. 2025. Randomized trial of a generative AI chatbot for mental health treatment. *NEJM AI*, 2(4):AIoa2400802.

Jesse Hoey, Tobias Schroeder, and Areej Alhothali. 2013. Affect control processes: Intelligent affective interaction using a partially observable Markov decision process. *Preprint*, arXiv:1306.5279. Version 2, revised April 2014.

Yuhang Hu, Jiong Lin, Judah Allen Goldfeder, Philippe M Wyder, Yifeng Cao, Steven Tian, Yunzhe Wang, Jingran Wang, Mengmeng Wang, Jie Zeng, et al. 2026a. Learning realistic lip motions for humanoid face robots. *Science Robotics*, 11(110):eadx3017.

Yulin Hu, Zimo Long, Jiahe Guo, Xingyu Sui, Xing Fu, Weixiang Zhao, Yanyan Zhao, and Bing Qin. 2026b. OP-Bench: Benchmarking over-personalization for memory-augmented personalized conversational agents. In *Proceedings of the 2026 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Rongjie Huang, Mingze Li, Dongchao Yang, Jiatong Shi, Xuankai Chang, Zhenhui Ye, Yuning Wu, Zhiqing Hong, Jiawei Huang, Jinglin Liu, Yi Ren, Yuexian Zou, Zhou Zhao, and Shinji Watanabe. 2024. AudioGPT: Understanding and Generating Speech, Music, Sound, and Talking Head. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 23802–23804.

Sooyeon Jeong, Laura Aymerich-Franch, Kika Arias, Sharifa Alghowinem, Agata Lapedriza, Rosalind W. Picard, Hae Won Park, and Cynthia Breazeal. 2023. Deploying a robotic positive psychology coach to improve college students' psychological well-being. *User Modeling and User-Adapted Interaction*, 33(2):571–615.

Bowen Jiang, Zhuoqun Hao, Young-Min Cho, Bryan Li, Yuan Yuan, Sihao Chen, Lyle Ungar, Camillo J. Taylor, and Dan Roth. 2025. Know me, respond to me: Benchmarking LLMs for dynamic user profiling and personalized responses at scale. In *Conference on Language Modeling (COLM)*.

Patrik N. Juslin and Petri Laukka. 2003. Communication of emotions in vocal expression and music performance: Different channels, same code? *Psychological Bulletin*, 129(5):770–814.

Daniel Kahneman. 2003. A perspective on judgment and choice: mapping bounded rationality. *The American psychologist*, 58 9:697–720.

Taekyung Ki, Sangwon Jang, Jaehyeong Jo, Jaehong Yoon, and Sung Ju Hwang. 2026. Avatar forcing: Real-time interactive head avatar generation for natural conversation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18074–18084.

Taras Kucherenko, Rajmund Nagy, Youngwoo Yoon, Jieyeon Woo, Teodor Nikolov, Mihail Tsakov, and Gustav Eje Henter. 2023. The GENEA challenge 2023: A large-scale evaluation of gesture generation models in monadic and dyadic settings. In *Proceedings of the International Conference on Multimodal Interaction (ICMI)*, pages 792–801.

Julian Leff, Geoffrey Williams, Mark A. Huckvale, Maurice Arbuthnot, and Alex P. Leff. 2013. Computer-assisted therapy for medication-resistant auditory hallucinations: proof-of-concept study. *British Journal of Psychiatry*, 202(6):428–433.

Stephen C. Levinson. 2015. Turn-taking in Human Communication – Origins and Implications for Language Processing. *Trends in Cognitive Sciences*, 20(1):6–14.

Han Li, Renwen Zhang, Yi-Chieh Lee, Robert E. Kraut, and David C. Mohr. 2023. Systematic review and meta-analysis of AI-based conversational agents for promoting mental health and well-being. *npj Digital Medicine*, 6:236.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A Diversity-Promoting Objective Function for Neural Conversation Models. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 110–119.

Kun Li, Pengyu Liu, Dan Guo, Fei Wang, Zhiliang Wu, Hehe Fan, and Meng Wang. 2025. Mmad: Multi-label micro-action detection in videos. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 13225–13236. IEEE.

Bill Yuchen Lin, Yicheng Fu, Karina Yang, Faeze Brahman, Shiyu Huang, Chandra Bhagavatula, Prithviraj Ammanabrolu, Yejin Choi, and Xiang Ren. 2023. SwiftSage: A Generative Agent with Fast and Slow Thinking for Complex Interactive Tasks. In *neural information processing systems*, pages 23813–23825.

Guan-Ting Lin, Prashanth Gurunath Shivakumar, Ankur Gandhe, Chao-Han Huck Yang, Yile Gu, Shalini

Ghosh, Andreas Stolcke, Hung yi Lee, and Ivan Bulyko. 2024. Paralinguistics-Enhanced Large Language Modeling of Spoken Dialogue. In *IEEE International Conference on Acoustics Speech and Signal Processing*, pages 10316–10320.

Zhaojiang Lin, Andrea Madotto, Jamin Shin, Peng Xu, and Pascale Fung. 2019. MoEL: Mixture of Empathetic Listeners. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 121–132.

Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie Huang. 2021a. Towards emotional support dialog systems. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3469–3483. Association for Computational Linguistics.

Xin Liu, Henglin Shi, Haoyu Chen, Zitong Yu, Xiaobai Li, and Guoying Zhao. 2021b. iMiGUE: An identity-free video dataset for micro-gesture understanding and emotion analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10626–10637.

Xudong Lu, Xueying Li, Annan Wang, Yang Bo, Jinpeng Chen, Zengliang Li, Nianzu Yang, Rui Liu, Xue Yang, Jingwen Hou, and Hongsheng Li. 2026. OmniInteract: Benchmarking real-world streaming interaction for real-time omnimodal assistants. In *Proceedings of the 2026 Conference on Empirical Methods in Natural Language Processing*, pages 11272–11293.

Gale M. Lucas, Jonathan Gratch, Aisha King, and Louis-Philippe Morency. 2014. It's only a computer: Virtual humans increase willingness to disclose. *Computers in Human Behavior*, 37:94–100.

Adyasha Maharana, Dong-Ho Lee, Sergey Tulyakov, Mohit Bansal, Francesco Barbieri, and Yuwei Fang. 2024. Evaluating very long-term conversational memory of LLM agents. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13851–13870. Association for Computational Linguistics.

Amrita Mazumdar, Seonwook Park, Rajarshi Roy, Nikhil Srihari, Shengze Wang, Yuhao Zhou, Julia Wang, Koki Nagano, and Shalini De Mello. 2026. VideoFDB: Evaluating full-duplex vision-speech capabilities in conversational agents. *Preprint*, arXiv:2605.30256. Version 2, July 2026.

Walter Mischel and Yuichi Shoda. 1995. A cognitive-affective system theory of personality: Reconceptualizing situations, dispositions, dynamics, and invariance in personality structure. *Psychological Review*, 102(2):246–268.

Gyeongsik Moon, Takaaki Shiratori, and Shunsuke Saito. 2024. Expressive whole-body 3D Gaussian avatar. In *European Conference on Computer Vision (ECCV)*.

Louis-Philippe Morency, Giota Stratou, David DeVault, Arno Hartholt, Margaux Lhommet, Gale Lucas, Fabrizio Morbini, Kallirroi Georgila, Stefan Scherer, Jonathan Gratch, Stacy Marsella, David Traum, and Albert Rizzo. 2015. SimSensei demonstration: A perceptive virtual human interviewer for healthcare applications. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, pages 4307–4308.

Koki Nagano, Hongyu Liu, Seonwook Park, Tianye Li, Amrita Mazumdar, Christian Jacobsen, Shengze Wang, Michael Stengel, Rajarshi Roy, Ka Chun Cheung, Simon See, and Shalini De Mello. 2026. Dyaplex: Full-duplex speech-motion model for dyadic interaction. In *Advances in Neural Information Processing Systems*.

Rajmund Nagy, Silvia Arellano García, Hendric Voss, Mihail Tsakov, Taras Kucherenko, Youngwoo Yoon, and Gustav Eje Henter. 2026. The GENEA challenge 2026: A large-scale disentangled evaluation of speech-driven gesture generation on the Seamless Interaction dataset. *Preprint*, arXiv:2608.10839.

Allen Newell, J. C. Shaw, and Herbert A. Simon. 1958. Elements of a theory of human problem solving. *Psychological Review*, 65(3):151–166.

Evonne Ng, Hanbyul Joo, Liwen Hu, Hao Li, Trevor Darrell, Angjoo Kanazawa, and Shiry Ginosar. 2022. Learning to listen: Modeling non-deterministic dyadic facial motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20363–20373.

Evonne Ng, Javier Romero, Timur Bagautdinov, Shaojie Bai, Trevor Darrell, Angjoo Kanazawa, and Alexander Richard. 2024. From Audio to Photoreal Embodiment: Synthesizing Humans in Conversations. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition/Proceedings - IEEE Computer Society Conference on Computer Vision and Pattern Recognition/Proceedings*, pages 1001–1010.

Tu Anh Nguyen, Eugene Kharitonov, Jade Copet, Yossi Adi, Wei-Ning Hsu, Ali Elkahky, Paden Tomasello, Robin Algayres, Benoît Sagot, Abdelrahman Mohamed, and Emmanuel Dupoux. 2023. Generative spoken dialogue language modeling. *Transactions of the Association for Computational Linguistics*, 11:250–266.

OpenAI. 2023. GPT-4 technical report. *Preprint*, arXiv:2303.08774.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *Preprint*, arXiv:2203.02155.

Hae Won Park, Mirko Gelsomini, Jin Joo Lee, and Cynthia Breazeal. 2017. Telling stories to robots. In *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 100–108.

Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (UIST)*, pages 1–22.

Ziqiao Peng, Yanbo Fan, Haoyu Wu, Xuan Wang, Hongyan Liu, Jun He, and Zhaoxin Fan. 2025. DualTalk: Dual-speaker interaction for 3D talking head conversations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Stephen Porter and Leanne ten Brinke. 2008. Reading between the lies: Identifying concealed and falsified emotions in universal facial expressions. *Psychological Science*, 19(5):508–514.

Kexin Quan, Zijian Ding, Jiaye Yong, Qinshi Zhang, Dong Wang, and Jessie Chin. 2026. PAIR: Perceptual affective inference and regulation in a real-time multimodal conversational agent. *Preprint*, arXiv:2610.07523.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5370–5381.

Carl R. Rogers. 1957. The necessary and sufficient conditions of therapeutic personality change. *Journal of Consulting Psychology*, 21(2):95–103.

Frank Rosenblatt. 1958. The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, 65(6):386–408.

Rajarshi Roy, Jonathan Raiman, Sang-gil Lee, Teodor-Dumitru Ene, Robert Kirby, Sungwon Kim, Jaehyeon Kim, and Bryan Catanzaro. 2026. Personaplex: Voice and role control for full duplex conversational speech models. In *ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 16137–16141. IEEE.

David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. 1986. Learning representations by back-propagating errors. *Nature*, 323(6088):533–536.

Harvey Sacks, Emanuel A. Schegloff, and Gail Jefferson. 1974. A simplest systematics for the organization of turn-taking for conversation. *Language*, 50(4):696–735.

Emanuel A. Schegloff. 2000. Overlapping talk and the organization of turn-taking for conversation. *Language in Society*, 29(1):1–63.

Klaus R. Scherer. 2003. Vocal communication of emotion: A review of research paradigms. *Speech Communication*, 40(1–2):227–256.

Klaus R. Scherer. 2009. The dynamic architecture of emotion: Evidence for the component process model. *Cognition and Emotion*, 23(7):1307–1351.

Marc Schröder, Elisabetta Bevacqua, Roddy Cowie, Florian Eyben, Hatice Gunes, Dirk Heylen, Mark ter Maat, Gary McKeown, Sathish Pammi, Maja Pantic, Catherine Pelachaud, Björn Schuller, Etienne de Sevin, Michel Valstar, and Martin Wöllmer. 2012. Building autonomous sensitive artificial listeners. *IEEE Transactions on Affective Computing*, 3(2):165–183.

Iulian Vlad Serban, Alessandro Sordoni, Yoshua Bengio, Aaron C. Courville, and Joëlle Pineau. 2016. Building End-To-End Dialogue Systems Using Generative Hierarchical Neural Network Models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30.

Ashish Sharma, Inna Wanyin Lin, Adam S. Miner, David C. Atkins, and Tim Althoff. 2023. Human–AI collaboration enables more empathic conversations in text-based peer-to-peer mental health support. *Nature Machine Intelligence*, 5(1):46–57.

Ashish Kumar Sharma, Adam S. Miner, David C. Atkins, and Tim Althoff. 2020. A Computational Approach to Understanding Empathy Expressed in Text-Based Mental Health Support. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5263–5276.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. Reflexion: language agents with verbal reinforcement learning. In *neural information processing systems*, pages 8634–8652.

Herbert A. Simon. 1956. Rational choice and the structure of the environment. *Psychological Review*, 63(2):129–138.

Gabriel Skantze. 2020. Turn-taking in Conversational Systems and Human-Robot Interaction: A Review. *Computer Speech & Language*, 67:101178.

Micol Spitale, Minja Axelsson, and Hatice Gunes. 2025. VITA: A multi-modal LLM-based system for longitudinal, autonomous and adaptive robotic mental well-being coaching. *ACM Transactions on Human-Robot Interaction*, 14(2):1–28.

Elizabeth C. Stade, Shannon Wiltsey Stirman, Lyle Ungar, Cody L. Boland, H. Andrew Schwartz, David B. Yaden, João Sedoc, Robert J. DeRubeis, Robb Willer, and Johannes C. Eichstaedt. 2024. Large language models could change the future of behavioral healthcare: a proposal for responsible development and evaluation. *npj Mental Health Research*, 3(1):12.

Laura Stafford, Cynthia S. Burggraf, and William F. Sharkey. 1987. Conversational memory: The effects of time, recall, mode, and memory expectancies on remembrances of natural conversations. *Human Communication Research*, 14(2):203–229.

Tanya Stivers, N. J. Enfield, Penelope Brown, Christina Englert, Makoto Hayashi, Trine Heinemann, Gertie Hoymann, Federico Rossano, Jan Peter de Ruiter, Kyung-Eun Yoon, and Stephen C. Levinson. 2009. Universals and cultural variation in turn-taking in conversation. *Proceedings of the National Academy of Sciences*, 106(26):10587–10592.

Bandhav Veluri, Benjamin N. Peloquin, Bokai Yu, Hongyu Gong, and Shyamnath Gollakota. 2024. Beyond Turn-Based Interfaces: Synchronous LLMs as Full-Duplex Dialogue Agents. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 21390–21402.

Oriol Vinyals and Quoc V. Le. 2015. A neural conversational model. *Preprint*, arXiv:1506.05869.

Jingwen Wang, Jiao Li, Yuhang Qian, Yuyin Wang, and Min Zhang. 2026a. Mind the divergence: Divergence-guided reliability-aware fusion and adaptive LLM prompting for conversational emotion recognition. *Preprint*, arXiv:2611.08860.

Pengyu Wang, Chenkun Tan, Shaojun Zhou, Qirui Zhou, Yanxin Chen, Xingyang He, Huazheng Zeng, Jijun Cheng, Chenghao Wang, Xiaomeng Qian, et al. 2026b. Moss-vl technical report. *arXiv preprint arXiv:2608.15045*.

Taorui Wang, Wei Xia, Hui Ma, Zijia Song, Jiayu Zhang, Zeheng Wang, Yong Xu, and Zitong Yu. 2026c. GMoT: Gated motion-aware tokenization for fine-grained micro-gesture video reasoning with multimodal LLMs. *Preprint*, arXiv:2607.16322. Version 1. Accepted to ACM MM 2026.

Yuxiang Wang, Shengbo Cai, Zhixian Zhao, Yingda Shen, Yuancheng Wang, Xueyao Zhang, Liqiang Zhang, Teddy Sun, Steve Yves, and Zhizheng Wu. 2026d. Jarvis: Coordination deserves a brain of its own in agentic interactive models. *Preprint*, arXiv:2610.23804.

Nigel Ward and Wataru Tsukahara. 2000. Prosodic features which cue back-channel responses in English and Japanese. *Journal of Pragmatics*, 32(8):1177–1207.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. 2022. Emergent abilities of large language models. *Preprint*, arXiv:2206.07682.

Joseph Weizenbaum. 1966. ELIZA—a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 9(1):36–45.

J. D. Williams and S.J. Young. 2006. Partially observable Markov decision processes for spoken dialog systems. *Computer Speech & Language*, 21(2):393–422.

Di Wu, Hongwei Wang, Wenhao Yu, Yuwei Zhang, Kai-Wei Chang, and Dong Yu. 2025. LongMemEval: Benchmarking chat assistants on long-term interactive memory. In *The Thirteenth International Conference on Learning Representations*.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, Ahmed Hassan Awadallah, Ryen W. White, Doug Burger, and Chi Wang. 2023. AutoGen: Enabling next-gen LLM applications via multi-agent conversation. *Preprint*, arXiv:2308.08155.

Hong-Xia Xie, Ling Lo, Hong-Han Shuai, and Wen-Huang Cheng. 2022. An overview of facial micro-expression analysis: Data, methodology and challenge. *IEEE Transactions on Affective Computing*, 14(3):1857–1875.

Zhifei Xie, Zihang Liu, Ze An, Xiaobin Hu, Yue Liao, Ziyang Ma, Dongchao Yang, Mingbao Lin, Deheng Ye, Shuicheng Yan, et al. 2026. Audio interaction model. *arXiv preprint arXiv:2606.05121*.

Jing Xu, Arthur Szlam, and Jason Weston. 2022. Beyond goldfish memory: Long-term open-domain conversation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5180–5197, Dublin, Ireland. Association for Computational Linguistics.

Wujiang Xu, Zujie Liang, Kai Mei, Hang Gao, Juntao Tan, and Yongfeng Zhang. 2025. A-Mem: Agentic Memory for LLM Agents. In *neural information processing systems*, pages 20004–20031.

Yuting Yan, Shihao Xu, Junhao Yu, Mingcong Zuo, Lu Chen, Nan Xiang, Haiyang Geng, Dongjie Tao, and Minghao Wang. 2026. PsyEvo: A personalized counseling agent that self-evolves at test time. *Preprint*, arXiv:2610.02885.

Shu-wen Yang, Ming Tu, Andy T. Liu, Xinghua Qu, Hung-yi Lee, Lu Lu, Yuxuan Wang, and Yonghui Wu. 2026. ParaS2S: Benchmarking and aligning spoken language models for paralinguistic-aware speech-to-speech interaction. In *International Conference on Learning Representations*.

Steve Young, Milica Gašić, Blaise Thomson, and Jason D. Williams. 2013. POMDP-based statistical spoken dialog systems: A review. *Proceedings of the IEEE*, 101(5):1160–1179.

Wenyi Yu, Siyin Wang, Terumi Chiba, Xianzhao Chen, Xiaohai Tian, Jun Zhang, Lu Lu, and Chao Zhang. 2026. SALMONN-duo: Adaptive dual-system coordination for full-duplex voice agents. *Preprint*, arXiv:2609.34247.

Bowen Zeng, Peipei Song, Weidong Chen, Shengeng Tang, Song Ye, Yuanhong Zhong, Beier Zhu, and Xun Yang. 2026. Multi-faceted evaluation and mitigation of emotion hallucinations in MLLMs. *Preprint*, arXiv:2609.11154. Version 1.

Dong Zhang, Shimin Li, Xin Zhang, Jun Zhan, Pengyu Wang, Yaqian Zhou, and Xipeng Qiu. 2023a. SpeechGPT: Empowering large language models with intrinsic cross-modal conversational abilities. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15757–15773.

Hang Zhang, Xin Shane Li, and Lidong Bing. 2023b. Video-LLaMA: An Instruction-tuned Audio-Visual Language Model for Video Understanding. In *Conference on Empirical Methods in Natural Language Processing: System Demonstrations (EMNLP)*, pages 543–553.

Weizhi Zhang, Xinyang Zhang, Chenwei Zhang, Liangwei Yang, Jingbo Shang, Zhepei Wei, Henry Peng Zou, Zijie Huang, Zhengyang Wang, Yifan Gao, et al. 2026. Personaagent: Bridging memory and action for personalized llm agents. In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 26421–26439.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2020. DIALOGPT : Large-Scale Generative Pre-training for Conversational Response Generation. In *Annual Meeting of the Association for Computational Linguistics: System Demonstrations (ACL Demo)*, pages 270–278.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-Bench and chatbot arena. In *Advances in Neural Information Processing Systems*, pages 46595–46623.

Wanjun Zhong, Lianghong Guo, Qiqi Gao, He Ye, and Yanlin Wang. 2024. MemoryBank: Enhancing large language models with long-term memory. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19724–19731.

Li Zhou, Jianfeng Gao, Di Li, and Heung-Yeung Shum. 2020. The Design and Implementation of XiaoIce, an Empathetic Social Chatbot. *Computational Linguistics*, 46(1):53–93.

Li Zhou, Lutong Yu, You Lyu, Yihang Lin, Zefeng Zhao, Junyi Ao, Yuhao Zhang, Benyou Wang, and Haizhou Li. 2026. Echomind: An interrelated multi-level benchmark for evaluating empathetic speech language models. *Preprint*, arXiv:2510.22758.

Mohan Zhou, Yalong Bai, Wei Zhang, Ting Yao, and Tiejun Zhao. 2025. Interactive conversational head generation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(8):6673–6686.

Yongming Zhu, Longhao Zhang, Zhengkun Rong, Tianshu Hu, Shuang Liang, and Zhipeng Ge. 2025. INFP: Audio-Driven Interactive Head Generation in Dyadic Conversations. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition/Proceedings - IEEE Computer Society Conference on Computer Vision and Pattern Recognition/Proceedings*, pages 10667–10677.

Mingchen Zhuge, Changsheng Zhao, Dylan Ashley, Wenyi Wang, Dmitrii Khizbullin, Yunyang Xiong, Zechun Liu, Ernie Chang, Raghuraman Krishnamoorthi, Yuandong Tian, Yangyang Shi, Vikas Chandra, and Jürgen Schmidhuber. 2024. Agent-as-a-judge: Evaluate agents with agents. *Preprint*, arXiv:2410.10934.

A Extended Literature Review

This appendix expands Section 2, following the stages of Figure 2. For each stage we outline its representative systems, the psychology behind it, and its cues. We favor classic, widely cited, and recent work in established venues. Table 1 summarizes the psychological perspective that each stage drew on, the cues it read, and what it missed.

Text dialogue. ELIZA matched keywords and reflected a user's words back (Weizenbaum, 1966), much as a client-centered therapist mirrors a client (Rogers, 1957). Neural models later generated replies directly from data (Vinyals and Le, 2015; Serban et al., 2016), learned to avoid bland responses (Li et al., 2016), and scaled through pretraining on conversations (Zhang et al., 2020). Instruction tuning with human feedback made such models broadly useful assistants (Ouyang et al., 2022), and strong models now serve as judges of chat quality (Zheng et al., 2023). Psychologically, these systems build common ground through words alone (Clark and Schaefer, 1989). The only affective cue they can read is the sentiment of the text.

Empathetic and supportive dialogue. Datasets grounded in emotional situations taught agents to recognize and acknowledge feelings (Rashkin et al., 2019), and models mixed emotion-specific listeners to respond in kind (Lin et al., 2019). Social chatbots were designed for long-term emotional connection (Zhou et al., 2020), and empathy in peer support was measured and modeled (Sharma et al., 2020). Emotional support conversation drew on Helping Skills Theory (Liu et al., 2021a), and human–AI collaboration made peer support more empathic (Sharma et al., 2023). Chatbots for mental health have been tested in randomized trials (Fitzpatrick et al., 2017; Heinz et al., 2025),

Stage	Psychological lens	Key sources	Cues it reads	What it leaves out
Text dialogue	Client-centered reflection; common ground	Rogers (1957); Clark and Schaefer (1989)	Sentiment of words	How things are said
Empathetic & supportive dialogue	Empathy as a predictor of outcome; helping skills	Davis (1983); Elliott et al. (2011)	Stated emotions, emotion labels	Unspoken signals
Turn-based spoken dialogue	Acoustic profiles of vocal emotion	Banse and Scherer (1996); Juslin and Laukka (2003); Scherer (2003)	Overall tone of a finished turn	Moment-to-moment change
Full-duplex speech	Turn-taking and overlap in conversation	Sacks et al. (1974); Stivers et al. (2009); Levinson (2015)	Timing, overlap, backchannels	Who the person is
Memory & asynchronous agents	Dual-process cognition; working and episodic memory	Kahneman (2003); Evans and Stanovich (2013); Baddeley (2000)	Stated facts and preferences	How a person usually behaves
Full-duplex audiovisual interaction	Nonverbal behavior; thin slices; mimicry	Ekman and Friesen (1969); Ambady and Rosenthal (1992); Chartrand and Bargh (1999)	Basic expressions, head pose, gestures	Brief, person-specific cues
Our position	Lens model; appraisal; if-then signatures	Brunswik (1956); Scherer (2009); Mischel and Shoda (1995)	Subtle cues against a person's baseline	—

Table 1: The psychology behind each stage of Figure 2. Each stage reads a richer channel, yet its cues stay coarse until they are judged against the person, which our position adds.

and a meta-analysis reports benefits for distress (Li et al., 2023). Chatbot answers can even be rated more empathic than physicians' (Ayers et al., 2023). These systems rest on the finding that therapist empathy predicts outcome (Elliott et al., 2011; Davis, 1983), but they read empathy from stated emotions and words, not from how something is said.

Turn-based spoken dialogue. Statistical spoken dialogue systems modeled uncertainty about the user's goal as a partially observable process (Williams and Young, 2006; Young et al., 2013). Large language models later coordinated speech and audio tools (Huang et al., 2024), took speech directly as input and output (Zhang et al., 2023a), and streamed spoken replies in real time (Fang et al., 2025). Some models now treat paralinguistic cues such as sentiment as part of the dialogue (Lin et al., 2024; Yang et al., 2026). Psychology long showed that emotions have distinct acoustic profiles (Banse and Scherer, 1996; Juslin and Laukka, 2003; Scherer, 2003). Turn-based systems, however, hear the voice only after a turn ends and summarize its tone as a single whole.

Full-duplex speech. Human conversation keeps gaps short and overlaps brief across languages (Sacks et al., 1974; Stivers et al., 2009), and overlap and backchannels follow systematic rules (Schegloff, 2000; Ward and Tsukahara, 2000). Turn-taking has been argued to shape language processing itself (Levinson, 2015). Conversational systems first predicted turn shifts (Skantze, 2020; Ekstedt and Skantze, 2022). Generative spoken dialogue models then learned both speakers' streams (Nguyen et al., 2023), and full-duplex models now listen while they speak (Défossez et al., 2024; Veluri et al., 2024). Their cues are timing, overlap, and backchannels, richer than the tone of a turn but still read without regard to the person.

Memory and asynchronous agents. Dialogue agents began to remember across sessions (Xu et al., 2022), and long-term memory was built for personalization (Zhong et al., 2024) and tested over long, multi-session conversations (Maharana et al., 2024; Wu et al., 2025). Agents simulate believable behavior by reflecting on stored memories (Park et al., 2023), improve through self-reflection (Shinn et al., 2023), and organize memory on their own (Xu et al., 2025). Some pair fast and slow modules (Lin et al., 2023), echoing dual-process accounts (Kahneman, 2003; Evans and Stanovich, 2013) and working-memory models (Baddeley, 2000). These agents remember the facts and preferences that users state, but rarely how a particular person usually behaves.

Full-duplex audiovisual interaction. Sociable robots expressed and read emotion (Breazeal, 2003), and virtual humans sensed nonverbal behavior in real time (Schröder et al., 2012; Morency et al., 2015) and sustained long-term relationships (Bickmore and Picard, 2005). Learning-based models generated listener motion (Ng et al., 2022) and photorealistic conversational avatars (Ng et al., 2024), while audio-visual language models began to understand video (Zhang et al., 2023b). Recent work merges listening and speaking into one interactive stream. Heads and avatars now switch between the two roles within a conversation (Zhou et al., 2025; Zhu et al., 2025; Peng et al., 2025), respond in real time (Guo et al., 2025; Ki et al., 2026), and generate speech and motion to-

gether in full duplex (Nagano et al., 2026). Omnimodal models listen, watch, and speak at once (Cui et al., 2026), and benchmarks now test streaming audio-visual interaction (Lu et al., 2026; Mazumdar et al., 2026). Psychology shows that thin slices of nonverbal behavior predict social judgments (Ambady and Rosenthal, 1992) and that people nonconsciously mirror one another (Chartrand and Bargh, 1999). These agents read basic expressions, head pose, and gestures as the user speaks, yet these cues remain coarse compared with the subtle, person-specific cues that our position concerns.

B Bibliometric Analysis

Figure 3 is a coarse, keyword-based estimate, not a systematic review. We queried OpenAlex[1] on 13 November 2026. For each topic, we counted works published between 2016 and 2026 whose title or abstract matches the topic's query, using the `title_and_abstract.search` filter grouped by publication year. We then summed the yearly counts into three periods: 2016–2019, 2020–2022, and 2023–2026, where 2026 is a partial year.

Table 2 lists every query, and Table 3 gives the exact search URL for each, so anyone can reproduce the counts directly. Each query combines term blocks in two steps. Terms inside a block are joined with OR, so a work matches the block if its title or abstract contains any one of the terms. Blocks are then joined with AND, so a work must match every block in the query. Multi-word terms are searched as phrases, and OpenAlex matching is case-insensitive and applies stemming.

Every row of Figure 3 is an independent query, and no row is combined with the rows above it. The two rows starting with "+" are the only exception, and they appear only among the intersections. Each of them adds one block to the query of the row directly above it, so "+ psychological support" extends "Subtle cues in dialogue" with PSYCH, and "+ user memory" further adds MEMORY.

Keyword counts trade precision for coverage. Some papers match a term without studying the topic, and relevant papers may use other terms. We therefore read every paper in the smallest intersection. None of its three matches clearly combines all four ingredients. The closest pairs facial micro-expressions with an avatar for counseling-style conversation but shows no per-client memory (Ghadekar et al., 2025), and the other two are clinical assessment or general writing. Precision also varies across topics, and OpenAlex indexing has grown sharply since 2025, so absolute counts and growth rates should not be compared directly. The figure should be read as a comparison of orders of magnitude across topics, not as an exact census.

[1] `https://openalex.org`, accessed through its public API (`https://docs.openalex.org`).

C Platform Interface and Functions

This appendix shows an early prototype of the platform in Section 4 and Figure 5. It illustrates the design and has not been evaluated with clients or experts. All names and accounts in the screenshots are fictional test data created for this demonstration.

Client workspace. Figure 6 shows the client side. The conversation view places the avatar next to a running transcript, and the client turns the microphone and camera on or off at any time. Turning off the camera ends the call, so the agent never watches a client who has not chosen to be seen. A focus view keeps only the avatar and the call controls for clients who prefer fewer distractions. Under preferences and privacy, the client can review, correct, or stop the agent's understanding of them, which is the client's side of the evolving model in Section 3. Professional support is a separate page that the client opens by choice. The client can talk, review records, and adjust preferences without any expert, and every proposal from an expert waits on this page until the client accepts or declines it explicitly, so nothing changes without consent.

Expert workspace. Figure 7 shows the expert side. The client list contains only assigned clients who have agreed to share, and each client page groups the expert's tasks into planning, understanding, and shared records. Support plans start as drafts that need human review and carry a review date, so that no plan reaches a client silently. The expert works between sessions and has no control in the live conversation, which matches the expert's role as the slowest clock of PsychEvolve.

Administrator workspace. Administrators manage accounts, usage limits, approved support methods, releases, and an emergency stop. They see the metadata of sessions, such as when a session took place and how long it lasted, but never its content. We omit screenshots of this workspace because it serves platform operations only and is never shown to clients or to the experts who support them.

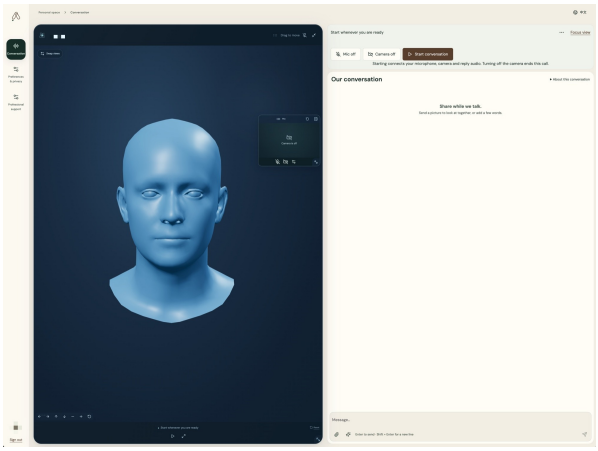
Block	Terms, joined with OR
AGENT	chatbot, conversational agent, dialogue system, dialog system, dialogue agent
DIALOGUE	dialogue, dialog, conversation, conversational, chatbot
SPOKEN	spoken dialogue, spoken dialog, speech language model, voice assistant, spoken conversational
FULLDUPLEX	full-duplex
EMBODIED	embodied conversational agent, talking head, listening head, digital human, virtual human, conversational avatar, social robot
EMOTION	emotion, emotional, empathy, empathetic, affective
CUES	micro-expression, microexpression, micro-gesture, micro-action, gaze estimation, prosody recognition, prosodic cues, facial action unit, backchannel
PSYCH	psychological support, mental health, emotional support, well-being, wellbeing, depression, anxiety, therapy, counseling, counselling
MEMORY	personalization, personalized, long-term memory, user memory, user profile, user modeling, lifelong

Topic in Figure 3	Query, blocks joined with AND
Empathetic text dialogue	AGENT AND EMOTION
Affective spoken dialogue	SPOKEN AND (emotion OR emotional OR paralinguistic OR prosody OR affective)
Full-duplex dialogue	FULLDUPLEX AND (dialogue OR dialog OR conversation OR conversational OR spoken language model)
Avatar and robot agents	EMBODIED AND (conversation OR dialogue OR dialog OR interaction)
Subtle-cue perception	CUES
User memory in dialogue	AGENT AND MEMORY
Psych. support agents	AGENT AND PSYCH
Subtle cues in full duplex	CUES AND FULLDUPLEX
Subtle cues in dialogue	CUES AND DIALOGUE
+ psychological support	CUES AND DIALOGUE AND PSYCH
+ user memory	CUES AND DIALOGUE AND PSYCH AND MEMORY

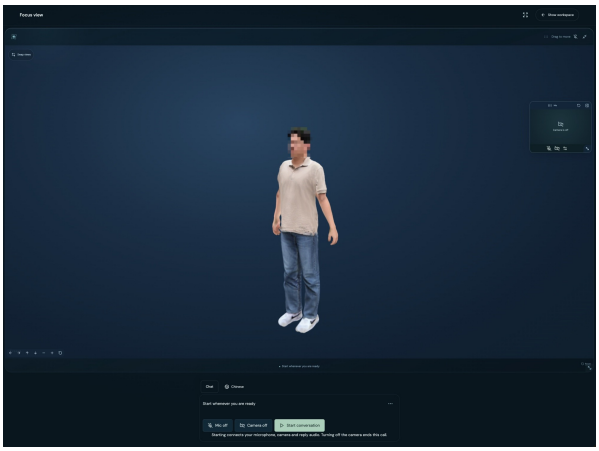
Table 2: Keyword queries behind Figure 3. Left: term blocks. Right: one query per topic, grouped as in the figure.

Topic in Figure 3	OpenAlex search URL (retrieved 13 November 2026)
Empathetic text dialogue	https://openalex.org/works?filter=title_and_abstract.search:%28chatbot+OR+%22conversational+agent%22+OR+%22dialogue+system%22+OR+%22dialog+system%22+OR+%22dialogue+agent%22%29+AND+%28emotion+OR+emotional+OR+empathy+OR+empathetic+OR+affective%29,publication_year:2016-2026
Affective spoken dialogue	https://openalex.org/works?filter=title_and_abstract.search:%28%22spoken+dialogue%22+OR+%22spoken+dialog%22+OR+%22speech+language+model%22+OR+%22voice+assistant%22+OR+%22spoken+conversational%22%29+AND+%28emotion+OR+emotional+OR+paralinguistic+OR+prosody+OR+affective%29,publication_year:2016-2026
Full-duplex dialogue	https://openalex.org/works?filter=title_and_abstract.search:%22full-duplex%22+AND+%28dialogue+OR+dialog+OR+conversation+OR+conversational+OR+%22spoken+language+model%22%29,publication_year:2016-2026
Avatar and robot agents	https://openalex.org/works?filter=title_and_abstract.search:%28%22embodied+conversational+agent%22+OR+%22talking+head%22+OR+%22listening+head%22+OR+%22digital+human%22+OR+%22virtual+human%22+OR+%22conversational+avatar%22+OR+%22social+robot%22%29+AND+%28conversation+OR+dialogue+OR+dialog+OR+interaction%29,publication_year:2016-2026
Subtle-cue perception	https://openalex.org/works?filter=title_and_abstract.search:%28%22micro-expression%22+OR+microexpression+OR+%22micro-gesture%22+OR+%22micro-action%22+OR+%22gaze+estimation%22+OR+%22prosody+recognition%22+OR+%22prosodic+cues%22+OR+%22facial+action+unit%22+OR+backchannel%29,publication_year:2016-2026
User memory in dialogue	https://openalex.org/works?filter=title_and_abstract.search:%28chatbot+OR+%22conversational+agent%22+OR+%22dialogue+system%22+OR+%22dialog+system%22+OR+%22dialogue+agent%22%29+AND+%28personalization+OR+personalized+OR+%22long-term+memory%22+OR+%22user+memory%22+OR+%22user+profile%22+OR+%22user+modeling%22+OR+lifelong%29,publication_year:2016-2026
Psych. support agents	https://openalex.org/works?filter=title_and_abstract.search:%28chatbot+OR+%22conversational+agent%22+OR+%22dialogue+system%22+OR+%22dialog+system%22+OR+%22dialogue+agent%22%29+AND+%28%22psychological+support%22+OR+%22mental+health%22+OR+%22emotional+support%22+OR+%22well-being%22+OR+wellbeing+OR+depression+OR+anxiety+OR+therapy+OR+counseling+OR+counselling%29,publication_year:2016-2026
Subtle cues in full duplex	https://openalex.org/works?filter=title_and_abstract.search:%28%22micro-expression%22+OR+microexpression+OR+%22micro-gesture%22+OR+%22micro-action%22+OR+%22gaze+estimation%22+OR+%22prosody+recognition%22+OR+%22prosodic+cues%22+OR+%22facial+action+unit%22+OR+backchannel%29+AND+%22full-duplex%22,publication_year:2016-2026
Subtle cues in dialogue	https://openalex.org/works?filter=title_and_abstract.search:%28%22micro-expression%22+OR+microexpression+OR+%22micro-gesture%22+OR+%22micro-action%22+OR+%22gaze+estimation%22+OR+%22prosody+recognition%22+OR+%22prosodic+cues%22+OR+%22facial+action+unit%22+OR+backchannel%29+AND+%28dialogue+OR+dialog+OR+conversation+OR+conversational+OR+chatbot%29,publication_year:2016-2026
+ psychological support	https://openalex.org/works?filter=title_and_abstract.search:%28%22micro-expression%22+OR+microexpression+OR+%22micro-gesture%22+OR+%22micro-action%22+OR+%22gaze+estimation%22+OR+%22prosody+recognition%22+OR+%22prosodic+cues%22+OR+%22facial+action+unit%22+OR+backchannel%29+AND+%28dialogue+OR+dialog+OR+conversation+OR+conversational+OR+chatbot%29+AND+%28%22psychological+support%22+OR+%22mental+health%22+OR+%22emotional+support%22+OR+%22well-being%22+OR+wellbeing+OR+depression+OR+anxiety+OR+therapy+OR+counseling+OR+counselling%29,publication_year:2016-2026
+ user memory	https://openalex.org/works?filter=title_and_abstract.search:%28%22micro-expression%22+OR+microexpression+OR+%22micro-gesture%22+OR+%22micro-action%22+OR+%22gaze+estimation%22+OR+%22prosody+recognition%22+OR+%22prosodic+cues%22+OR+%22facial+action+unit%22+OR+backchannel%29+AND+%28dialogue+OR+dialog+OR+conversation+OR+conversational+OR+chatbot%29+AND+%28%22psychological+support%22+OR+%22mental+health%22+OR+%22emotional+support%22+OR+%22well-being%22+OR+wellbeing+OR+depression+OR+anxiety+OR+therapy+OR+counseling+OR+counselling%29+AND+%28personalization+OR+personalized+OR+%22long-term+memory%22+OR+%22user+memory%22+OR+%22user+profile%22+OR+%22user+modeling%22+OR+lifelong%29,publication_year:2016-2026

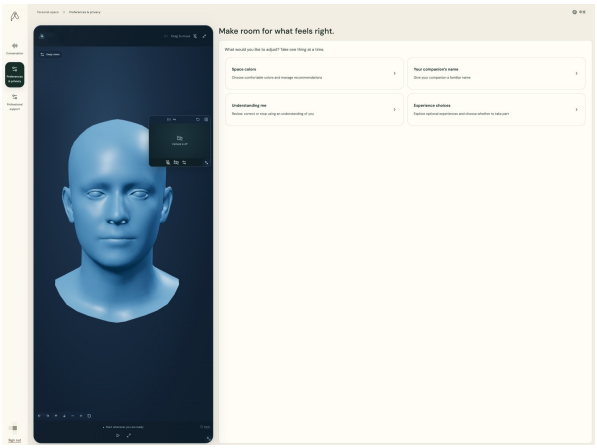
Table 3: The exact OpenAlex request behind each row of Figure 3. Opening a URL shows the matching works and their total in the OpenAlex search interface. Counts may drift slightly as OpenAlex updates its index.



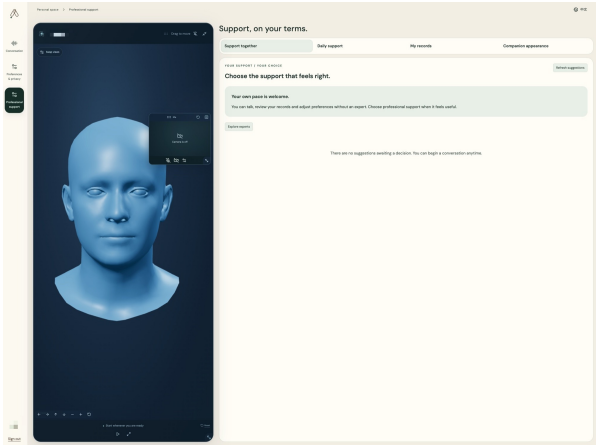
(a) Conversation workspace



(b) Focus view with a full-body avatar



(c) Preferences and privacy



(d) Professional support, chosen by the client

Figure 6: Client workspace of the prototype. (a) The client starts a conversation with the avatar and controls the microphone and camera. (b) A focus view hides everything but the avatar and the call controls. (c) The client reviews, corrects, or stops the agent's understanding of them. (d) Expert support is optional and begins only when the client chooses it. The face in (b) is pixelated for anonymity.

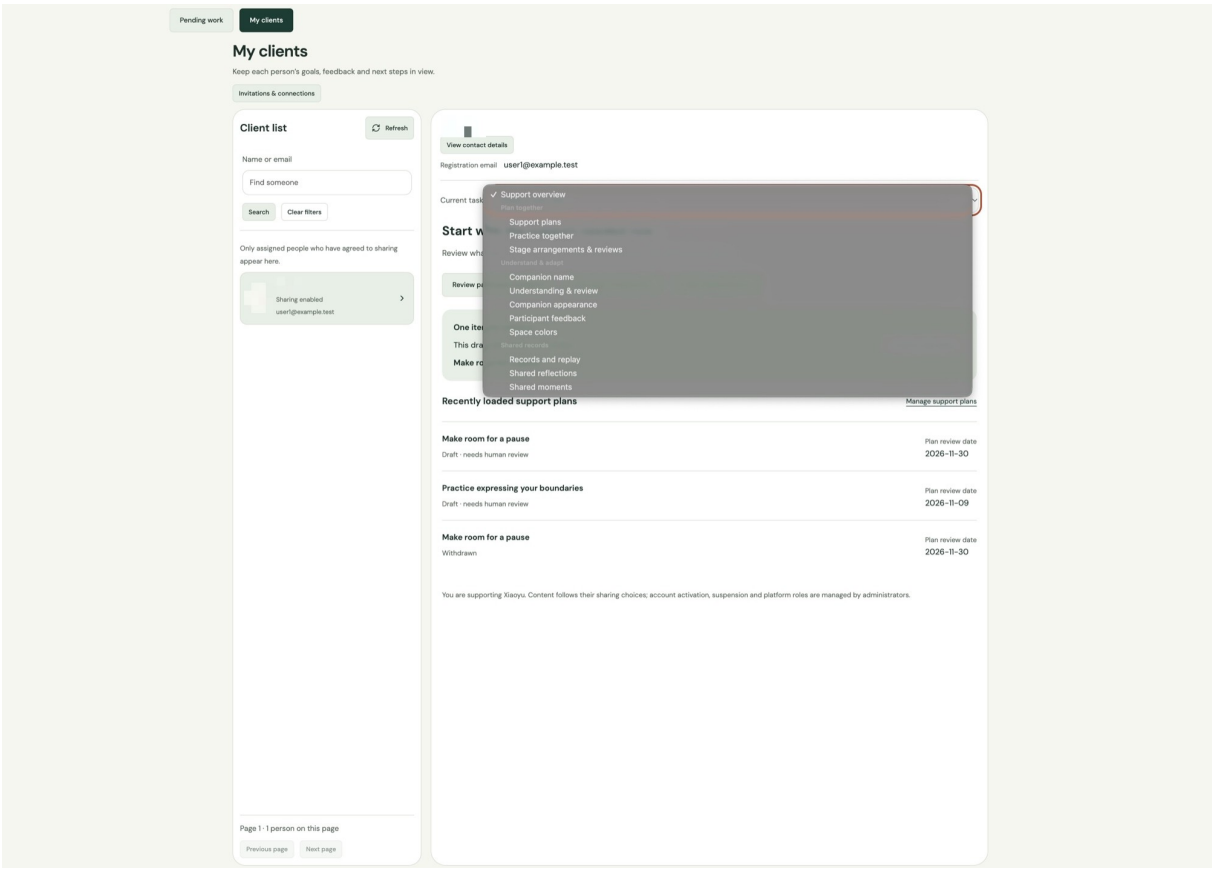


Figure 7: Expert workspace of the prototype. The expert sees only assigned clients who agreed to share, and works on support plans and shared understanding between sessions.