

Trust-by-Design in an Avatar Advising Chatbot: Per-Signal Endorsement

How twenty embedded trust signals were endorsed by their users

Author details withheld for double-anonymized review. A separate title page carries the author list, affiliations, and corresponding-author details.

Computers and Education: Artificial Intelligence (Elsevier). Submission draft.

Highlights

- Twenty trust signals across five layers are built into one advising chatbot.
- A background-recognition layer carries programme knowledge into every reply.
- Each signal was endorsed by 83.3 to 100 per cent of the users it applied to.
- Six of twenty signals clear the pre-set 0.80 rule; fourteen remain undetermined.
- All users reported trust while signals varied, so endorsement and trust are distinct.

Abstract

Trust-relevant mechanisms in conversational agents have mostly been studied one at a time, and it is rarely asked whether the cues a designer embeds are noticed by the people they are built for. This study implements a trust-by-design approach in a multimodal avatar advising chatbot for international undergraduates and measures, signal by signal, how far the design is endorsed by its users. Twenty trust signals were specified across five layers, covering chatbot identity, response quality, interaction naturalness, operational-policy transparency, and a background-recognition layer that carries what the programme knows about a student into the reply, with each signal tied to a stated basis, an implementation location, and a survey item. The system pairs retrieval-augmented generation with an embodied avatar and runs face-to-face, speech and text modes from a single backend. It was deployed in a business programme and evaluated with thirty-nine respondents. Every one of the twenty signals was reported by between 83.3 and 100 per cent of the respondents to whom it applied, and the five layers sit within six percentage points of one another, so endorsement was not concentrated in conspicuous cues at the expense of the rest. Under a conservative rule requiring the lower bound of a Wilson interval to exceed 0.80, six signals met the design target; at this sample size the remaining fourteen cannot be resolved in either direction. Overall trust was reported by every respondent while individual signals varied, so an aggregate trust judgement cannot show that the parts are working.

Keywords: Trust-by-design; Conversational agents; Embodied avatar; Retrieval-augmented generation; International students; Signal endorsement

1. Introduction

1.1 Background

Conversational agents have been applied across a range of functions in higher education, and a scoping review of 66 studies documents both their versatility and the benefits and challenges reported to date (Pereira et al., 2023). Large language models have since made advising assistants more capable. A recent system combines retrieval over institutional documents with email-based human-in-the-loop interaction to help students navigate university policy (Tamascelli et al., 2025), and case studies from universities in North America, Europe, Asia, and Africa report chatbots used for academic advising, personalized learning, and administrative support (Kalashi et al., 2025).

Academic advising is a natural target for this kind of support. Student access to advising is unevenly distributed, and Soria (2023) reported disparities in undergraduates' access to academic advising during the COVID-19 pandemic. Advisors themselves are under strain, with Soria et al. (2023) documenting burnout among academic advisors. An always-available assistant does not resolve either condition on its own, but it can extend the hours and the volume at which routine advising questions are answered.

Whether students use such a system depends on whether they trust it. Trust is a key influence on whether people rely on automation, particularly where complexity or unanticipated situations make a complete understanding of the system impractical (Lee and See, 2004). Trust in artificial intelligence rests on both cognitive and emotional bases, and the form of AI representation is an antecedent of each (Glikson and Woolley, 2020). In education specifically, Ayanwale and Molefi (2024) reported a direct path from perceived trust to undergraduates' intention to use a chatbot, while Polyportis and Pahos (2025) placed trust as an antecedent of students' attitude toward a large language model chatbot.

These considerations are sharpened for international students, who access university services across language and institutional boundaries. Language and institutional barriers to service access are documented across settings, and the difficulties do not necessarily ease with time on campus (Hiratsuka, 2016; Newton et al., 2021; Tajvar et al., 2024). Section 2.4 returns to this evidence. Conversational AI is already used by this population as an informal first line of support (Nourian et al., 2026).

Two technical developments make a trust-oriented advising agent feasible. Retrieval-augmented generation conditions a model on passages drawn from an external index rather than on parametric memory alone, and produces more specific, diverse, and factual language (Lewis et al., 2020), and placing retrieval in the loop reduces hallucination in knowledge-grounded dialogue (Shuster et al., 2021). Agent representation raises social presence: increasing the anthropomorphic level of an agent's visual representation significantly increased perceived social presence (Li et al., 2023), and humanoid embodiment with human voice raised social presence, which in turn enhanced trusting beliefs (Qiu and Benbasat, 2009). In practice, however, these mechanisms are usually

introduced and evaluated one at a time. It remains unclear whether they can be integrated into a single system by design, and whether each intended trust signal is actually endorsed by users.

1.2 Research problem and research questions

Prior work has largely examined trust-relevant mechanisms one at a time. Retrieval augmentation has been evaluated for factual grounding (Lewis et al., 2020; Shuster et al., 2021), agent representation for its effect on social presence (Qiu and Benbasat, 2009; Li et al., 2023; Liew et al., 2017), and chatbot trust for its structure and predictors (Glikson and Woolley, 2020; Ng and Zhang, 2025). That each mechanism contributes in isolation does not establish that a system combining them will be trusted as a whole. Two problems follow. The first is whether such mechanisms can be organized and implemented as a coherent design rather than accumulated ad hoc. The second is whether each implemented signal is actually endorsed by the user.

The second problem has a basis in credibility research. Prominence-interpretation theory holds that a user evaluates an element only after noticing it, so an element that goes unremarked does not enter the judgement (Fogg, 2003). Trust seals, though placed deliberately as assurance cues, were frequently overlooked or misread by users (Kirlappos et al., 2012), and an empirical test of the theory found that prominence and interpretation together, rather than either alone, distinguished how users assessed website credibility (George et al., 2016). These findings come from website credibility rather than from conversational agents, and we import them as a design principle rather than as established results for agents. They nevertheless motivate measuring endorsement signal by signal instead of inferring it from an aggregate trust score. Throughout this paper we use endorsement for the share of eligible respondents who agreed that a signal was present and had its intended character. Endorsement is not equivalent to noticing, and Section 4.1 sets out why.

Advising for international students adds a third problem. Beyond chatbot identity, response quality, interaction naturalness, and operational-policy transparency, these users may be sensitive to whether a system recognizes their background, their prior visits, and their situation. Evidence here is thin. Personalization has been associated with trust in an agent in a small study, where explanations were adapted within a session from the user's measured trust, workload, and performance (Verhagen et al., 2023), and it has been applied to the design of motivational messages that adapt between sessions from a patient's symptom trend and initial trust (Tielman et al., 2019). In both cases the system adapts to a state it has measured during use. We are not aware of work that specifies recognition of a user's enrolment background, carried across sessions, as a distinct trust layer, or that measures whether such recognition is endorsed by users.

A prior deployment of the same design approach in a domestic setting reported that subtle cues were noticed less often than overt ones (Authors, 2026). Whether that pattern recurs in a different cohort has not been tested. Replication is uncommon in this literature (Hornbæk et al., 2014; Wilson et al., 2011, 2012, 2013, 2014), and Section 2.5 sets out what follows from that.

This study accordingly sets three research questions. RQ1 concerns system construction, RQ2 concerns how far that construction is endorsed by users, and RQ3 concerns whether that endorsement aligns with overall trust; the three are examined together.

RQ1. Can a trust-by-design avatar advising chatbot, extended with a background-recognition layer, be designed and deployed for international university students?

RQ2. To what extent is each embedded trust signal endorsed by users, and does endorsement differ systematically between element-borne and behaviour-borne signals?

RQ3. How does per-signal endorsement relate to overall trust, and how well are the background-recognition signals endorsed by users?

RQ2 and RQ3 are addressed through descriptive analysis of an exploratory cross-sectional survey, and causal inference among variables is left to future work. The case domain is major introduction and academic advising for international undergraduate students in a business program.

1.3 Contributions and organization of the paper

This study implements a trust-by-design approach in a multimodal avatar advising chatbot and examines how far its signals are endorsed by users. It contributes in four respects: an analytical framework, a system architecture, an empirical case, and a finding on the relation between endorsement and trust.

First, it proposes a trust-by-design framework of twenty trust signals organized into five layers, covering chatbot identity, response quality, interaction naturalness, operational-policy transparency, and background recognition. Personalization has been studied where the system adapts to a state it measures during use, but recognition of a user's enrolment background, carried across sessions and surfaced back to the user, has not been specified or measured as a distinct trust layer, and this study is the first to do so.

Second, it reports the design and implementation of a multi-tier architecture that pairs retrieval-augmented generation with an embodied avatar, operates face-to-face, speech, and text modalities from a single backend, and keeps user data on institutional servers rather than external services.

Third, it reports an empirical deployment with international undergraduate students and provides per-signal endorsement for every implemented signal. Because the same design approach and the same measurement procedure were used in an earlier domestic deployment (Authors, 2026), the study also tests whether the reported gap between overt and subtle cues recurs in a different cohort, which addresses the scarcity of replication in this literature.

Fourth, it reports that per-signal endorsement and overall trust do not move together. Several signals were endorsed by noticeably fewer users than others, while overall trust was reported by every respondent. Endorsing a designed signal and trusting the system therefore appear to be distinct outcomes, which has consequences for how trust-oriented systems are evaluated.

The remainder of the paper is organized as follows. Section 2 reviews prior work on chatbot trust, retrieval-augmented generation, agent representation, and AI for international students. Section 3 describes the system design and the five-layer, twenty-signal framework. Section 4 describes the survey instrument, the deployment, and the analysis procedure. Section 5 reports the sample, per-signal endorsement, the layer-level comparison, the background-recognition layer, and free-text responses. Section 6 discusses the endorsement gap, the relation between endorsement and trust, limitations, and directions for future work.

2. Related Work

2.1 Trust in AI and chatbots

Trust is generally understood as a willingness to accept vulnerability toward another party on the expectation that the party will act in one's interest. Applied to automated systems, Lee and See (2004) treated trust as a key influence on whether people rely on automation, particularly where complexity or unanticipated situations make complete understanding impractical, and argued that the design objective is appropriate reliance rather than maximal trust.

Research on artificial intelligence has since separated the bases of that judgement. Glikson and Woolley (2020) distinguished cognitive trust, built on cues such as reliability, tangibility, transparency, and immediacy behaviours, from emotional trust, and identified the form of AI representation as an antecedent of both. Lukyanenko et al. (2022) observed that trust-in-AI research remains fragmented without a common conceptual foundation, and proposed a foundational framework in response.

Chatbot-specific reviews have organized the field empirically. Ng and Zhang (2025) grouped trust predictors into user-related, machine-related, interaction-related, social, and contextual factors, and mapped trust outcomes across affective, relational, behavioural, cognitive, and psychological domains. They also noted that the literature is dominated by cross-sectional designs and called for longitudinal work on the dynamics of trust. In a qualitative interview study of chatbot users, Wang et al. (2026) found that human-like cues and perceived reliability both emerged as themes in users' accounts of trust.

In educational settings the evidence is more specific but also more heterogeneous. Ayanwale and Molefi (2024) reported a direct path from perceived trust to undergraduates' behavioural intention to use a chatbot, alongside relative advantage, compatibility, and trialability. Polyportis and Pahos (2025) placed trust as an antecedent of students' attitude rather than of intention directly, and Shahzad et al. (2024) modelled perceived trust as a moderator between awareness and perceived usefulness. Bilquise et al. (2024), surveying 207 students at two universities, found that perceived ease of use and social influence, rather than perceived usefulness or trust, drove intention to adopt an academic advising chatbot. Trust therefore appears in different positions across these models, and all of them measure intention rather than sustained use.

Two implications follow for the present study. Because trust rests on both cognitive and affective bases, a trust-oriented design has to address factual accuracy and the interpersonal qualities of the interface together. And because these frameworks describe trust as an outcome of cues that users perceive, a designed signal can contribute only once it is noticed, which is why this study measures endorsement signal by signal.

2.2 Retrieval-augmented generation and factual grounding

Grounding a model's output in retrievable sources is the standard technical response to unreliable generation. Lewis et al. (2020) introduced retrieval-augmented generation, which conditions a sequence-to-sequence model

on passages drawn from an external non-parametric index rather than on parametric memory alone, and reported that it produces more specific, diverse, and factual language than a purely parametric baseline. Shuster et al. (2021) then showed that placing retrieval in the loop substantially reduces hallucination in knowledge-grounded dialogue while retaining conversational ability.

Subsequent work has tested the approach in narrower settings. Xiong et al. (2024) benchmarked retrieval-augmented generation for medicine, and Béchar and Marquez Ayala (2024) reported reduced hallucination in structured output generation for an enterprise task. The available evidence is therefore domain-bounded. It covers dialogue models of that period, medical question answering, and structured enterprise output, and does not establish a general reduction in hallucination for contemporary instruction-tuned models. For an advising agent the relevant implication is narrower and safer, namely that answers tied to an institutional corpus are less likely to be fabricated than answers generated from parametric memory alone.

In the present system, retrieval serves the cognitive basis of trust identified in Section 2.1. It is treated as one implemented trust signal among others rather than as a sufficient condition for trust.

2.3 Agent representation and social presence

How an agent is represented to the user affects how present it feels. Qiu and Benbasat (2009) found in a laboratory experiment that humanoid embodiment together with human voice significantly raised perceived social presence, which in turn enhanced trusting beliefs and intention to use the agent as a decision aid. Li et al. (2023) manipulated the anthropomorphic level of a health care agent's visual representation and found that a higher level significantly increased perceived social presence; in their path model, social presence and mindless anthropomorphism predicted trust, while privacy concerns predicted it negatively. Liew et al. (2017) added a non-interactive talking avatar to an e-commerce site and found that perceived social presence improved relative to the same site without an avatar.

Three qualifications matter. First, the repeated finding across these studies is that representation raises social presence, and the effect on trust is mediated rather than direct. Second, the effect is not uniformly positive. In Liew et al. (2017) the credibility and patronage effects were moderated by gender, and Kim et al. (2024) found that an affectively expressive virtual agent reduced perceived social presence and trustworthiness in a group discussion task, while the agent's task engagement yielded the most preferred configuration. Third, embodiment is not the only relational lever. Bickmore and Picard (2005) varied relational rather than embodied qualities and found that an agent using social-emotional and relationship-building dialogue was respected, liked, and trusted more than an equivalent task-oriented agent over four weeks of use.

Embodiment is therefore adopted in this study as a signal expected to raise social presence, with trust treated as a downstream and mediated outcome rather than an assumed consequence.

2.4 AI and international students

International students access university services through additional linguistic and institutional filters. Hiratsuka (2016) described the Japanese language as an organizational barrier to international students' access to campus services, and Newton et al. (2021), in a qualitative study of 21 students at an Australian university, identified informational and cultural obstacles alongside opportunities for improved support. Perry et al. (2020) compared international and domestic students at one university on their awareness, use, and valuation of student services and found the gap to lie in awareness rather than provision, so that making an existing service known mattered more than adding another; that is the same distinction between what is provided and what users report back, which this study measures signal by signal. Tajvar et al. (2024) surveyed 165 international students at a single Iranian medical university and found that total challenge scores differed significantly by length of study (one-way ANOVA, $p < 0.001$), with those enrolled more than three years scoring highest ($M = 75.1$) and those enrolled under a year the lowest ($M = 47.0$, $n = 11$); the study enrolled only students who had completed their first year, so it does not contrast newcomers with continuing students. These studies concern access to services and information broadly, and none of them isolates academic advising.

Generative AI has entered this space quickly. Zhu and Yang (2026) examined generative AI as an informational, emotional, and social support channel in international students' cross-cultural adaptation, and Nourian et al. (2026) reported in a mixed-methods study that international students in the United States already use conversational AI as a first-aid resource within a fragmented support ecosystem, while identifying boundaries to what they will delegate to it. At a broader level, a review of generative AI in language education positions such tools as a means of narrowing equity gaps through real-time feedback and adaptation to individual learner needs (Weng and Fu, 2025).

This population therefore combines high need with high sensitivity to whether a system can be relied upon, which makes it a demanding setting in which to test whether designed trust signals are endorsed by users.

2.5 The need for a trust-by-design approach

The preceding sections describe a literature organized by mechanism. Trust research specifies what trust rests on, retrieval research addresses factual grounding, work on agent representation addresses social presence, and work on international students describes the access conditions under which such a system would be used. What is missing is a system that embeds these mechanisms deliberately, as layers of a single design, together with evidence on whether each embedded signal is endorsed by users.

Two further gaps shape this study. First, personalization has been linked to trust and to engagement in agent research, but the adaptation studied is driven by a state the system measures during use, whether within a session from measured trust, workload, and performance (Verhagen et al., 2023) or between sessions from a symptom trend (Tielman et al., 2019), and the most direct measurement of trust in the agent comes from a small study (Verhagen et al., 2023). Recognition of a user's enrolment background, carried across sessions and surfaced back to the user, has not to our knowledge been specified as a distinct trust layer or measured for whether users endorse it. Second, replication is rare in this literature. A survey of publications in human-computer interaction

found roughly 3% to be replications (Hornbæk et al., 2014), and the field has repeatedly called for more (Wilson et al., 2011, 2012, 2013, 2014). A design approach evaluated in one deployment therefore has limited standing until it is applied again in a different setting.

This study responds to both. It implements a trust-by-design system whose signals are specified layer by layer, adds a background-recognition layer for which no precedent has been established, measures whether each signal is endorsed, and does so in a second deployment with a population different from the one in which the approach was first evaluated.

3. System Design

3.1 Architecture overview

The assistant is organized into four tiers, shown in Fig. 1. The first tier is the browser client, a single-page application that renders the advising interface, holds the avatar session, and captures speech and text input. The second tier is an edge proxy that receives the client's chat request and forwards it to the inference host; it holds no conversation state and stores nothing. The third tier is the inference host, a self-managed server that runs the retrieval pipeline and the language model. The fourth tier is the institutional server, which authenticates users, records chat logs, and stores survey responses in the institution's own database.

The separation is deliberate and follows the design intent stated in Section 1. Student identifiers, conversation logs, and survey records are written only on the institutional tier, so no user data accumulates with an external service provider. Generation runs on the self-managed inference host rather than a commercial API, which keeps the model and the corpus under institutional control. The tiers communicate over HTTP only, so a change inside one tier does not require changes in the others.

3.2 Retrieval pipeline

Answers are grounded in a curated corpus of the program's public information. The corpus holds forty-three question-and-answer chunks covering the program overview, admission, curriculum, career tracks, and faculty. Each chunk carries an identifier, a section label, a question, and an answer.

At build time each chunk is embedded with a multilingual embedding model and the vectors are stored with the chunk text. At query time the user's message is embedded with the same model, the vectors are L2-normalized, and cosine similarity is computed against the corpus. The three highest-scoring chunks are retrieved, and any hit scoring below 0.40 is discarded. When no chunk clears the threshold, no source text is injected into the prompt, and the assistant is instructed to say that it does not have the information rather than to answer from parametric memory. This threshold is the system's main guard against fabricated answers on off-topic questions, and it is the implementation behind the retrieval-grounding and limitation-admission signals described in Section 3.4.

Generation is performed by an open-weight instruction-tuned model served locally on the inference host. Because both retrieval and generation run on institution-managed hardware, student questions are not transmitted to an external inference provider.

3.3 Multimodal interface

The advising page is divided into a video interaction area and a conversation area, shown in Fig. 2. The video area holds the avatar stage, the session start and end control, the mode switch, and a control that interrupts the avatar's speech. The conversation area holds the message thread, a set of suggested questions, and an input dock with a microphone control, a text field, and a send action.

Three interaction modes run from the same backend. In face-to-face mode the student speaks and receives a spoken reply from the avatar; in speech mode the exchange is voice only; in text mode it is keyboard only. A student can switch modes within a session. Because all three modes pass through the same retrieval and generation path, the factual content of a reply does not depend on the mode chosen. Speech output is normalized before synthesis so that acronyms and program names are pronounced consistently.

3.4 The five-layer, twenty-signal framework

The trust signals implemented in the system are organized into five layers. The first four follow the formulation used in the earlier domestic deployment (Authors, 2026): chatbot identity, response quality, interaction naturalness, and operational-policy transparency. The fifth layer, background recognition, is introduced here.

Each signal is specified so that three things are tied to one identifier: the reason the signal is expected to matter, the place in the system where it is implemented, and the survey item that records whether users endorsed it. Organizing the signals into layers rather than a flat list follows an integrative review of human-AI dialogue which found that user judgements of a conversational agent converge on a small number of distinguishable levels rather than a single quality (Marconi et al., 2026), so the grouping used here is not merely an authoring convenience. Fig. 2 shows where the signals surface in the interface. Layer one covers the avatar's likeness to a named instructor, the display of institutional identity, and the explicit statement that the assistant is an AI. Layer two covers retrieval grounding, admission of limitations, tone, and format consistency. Layer three covers response pacing, guarding against the system transcribing its own speech, user-initiated interruption, avatar lip-sync and gesture, and mode switching. Layer four covers the privacy consent screen, browsing without an account, greeting form, visit counting, and speech normalization.

The fifth layer follows from the setting. At sign-up the system collects the student's country and interest track, and these are available to the assistant when it composes a reply, which supports the background-awareness and background-tailored signals. Visit days are counted per account, and from the second distinct visit day a returning-user greeting is shown; first-time visitors see nothing. This layer is treated as new rather than as an extension of existing personalization work, because prior studies adapt to a state the system measures during use, whether stated preferences or observed behaviour, whereas the signals here depend on enrolment information the user supplied once and never restated.

Grounding is stated at the level of the layer, not the individual signal. What the literature establishes is a mechanism: that an embodied representation raises social presence, that retrieval grounding reduces unsupported generation, that privacy concern depresses trust. It does not establish that a particular threshold, a particular greeting, or a particular button is the right way to realize that mechanism. Those are engineering choices made here, and attaching a citation to each of them would misstate what the cited work shows. Table 1 therefore gives the grounding per layer, and Table 2 gives the signals as what they are, the concrete form each layer takes in this deployment.

The grounding is transferred, and the transfer should be stated. The retrieval evidence comes from pre-LLM sequence-to-sequence models on open-domain benchmarks, not from an instruction-tuned assistant answering institutional questions. The calibration guidance comes from supervisory control of safety-critical automation, not from conversation. The social-presence findings come from a laboratory shopping task and from a video-vignette study in which participants watched an interaction rather than held one. Each mechanism is used here as a design principle carried into a new setting, not as a result that has been shown to hold in this setting.

Two things follow from splitting the tables this way. The signals are presented as design decisions, which is what they are, and the reader can see which parts of the system rest on published mechanisms and which rest on our judgement. And the fifth layer's empty grounding cell is not a weakness in the table but the statement of the gap: the signals in L5 depend on information carried between sessions, and we found no prior work that treats that as a trust signal or measures it. The survey item for each signal is listed in the appendix.

4. Method

4.1 Survey instrument

Perceptions were collected with a questionnaire delivered in the browser as a modal dialogue at the end of a session. It has four parts. Part I records the respondent's country, intended track, age band, gender, and MBTI type, the last of these optional. The personality-type field was not used in any analysis reported here. Part II presents the twenty trust-signal items, grouped under five headings that correspond to the layers of the framework. Part III presents a single overall-trust item. Part IV presents two open-ended prompts, one asking what felt most trustworthy and one asking what felt least trustworthy or awkward.

Each of the twenty-one closed items is answered with one of two buttons. There is no midpoint, no "not sure", and no explicit non-response option, and the form cannot be submitted until every applicable item has been answered. The binary format was chosen to keep the response burden low enough that students would complete the questionnaire immediately after a session rather than abandon it. The cost of that choice is that the instrument cannot register degree, so a signal that was noticed faintly and one that was noticed clearly are recorded identically. Section 6 returns to this limitation.

The items are worded as first-person statements about the session just completed, and the respondent indicates agreement. An item therefore records agreement that a signal was present and had its intended character, rather than an isolated report of having noticed it. We use endorsement throughout for the proportion of eligible respondents who agreed, and it should be read in that sense. Agreement of this kind bundles noticing with appraisal: a respondent who noticed the avatar's speech but judged it poorly synchronised is recorded in the same way as one who did not attend to it at all. The measure is therefore not a conservative estimate of noticing and should not be read as one. Agreement can be raised by acquiescence, or by a respondent inferring that a signal must have been present, and it can be lowered by a respondent who did register the signal but judged it unfavourably. The two directions do not cancel, and this instrument cannot separate them. Several items are moreover worded so that they record appraisal, preference, or effect rather than registration: F4 asks whether answers felt accurate rather than whether retrieved material was present, F12 asks whether the respondent liked being able to switch modes, and F14 asks whether browsing without an account made them feel more at ease. Results for those items are read as endorsement of the feature, not as evidence that the underlying mechanism was noticed. Of the twenty items, five ask directly whether a signal was present or registered; the remaining fifteen ask about an appraisal, a preference, or an effect, with one of those fifteen mixing the two. Appendix A records which is which for every item.

Five items are conditional. Two concern voice output and are presented only to respondents whose recorded interaction included the avatar or speech mode; one concerns the avatar's visible behaviour and is presented only when the avatar mode was recorded; and two concern recognition across visits and are presented only from the second distinct visit day. Items that are not applicable are not displayed, are not required for submission, and are

stored as missing rather than as a negative response. Because the mode record is held in the browser, a respondent whose record was empty was treated as eligible for the voice and avatar items, so the eligible group for those three items may include students who did not use the mode in question. Section 4.3 states how missing responses are handled, and Section 6 treats the consequences of this fallback.

Item numbering shown on screen is assigned in sequence over the items a given respondent actually sees, so the same signal may appear under different numbers for different respondents. Signals are therefore identified in this paper by their framework identifiers rather than by the numbers respondents saw. The full instrument, with the wording of every item and the option lists, is reproduced as Table A.1 in Appendix A.

4.2 Deployment and participants

The assistant was deployed for an undergraduate business administration programme at a private university in the Republic of Korea, and was reached over the programme's public web address. Data were collected between 5 and 13 June 2026. A survey entry point had been present in the site footer since 1 June, and responses were gathered mainly during class sessions in which students were introduced to the assistant and invited to try it. Participation was voluntary and no incentive was offered. Respondents were international undergraduate students in the programme.

Forty-one questionnaires were submitted. The instrument recorded the time each respondent spent in the form, and two submissions completed in under sixty seconds were excluded as inattentive. Both fell on 9 June, at forty-five and forty-six seconds; the fastest retained response took sixty-seven seconds, so the exclusion did not depend on a marginal judgement. The analysed sample is thirty-nine responses, with a median completion time of two minutes and thirty-three seconds. No further exclusions were applied.

The study collected feedback on a departmental teaching tool. The questionnaire asked for no name, contact detail, or other direct identifier; it recorded country of origin, intended track, age band, year of study, gender, and an optional personality-type field alongside the closed items. The ethics statement in the back matter sets out the basis on which the study proceeded without institutional review board review.

Two features of the deployment constrain what the sample can support, and we state them here rather than in the limitations alone because they bear on how the results in Section 5 should be read.

First, opening the questionnaire required no minimum amount of conversation. There was no turn count, no elapsed-time condition, and no check that the assistant had been used at all, so a visitor could reach the survey from the footer without having held a session. Thirty-nine responses arrived from thirty-six distinct browser sessions, and the instrument carried no safeguard against a second submission from the same browser.

Second, the presentation of the system changed on 10 June, in the middle of the collection window. The advising page began to open in avatar mode rather than text mode, the avatar option was marked as recommended, a short usage guide and a sign-in prompt were added, a survey link was placed directly beneath the conversation,

and an introductory screen was added ahead of the questionnaire. The retrieval settings, the prompt, the avatar identity, the returning-visitor logic and the survey items themselves were not altered. The change is therefore one of presentation rather than of the advising system, but it is not inert: making avatar mode the default alters how many respondents encountered the avatar at all, and the sign-in prompt stated that signing in lets the avatar greet the student by name. Twenty-two of the retained responses precede 10 June and seventeen follow it, and no response was recorded on that day, so each response can be assigned to one arrangement without ambiguity. Section 5 reports the two periods separately where the distinction bears on a result.

4.3 Analysis procedure

The analysis is descriptive. It reports, for each signal, the proportion of respondents who agreed, together with an interval estimate, and compares those proportions across the layers and across a classification of the signals fixed in advance. No causal relation among variables is estimated, and no model is fitted.

Endorsement is defined as the number of agreements divided by the number of respondents to whom the item was presented. Items that were not applicable to a respondent are excluded from that item's denominator rather than counted as disagreement. Applying a uniform denominator of thirty-nine would treat a question a student never saw as a signal they did not endorse, which would understate the conditional signals for a reason that has nothing to do with perception. The denominator is therefore thirty-nine for fifteen signals, thirty-seven for two, thirty-six for one, and six for two.

Interval estimates use the Wilson score method at the ninety-five per cent level. With thirty-nine respondents and proportions near the upper end of the scale, the normal approximation produces intervals that are too narrow and can extend above one; the Wilson interval is bounded and behaves properly for small samples and extreme proportions, which is the situation here for most signals. Intervals on the two signals with six eligible respondents are correspondingly wide, and we do not interpret those two point estimates as if they were comparable in precision to the rest.

Reporting an interval rather than a test follows the position that a study of this size is better used to estimate a quantity than to decide a hypothesis about it (Anderson and Maxwell, 2016), and that a claim about whether a proportion sits above or below a stated bound is properly settled by where the interval falls rather than by a rejected null (Lakens et al., 2018).

The design set a target that each signal should be endorsed by at least eighty per cent of the users to whom it applied. A signal is counted as having met that target only when the lower bound of its interval lies above 0.80. Requiring the whole interval to clear the threshold, rather than the point estimate alone, is the conservative reading and is fixed here before the results are presented. Layer values are reported as the mean of the endorsement of the signals belonging to that layer, so each signal contributes equally regardless of how many respondents answered it.

To examine whether endorsement differs between signals that are carried by a discrete interface element and signals that are carried by a property of the system's behaviour, the twenty signals were assigned to two groups before the endorsement values were examined. A signal is classed as element-borne when it is delivered by a distinct thing the user can read or act on: a label, a screen, a control, a message. A signal is classed as behaviour-borne when it has no dedicated element and is instead a quality distributed across the interaction, typically apparent only when it degrades. The assignment yields ten signals in each group and is listed in the appendix. This distinction is the one the earlier deployment used to account for its results (Authors, 2026), and fixing it in advance is what allows the present study to test that account rather than restate it. Because thirty-nine respondents provide little power to detect a difference between two groups of proportions, the comparison is presented as two sets of estimates with their intervals and is not accompanied by a significance test; the intervals are reported so that the reader can see directly how far they overlap.

The overall-trust item is reported with the same caution. A single closed item invites a ceiling, and where more than fifteen per cent of respondents select the highest available option the instrument is generally taken to have lost the resolution needed to relate that measure to anything else (Terwee et al., 2007). Section 5.4 reports where this study stands against that criterion.

One subgroup analysis follows from Section 4.2. Signals are reported separately for the periods before and after the presentation change of 10 June, since the design changed mid-collection. The two periods are small, and the split is presented to qualify the pooled values rather than to support a claim of difference.

5. Results

5.1 Sample

Thirty-nine retained submissions were analysed; the instrument did not verify unique participants, and the thirty-nine submissions arrived from thirty-six distinct browser sessions. Nationality was self-reported and was not an inclusion criterion: the questionnaire was open to users of the programme's site, and thirty-six gave their nationality as Nepal, two as Korea, and one as India, so the sample is dominated by a single national group and should not be read as representative of international students generally. Thirty respondents intended the business track, seven the artificial-intelligence track, and two were undecided. Twenty-two were aged between eighteen and twenty-one and fifteen between twenty-two and twenty-five, with one respondent in each of the two older bands. Twenty-one were men and eighteen were women. Completion took a median of two minutes and thirty-three seconds and a mean of two minutes and fifty-six seconds, with the shortest retained response at sixty-seven seconds and the longest at just under six minutes.

Twenty-three of the thirty-nine reached the questionnaire without a recorded interaction mode, which as Section 4.1 notes made them eligible for the voice and avatar items by default.

5.2 Per-signal endorsement

Endorsement was high across the twenty signals and varied within a narrow band. The lowest value was 83.3 per cent, for the avatar's lip movements and gestures, and the highest was 100 per cent, for the two conditional signals answered by six respondents each. Fourteen of the twenty signals fell between 83 and 95 per cent. Overall trust was reported by every respondent. Table 3 reports the endorsement of each signal with its Wilson interval, and Fig. 3 plots those values against the two reference levels.

Applying the decision rule fixed in Section 4.3, six of the twenty signals met the design target: institutional identity, warm tone, format consistency, latency pacing, guest browsing, and natural greeting. The remaining fourteen did not, and for every one of them the point estimate lies above 0.80 while the lower bound of the interval does not. This is a statement about the precision available from thirty-nine respondents rather than a finding that those signals failed. With this sample size an interval on a proportion near 0.90 is roughly twenty points wide, so a signal would have to be endorsed by almost every eligible respondent to clear the threshold on the conservative rule. The two conditional signals answered by six respondents illustrate the point in its extreme form: both were endorsed by every eligible respondent, and both have intervals running from 61 per cent to 100 per cent, so neither can be said to have met the target.

5.3 Element-borne and behaviour-borne signals

The classification fixed in Section 4.3 divides the twenty signals into ten carried by a discrete interface element and ten carried by a property of the system's behaviour. Element-borne signals were endorsed by a mean of 93.6 per cent of eligible respondents and behaviour-borne signals a mean of 89.7 per cent, a difference of 3.9

percentage points in the expected direction. Pooling responses within each group gives 92.3 per cent for element-borne signals, with an interval from 88.9 to 94.7, and 89.8 per cent for behaviour-borne signals, with an interval from 86.4 to 92.5. These intervals overlap substantially, and the difference between the groups cannot be distinguished from sampling variation at this sample size.

One pattern is nevertheless consistent. The three signals with the lowest endorsement are all behaviour-borne: the avatar's lip-sync and gestures at 83.3 per cent, the echo guard at 83.8 per cent, and background-tailored replies at 84.6 per cent. Three further signals tie at 87.2 per cent, two of them behaviour-borne and one element-borne. No element-borne signal falls below 87.2 per cent. The ordering therefore reproduces the direction reported in the earlier domestic deployment (Authors, 2026), where cues distributed across the interaction were endorsed less often than cues presented as explicit labels; because that study used the same agreement instrument, its result is restated here in the same terms. The magnitude is much smaller. The earlier study reported a gap of 27.5 percentage points between its highest and lowest layers (Authors, 2026); the corresponding spread here is 6.0 points, between the operational-policy layer at 94.2 per cent and the interaction-naturalness layer at 88.3 per cent. We treat this as a directional replication with substantial attenuation, not as a confirmation of the earlier effect size.

5.4 Layers and overall trust

Layer means were 91.5 per cent for chatbot identity, 92.9 per cent for response quality, 88.3 per cent for interaction naturalness, 94.2 per cent for operational-policy transparency, and 91.5 per cent for background recognition; Fig. 4 shows the layer means. The layers are therefore close together, and the ordering places interaction naturalness last. Layer means do not track the share of behaviour-borne signals, however: response quality has the highest such share, three of four, yet stands second.

Overall trust was reported by all thirty-nine respondents. This ceiling has a direct methodological consequence: with no variance in the outcome, the relation between per-signal endorsement and overall trust cannot be estimated in this sample, and the question posed in RQ3 cannot be answered as a covariation. What can be said is descriptive and, in its way, more pointed. Every respondent reported trusting the assistant, while fourteen of the twenty signals did not meet the design target and three signals were endorsed by fewer than eighty-five per cent of those they applied to. Endorsing the individual designed signals and arriving at a favourable overall judgement therefore came apart in this deployment. A system can be trusted overall while a substantial share of the specific mechanisms intended to earn that trust go unremarked.

5.5 The background-recognition layer

The fifth layer is realized at two levels, and they are worth separating because the survey items speak mainly to the first. At the level of the programme, the retrieval corpus is written for this student population: of its forty-three chunks, nine address international students explicitly, covering English-medium instruction, the regional career advantage, and the practical concerns of studying in Korea. Any student who asks about those topics

receives an answer composed for someone in their situation, whether or not they hold an account. At the level of the individual, the country and interest track recorded at sign-up are available to the assistant when it composes a reply, and the returning-visitor greeting appears from the second distinct visit day.

Measured endorsement was 89.7 per cent for background awareness and 84.6 per cent for background-tailored replies. Background awareness falls inside the range of the other thirteen non-conditional signals, which runs from 87.2 to 97.4 per cent, while background-tailored replies at 84.6 per cent sit below all of them and are the lowest of the non-conditional signals. Neither clears the design target on the conservative rule. The returning-user greeting was endorsed by all six eligible respondents, with the wide interval that six observations imply.

These two items ask whether the assistant seemed aware of the respondent's background and whether answers felt directed at someone like them. Both are satisfied by programme-level composition as much as by account-level personalization, so the values above should be read as evidence about the layer as a whole rather than about the account-level mechanism in particular. Section 6.4 returns to what would be needed to separate the two.

5.6 The presentation change of 10 June

Twenty-two responses precede the presentation change and seventeen follow it. Most signals move little between the two periods. Three move downward, two of them by more than fifteen percentage points: smooth mode transition from 95.5 to 76.5 per cent and the privacy consent screen from 100 to 82.4 per cent; the natural greeting falls by 11.8 points, from 100 to 88.2 per cent. None of these three interface elements was itself altered on 10 June.

A plausible reading is that the change altered which parts of the interface respondents encountered rather than how those parts behaved. Opening the advising page in avatar mode rather than text mode, and marking avatar as recommended, makes it less likely that a respondent will have exercised the mode switch, which is what the mode-transition item asks about. The added in-conversation survey link also shortens the path from the conversation to the questionnaire, so respondents in the later period may have reached the survey without passing through the sign-up and entry screens that carry the consent notice and the greeting. Both readings are post hoc, the subgroups are small, and the intervals are wide; we report the differences because the design changed mid-collection and it would be misleading to present only pooled values, not because the comparison identifies an effect.

5.7 Free-text responses

Fifteen respondents wrote something in the positive prompt and fourteen in the negative prompt. Many entries are brief acknowledgements rather than descriptions, and several of the negative entries state that nothing felt untrustworthy, so the material supports illustration rather than analysis.

Among the substantive positive entries, respondents most often named the clarity and usefulness of the answers, and two named the assistant's handling of what it did not know. One wrote that the system provides many kinds of information and that when the assistant does not know something it will guide the user, which describes the limitation-admission signal as intended. Among the substantive negative entries, response latency was named twice. Two entries describe the boundary of the corpus from the user's side: one respondent reported that when a question fell outside what the assistant knew it did not give a direct answer, and another that responses were sometimes repetitive or lacked specific detail. One entry noted that the voice was somewhat loud. These correspond to the retrieval threshold behaviour described in Section 3.2 and to the pacing and speech-output signals, and they indicate that the same mechanism can be experienced as honesty by one respondent and as evasion by another.

6. Discussion and Conclusion

6.1 What the results show

Two commitments shape how the results below are read. Each signal is weighted equally within its layer and each layer equally within the design, so no cue is treated as more diagnostic than another before the data are seen; equal weighting is a defensible default when the relative importance of predictors is not established in advance, and often outperforms weights fitted on small samples (Dawes, 1979). And the endorsement values are reported signal by signal rather than as a single index, because an aggregate would conceal the variation that Section 6.3 turns on.

The three research questions can now be addressed. RQ1 asked whether a trust-by-design avatar advising assistant, extended with a background-recognition layer, could be designed and deployed for international university students. Section 3 reports such a system: four tiers with retrieval and generation on institution-managed hardware, three interaction modes over one backend, and twenty signals specified layer by layer with their implementation locations. The system ran with real students over a nine-day window and produced the records analysed here. In that sense RQ1 is answered affirmatively, with the qualification that construction is not the same as demonstrated effectiveness.

RQ2 asked how far each signal was endorsed by users and whether endorsement differs systematically between signals presented as discrete elements and signals carried by system behaviour. Endorsement was high and narrowly spread, between 83.3 and 100 per cent. Six of twenty signals met the design target under the conservative rule fixed in advance, and the failure of the remaining fourteen to meet it is largely a matter of interval width at this sample size rather than of low endorsement. On the second part, the difference between element-borne and behaviour-borne signals runs in the expected direction at 3.9 percentage points, the three least-endorsed signals are all behaviour-borne with three further signals tied at 87.2 per cent, and the confidence intervals for the two groups overlap. The pattern reported in the earlier deployment (Authors, 2026) is reproduced in direction and strongly attenuated in size.

RQ3 asked how per-signal endorsement relates to overall trust and how far the background-recognition signals were endorsed. The background signals were endorsed by 89.7 and 84.6 per cent of respondents, and the returning-visitor greeting was endorsed by all six who were eligible for it, so the layer performed in the same band as the four it was added to. The first part cannot be answered as framed: overall trust was reported by every respondent, and with no variance in the outcome no covariation is estimable. What can be said is that the two did not move together, which Section 6.3 takes up.

6.2 Implementation and endorsement

Two questions organize this study, and the results answer both directly. The first is whether the trust literature can be turned into a system rather than left as a set of separate findings. Twenty signals were specified across five

layers, each with a stated basis at the layer level and a named implementation location, and all twenty were built and deployed. Section 3.4 lists them, and the appendix carries the survey item that measures each. The count is the contribution here: not one mechanism evaluated in isolation, but twenty carried simultaneously by one running system with real students.

The second is whether those signals were endorsed by the people they were built for. They were, and evenly. Every signal was reported by between 83.3 and 100 per cent of the respondents to whom it applied, the five layers sit between 88.3 and 94.2 per cent, and no signal fell into the range where a designer would conclude that the implementation had failed to surface. For a system carrying twenty simultaneous signals, that is the substantive result: the design did not have to trade one signal against another to be endorsed.

The endorsement values are nevertheless tighter than they look. Under the rule fixed in Section 4.3, six of twenty signals clear the 0.80 design target, for the reason of interval width given in Section 5.2. A study of this size can establish that endorsement is high and even; it cannot certify individual signals against a threshold. That is a statement about precision, and the remedy is sample size rather than redesign.

6.3 Endorsement and trust are distinct outcomes

Overall trust was reported by every respondent, while fourteen of twenty signals did not clear the design target and the least-endorsed signal was not endorsed by roughly one respondent in six. The two measures therefore did not move together: the global judgement was uniform while agreement about the specific parts built to earn it was not.

Two readings are available and this study cannot separate them. Trust may rest on a small number of signals endorsed by nearly everyone, with the response-quality signals among the candidates, since two of the six that cleared the target belong to that layer, as do two from the operating-policy layer. Alternatively a single binary item asked immediately after a session may measure something closer to satisfaction with a system just introduced in class, in which case its ceiling reflects the setting rather than the design. What follows for practice holds under either reading: an aggregate trust score is not evidence that the individual designed signals are working, and the two should be reported separately, as they are here.

A third possibility bears on both readings. Agreement-format items draw more agreement from respondents in cultures scoring higher on collectivism and power distance, and administering the instrument in a second language moves responses in the same direction (Harzing, 2006). The present sample is drawn almost entirely from one such setting and answered in English, so the uniformity of the overall-trust item cannot be attributed to the design alone.

6.4 Limitations

The sample is small and narrow. Thirty-nine respondents give the interval width noted in Section 5.2, which is why most signals cannot be resolved against the 0.80 target in either direction. Thirty-six of the thirty-nine gave the

same nationality, so the results describe one cohort in one programme and not international students in general. Participation was voluntary and took place around class sessions, which favours students already disposed toward the system.

The instrument has three specific weaknesses. It records agreement rather than perception, as Section 4.1 sets out, so a signal that was noticed but judged unfavourably is recorded in the same way as one that went unnoticed. It is binary, so it cannot register degree and cannot distinguish a signal noticed faintly from one noticed clearly. And its single overall-trust item was at ceiling, which removed the outcome variance RQ3 required.

The deployment constrains inference further. Opening the questionnaire required no minimum amount of conversation, so some respondents may have answered about a system they had barely used; the instrument carried no safeguard against a second submission from the same browser, and thirty-nine responses arrived from thirty-six browser sessions. The eligibility record for the voice and avatar items was held in the browser and defaulted to eligible when absent, so twenty-three of the respondents counted in the denominators of three signals have no recorded use of the mode in question. A complete-case check restricted each of those three signals to respondents with a recorded use of the relevant mode. Endorsement moved in both directions: from 83.8 to 78.6 per cent for the echo guard, from 83.3 to 84.6 for lip-sync and gesture, and from 89.2 to 100 for pronunciation normalization. No signal changed its status against the design target under either rule. The complete-case denominators are thirteen or fourteen, so those intervals are wider than the ones reported in Section 5.2; the check bounds the fallback rule rather than replacing it. The presentation of the system changed mid-collection, and while retrieval, prompt and avatar configuration were untouched, the default mode and the route into the survey were not.

A further constraint follows from the design itself. Browsing without an account is one of the twenty signals (F14) and was endorsed by 94.9 per cent of respondents, among the six that met the design target. The deployment accordingly did not require sign-in, and six of the thirty-nine respondents held an account when they answered. For the remaining thirty-three the account-level mechanism of the background-recognition layer, which carries the country and interest track recorded at sign-up into the reply, was not active, while the programme-level mechanism described in Section 5.5 was: nine of the forty-three retrieval chunks are written for international students and are returned irrespective of account state. The two background items were therefore answered by a majority for whom only one of the two mechanisms could have operated. Those responses are reported as recorded, and no attribution to either mechanism is made here; Section 6.5 sets out the logging that would allow the two to be separated.

The interface carries a constraint of the same kind. Mode switching is signal F12 and was endorsed by 87.2 per cent, short of the design target. The three modes share one retrieval and generation path, so the factual content of a reply does not depend on the mode chosen, but they do not share conversational state: a student who moves from face-to-face to speech or text begins the new mode without the exchange that preceded it. This was the

limit of what the implementation reached at the time of deployment rather than a property of the design, and it is recorded here as a technical limitation and carried into Section 6.5 as a development task.

The narrowness of the sample cuts in two directions. A cohort drawn almost entirely from Nepal is a poor basis for claims about international students generally. It is also, however, a population that this literature has hardly reached: samples in behavioural science, and in human-computer interaction specifically, are drawn overwhelmingly from Western, educated, industrialized, rich and democratic settings, and generalizations built on them have repeatedly failed to hold elsewhere (Henrich et al., 2010; Linxen et al., 2021). The one comparison that speaks to this cohort directly found that Nepali and British respondents differed in what drove their adoption of a generative assistant, with the same technology-acceptance constructs carrying different weight in each country (Budhathoki et al., 2024), which is the kind of divergence a single-country sample cannot reveal. What is reported here is therefore narrow in the ordinary sense and unusual in its constituency, and the two should be stated together rather than one standing in for the other. Following the recommendation that empirical papers state the participants, materials and contexts to which a result is intended to apply (Simons et al., 2017), the findings are offered as applying to South and East Asian undergraduates using an advising assistant in a second language, in a programme whose retrieval corpus was written for them.

Finally, the study is descriptive throughout. No causal claim is made about the effect of any signal on trust, and the layer structure is a design scheme rather than a validated measurement model; no factor analysis was performed and none is warranted at this sample size.

6.5 Future work

The clearest next step is sample size. A deployment of two to three hundred respondents would narrow those intervals enough to certify individual signals and to test the element-borne and behaviour-borne difference, which is 3.9 percentage points here and would need a sample several times larger to detect reliably.

The instrument should also be strengthened. Items should be paired with a graded response so that degree is recoverable, with a subset reverse-coded so that undifferentiated responding can be detected directly. Overall trust should be measured with a multi-item scale rather than a single binary item, which would restore the outcome variance this study lacked.

For the background-recognition layer specifically, the programme-level and account-level mechanisms described in Section 5.5 should be separated by design. Requiring sign-in, or logging per session whether account context was actually transmitted, would allow endorsement to be reported for each mechanism rather than for the layer as a whole. The same logging would benefit the conditional signals, whose eligibility here was inferred from a browser-held record.

Continuity across modes should also be closed. Carrying conversational state through a mode switch, so that a student who moves from face-to-face to text resumes rather than restarts, would remove the discontinuity recorded in Section 6.4 and is a precondition for asking whether the choice of mode itself affects endorsement.

Separating the two mechanisms would also let the background-recognition layer be compared with work that manipulates who the agent appears to be rather than what it knows, where avatar identity has been shown to shift epistemic trust and the uptake of an agent's answers (Anthis and Kyriakidou-Zacharoudiou, 2026). Endorsement measured per signal and trust manipulated by identity are complementary handles on the same question, and a design that carried both would say more than either alone.

Beyond measurement, the replication should be extended. The direction of the element-borne and behaviour-borne difference has now been observed in two deployments with different cohorts, languages, and domains, while the magnitude differed substantially, and a third deployment powered for a difference of a few percentage points would establish which magnitude is the stable one.

6.6 Conclusion

This study set out to do two things and reports on both. It turned a literature organized by separate mechanisms into a single working system: twenty trust signals across five layers, each with a stated basis and a named implementation location, extended with a background-recognition layer that carries what the programme knows about a student into the reply. And it measured, signal by signal, whether what was built was endorsed by the people it was built for.

It did. Every one of the twenty signals was reported by between 83.3 and 100 per cent of the respondents to whom it applied, and the five layers sit within six percentage points of one another. Endorsement was not concentrated in the conspicuous signals at the expense of the rest, which is the outcome a layered design is meant to produce and the one that had not previously been demonstrated for a system carrying this many signals at once.

Two qualifications travel with that result. Thirty-nine respondents give intervals wide enough that only six signals can be certified against the 0.80 design target, so the finding is that endorsement is high and even, not that each signal individually meets a standard. And overall trust was reported by every respondent while individual signals varied, so the aggregate judgement cannot be used as evidence that the parts are working. Reporting the two separately, as this paper does, is what allows either to be read honestly.

Table 1. Grounding of each layer.

Layer	What the layer concerns	Grounding
L1 Chatbot identity	Who the student is talking to, and whether that is disclosed	An embodied humanoid representation raises social presence, which in turn raises trusting beliefs (Qiu and Benbasat, 2009; Li et al., 2023)
L2 Response quality	Whether answers are grounded in institutional sources and bounded when they are not	Conditioning generation on retrieved documents yields more factual output and less hallucination (Lewis et al., 2020; Shuster et al., 2021). Conveying the limits of an automated system's competence supports appropriately calibrated rather than maximal trust (Lee and See, 2004)
L3 Interaction naturalness	Whether the exchange behaves like a conversation	Raising the level of animation, from no image to a static portrait to an animated avatar, raises perceived social presence (Li et al., 2023)
L4 Operational-policy transparency	What the system does with the student's data and access	Privacy concern is a negative determinant of trust in a conversational agent (Li et al., 2023). Relational dialogue behaviour sustains a user-agent relationship over repeated sessions (Bickmore and Picard, 2005). Trust predicts behavioural intention toward educational AI (Ayanwale and Molefi, 2024)
L5 Background recognition	Whether the system carries what it knows about the student across sessions	No established precedent. Prior personalization adapts to a state the system measures during use (Tielman et al., 2019; Verhagen et al., 2023)

Table 2. The twenty signals and where each is implemented.

ID	Layer	Signal	Implemented as
F1	L1	Full digital twin	Photo-based avatar session rendered in the video stage
F2	L1	Institutional identity	Program identity shown in the top bar and page framing
F3	L1	AI disclosure	System prompt and interface state that the assistant is an AI
F4	L2	Retrieval grounding	Top-3 chunks above cosine 0.40 injected into the prompt
F5	L2	Limitation admission	Below threshold, no source is injected and the assistant says it lacks the information
F6	L2	Warm tone	Tone specified in the system prompt
F7	L2	Format consistency	Fixed answer structure in the system prompt
F8	L3	Latency pacing	Streamed response delivery
F9	L3	Multi-layer echo guard	Speech capture suppressed while the avatar is speaking
F10	L3	Speech interrupt control	Control that stops avatar speech mid-utterance

F11	L3	Avatar lip-sync and gesture	Avatar video with synchronized speech
F12	L3	Smooth mode transition	Face-to-face, speech and text modes over one backend
F13	L4	Privacy consent UI	Consent presented at sign-up
F14	L4	Guest browsing	Program pages usable without an account
F15	L4	Natural greeting	Greeting phrasing on entry
F16	L4	Visit-count tracking	Distinct visit days counted per account
F17	L4	TTS pronunciation normalization	Text normalized before speech synthesis
F18	L5	Background awareness	Country and interest track from sign-up available when composing a reply
F19	L5	Background-tailored reply	Reply adapted to the stored background
F20	L5	Returning-user recognition	Greeting shown from the second distinct visit day

Table 3. Per-signal endorsement with Wilson 95% confidence intervals (N = 39).

Layer means and the layer gap are computed on unrounded proportions.

ID	Layer	Signal	Class	Endorsement %	95% CI	n	Target
F1	L1	Full digital twin	behaviour	87.2	[73.3, 94.4]	39	
F2	L1	Institutional identity	element	97.4	[86.8, 99.5]	39	yes
F3	L1	AI disclosure	element	89.7	[76.4, 95.9]	39	
F4	L2	Retrieval grounding	behaviour	87.2	[73.3, 94.4]	39	
F5	L2	Limitation admission	element	89.7	[76.4, 95.9]	39	
F6	L2	Warm tone	behaviour	97.4	[86.8, 99.5]	39	yes
F7	L2	Format consistency	behaviour	97.4	[86.8, 99.5]	39	yes
F8	L3	Latency pacing	behaviour	97.4	[86.8, 99.5]	39	yes
F9	L3	Multi-layer echo guard	behaviour	83.8	[68.9, 92.3]	37	
F10	L3	Speech interrupt control	element	89.7	[76.4, 95.9]	39	
F11	L3	Avatar lip-sync and gesture	behaviour	83.3	[68.1, 92.1]	36	
F12	L3	Smooth mode transition	element	87.2	[73.3, 94.4]	39	
F13	L4	Privacy consent UI	element	92.3	[79.7, 97.3]	39	
F14	L4	Guest browsing	element	94.9	[83.1, 98.6]	39	yes
F15	L4	Natural greeting	element	94.9	[83.1, 98.6]	39	yes
F16	L4	Visit-count tracking	element	100.0	[61.0, 100.0]	6	
F17	L4	TTS pronunciation norm.	behaviour	89.2	[75.3, 95.7]	37	
F18	L5	Background awareness	behaviour	89.7	[76.4, 95.9]	39	
F19	L5	Background-tailored reply	behaviour	84.6	[70.3, 92.8]	39	
F20	L5	Returning-user recognition	element	100.0	[61.0, 100.0]	6	
F21	-	Overall trust	-	100.0	[91.0, 100.0]	39	

Target = the lower bound of the Wilson interval lies above 0.80 (decision rule fixed in Section 4.3). Class: element = element-borne, behaviour = behaviour-borne.

Appendix A. Full survey instrument

Each closed item was answered with one of two buttons (Agree / Disagree); there was no midpoint and no explicit non-response option. Part I recorded country, intended track, age band, gender, and MBTI type (the last optional) and is not reproduced here. The university name in item F2 is masked for anonymized review and was shown to respondents in full. The Records column states what each item asks the respondent to confirm: the presence or registration of the signal, or an appraisal, a preference, or an effect. Five items are of the first kind.

Table A.1. Full survey instrument.

Survey item (as presented)	ID	Records
Talking with this assistant felt like talking to the program's real avatar host.	F1	Appraisal
This assistant felt like an official tool of [university name].	F2	Appraisal
The assistant made it clear that it is an AI, not a real person.	F3	Presence
The assistant's answers felt accurate and fact-based.	F4	Appraisal
When it did not know something, the assistant honestly said so or suggested contacting the program staff.	F5	Presence
The assistant's tone felt warm and friendly.	F6	Appraisal
The assistant's answers were consistent and professional in style.	F7	Appraisal
The pacing of the conversation (time from my question to its reply) felt natural.	F8	Appraisal
During voice conversation, my voice and the assistant's voice never overlapped or cut off awkwardly. [conditional]	F9	Presence
When I wanted to interrupt while the assistant was speaking, I was able to (or felt I could).	F10	Mixed
The avatar's lip movements and expressions looked natural. [conditional]	F11	Appraisal
I liked being able to switch freely between Avatar, Voice, and Text modes.	F12	Preference
At sign-up, the data-handling notice (what is collected, why, and for how long) was presented clearly.	F13	Appraisal
Being able to browse without logging in made me feel more at ease.	F14	Effect
I liked that the assistant greeted me with a natural, friendly welcome message.	F15	Preference
It felt friendly that the assistant remembered how many times I had visited. [conditional]	F16	Effect
The assistant read out abbreviations (AI, GPT, etc.) naturally when speaking. [conditional]	F17	Appraisal
The assistant seemed aware of my background and interests.	F18	Presence
The answers felt tailored to someone like me, rather than generic.	F19	Appraisal
I felt the site recognized me as a returning or known visitor. [conditional]	F20	Presence
Overall, I felt I could trust this assistant.	F21	Overall

What moment or feature felt most trustworthy? [free text]	-	
What felt least trustworthy or awkward? [free text]	-	

CRedit authorship contribution statement

Author 1: Conceptualization, Methodology, Software, Formal analysis, Investigation, Data curation, Writing – original draft, Visualization. Author 2: Software, Investigation, Data curation, Writing – review & editing. Author 3: Conceptualization, Methodology, Writing – review & editing, Supervision. Author 4: Conceptualization, Methodology, Writing – review & editing, Supervision, Project administration. Author names are withheld for double-anonymized review and appear on the separate title page.

Ethics statement

The study collected feedback on a departmental teaching tool from undergraduate students in the programme that operated it. At the entry to the questionnaire respondents were shown a notice describing it as an anonymous research survey of about two minutes that collects no personally identifying data, and stating that their answers support the authors' research. Participation was voluntary, no incentive was offered, and a respondent could close the questionnaire at any point without submitting. The questionnaire asked for no name, contact detail, or other direct identifier. Six of the thirty-nine analysed respondents held a service account when they answered, so those responses carry an account reference; the analysis was carried out on the response table alone, the tables were not joined, and no participant is identified in this article or in any accompanying material. The personality-type field is withheld from any shared file. Review by an institutional review board was not sought. Under Article 13 of the Enforcement Rule of the Bioethics and Safety Act of the Republic of Korea, research that neither collects nor records personally identifiable information, and that records no sensitive information, is not subject to review, and the study proceeded on that basis.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request. The personality-type field is withheld from any shared file: it was supplied by eight of the thirty-nine respondents in seven distinct types, and on its own would identify most of them.

Funding

This work was externally funded. The funding body and the grant number identify the authors' institution and region, so they are given on the separate title page and are withheld here for double-anonymized review.

Declaration of generative AI and AI-assisted technologies in the manuscript preparation process

During the preparation of this work the authors used Google Gemini in order to generate the schematic illustration presented in Fig. 2, and Anthropic Claude in order to draft and revise sections of the manuscript, to check the internal consistency of reported figures against the underlying tables, and to format the reference list. After using these tools the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

References

- Anderson, S. F., & Maxwell, S. E. (2016). There's more than one way to conduct a replication study: Beyond statistical significance. *Psychological Methods*, 21(1), 1–12. <https://doi.org/10.1037/met0000051>
- Anthis, Z., & Kyriakidou-Zacharoudiou, A. (2026). Face value: How avatar identity shapes epistemic trust in AI-mediated learning. *Computers and Education: Artificial Intelligence*, 10, Article 100610. <https://doi.org/10.1016/j.caeai.2026.100610>
- Authors (2026). [Self-citation. Title, journal, volume, pages and DOI are withheld for double-anonymized review and have been supplied to the editor.]
- Ayanwale, M. A., & Molefi, R. R. (2024). Exploring intention of undergraduate students to embrace chatbots: From the vantage point of Lesotho. *International Journal of Educational Technology in Higher Education*, 21(1), Article 20. <https://doi.org/10.1186/s41239-024-00451-8>
- Béchar, P., & Marquez Ayala, O. (2024). Reducing hallucination in structured outputs via Retrieval-Augmented Generation. In Y. Yang, A. Davani, A. Sil, & A. Kumar (Eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 6: Industry Track)* (pp. 228–238). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2024.naacl-industry.19>
- Bickmore, T. W., & Picard, R. W. (2005). Establishing and maintaining long-term human-computer relationships. *ACM Transactions on Computer-Human Interaction*, 12(2), 293–327. <https://doi.org/10.1145/1067860.1067867>
- Bilquise, G., Ibrahim, S., & Salhie, S. M. (2024). Investigating student acceptance of an academic advising chatbot in higher education institutions. *Education and Information Technologies*, 29(5), 6357–6382. <https://doi.org/10.1007/s10639-023-12076-x>
- Budhathoki, T., Zirar, A., Njoya, E. T., & Timsina, A. (2024). ChatGPT adoption and anxiety: A cross-country analysis utilising the unified theory of acceptance and use of technology (UTAUT). *Studies in Higher Education*, 49(5), 831–846. <https://doi.org/10.1080/03075079.2024.2333937>
- Dawes, R. M. (1979). The robust beauty of improper linear models in decision making. *American Psychologist*, 34(7), 571–582. <https://doi.org/10.1037/0003-066X.34.7.571>
- Fogg, B. J. (2003). Prominence-interpretation theory: Explaining how people assess credibility online. In *CHI '03 Extended Abstracts on Human Factors in Computing Systems* (pp. 722–723). ACM Press. <https://doi.org/10.1145/765891.765951>
- George, J. F., Mirsadikov, A., & Mennecke, B. E. (2016). Website credibility assessment: An empirical investigation of prominence-interpretation theory. *AIS Transactions on Human-Computer Interaction*, 8(2), 40–57. <https://doi.org/10.17705/1thci.00078>
- Glikson, E., & Woolley, A. W. (2020). Human trust in artificial intelligence: Review of empirical research. *Academy of Management Annals*, 14(2), 627–660. <https://doi.org/10.5465/annals.2018.0057>
- Harzing, A. W. (2006). Response styles in cross-national survey research. *International Journal of Cross Cultural Management*, 6(2), 243–266. <https://doi.org/10.1177/1470595806066332>
- Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences*, 33(2–3), 61–83. <https://doi.org/10.1017/S0140525X0999152X>
- Hiratsuka, H. (2016). Japanese language as an organizational barrier for international students to access to university services: A case of Aoyama Gakuin University. *Journal of International Students*, 6(4), 956–966. <https://doi.org/10.32674/jis.v6i4.328>
- Hornbæk, K., Sander, S. S., Bargas-Avila, J. A., & Simonsen, J. G. (2014). Is once enough? On the extent and content of replications in human-computer interaction. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '14)* (pp. 3523–3532). ACM. <https://doi.org/10.1145/2556288.2557004>

- Kalashi, K., Montazer, G. A., Saed, S., & Olyaei, S. (2025). Adoption of chatbots in higher education: Global insights and recommendations for Iran. In 2025 12th International and the 18th National Conference on E-Learning and E-Teaching (ICeLeT) (pp. 1–6). IEEE. <https://doi.org/10.1109/ICeLeT66022.2025.11025265>
- Kim, H., Han, B., Kim, J., Syawaludin, M. F., Kim, G. J., & Hwang, J.-I. (2024). Engaged and affective virtual agents: Their impact on social presence, trustworthiness, and decision-making in the group discussion. In Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24) (pp. 1–17). ACM, New York, Article 149. <https://doi.org/10.1145/3613904.3642917>
- Kirlappos, I., Sasse, M. A., & Harvey, N. (2012). Why trust seals don't work: A study of user perceptions and behavior. In S. Katzenbeisser, E. Weippl, L. J. Camp, M. Volkamer, M. Reiter, & X. Zhang (Eds.), Trust and Trustworthy Computing (TRUST 2012), Lecture Notes in Computer Science, vol. 7344 (pp. 308–324). Springer. https://doi.org/10.1007/978-3-642-30921-2_18
- Lakens, D., Scheel, A. M., & Isager, P. M. (2018). Equivalence testing for psychological research: A tutorial. *Advances in Methods and Practices in Psychological Science*, 1(2), 259–269. <https://doi.org/10.1177/2515245918770963>
- Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors*, 46(1), 50–80. https://doi.org/10.1518/hfes.46.1.50_30392
- Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, & H. Lin (Eds.), *Advances in Neural Information Processing Systems*, vol. 33 (pp. 9459–9474). Curran Associates.
- Li, Q., Luximon, Y., & Zhang, J. (2023). The influence of anthropomorphic cues on patients' perceived anthropomorphism, social presence, trust building, and acceptance of health care conversational agents: Within-subject web-based experiment. *Journal of Medical Internet Research*, 25, Article e44479. <https://doi.org/10.2196/44479>
- Liew, T. W., Tan, S.-M., & Ismail, H. (2017). Exploring the effects of a non-interactive talking avatar on social presence, credibility, trust, and patronage intention in an e-commerce website. *Human-centric Computing and Information Sciences*, 7(1), Article 42. <https://doi.org/10.1186/s13673-017-0123-4>
- Linxen, S., Sturm, C., Brühlmann, F., Cassau, V., Opwis, K., & Reinecke, K. (2021). How WEIRD is CHI? Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, 1–14. <https://doi.org/10.1145/3411764.3445488>
- Lukyanenko, R., Maass, W., & Storey, V. C. (2022). Trust in artificial intelligence: From a foundational trust framework to emerging research opportunities. *Electronic Markets*, 32(4), 1993–2020. <https://doi.org/10.1007/s12525-022-00605-4>
- Marconi, L., Longo, L., & Cabitza, F. (2026). Assessing interaction quality in human–AI dialogue: An integrative review and multi-layer framework for conversational agents. *Machine Learning and Knowledge Extraction*, 8(2), Article 28. <https://doi.org/10.3390/make8020028>
- Newton, D. C., Tomin, A. J., & LaMontagne, A. D. (2021). Exploring the challenges and opportunities for improving the health and wellbeing of international students: Perspectives of international students. *Journal of the Australian and New Zealand Student Services Association*, 29(1), 18–34. <https://doi.org/10.30688/janzssa.2021.1.02>
- Ng, S. W. T., & Zhang, R. (2025). Trust in AI chatbots: A systematic review. *Telematics and Informatics*, 97, Article 102240. <https://doi.org/10.1016/j.tele.2025.102240>
- Nourian, L., Callis, A., Patterson, S., Miao, J., Heard, J., & Tigwell, G. W. (2026). Understanding how international students in the U.S. are using conversational AI to support cross-cultural adaptation. *arXiv:2605.15127*. <https://doi.org/10.48550/arXiv.2605.15127> (preprint).
- Pereira, D. S. M., Falcão, F., Costa, L., Lunn, B. S., Pêgo, J. M., & Costa, P. (2023). Here's to the future: Conversational agents in higher education — a scoping review. *International Journal of Educational Research*, 122, Article 102233. <https://doi.org/10.1016/j.ijer.2023.102233>

- Perry, C. J., Lausch, D., McKim, C. A., & Weatherford, J. (2020). Knowledge, use, and perceived value of university student services. *Journal of International Students*, 10(3), 613–628. <https://doi.org/10.32674/jis.v9i4.269>
- Polyportis, A., & Pahos, N. (2025). Understanding students' adoption of the ChatGPT chatbot in higher education: The role of anthropomorphism, trust, design novelty and institutional policy. *Behaviour & Information Technology*, 44(2), 315–336. <https://doi.org/10.1080/0144929X.2024.2317364>
- Qiu, L., & Benbasat, I. (2009). Evaluating anthropomorphic product recommendation agents: A social relationship perspective to designing information systems. *Journal of Management Information Systems*, 25(4), 145–182. <https://doi.org/10.2753/MIS0742-1222250405>
- Shahzad, M. F., Xu, S., & Javed, I. (2024). ChatGPT awareness, acceptance, and adoption in higher education: The role of trust as a cornerstone. *International Journal of Educational Technology in Higher Education*, 21(1), Article 46. <https://doi.org/10.1186/s41239-024-00478-x>
- Shuster, K., Poff, S., Chen, M., Kiela, D., & Weston, J. (2021). Retrieval augmentation reduces hallucination in conversation. In *Findings of the Association for Computational Linguistics: EMNLP 2021* (pp. 3784–3803). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2021.findings-emnlp.320>
- Simons, D. J., Shoda, Y., & Lindsay, D. S. (2017). Constraints on generality (COG): A proposed addition to all empirical papers. *Perspectives on Psychological Science*, 12(6), 1123–1128. <https://doi.org/10.1177/1745691617708630>
- Soria, K. M. (2023). Disparities in college students' access to academic advising during the COVID-19 pandemic. *NACADA Journal*, 43(1), 17–30. <https://doi.org/10.12930/NACADA-22-27>
- Soria, K. M., Kokenge, E., Heath, C. A., Standley, E. C., Wilson, S. J. F., Connley, J. R., & Agramon, A. I. (2023). Factors associated with academic advisors' burnout. *NACADA Journal*, 43(2), 105–120. <https://doi.org/10.12930/NACADA-23-14>
- Tajvar, M., Ahmadizadeh, E., Sajadi, H. S., & Shaqura, I. I. (2024). Challenges facing international students at Iranian universities: A cross-sectional survey. *BMC Medical Education*, 24(1), Article 210. <https://doi.org/10.1186/s12909-024-05167-x>
- Tamascelli, M., Bunch, O., Fowler, B., Taeb, M., & Cohen, A. (2025). Academic advising chatbot powered with AI agent. In *Proceedings of the 2025 ACM Southeast Conference (ACMSE 2025)* (pp. 195–202). ACM. <https://doi.org/10.1145/3696673.3723065>
- Terwee, C. B., Bot, S. D., de Boer, M. R., van der Windt, D. A., Knol, D. L., Dekker, J., Bouter, L. M., & de Vet, H. C. (2007). Quality criteria were proposed for measurement properties of health status questionnaires. *Journal of Clinical Epidemiology*, 60(1), 34–42. <https://doi.org/10.1016/j.jclinepi.2006.03.012>
- Tielman, M. L., Neerincx, M. A., & Brinkman, W.-P. (2019). Design and evaluation of personalized motivational messages by a virtual agent that assists in post-traumatic stress disorder therapy. *Journal of Medical Internet Research*, 21(3), Article e9240. <https://doi.org/10.2196/jmir.9240>
- Verhagen, R. S., Neerincx, M. A., Parlar, C., Vogel, M., & Tielman, M. L. (2023). Personalized agent explanations for human-agent teamwork: Adapting explanations to user trust, workload, and performance (extended abstract). In A. Ricci, W. Yeoh, N. Agmon, & B. An (Eds.), *Proceedings of the 22nd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2023)* (pp. 2316–2318). IFAAMAS.
- Wang, S., Fatima, N., Shahbaz, M., & Asif, M. (2026). Building user trust in AI chatbots for customer service through human-like cues and perceived reliability. *Scientific Reports*, 16(1), Article 7860. <https://doi.org/10.1038/s41598-026-38179-2>
- Weng, Z., & Fu, Y. (2025). Generative AI in language education: Bridging divide and fostering inclusivity. *International Journal of Technology in Education*, 8(2), 395–420. <https://doi.org/10.46328/ijte.1056>
- Wilson, M. L., Mackay, W. E., Chi, E. H., Bernstein, M. S., Russell, D. M., & Thimbleby, H. W. (2011). RepliCHI — CHI should be replicating and validating results more: Discuss. In *CHI '11 Extended Abstracts on Human Factors in Computing Systems* (pp. 463–466). ACM. <https://doi.org/10.1145/1979742.1979491>

- Wilson, M. L., Mackay, W. E., Chi, E. H., Bernstein, M. S., & Nichols, J. (2012). RepliCHI SIG: From a panel to a new submission venue for replication. In CHI '12 Extended Abstracts on Human Factors in Computing Systems (pp. 1185–1188). ACM. <https://doi.org/10.1145/2212776.2212419>
- Wilson, M. L., Resnick, P., Coyle, D., & Chi, E. H. (2013). RepliCHI: The workshop. In CHI '13 Extended Abstracts on Human Factors in Computing Systems (pp. 3159–3162). ACM. <https://doi.org/10.1145/2468356.2479636>
- Wilson, M. L., Chi, E. H., Reeves, S., & Coyle, D. (2014). RepliCHI: The workshop II. In CHI '14 Extended Abstracts on Human Factors in Computing Systems (pp. 33–36). ACM. <https://doi.org/10.1145/2559206.2559233>
- Xiong, G., Jin, Q., Lu, Z., & Zhang, A. (2024). Benchmarking retrieval-augmented generation for medicine. In L.-W. Ku, A. Martins, & V. Srikumar (Eds.), Findings of the Association for Computational Linguistics: ACL 2024 (pp. 6233–6251). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2024.findings-acl.372>
- Zhu, X., & Yang, J. H. (2026). Exploring the role of generative AI in international students' cross-cultural adaptation: A stimulus-organism-response framework approach. *International Journal of Educational Development*, 122, Article 103568. <https://doi.org/10.1016/j.ijedudev.2026.103568>

Figures

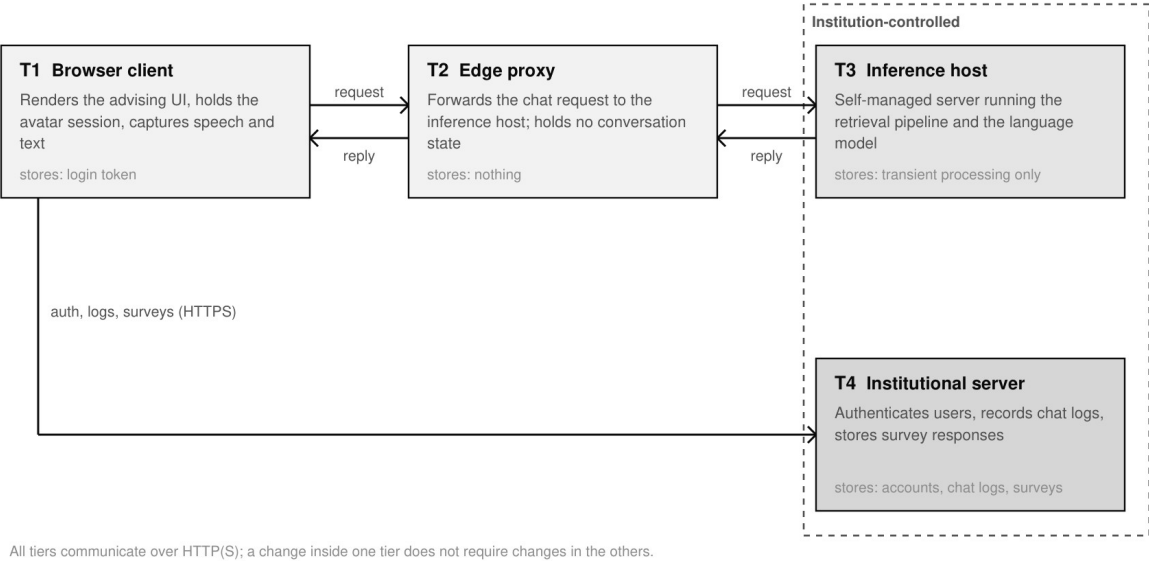


Fig. 1. Four-tier system architecture.

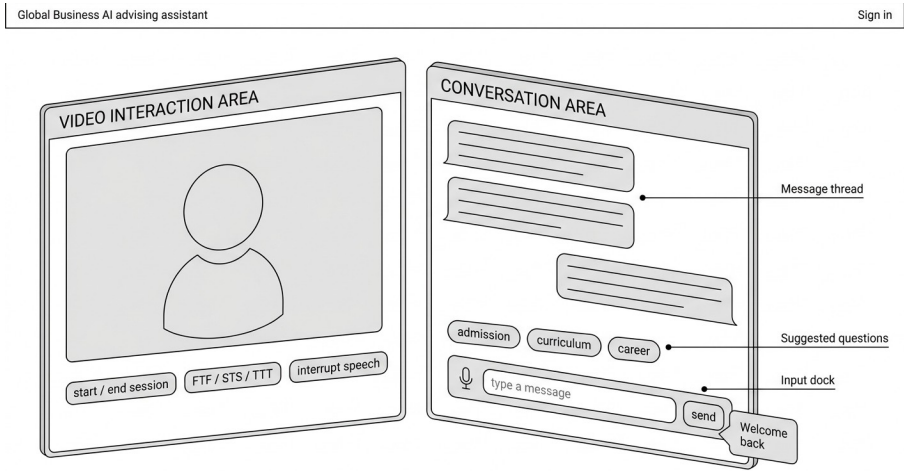
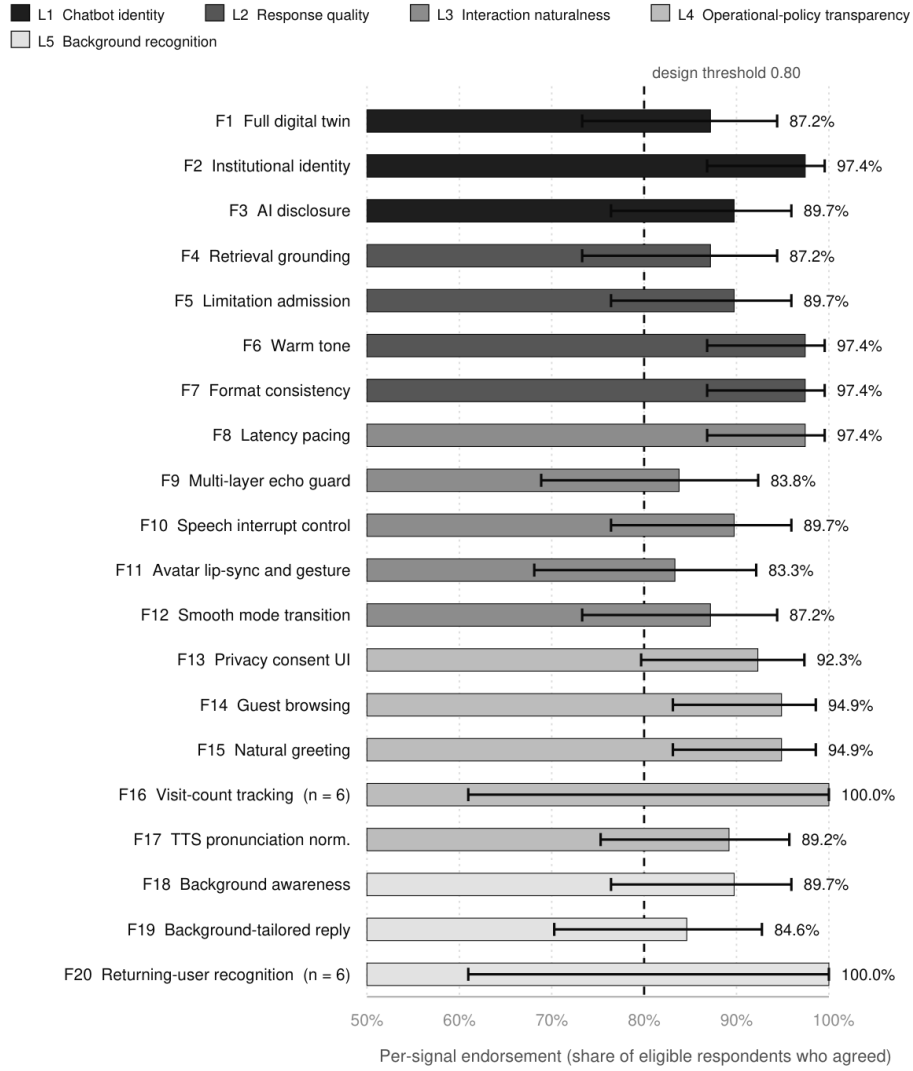
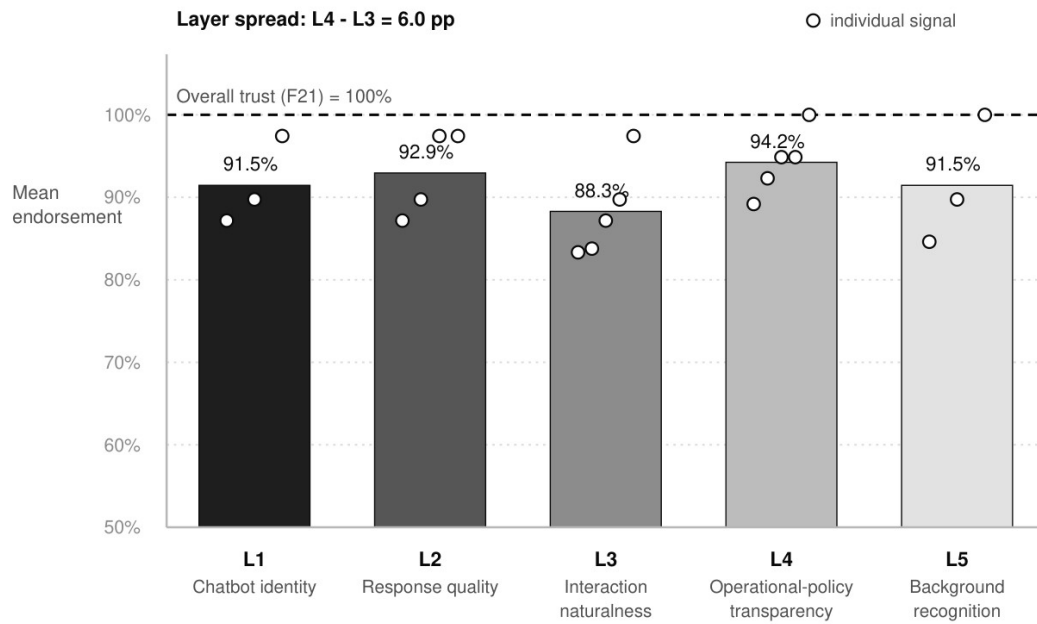


Fig. 2. Structure of the advising interface. This schematic illustration was generated with Google Gemini and was reviewed and edited by the authors.



Bars show point estimates; horizontal lines show Wilson 95% confidence intervals. Signals marked (n = 6) are conditional; all others have n = 36-39. Only a signal whose interval lies entirely above 0.80 clears the threshold.

Fig. 3. Per-signal endorsement with Wilson 95% confidence intervals.



Bars show the mean endorsement of the signals in each layer; open circles show the individual signals.
Two conditional signals (F16, F20; eligible n = 6) are included and both were endorsed by 100%, which lifts L4 and L5.

Fig. 4. Layer-mean endorsement.